School of Doctoral Studies in Biological Sciences
University of South Bohemia in České Budějovice
Faculty of Science

Molecular Evolution of Flaviviral Genes

Ph. D. thesis

Mgr. Jiří Černý

Supervisors: **Doc. RNDr. Daniel Ružek, Ph.D** (1,2) & **Prof. RNDr. Libor Grubhoffer, CSc.** (1,3)

(1) Institute of Parasitology, Biology Centre, Academy of Science of the Czech Republic
(2) Veterinary Research Institute
(3) Faculty of Science, University of South Bohemia

České Budějovice 2015

This thesis should be cited as:
Černý J., 2015: Molecular Evolution of Flaviviral Genes. Ph.D. Thesis Series No 9. University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 220 pp.

Annotation

Flaviviruses are important human and veterinary pathogens causing tens thousands of deaths annually. Despite Flavivirus life cycle, their molecular biology, pathogenesis, biochemistry of their proteins etc. are intensively studied their evolution is somehow out of scope of actual research. This thesis describes mechanisms standing behind evolution of flaviviral genes and their relationship to other viral and cellular proteins. Special attention is paid to (flavi)viral polymerases, as they were showed to be the most suitable marker for studies on evolution of RNA viruses.

Declaration [in Czech]

Prohlašuji, že svoji dizertační práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.
Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své dizertační práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajobě kvalifikační práce.
Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích dne 19. 10. 2015

……
Jiří Černý

This thesis originated from a partnership of Faculty of Science, University of South Bohemia, and Institute of Parasitology, Biology Centre of the ASCR, supporting doctoral studies in the Molecular and Cell Biology and Genetics study program.

Přírodovědecká fakulta
Faculty of Science

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Financial support

Czech Science Foundation (P302/12/2490)
Grant Agency of University of South Bohemia (155/2013/P, 04-026/2011/P)
Seventh Framework Programme ANTIGONE (278976)
Ministry of Education, Youth and Sports of the Czech Republic, NPU I program (LO1218)

Acknowledgements
I would like to express my gratitude to both my supervisors' doc. Daniel Růžek and prof. Libor Grubhoffer for patient help and guidance whenever I need it. My thanks belong also to all members of Laboratory of Molecular Ecology of Vectors and Vector-borne Diseases and Laboratory of Arbovirology as well as whole to Budweis Tick Group for support. Special thanks belong to prof. Rolf Hilgenfeld and dr. Naoki Sakai from the University of Lübeck, Germany and to dr. Manfred Weidmann and dr. Meik Dilcher from the University of Göttingen, Germany for supervising me during my internship in their laboratories. I would also like to thanks to my parents for their support of all kinds. Finally, I would like to express many thanks to my beloved wife Barča for her help, love and patience and to my daughter Ema for all hours of her sleep which allowed me to finish this work.

Dedication:
I would like to dedicate this work to my grandparents Anna and Štefan Knotkovi, who left us suddenly in August 2014. Granny, grandpa, I will never forget to you.

List of papers and author's contribution

1) **Jiří Černý**; Barbora Černá Bolfíková; James J. Valdés; Libor Grubhoffer; Daniel Růžek: Evolution of tertiary structure of viral RNA dependent polymerases, PLoS ONE (IF 3.234), 2014, doi: 10.1371/journal.pone.0096070
Jiří Černý selected candidates for analysis, prepared alignments, evaluated results and wrote manuscript.

2) Petra Formanová; **Jiří Černý**; Barbora Černá Bolfíková; James J. Valdés; Irina Kozlova; Yuri Dzhioev; Daniel Růžek: Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients. Ticks and Tick-Borne Diseases (IF 2.718), 2015, 6(1):38-46. doi:10.1016/j.ttbdis.2014.09.002
Jiří Černý produced homologue structural models of TBEV proteins and participated on aligning of TBEV sequences and on interpretation of results.

3) **Jiří Černý**; Barbora Černá Bolfíková; Paolo M. de A. Zanotto, Libor Grubhoffer, Daniel Růžek: A deep phylogeny of viral and cellular right-hand polymerases. Infection, Genetics and Evolution (3.015), 2015 Sep 30. pii: S1567-1348(15)00402-5. doi: 10.1016/j.meegid.2015.09.026.

Jiří Černý selected candidates for analysis, prepared alignments, evaluated results and wrote manuscript.

4) **Jiří Černý**, Martin Selinger, Martin Palus, Zuzana Vavrušková, Hana Tykalová, Lesley Bell-Sakyi, Libor Grubhoffer, Daniel Růžek: Expression of a second open reading frame present in the genome of tick-borne encephalitis virus strain Neudoerfl is not detectable in infected cells. (manuscript)
Jiří Černý selected candidates for analysis, prepared alignments, evaluated results and wrote manuscript.

5) **Jiří Černý**, Barbora Černá Bolfíková, Libor Grubhoffer, Daniel Růžek: Genomes of viruses classified in genus Flavivirus (family Flaviviridae) evolved via multiple recombination events. (manuscript)
Jiří Černý selected candidates for analysis, prepared alignments, evaluated results and wrote manuscript.

Contents

1. INTRODUCTION

Viruses classified within genus Flavivirus (family *Flaviviridae*) are important human and veterinary pathogens. Great amount of work and incredible financial expenses were given to fight with the threat which flaviviruses pose. During last decades numerous important discoveries were done in many fields of flavivirus biology as biochemistry of flaviviral proteins, flavivirus-host cell interaction, pathology of flavivirus infection, anti-flavivirus immunology, epidemiology of flaviviruses, etc. (Coutard & Canard 2010; Coutard et al. 2008; Randolph & team 2010).

Nevertheless, molecular mechanisms standing behind evolution flaviviral as well as all viral genes are still neglected. Despite, understanding of these mechanisms is very important in construction of effective antiviral drugs, precise modeling of viral epidemics etc.

This work is focused on molecular evolution of flaviviral genes and genomes. It summarizes my original results and gives them in context of recent knowledge in this field.

1.1 Flaviviruses

1.1.1 Ecology and epidemiology of flaviviruses

Vast majority of flaviviruses are arboviruses (arthropod-borne viruses). They are transmitted from one vertebrate host to another by blood-sucking vectors, mostly mosquitoes or ticks. Remaining flaviviruses infects either only mosquitoes (Cell fusing agent virus – CFAV, etc.) or they were isolated only from vertebrate hosts and their vectors remain elusive (Entebbe bat virus – EBV, Modoc virus – MODV, Rio Bravo virus – RBV, etc.) (Gould et al. 2001; Gould et al. 2004).

Arthropod-borne flaviviruses may be divided into four main groups according to ecological niche they occupy. The first group consists of mosquito-borne flaviviruses transmitted by *Culex* mosquitoes to bird hosts. The most important representatives of this group are Japanese encephalitis virus (JEV), West Nile virus (WNV), St. Louise Encephalitis virus (SLEV), Murray Valley encephalitis virus (MVEV) etc. (Le Flohic et al. 2013; Petersen et al. 2013). The second group

is formed by mosquito-borne flaviviruses transmitted by *Aedes* mosquitoes to primate hosts. This group is represented by four serotypes of Dengue virus (DENV), Yellow fever virus (YFV) etc. (Beck et al. 2013; Messina et al. 2014). The third group consists of viruses transmitted from ticks to mammals such as Tick-borne encephalitis virus (TBEV), Omsk hemorrhagic fever virus (OHFV), and Louping ill virus (LIV). The fourth group includes flaviviruses transmitted from ticks to sea birds. Maeban virus (MEAV) and Tyuleniy virus (TYUV) are representatives of this group (Gritsun et al. 2003).

Despite these well-defined niches, most flaviviruses are promiscuous infecting various hosts and vectors. Humans are usually death-end hosts in flavivirus transmission cycle. Only small subset of flaviviruses (as DENV, YFV) can establish successful urban transmission cycle being transmitted by an appropriate vector from human to human (Diaz et al. 2012; Durbin et al. 2013a; Monath 2001).

Affiliation of flaviviruses to ecological niches also determines clinical signs of their infection in humans. *Culex* transmitted mosquito-borne flaviviruses and mammals infecting tick-borne flaviviruses cause usually viral encephalites (Gritsun et al. 2003; Knox et al. 2012; Unni et al. 2011) with OHFV and Kyasanur forest disease virus (KFDV) being an exception as they cause hemorrhagic fevers (Růžek et al. 2010). *Aedes* transmitted mosquito borne flaviviruses cause usually hemorrhagic fevers (Bäck & Lundkvist 2013; Gardner & Ryman 2010). Sea birds infecting tick-borne flaviviruses usually do not usually cause any clinical symptoms in humans (Dietrich et al. 2011; Gritsun et al. 2003).

From the medical point of view, DENV1-4, YFV, JEV, WNV, TBEV, SLEV, and MVEV belong among the most important flaviviruses endangering people in large areas continuously for a long time (Gould & Solomon 2008) (Figure 1). Apart these, there exist flaviviruses such as OHFV (Růžek et al. 2010), Alkhurma virus (ALKV) (Charrel et al. 2001), KFDV (Venugopal et al. 1994) etc. which emerge unexpectedly causing small but deadly epidemics (Figure 1). All together flaviviruses stand behind tens thousands of human deaths and billions euros of economical loses annually (Gould & Solomon 2008).

Almost whole human population lives in areas where at least one flaviviral species is endemic (Gould & Solomon 2008). Plus many flaviviruses recently expanded their endemic areas being introduced to novel loci either on new continents or to areas with higher altitude or latitude (Casati et al. 2006; Deardorff et al. 2013). Due to this reasons, flaviviruses pose extremely important threat to public and animal health. Moreover, flaviviruses have high zoonotic potential promiscuously infecting various hosts and vectors including important domestic animals. It brings them in close proximity of humans making human infections quite easy. Therefore, one-health strategy unifying human and animal health surveillance with careful ecological, epidemiological and evolutionary studies is needed to control, successfully predict and fight with possible future flaviviral outbreaks.

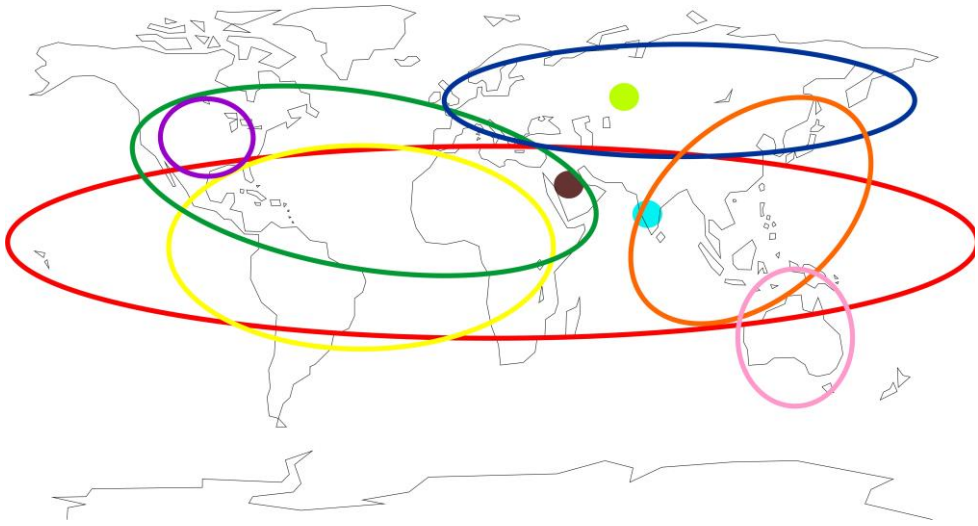


Figure 1 - Geographic distribution of medically important flaviviruses: The geographic distribution of the most medically important flaviviruses is shown by circles (DENV by red, YFV by yellow, WNV by green, TBEV by blue, JEV by orange, MVEV by pink, and SLEV by violet). The geographic locations of emerging flaviviruses endemic only on small geographic areas are indicated dots (OHFV by green, ALKV by brown, and KFV by azure).

1.1.2 Molecular biology of flaviviruses

Flavivirus are enveloped viruses. Their particles are spherical, about 50nm in diameter. Particles has icosahedral symmetry with triangulation number T=3 (Figure 2A) (Huiskonen & Butcher 2007).

Flaviviral genome is not fragmented. It consists of one single-stranded RNA molecule of positive polarity (+ssRNA) roughly 11,000nt long. Flaviviral genomic RNA is terminated by 7-methylguanosine cap at its 5' end but it lacks polyA tail at its 3' end. It contains a single open reading frame embedded by two untranslated regions. Flaviviral genomic RNA serves also as an mRNA being translated into a single polyprotein. This polyprotein is co- and posttranslationally processed by viral and cellular proteases into ten major flaviviral proteins: three structural (C, M, and E) and seven nonstructural (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) proteins (Figure 2B) (Harris et al. 2006). Apart the major proteins, genomic RNA of some flaviviruses encodes for minor proteins such as NS1' (Melian et al. 2010) and WARF4 (Faggioni et al. 2012).

Flaviviral structural proteins form viral particles, in which genomic RNA is surrounded by a virus core formed by a highly positively charged C protein (Pong et al. 2011). Precise molecular mechanism of RNA encapsidation is still unknown. Virus core is surrounded by envelope formed by proteins M and E (Yu et al. 2008). Protein M has chaperon function. It is produced by cleavage from preM during virus maturation in Golgi apparatus (Stadler et al. 1997). preM precursor prevents bounding of immature viral particles during exocytosis (Junjhon et al. 2010). Flaviviral E protein, the dominant part of flavivirus envelope, is responsible for receptor recognition and virus-cell membrane joining. In neutral pH, protein E forms dimers (Rey et al. 1995). After its N-terminal domain binds cellular receptor, virus particle is internalized and transported to the late endosome. In low pH of the late endosome, E protein undergoes reassortment forming trimmers necessary for virus-cell membrane joining (Bressanelli et al. 2004; Modis et al. 2004).

While structural proteins form flavivirus particle, nonstructural proteins catalyze individual steps in flavivirus replication cycle and modulate host immune response against the virus. The largest flaviviral protein NS5 has two domains (Davidson 2009). The C-terminal domain bears polymerase activity

and has major role in replication of flaviviral genome (Tan et al. 1996) while N-terminal domain bears methyltransferase activity and is responsible for methylation of flavivirus cap (Egloff et al. 2002). The second largest flaviviral protein is NS3 protein. It also contains two domains. The C-terminal domain is an ATP dependent helicase and an RNA phosphatase tightly cooperating with the NS5 protein on replication and capping of flaviviral genome (Utama et al. 2000). The N-terminal part of NS3 protein acts in cooperation with NS2B protein as a viral protease (Chambers et al. 1990). Precise role of the last soluble flaviviral protein, NS1 protein, is still not understood. It seems its secreted form modulates anti-virus host response, while its endoplasmic reticulum bound form participates on genome replication (Muller & Young 2013). Role of NS2A, NS4A, and NS4B proteins remains elusive. As these transmembrane proteins are necessary for flavivirus replication, it is speculated that they may be involved in formation of replication machinery (Yu et al. 2013).

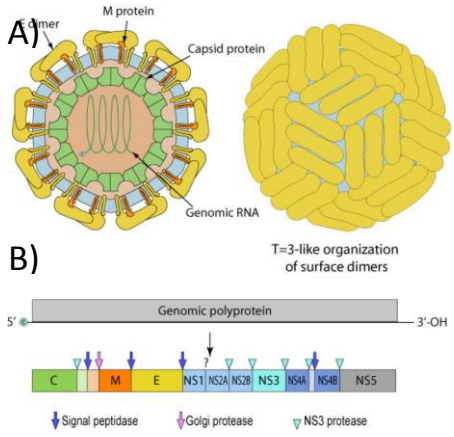


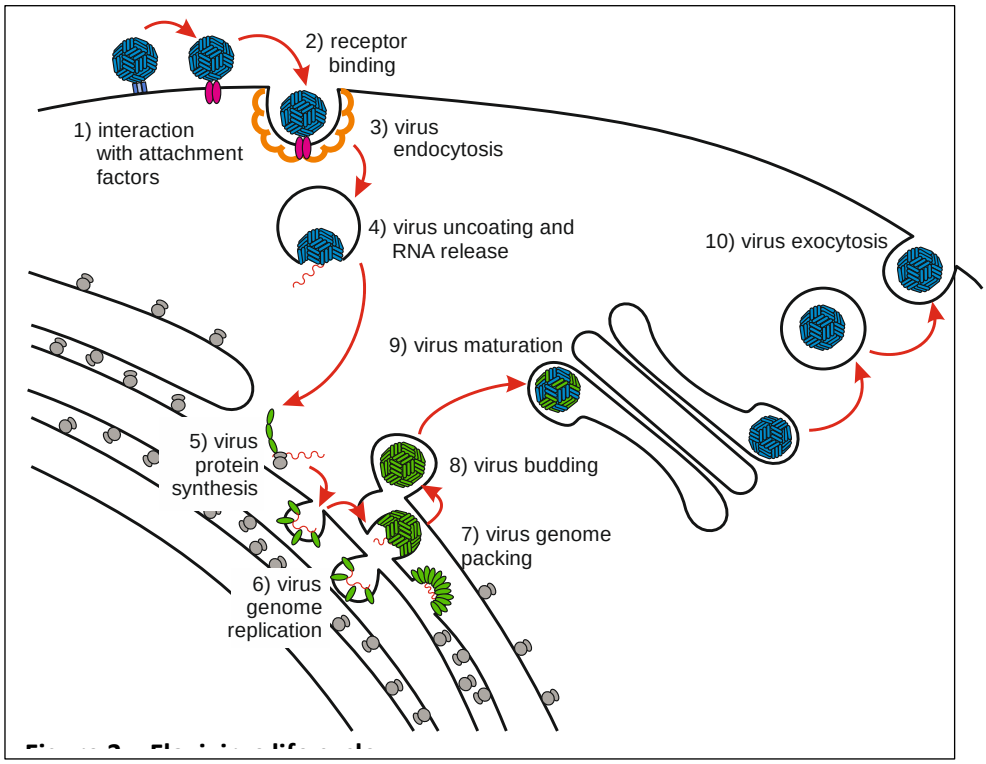
Figure 2 - Structure of flaviviral virion and genome: A) Flavivirus virion has triangulation number T=3. It consists of a core formed by protein C and envelope developed from host cell membranes containing viral proteins E and M. Flavivirus virion has approximately 50nm in diameter. B) Flavivirus genome is approximantely 11.000nt long. It is terminated by 7mG cap at the 5´end but it lacks polyA tail at the 3´end. The genome encode single open reading frame which is co- and posttranslationally cleaved by viral and cellular proteases into ten major viral proteins (C, M, E, NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5). The figure was created according ViralZone (http://viralzone.expasy.org/).

Flavivirus life cycle starts by attachment of the viral envelope protein E to host receptors followed by virus internalization into the host cell by clathrin-mediated endocytosis (Perera-Lecoin et al. 2014). After endocytosis, flaviviral

membrane fuses with the membrane of late endosome realizing viral RNA genome into cytoplasm (Stiasny et al. 2004). The flaviviral +ssRNA genomic is immediately translated into a viral polyprotein cleaved by viral and cellular proteases on all structural and non-structural proteins (Bera et al. 2007). When sufficient amount of viral proteins is produced, replication takes place in virus induced replication factories derived from endoplasmic reticulum (Paul & Bartenschlager 2013; Westaway et al. 1997). Later, virus assembly occurs. Virions bud into the endoplasmic reticulum, are transported to the Golgi apparatus where they maturate, and then exit the host cell via the secretory pathway (Apte-Sengupta et al. 2014; Junjhon et al. 2014; Welsch et al. 2009)(Figure 3).

1.1.3 Evolution of flaviviruses

Genus Flavivirus pose a monophyletic group within the family *Flaviviridae* (Venugopal et al. 1994). As other genera within the family *Flaviviridae* (Hepacivirus, Pestivirus, and Pegivirus) are not transmitted by any arthropod vector, viruses classified in genus Flavivirus evolved most probably from non-



vectored vertebrate viruses (Gould et al. 2003). Recent molecular clock studies showed that genus Flavivirus appeared 120 000 years ago in Africa (Pettersson & Fiz-Palacios 2014). This is in contrast with older studies postulating that genus Flavivirus emerged after the last glaciation maximum before 10 000 years (Gould et al. 2003; Zanotto et al. 1996b). After their emergence, Flaviviruses were further dispersed to all continents except continental Antarctica. The most probable vector of this dispersal are migratory animals (mostly birds) (Pettersson & Fiz-Palacios 2014). Currently, in anthropocene, flaviviruses expand mostly due to human activities. Slave trade stand behind introduction of YFV into Americas in 16th century (Bryant et al. 2007), while used tire trade probably caused introduction of WNV into USA in 1999 (Murray et al. 2010).

Shortly after its emergence, genus Flavivirus divided according occupied vector-host associated niches on tick- and mosquito-borne virus groups (Gould et al. 2003). Molecular clock dating shows that this split happened some 50 000 years ago (Pettersson & Fiz-Palacios 2014). Further speciation lead to establishment of ecologically separated groups described above in chapter 1.1.1. Ecology and epidemiology of flaviviruses. Evolutionary relations between viruses classified in genus Flavivirus are shown in Figure 4.

Mosquito-borne flaviviruses form two evolutionary and ecologically distinct groups (group I and II) (Gould et al. 2003). These two groups are separated by a group of flaviviruses with unknown vector. Group I includes viruses associated with *Aedes* (DENV1-4 etc.) and *Culex* mosquitoes as vector (JEV, WNV, MVEV, etc.). Group II associates only *Aedes* mosquito vectored viruses (YFV etc.).

Both groups of mosquito-borne flaviviruses are evolving in fast and discontinuous manner (Gould et al. 2003). It is due to feeding habits of mosquitoes which can feed many times on many hosts during their replication cycle giving the virus more opportunities to infect new hosts. This is apparent also from the evolutionary tree of mosquito-borne flaviviruses, which has balanced appearance (Zanotto et al. 1996b).

In contrast to mosquito-borne flaviviruses, evolution of tick-borne flaviviruses is rather slow, continuous and clinal. In tick-borne flaviviruses, evolution was 2.5 times slower than in the case of mosquito-borne flaviviruses and there can be tracked direct correlation between genetic and geographical distance (Shiu et

al. 1991; Zanotto et al. 1995). It shows that spread of these viruses is slow and it is not influenced by migratory birds or international trade (Gould et al. 2003).

Despite geographic distribution of many flavivirus species overlaps, genetic data shows that recombination did not play a role in evolution of mosquito- or tick-borne flaviviruses. Phylogenetic trees produced on the base of any flaviviral gene are almost identical (Gould et al. 2003). Some recombination signal was observed in SLEV (Gaunt et al. 2001) but further extensive reevaluation led to rejection of this hypothesis (Baillie et al. 2008). It is good news as it opens a way to production of safe life attenuated hybrid vaccines (Durbin et al. 2013b; Wang et al. 2014).

In absence of recombination, the leading strength forming flaviviral genes is slow adaptive evolution. As flaviviruses are arboviruses transmitted between arthropod and vertebrate hosts, their proteins have to fulfill their role in both types of hosts. Flaviviruses circulate in nature as a quasispecies mix (Chin-inmanu et al. 2012). Some of these quasispecies are more suitable for replication in vector cells while other are better adopted to host cells. Serial passaging flaviviruses on host cell cultures can lead to selection of strains more suitable for replication in host cells. Such strains exhibit changes in genome, which cause their increased pathogenicity (Růzek et al. 2008). Role of several mutations on increased/decreased pathogenicity was already described for many flaviviruses (Brault et al. 2007; Lin et al. 2014; Růzek et al. 2008; Tajima et al. 2010; Yamaguchi et al. 2011).

1.2 Evolution of viral proteins, genes and genomes

1.2.1 Types of viral proteins from the evolutionary biology point of view

From the evolutionary point of view, viral proteins can be divided into five classes in three groups (Koonin et al. 2006). Group I consist of virus genes with readily detectable homologs in cellular life forms. This group contains either proteins which were recently incorporated into viral genomes (class 1) or proteins which were adopted relatively long time ago (class 2). Proteins in class

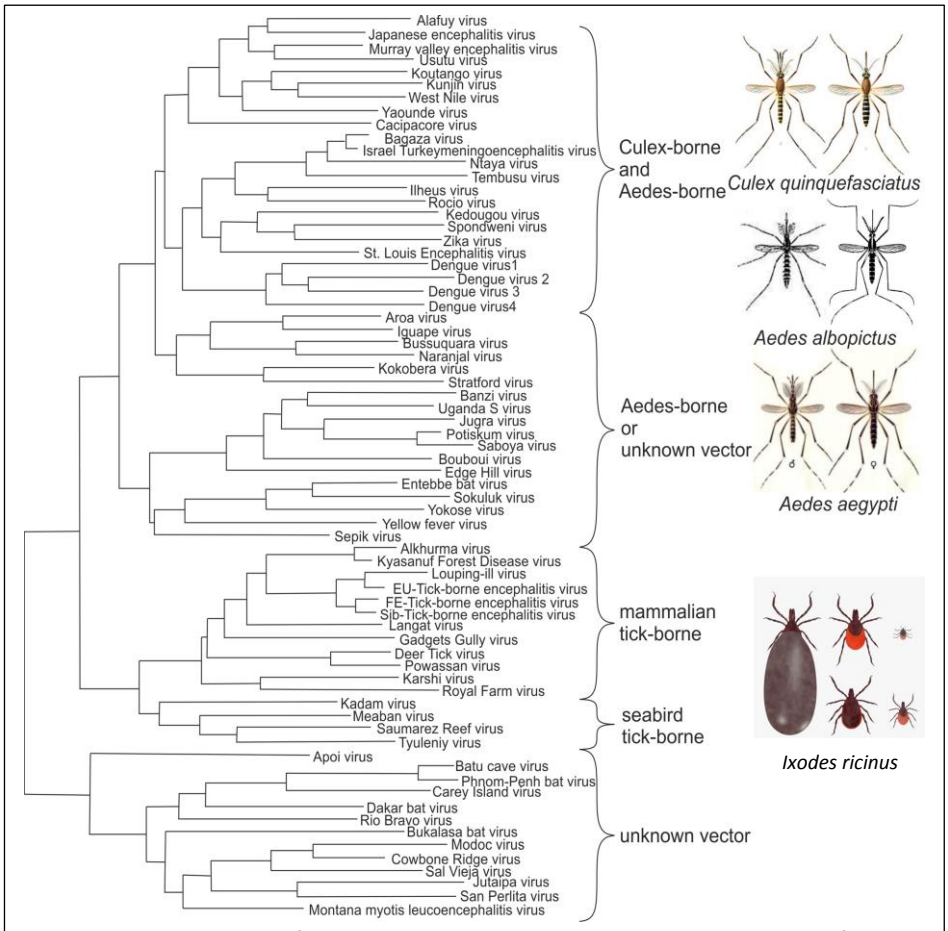


Figure 4 – Evolution of viruses within genus Flavivirus: Evolution of genus Flavivirus was reconstructed by maximum likelihood using a fragment of NS5 protein. Figure was adopted from (Gould et al. 2003). Right panel shows some of the most typical vectors of flaviviruses.

1 have close cellular homologues and are typical only for a narrow group of viruses. Proteins in class 2 have more distant cellular homologues but are typical for wider group of viruses (Koonin et al. 2006). Group II includes virus-specific genes. These can be either specific for a narrow (class 3 – ORFans) or relatively wide group of viruses (class 4). Group III (Class 5) consist of so called viral hallmark genes. These genes have only extremely distant cellular homologues but they share high homology across many very diverse groups of viruses (Koonin et al. 2006).

Table 1 – Evolutionary division of viral genes:

	Group I		Group II		Group III
	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**
Poliovirus example	---	3C (chimotripsin-like protease)	3A (unknown function)	Vpg (genome linked protein)	VP1-VP4 (jelly-roll capsid protein), 3D (RNA polymerase)
Alphavirus example	nsP3 (virus replication)	nsP2 (protease), nsP1 (methyl-transferase)	CP (protease)	E1 (envelope protein)	nsP4 (RNA polymerase)
Flaviviral genes	NS5Met	NS3Pro, NS3Hel	M, C, NS2A, NS2B, NS4A, NS4B	E	NS5Pol (RNA polymerase)

1.2.2 Adaptive evolution of viral genes

RNA virus encoded RNA-dependent polymerases miss the proofreading activity. It leads to relatively high percentage of improperly incorporated nucleotides (mutation rate can reach up to 10^{-3}) (Nowak 1990; Ogata et al. 1991), which gives to RNA viruses very high evolutionary plasticity. Viruses bearing mutations increasing their fitness reach significant advantages in further replication cycles and therefore such mutations are fixed very fast in virus population. It leads to swift establishment of new virus strains (Cabanillas et al. 2013; Pickett et al. 2011; Smith et al. 2012).

The strongest selection pressure acts against the parts of viral proteins which are in contact with host immune system (surface epitopes of viral structural proteins etc.) (Carrillo et al. 1989; Nayak et al. 2014; Pandey et al. 2014). On the other hand, functions that are crucial for efficient virus reproduction have to be preserved (Krupovič & Bamford 2010). Therefore, proteins involved in important steps of the virus life cycle accumulate mutations slower and preserve higher degree of conservation (Krupovič & Bamford 2010). The most

conserved proteins among RNA viruses are polymerases, helicases, proteases and methyltransferases (Koonin & Dolja 1993).

Contrary to the primary structure, the tertiary structure of most proteins sharing a common evolutionary origin remains conserved even after the very significant changes in their primary sequence (Holm & Sander 1996; Illergård et al. 2009). It is reached by a high plasticity of interactions among various amino acid residues. Particular interaction may be achieved in a variety of ways (hydrogen bonding, stacking interactions of aromatic residues, hydrophobic interactions, etc.) without substantial changes in the protein fold (Illergård et al. 2009). The most conserved part of the protein is usually the core structure essential for protein function. The core is often surrounded by less conserved structures modifying the protein function. Changes in these additional structures often lead to minor changes in protein character (e. g., different substrate specificity or interacting partners), but the major protein function remains unchanged.

The evolutionary stability of protein tertiary structures can be used to reconstruct the evolutionary relationships of distantly related proteins (Mönttinen et al. 2014; Scheeff & Bourne 2005). This is similar to the paleontological approach where evolution of dinosaurs is deduced only from the similarities in the structure of their bones. In our approach, protein tertiary structures are such bones, while protein sequences pose “dinosaur meat” which is not preserved. One of the possible approaches how to use similarities in protein structure is to create a character matrix quantifying morphological features of studied proteins and use it for a phylogenetic analysis (Ravantti et al. 2013; Scheeff & Bourne 2005). This approach allows studying evolutionary history much deeper than if only sequence information is used.

1.2.3 *De novo* evolution of viral genes

As viral sequences are changing quickly, there is also a huge potential for formation of new genes *de novo* via development of new open reading frames. This process is very beneficial for viruses as it gives them high coding capacity as one sequence can encode more proteins in more reading frames. The best example of such intensive usage of coding capacity is Hepatitis B virus (Glebe & Bremer 2013).

De novo developed genes can arise in three different ways: i) They can be formed in noncoding regions such as intergenic regions (Li et al. 2010), introns (Sorek 2007), and 5' or 3' untranslated regions (Crowe et al. 2006). ii) They can arise in already coding regions by "overprinting" (Sabath et al. 2012). iii) They can be produced by ribosome frameshifting in already coding regions (Faggioni et al. 2012; Melian et al. 2010).

New *de novo* evolved genes from the last two groups significantly affect original genes. Genes from the second group usually reduce expression of original genes (Kozak 2002), while genes from the third group compete for nucleic acid sequence with original gene (Sabath et al. 2012). If the function of such *de novo* evolved genes becomes crucial for virus reproduction it may lead to „extinction" of original gene (Sabath et al. 2012).

1.2.4 Evolution of viral genomes

Viral genomes also evolve very rapidly. Apart classical accumulation of mutations (genetic shift), it is caused by recombination (genetic drift). In such case, recombination usually takes place within two very closely related viruses (usually the same virus species or genus). Importance of recombination for evolution of viral genomes was shown for numerous viruses with segmented genome such as influenza virus (Martcheva 2012) or birnaviruses (Gibrat et al. 2013) etc., and also with non-segmented genome such coronaviruses (Graham & Baric 2010). Nevertheless, numerous viruses, such as flaviviruses, seem not to use recombination in evolution of their genome in present-days (Baillie et al. 2008).

Recombination between viruses and their cellular hosts is also well documented. Incorporation of virus genome into the host DNA is one step in the replication strategy for some groups of viruses such as retroviruses (Matsuoka & Yasunaga 2013). These viruses can also stole part of host genome and incorporate it into the viral genome (Maeda et al. 2008). Nevertheless, incorporation of a part of virus nucleic acid into the host genome is not restricted only on viruses which have this step in their life cycle. It can occasionally happen also in other RNA only viruses. It was documented for viruses within families *Flaviviridae* (Cook et al. 2006; Crochu et al. 2004; Roiz et al. 2009), *Arenaviridae* (Geuking et al. 2009; Klenerman et al. 1997), *Dicistroviridae* (Maori et al. 2007), and *Potyviridae* (Tanne & Sela 2005). This is

most probably caused by recombination between viral RNA and activated cellular retrotransposome (Geuking et al. 2009).

Despite everything written above, there is only a limited number of information about the role of recombination in evolution of viral genomes. Comparison of housekeeping genes (polymerase, helicase, protease, and methyltransferase) from many viral families showed that these viral genes are organized in conserved modules surrounded by less conserved shell formed by other proteins. These modules are organized differently in different viral groups showing that virus genome reorganization and recombination between remote groups of viruses is considered to be one of the major factors of virus evolution (Koonin & Dolja 1993). Nevertheless this problematic is not studied intensively in current days and therefore major mechanisms standing behind formation of viral genomes are still not understood.

1.3 Evolution of viral polymerases and what does it says about evolution of life

1.3.1 Evolution of viral polymerases

Viral RNA-dependent polymerases are the only universally conserved protein of RNA viruses. Genes coding for viral RNA-dependent polymerases were found in all non-satellite RNA viruses and RNA viruses reproducing via a DNA intermediate (Baltimore 1971). Moreover, viral RNA-dependent polymerases display the highest degree of conservation among all viral proteins.

All viral RNA-dependent polymerases contain seven typical sequence motifs (G, F, A, B, C, D and E) (Bruenn 2003; Poch et al. 1989) that incorporate conserved amino acid residues crucial for polymerase function (Gohara et al. 2000; Korneeva & Cameron 2007). Moreover, all viral RNA-dependent polymerases share remarkable structural homology. Their structures resemble a right hand with subdomains called fingers, palm and thumb (Ferrer-Orta et al. 2006; Hansen et al. 1997; Ng et al. 2008; Shatskaya & Dmitrieva 2013). The palm subdomain is structurally well conserved among all viral RNA-dependent polymerases. Finger and thumb subdomains are more variable. They can be fully aligned only among RNA-dependent RNA polymerases of +ssRNA viruses (Ferrer-Orta et al. 2006). The most viral RNA-dependent polymerases

accommodate seven conserved structural motifs (homomorphs) equivalent to conserved sequence motifs (Lang et al. 2013).

Unfortunately, sequence similarity alone was shown to be too low to produce an accurate sequence alignment for further phylogenetic analysis of viral RNA-dependent polymerases using traditional phylogenetic approaches. Therefore it was suggested that the similarities among viral RNA-dependent polymerases may be caused by convergent evolution (Zanotto et al. 1996a).

Hypothesis about convergent evolution of viral RNA-dependent polymerases may be challenged by several arguments. i) The viral RNA-dependent polymerases share seven conserved sequential collinearly arranged motifs; a phenomenon highly improbable to evolve via convergence (Poch et al. 1989). ii) The right hand conformation is not the only fold that can be adapted by RNA-dependent polymerases. For example, cellular RNA-dependent RNA polymerases participating in RNA interference accommodate double barrel conformations which is totally different form right/hand conformation but which can also provide functional fold (Salgado et al. 2006). iii) Conserved protein tertiary structure of all viral RNA-dependent polymerases can supplement missing information in highly diverged protein sequences and allowing us to study the evolution of extremely distantly related proteins (Aravind et al. 2002; Scheeff & Bourne 2005). iv) Modern bioinformatics approaches based on Bayesian analyses are more suitable for reconstruction of distant evolutionary relationships (Huelsenbeck & Ronquist 2001) which could be unnoticed in previous analyses.

1.3.2 Evolution of viruses from the perspective of evolution of viral polymerases

Virus evolution is an extremely complicated story. Viral genes and proteins evolve rapidly and closely related proteins may share only a low degree of sequence homology (Cabanillas et al. 2013; Pickett et al. 2011; Smith et al. 2012). Only a few viral proteins show sufficient conservation across different viral families to be suitable for phylogenetic studies. The most important are methyltransferases, proteases, helicases, polymerases, and jelly-roll capsid protein (Koonin & Dolja 1993; Rossmann & Johnson 1989) but only viral polymerases are present in all families of RNA viruses.

The sequential and structural similarities of virus RNA-dependent RNA polymerases qualify them for the role of a marker gene suitable for studying of RNA virus evolution and they were used in this role many times in history (Bruenn 1991; Dolja & Carrington 1992; Eickbush 1994; Goldbach et al. 1994; Gorbalenya et al. 2002; Koonin 1991; Koonin & Dolja 1993; Mönttinen et al. 2014; Poch et al. 1989; Ravantti et al. 2013; Ward 1993). As virus RNA polymerases were suggested to share too low sequential similarity to be used as a phylogenetic marker for virus evolution (Zanotto et al. 1996a), evolutionary relationships among more distant viral groups are reconstructed by other factors such as genome structure, virus particle organization, genome replication strategies or by combination of these factors in modern times (Ahlquist 2006; Bamford et al. 2005). This approach is very sensitive to artefacts originating from recombination and convergent evolution (Dolja & Koonin 2011; Pond et al. 2012; Scheel et al. 2013; Smith et al. 2013). Nevertheless, very recent phylogenetic studies show that insufficient sequence similarity in virus RNA polymerases may be overcame using information encoded in RNA polymerase structure (Mönttinen et al. 2014; Ravantti et al. 2013).

1.3.3 Evolution of life from the perspective of evolution of polymerases

Reconstruction of evolutionary history of cellular organisms (Archaea, Eubacteria, and Eukarya) is based on genes of the translation apparatus (Woese et al. 1990). Viruses do not encode any genes of translation apparatus. Therefore, they are *a priory* discriminated from deep-rooted phylogenetic studies and we have no idea about their phylogenetic relationships to cellular organisms (Forterre 2006b). Lack of quantitative phylogenetic data lead to formulation of “virus ocean” theory describing viruses as an ocean surrounding evolutionary tree of cellular organisms (Bamford 2003).

Virus origins is nowadays described by three hypotheses: (i) The virus-first hypothesis says that virus-like organism evolved in primordial soup before the primitive cells appeared (Prangishvili et al. 2001), (ii) the escape hypothesis postulate that viruses evolved from genes escaping cellular environment (Hendrix et al. 2000), and finally (iii) the reduction hypothesis assume that viruses originated from intracellular parasites by extreme simplification of their structure (Forterre 2005). All of these hypotheses have their plus and contras. Without a marker gene suitable for deep-rooted virus-cell evolutionary studies,

it is not possible to decide which one describes the virus-cell evolution in the most proper way. Finding a marker gene suitable for virus-cell evolutionary studies is a difficult task because of the enormous sequential differences between the hallmark cellular and viral proteins (Koonin et al. 2006).

Contrary to translation apparatus, which is not necessary for viruses kidnapping host proteosynthetic machinery, all replicationally independent life forms have to contain some form of replication apparatus. Therefore one would expect that genes of this apparatus may be used as universal phylogenetic markers. Unfortunately, genome wide comparisons studies have shown that there are two different replication apparats (Leipe et al. 1999). The first system is typical for Archaea, Eukarya, and vast majority of viruses, while the second one is used to replicate genomes of Eubacteria (Koonin 2006). Right hand polymerases such as viral RNA-dependent RNA or DNA polymerases, single subunit DNA-dependent RNA polymerases and DNA polymerases of families A, B, D, X, and Y form the key component of the first, archeo-eukaryotic replication apparatus, while DNA polymerases family C are responsible for replication of eubacterial genomes (Filée et al. 2002; Forterre 2006b). Archaeo-eukaryotic and Eubacterial replication systems share only a small number of proteins which do not play essential role in replication and which are most probably recent recombinants (Forterre 2006b; Koonin 2006).

Numerous theories describe possible evolution of this strange duality in such crucial biological aspect as replication (Filée et al. 2002; Forterre 2002; Forterre 2005; Forterre 2006a; Forterre 2006b; Koonin 2006; Koonin et al. 2006). Most probably, it will be never possible to decide which one of these theories is the right one but I would cline to the possibility that the archaeo-eukaryotic replication apparatus pose the original replication system while eubacterial replication apparatus evolved more recently probably after divergence of Eubacteria from the last universal common ancestor (Koonin 2006). This theory is supported by two indirect indications: i) The archaeo-eukaryotic replication apparatus is the most widely distributed system, despite eubacterial DNA polymerases are more effective enzymes than right-hand polymerases (Koonin 2006). Right-hand polymerases can be found even in some Eubacteria. ii) Absence of eubacterial replication apparatus among viruses (even that using Eubacteria as their hosts) indicates that this niche was already occupied when eubacterial replication apparatus appeared (Koonin 2006).

With limitations described above, genes of archaeo-eukaryotic replication apparatus can be used as markers for distant phylogeny namely to reconstruction of virus-cell evolutionary relationships. This approach was already used in the study focused on primases (Iyer et al. 2005).

Therefore right-hand polymerases may also be used as a marker gene to reconstruct virus-cell evolutionary relationships (Mönttinen et al. 2014). This protein superfamily consist of numerous protein families including viral RNA-dependent RNA and DNA polymerases. As viral RNA-dependent polymerases, also all proteins within the superfamily of the right-hand polymerases fold in a structure resembling right hand. They contain three subdomains called fingers, palm, and thumb (Hansen et al. 1997; Kohlstaedt et al. 1992; Ollis et al. 1985; Sousa et al. 1993). The palm subdomain responsible for nucleotide polymerization is the only conserved protein domain among all right hand polymerases. It folds into a RNA recognition motif (RRM). In contrast to eight conserved structure motifs, typical for viral RNA-dependent polymerases, all right-hand polymerases share only four collinearly succeeding conserved sequence motifs (A, B, C, and D) (Lang et al. 2013).

2. INTRODUCTION TO USED METHODS

All bioinformatic methods are very sensitive to production of various artifacts. Therefore it is very important to use these methods in proper way and always confront the obtained results with other available data. In this chapter I would like to discuss methods which were used in this work and explain why they were used.

2.1 Selection of samples involved in evolutionary studies

Selection of suitable samples is the crucial step in all evolutionary studies. Incomplete, biased, or improper sampling leads to misleading results (Plazzi et al. 2010). Therefore, it is very important to pay great attention to samples selection and to include all suitable samples into the study.

In my work, I always used various search approaches to screen for proteins of interest. If they were searched on the base of structural similarity, DaliServer was used to search the PBD database of protein structures (Holm & Rosenström 2010). If they were searched on the base of sequence similarity, simple BLAST (Altschul et al. 1990) algorithm was used to find near homologs, while PSI-BLAST (Altschul et al. 1997), HHpred (Söding et al. 2005) and HHblits (Remmert et al. 2012) were used in search for distantly homologous proteins.

Involvement of too many samples from one taxon into the study may also lead to biased results. Therefore I always used simple logical rules for limitations of representatives involved. These rules are detail described in individual publications.

2.2 Protein structure dependent sequence alignment

Evolutionary stability of protein structure may be used in aligning of extremely evolutionary diversified proteins sharing sequence similarity lower than 40%. These are very difficult to align using sequence information only (Holm & Sander 1996). Numerous algorithms using protein tertiary structure to align their sequence were developed for example CE (Shindyalov & Bourne 1998), DaliLite (Holm & Rosenström 2010), MUSTANG (Konagurthu et al. 2006), MAMMOTH (Ortiz et al. 2002), TopMatch (Sippl & Wiederstein 2012), UCSF Chimera MachMaker (Meng et al. 2006), PDB protein comparison tool (Prlic et al. 2010) etc. Unfortunately, vast majority of structure base sequence aligning

programs does not produce multiple alignments but only pair alignments. The algorithms producing multiple alignments are usually quite demanding on computational time (Notredame 2007).

Therefore we decided to use T-Coffee Expresso (Armougom et al. 2006). This program can be run either locally or it offers user friendly web interface. If Expresso is run on-line, all calculations are done on distant server. As output, the user will receive all results as well as log file reporting about all calculations during aligning processes, which can be used for future aligning process optimization. In my work, most calculations were run under default conditions. Structural information was used whenever it was available.

2.3 Manual quantification of protein structures

There are also other ways how to used evolutionary information encoded in protein tertiary structure apart using of structure based sequence alignment. One of them is selection of "morphological" markers in protein structure, which can be encoded in character matrix. Such matrix can be used in further phylogenetic studies (Aravind et al. 2002; Scheeff & Bourne 2005).

According to my knowledge, there is no freely available software which can do this morphological characterization automatically. Therefore all quantifications have to be done manually, which brings a risk of artifact introduction. This can be overcome by careful selection of characters which are quantified. I always tend to select characters which were used previously in literature for protein structure description. Moreover, comparison of phylogenetic trees calculated only either on the base of protein sequence or on the base of protein structure "morphological" description can show whether the quantified characters were properly selected. If yes, "morphological" description deepens the preciseness of resulting phylogenetic tree. If no, it brings only the bias into the analysis.

2.4 MrBayes and its advantages in reconstruction of distant phylogenies

Reconstruction of distant evolutionary relationships is often very difficult task even when the sequences are very well aligned. With increasing evolutionary distance, the number of informative sites in alignment is decreasing, while the number of saturated positions is increasing (Ho et al. 2005). Genetic saturation

poses an extreme problem for distance-based phylogenetic methods as it leads to underestimation of genetic distance (Van de Peer et al. 2002). Despite, distance-based phylogenetic methods were recently used to reconstruct evolution of viral proteins (Mönttinen et al. 2014; Ravantti et al. 2013).

Advanced phylogenetic methods such as maximum likelihood or Bayesian Framework are more suitable for reconstruction of evolutionary relationships of distantly related sequences (Douady et al. 2003). In most of our studies I used MrBayes program as it is the best currently available program for reconstruction of distant evolutionary relationships. Morover it is less prone to attract long branches using proper model and appropriate taxon sampling (Glenner et al. 2004; Huelsenbeck & Ronquist 2001).

3. DISCUSSION

3.1 Evolution of TBEV genes

3.1.1 Evolution of TBEV strains isolated from human patients

In our work described in publication called “Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients.” (Formanová et al. 2015) we sequenced a set of five European TBEV strains which were isolated from TBEV infected patients in 1953. Several mutations specific for patient isolated TBEV strains were pointed out but their precise role has to be elucidated in future.

One of these mutations was I3203S/T. It was detected in three of these five TBEV strains plus it known from the other human TBEV isolate, Lubljana_I (GenBank Access. No. JQ654701.1)(Fajs et al. 2012). This mutation may have a role in increased pathogenicity of these TBEV strains. Phylogenetic analysis showed that TBEV strains bearing I3203S/T mutation do not form a monophyletic clade but that they are phylogenetically mixed with tick-isolated TBEV strains. It shows that this mutation is repeatedly selected in human TBEV isolates, which may indicate its importance for TBEV pathogenicity.

I3203 is located on the surface of NS5 polymerase subdomain far from catalytic site. It may indicate that this mutation is important for interaction with a protein which somehow interferes with TBEV replication. Flaviviral NS5 protein interacts with numerous partners of viral and host origin which modulates virus infection. There are several ways how the NS5 interacting partners can interfere with flavivirus replication: i) They may modulate function of NS5 protein such as flaviviral NS3 protein (Kapoor et al. 1995; Tay et al. 2014; Yon et al. 2005), eIFIII protein (Tay et al. 2014), Hdj2 protein (Wang et al. 2011) etc. ii) They may modulate interaction between NS5 protein and flaviviral genomic RNA (García-Montalvo et al. 2004). iii) They may modulate NS5 localization (Hannemann et al. 2013; Tay et al. 2013). iv) They may modulate host antivirus response (Ashour et al. 2009; Hannemann et al. 2013; Khunchai et al. 2012). The way, how I3203S/T mutation influences TBEV replication, still have to be elucidated.

3.1.2 TuORF

Apart from the major proteins, many flaviviruses produce minor proteins and peptides. NS1' produced by Japanese encephalitis virus (JEV) (Blitvich et al. 1999; Melian et al. 2010) and WNV WARF4 (Faggioni et al. 2012) are well examples. Each minor protein is usually specific only for a narrow group of closely related flaviviruses and they are important for flavivirus propagation and host-flavivirus interaction (Melian et al. 2010).

Presence of a short upstream open reading frame (uORF) in 5' untranslated region (UTR) of some TBEV strains is well known (Chausov et al. 2010). Nevertheless, it was not determined whether this uORF codes for a peptide. In our work called "Expression of a second open reading frame present in the genome of tick-borne encephalitis virus strain Neudoerfl is not detectable in infected cells." (Černý et al.) we showed that uORF does not code for a peptide.

Neither immunofluorescence nor immunoblotting using anti-TuORF peptide antibodies were able to detect any expression of TuORF peptide. Moreover, this result was supported by evolutionary analyses, showing that TuORF sequence is under positive selection pressure, which shows, that there is no selection pressure leading to conservancy of any specific amino acid sequence.

The role of TBEV uORF (TuORF) remains elusive. It is possible that it somehow regulates expression of main TBEV open reading frame. Translation regulation by uORF is a well know and intensively studied process. In most cases uORF down regulates gene expression (Firth & Brierley 2012). The rate of down regulation depends on sequence context of uORF initiation codon, uORF length, and distance between uORF and major ORF (Ryabova et al. 2006). In the case of TBEV uORF, down regulation of main open reading frame would not be high. AUG codon initiating TuORF peptide expression is in suboptimal sequence context (acgTgc**AUG**C) which is far from optimal Kosak sequence (gccRccAUGG) (Kozak 1984; Kozak 1986). Also the length of uORF is rather short and distance between uORF and the major TBEV ORF is sufficient for possible translation reinitiation. But the precise role of TuORF has to be evaluated yet.

3.2 Overall perspective on evolution of viral genes

3.2.1 Evolution of viral and cellular polymerases

In articles “Evolution of tertiary structure of viral RNA-dependent polymerases” (Černý et al. 2014) and “A deep phylogeny of viral and cellular right-hand polymerases” (Černý et al. 2015) right-hand polymerases were used as a marker gene to study evolution of RNA viruses and virus-cell evolutionary relationships, respectively. We showed that polymerases of RNA viruses and reverse transcriptases of RNA viruses replicating via DNA intermediate form two sisterly evolutionary groups. Polymerases of +ssRNA viruses and dsRNA viruses are not phylogenetically separated which indicates that viruses may theoretically switch from +ssRNA to dsRNA genomes and *vice versa*. On the other hand, viral polymerases are clearly evolutionary separated from other cellular and DNA viral polymerases. It may indicate that RNA viruses pose and ancient life group which originated from entities parasitizing on RNA life forms during RNA world.

Suitability of right-hand polymerases to fulfill the role of maker gen in reconstruction of distant virus evolution was challenged recently (Bamford et al. 2005; Mönttinen et al. 2014; Ravantti et al. 2013). It was proposed that polymerases spread among cellular organisms and viruses via horizontal gene transfer. One of the most important arguments standing behind this statement is that distribution of viral RNA dependent polymerases and their evolutionary relationships do not follow Baltimore classification of viruses (Mönttinen et al. 2014; Ravantti et al. 2013). Therefore jelly-roll capsid protein was suggested as better evolutionary marker (Poranen & Bamford 2012).

The discrepancies between pattern of right-hand polymerases evolutionary history and Baltimore classification can be easily explained. Baltimore classification is an artificial classification (Baltimore 1971). Nature of virus genome does not have to follow evolution of viruses. Polymerases are very flexible enzymes which can work on various templates. RNA polymerases can easily replicate both ssRNA and dsRNA genomes without any important rearrangements (Frick et al. 2007; Steimer & Klostermeier 2012).

On the other hand, jelly-roll capsid protein is typical for picorna-like viruses (+ssRNA genome), *Microviridae*, *Parvoviridae* (both ssDNA), *Papylomaviridae*,

Polyomaviridae (both dsDNA), etc. (Poranen & Bamford 2012). In my opinion, jelly-roll capsid protein is an inappropriate candidate for a virus phylogenetic marker since viruses sharing a jelly-roll capsid protein are only distantly related and jelly-roll capsid protein is missing among many virus families closely related to these which code it. Polymerases are present in all groups of non-satellite RNA viruses and RNA viruses replicating via DNA intermediate (Baltimore 1971). Moreover, polymerases follow a lot of small sameness typical for related viruses. Well examples are cyclically permuted virus polymerases. They are present *Birnaviridae*, which are viruses with a segmented genome formed by dsRNA as well as in *Permutotetraviridae* which are viruses with non-segmented genome formed by ssRNA of positive polarity. Despite these two families seems to be unrelated on the first look, they share many similarities when explored closer. For example, genomes of viruses within both families are primed by a VPg protein and both virus families code for 2A-like proteases (Gorbalenya et al. 2002).

Facts described above show that polymerases may be very suitable marker for virus evolution. Further studies on evolutionary history of other viral proteins as well as search for possible ancient recombination between different virus classes and other marks of horizontal gene transfer may shed more light on search for marker gene suitable for virus evolution.

3.2.2 Evolutionary history of flaviviral genes

As described above, evolution of viral genomes is a complicated and yet not fully understood process. In our work called “Evolutionary history of flaviviral genes.” we tried to describe evolution of individual major flaviviral genes” (Černý et al.).

The results of the analysis shows that proteins C, M, NS1, NS2A, NS2B, NS4A, and NS4B are true flaviviral ORFans as they have no homologues in any other viral or cellular genes. Protein NS3 share the common evolutionary history within family *Flaviviridae*.

Protein E, member of Class II Fusion Proteins family, is typical for Flavivirus genus only. It does not have any homologue within Flaviviridae family, but it is related to togaviral envelope protein E1 and distantly also to proteins EFF1 from worm *Caenorhabditis elegans* and BRAFL from lancelet *Branchiostoma*

floridae. As homologues of protein E can be found in cellular organsms as well as in viruses, it is not clear if the E protein is cellular or viral origin. Nevertheless, extremely rare occurrence of protein E homologues in cellular life forms indicates that it has viral origin and it was adopted by some cellular organisms via horizontal gene transfer.

Methyltransferase domain of NS5 protein (NS5Met) does not have any other homologue in *Flaviviridae* family apart viruses within Flavivirus genus. It is a member of Ftsj-like methyltransferase protein family which includes viral as well as cellular methyltransferases. The most closely related proteins to flaviviral NS5Met are bacterial 23S rRNA methyltransferases. It indicates that flavivirus NS5Met was most probably recently reached from a cellular organisms. As the closest cellular homologues of flaviviral NS5Met are bacteria, it remains elusive how it was reached to flaviviral genome, but we can speculate that it happened during a co-infection of one host.

Similar results were obtained in work of Koonin and Dolja (Koonin & Dolja 1993). These results show that viral genomes are patchy structures which are developing via frequent recombination events. Even within one virus family, evolutionary history of many genes can be very diverse (Koonin & Dolja 2012). Also recombination of viral genomic RNA with host RNA molecules may be quite often. It was proven that viruses are able to incorporate host RNA into their virions (Routh et al. 2012a; Routh et al. 2012b; Routh et al. 2012c). It gives viruses a possibility to acquire new genes not only by adaptive and *de novo* evolution and virus-virus recombination but also by virus-host recombination.

4. CONCLUSIONS AND FUTURE PERSPECTIVES

During my PhD study I focused on various aspects of virus evolution such as TBEV evolution, genus flavivirus evolution, and virus-cell evolutionary history. The most important findings done during my research are as follow:

1) Genomes of five patient isolates of TBEV were sequenced. Novel mutation (I3203S/T) in NS5 polymerase subdomain of human TBEV isolates was discovered. It was proposed that it may play an important role in TBEV pathogenicity.
2) TBEV upstream open reading frame was characterized. It was showed that it does not code for any peptide.
3) Evolutionary history of viral and cellular polymerases was described. It was showed that polymerases may serve as suitable markers for reconstruction of RNA virus evolutionary history and virus-cell evolutionary relationships. Using polymerases as a marker gene we showed that RNA viruses are ancient life forms which originated in RNA world.
4) Evolutionary history of flaviviral genes was described. It was shown that flaviviral genome is patchy structure formed by multiple recombination events. Flavivirus specific proteins (C, M, NS1, NS2A, NS2B, NS4A, and NS4B), proteins of viral origin (NS3 and NS5Pol), and proteins of cellular origin (E and NS5Met) are present in Flavivirus genome.

These results show that there still remain many unsolved problems in flavivirus evolution. In near future I would like to focus mostly on:

1) Collection and sequencing of next patient isolated TBEV strains and their comparison with field isolated TBEV strains. It will help us in better characterization of loci on TBEV genome which are important for TBEV virulence.
2) Construction of TBEV strain with and without TuORF and their virological characterization with the special concern on virus replication measures, virus infectivity, neuroinvasiveness etc. These experiments will tell us more about the role of TuORF in TBEV life cycle.
3) Study of viral polymerases as markers of virus evolution. This will help us in better understanding of evolutionary relationships among RNA viruses. Moreover, high quality polymerase alignments produced

during this work will be used for *in silico* prediction of polymerases structures and screen for possible anti-viral compounds.

4) Study of RNA virus genome plasticity on more RNA viruses with non-fragmented genome. It will help us in better understanding of processes standing behind virus genome evolution.

I hope that this work showed importance of virus molecular evolution studies in better understanding of natural processes standing behind (re)emergence of flaviviruses which may pose serious medical and veterinary threats. It is sure that importance of virus evolution studies will grow and understanding of these processes together with careful continuous surveillance of possible viral threats on health concept will give us powerful tool in prediction and control of virus epidemics.

5. LITERATURE

Ahlquist, P. 2006. Parallels among positive-strand RNA viruses, reverse-transcribing viruses and double-stranded RNA viruses. Nat Rev Microbiol 4.371-82.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers & D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215.403-10.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller & D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25.3389-402.

Apte-Sengupta, S., D. Sirohi & R. J. Kuhn. 2014. Coupling of replication and assembly in flaviviruses. Curr Opin Virol 9.134-42.

Aravind, L., V. Anantharaman & E. V. Koonin. 2002. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins 48.1-14.

Armougom, F., S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas & C. Notredame. 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34.W604-8.

Ashour, J., M. Laurent-Rolle, P. Y. Shi & A. García-Sastre. 2009. NS5 of dengue virus mediates STAT2 binding and degradation. J Virol 83.5408-18.

Baillie, G. J., S. O. Kolokotronis, E. Waltari, J. G. Maffei, L. D. Kramer & S. L. Perkins. 2008. Phylogenetic and evolutionary analyses of St. Louis encephalitis virus genomes. Mol Phylogenet Evol 47.717-28.

Baltimore, D. 1971. Expression of animal virus genomes. Bacteriol Rev 35.235-41.

Bamford, D. H. 2003. Do viruses form lineages across different domains of life? Res Microbiol 154.231-6.

Bamford, D. H., J. M. Grimes & D. I. Stuart. 2005. What does structure tell us about virus evolution? Curr Opin Struct Biol 15.655-63.

Beck, A., H. Guzman, L. Li, B. Ellis, R. B. Tesh & A. D. Barrett. 2013. Phylogeographic reconstruction of African yellow fever virus isolates indicates recent simultaneous dispersal into east and west Africa. PLoS Negl Trop Dis 7.e1910.

Bera, A. K., R. J. Kuhn & J. L. Smith. 2007. Functional characterization of cis and trans activity of the Flavivirus NS2B-NS3 protease. J Biol Chem 282.12883-92.

Blitvich, B. J., D. Scanlon, B. J. Shiell, J. S. Mackenzie & R. A. Hall. 1999. Identification and analysis of truncated and elongated species of the flavivirus NS1 protein. Virus Res 60.67-79.

Brault, A. C., C. Y. Huang, S. A. Langevin, R. M. Kinney, R. A. Bowen, W. N. Ramey, N. A. Panella, E. C. Holmes, A. M. Powers & B. R. Miller. 2007. A single positively selected West Nile viral mutation confers increased virogenesis in American crows. Nat Genet 39.1162-6.

Bressanelli, S., K. Stiasny, S. L. Allison, E. A. Stura, S. Duquerroy, J. Lescar, F. X. Heinz & F. A. Rey. 2004. Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J 23.728-38.

Bruenn, J. A. 1991. Relationships among the positive strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases. Nucleic Acids Res 19.217-26.

—. 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. Nucleic Acids Res 31.1821-9.

Bryant, J. E., E. C. Holmes & A. D. Barrett. 2007. Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas. PLoS Pathog 3.e75.

Bäck, A. T. & A. Lundkvist. 2013. Dengue viruses - an overview. Infect Ecol Epidemiol 3.

Cabanillas, L., M. Arribas & E. Lázaro. 2013. Evolution at increased error rate leads to the coexistence of multiple adaptive pathways in an RNA virus. BMC Evol Biol 13.11.

Carrillo, E. C., E. R. Rojas, L. Cavallaro, M. Schiappacassi & R. Campos. 1989. Modification of foot-and-mouth disease virus after serial passages in the presence of antiviral polyclonal sera. Virology 171.599-601.

Casati, S., L. Gern & J. C. Piffaretti. 2006. Diversity of the population of Tick-borne encephalitis virus infecting Ixodes ricinus ticks in an endemic area of central Switzerland (Canton Bern). J Gen Virol 87.2235-41.

Chambers, T. J., R. C. Weir, A. Grakoui, D. W. McCourt, J. F. Bazan, R. J. Fletterick & C. M. Rice. 1990. Evidence that the N-terminal domain of nonstructural protein NS3 from yellow fever virus is a serine protease responsible for site-specific cleavages in the viral polyprotein. Proc Natl Acad Sci U S A 87.8898-902.

Charrel, R. N., A. M. Zaki, H. Attoui, M. Fakeeh, F. Billoir, A. I. Yousef, R. de Chesse, P. De Micco, E. A. Gould & X. de Lamballerie. 2001. Complete coding sequence of the Alkhurma virus, a tick-borne flavivirus causing severe hemorrhagic fever in humans in Saudi Arabia. Biochem Biophys Res Commun 287.455-61.

Chausov, E. V., V. A. Ternovoi, E. V. Protopopova, J. V. Kononova, S. N. Konovalova, N. L. Pershikova, V. N. Romanenko, N. V. Ivanova, N. P. Bolshakova, N. S. Moskvitina & V. B. Loktev. 2010. Variability of the tick-borne encephalitis virus genome in the 5' noncoding region derived

from ticks Ixodes persulcatus and Ixodes pavlovskyi in Western Siberia. Vector Borne Zoonotic Dis 10.365-75.

Chin-inmanu, K., A. Suttitheptumrong, D. Sangsrakru, S. Tangphatsornruang, S. Tragoonrung, P. Malasit, S. Tungpradabkul & P. Suriyaphol. 2012. Feasibility of using 454 pyrosequencing for studying quasispecies of the whole dengue viral genome. BMC Genomics 13 Suppl 7.S7.

Cook, S., S. N. Bennett, E. C. Holmes, R. De Chesse, G. Moureau & X. de Lamballerie. 2006. Isolation of a new strain of the flavivirus cell fusing agent virus in a natural mosquito population from Puerto Rico. J Gen Virol 87.735-48.

Coutard, B. & B. Canard. 2010. The VIZIER project: overview; expectations; and achievements. Antiviral Res 87.85-94.

Coutard, B., A. E. Gorbalenya, E. J. Snijder, A. M. Leontovich, A. Poupon, X. De Lamballerie, R. Charrel, E. A. Gould, S. Gunther, H. Norder, B. Klempa, H. Bourhy, J. Rohayem, E. L'Hermite, P. Nordlund, D. I. Stuart, R. J. Owens, J. M. Grimes, P. A. Tucker, M. Bolognesi, A. Mattevi, M. Coll, T. A. Jones, J. Aqvist, T. Unge, R. Hilgenfeld, G. Bricogne, J. Neyts, P. La Colla, G. Puerstinger, J. P. Gonzalez, E. Leroy, C. Cambillau, J. L. Romette & B. Canard. 2008. The VIZIER project: preparedness against pathogenic RNA viruses. Antiviral Res 78.37-46.

Crochu, S., S. Cook, H. Attoui, R. N. Charrel, R. De Chesse, M. Belhouchet, J. J. Lemasson, P. de Micco & X. de Lamballerie. 2004. Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of Aedes spp. mosquitoes. J Gen Virol 85.1971-80.

Crowe, M. L., X. Q. Wang & J. A. Rothnagel. 2006. Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. BMC Genomics 7.16.

Davidson, A. D. 2009. Chapter 2. New insights into flavivirus nonstructural protein 5. Adv Virus Res 74.41-101.

Deardorff, E. R., R. A. Nofchissey, J. A. Cook, A. G. Hope, A. Tsvetkova, S. L. Talbot & G. D. Ebel. 2013. Powassan virus in mammals, Alaska and New Mexico, U.S.A., and Russia, 2004-2007. Emerg Infect Dis 19.2012-6.

Diaz, L. A., F. S. Flores, A. Quaglia & M. S. Contigiani. 2012. Intertwined arbovirus transmission activity: reassessing the transmission cycle paradigm. Front Physiol 3.493.

Dietrich, M., E. Gómez-Díaz & K. D. McCoy. 2011. Worldwide distribution and diversity of seabird ticks: implications for the ecology and epidemiology of tick-borne pathogens. Vector Borne Zoonotic Dis 11.453-70.

Dolja, V V & J C. Carrington. 1992. Evolution of positive-strand RNA viruses. In *Seminars in Virology*.

Dolja, V. V. & E. V. Koonin. 2011. Common origins and host-dependent diversity of plant and animal viromes. Curr Opin Virol 1.322-31.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle & E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol 20.248-54.

Durbin, A. P., S. V. Mayer, S. L. Rossi, I. Y. Amaya-Larios, J. Ramos-Castaneda, E. Eong Ooi, M. Jane Cardosa, J. L. Munoz-Jordan, R. B. Tesh, W. B. Messer, S. C. Weaver & N. Vasilakis. 2013a. Emergence potential of sylvatic dengue virus type 4 in the urban transmission cycle is restrained by vaccination and homotypic immunity. Virology 439.34-41.

Durbin, A. P., P. F. Wright, A. Cox, W. Kagucia, D. Elwood, S. Henderson, K. Wanionek, J. Speicher, S. S. Whitehead & A. G. Pletnev. 2013b. The live attenuated chimeric vaccine rWN/DEN4Δ30 is well-tolerated and immunogenic in healthy flavivirus-naïve adult volunteers. Vaccine 31.5772-7.

Egloff, M. P., D. Benarroch, B. Selisko, J. L. Romette & B. Canard. 2002. An RNA cap (nucleoside-2'-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. Embo J 21.2757-68.

Eickbush, Thomas H. 1994. Origin and evolutionary relationships of retroelements. The evolutionary biology of viruses, ed. by S.S. Morse, 121-57: Raven Press, 1185 Avenue of the Americas, New York, New York 10036-2806, USA.

Faggioni, G., A. Pomponi, R. De Santis, L. Masuelli, A. Ciammaruconi, F. Monaco, A. Di Gennaro, L. Marzocchella, V. Sambri, R. Lelli, G. Rezza, R. Bei & F. Lista. 2012. West Nile alternative open reading frame (N-NS4B/WARF4) is produced in infected West Nile Virus (WNV) cells and induces humoral response in WNV infected individuals. Virol J 9.283.

Fajs, L., E. Durmiši, N. Knap, F. Strle & T. Avšič-Županc. 2012. Phylogeographic characterization of tick-borne encephalitis virus from patients, rodents and ticks in Slovenia. PLoS One 7.e48420.

Ferrer-Orta, C., A. Arias, C. Escarmís & N. Verdaguer. 2006. A comparison of viral RNA-dependent RNA polymerases. Curr Opin Struct Biol 16.27-34.

Filée, J., P. Forterre, T. Sen-Lin & J. Laurent. 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 54.763-73.

Firth, A. E. & I. Brierley. 2012. Non-canonical translation in RNA viruses. J Gen Virol 93.1385-409.

Formanová, P., J. Černý, B. Bolfíková, J. J. Valdés, I. Kozlova, Y. Dzhioev & D. Růžek. 2015. Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients. Ticks Tick Borne Dis 6.38-46.

Forterre, P. 2002. The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5.525-32.

—. 2005. The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie 87.793-803.
—. 2006a. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117.5-16.
—. 2006b. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. Proc Natl Acad Sci U S A 103.3669-74.
Frick, D. N., S. Banik & R. S. Rypma. 2007. Role of divalent metal cations in ATP hydrolysis catalyzed by the hepatitis C virus NS3 helicase: magnesium provides a bridge for ATP to fuel unwinding. J Mol Biol 365.1017-32.
García-Montalvo, B. M., F. Medina & R. M. del Angel. 2004. La protein binds to NS5 and NS3 and to the 5' and 3' ends of Dengue 4 virus RNA. Virus Res 102.141-50.
Gardner, C. L. & K. D. Ryman. 2010. Yellow fever: a reemerging threat. Clin Lab Med 30.237-60.
Gaunt, M. W., A. A. Sall, X. de Lamballerie, A. K. Falconar, T. I. Dzhivanian & E. A. Gould. 2001. Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. J Gen Virol 82.1867-76.
Geuking, M. B., J. Weber, M. Dewannieux, E. Gorelik, T. Heidmann, H. Hengartner, R. M. Zinkernagel & L. Hangartner. 2009. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science 323.393-6.
Gibrat, J. F., M. Mariadassou, P. Boudinot & B. Delmas. 2013. Analyses of the radiation of birnaviruses from diverse host phyla and of their evolutionary affinities with other double-stranded RNA and positive strand RNA viruses using robust structure-based multiple sequence alignments and advanced phylogenetic methods. BMC Evol Biol 13.154.
Glebe, D. & C. M. Bremer. 2013. The molecular virology of hepatitis B virus. Semin Liver Dis 33.103-12.
Glenner, H., A. J. Hansen, M. V. Sørensen, F. Ronquist, J. P. Huelsenbeck & E. Willerslev. 2004. Bayesian inference of the metazoan phylogeny; a combined molecular and morphological approach. Curr Biol 14.1644-9.
Gohara, D. W., S. Crotty, J. J. Arnold, J. D. Yoder, R. Andino & C. E. Cameron. 2000. Poliovirus RNA-dependent RNA polymerase (3Dpol): structural, biochemical, and biological analysis of conserved structural motifs A and B. J Biol Chem 275.25523-32.
Goldbach, R., J. Wellink, J. Verver, A. van Kammen, D. Kasteel & J. van Lent. 1994. Adaptation of positive-strand RNA viruses to plants. Arch Virol Suppl 9.87-97.
Gorbalenya, A. E., F. M. Pringle, J. L. Zeddam, B. T. Luke, C. E. Cameron, J. Kalmakoff, T. N. Hanzlik, K. H. Gordon & V. K. Ward. 2002. The palm

subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. J Mol Biol 324.47-62.

Gould, E. A., X. de Lamballerie, P. M. Zanotto & E. C. Holmes. 2001. Evolution, epidemiology, and dispersal of flaviviruses revealed by molecular phylogenies. Adv Virus Res 57.71-103.

—. 2003. Origins, evolution, and vector/host coadaptations within the genus Flavivirus. Adv Virus Res 59.277-314.

Gould, E. A., S. R. Moss & S. L. Turner. 2004. Evolution and dispersal of encephalitic flaviviruses. Arch Virol Suppl.65-84.

Gould, E. A. & T. Solomon. 2008. Pathogenic flaviviruses. Lancet 371.500-9.

Graham, R. L. & R. S. Baric. 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol 84.3134-46.

Gritsun, T. S., P. A. Nuttall & E. A. Gould. 2003. Tick-borne flaviviruses. Adv Virus Res 61.317-71.

Hannemann, H., P. Y. Sung, H. C. Chiu, A. Yousuf, J. Bird, S. P. Lim & A. D. Davidson. 2013. Serotype-specific differences in dengue virus non-structural protein 5 nuclear localization. J Biol Chem 288.22621-35.

Hansen, J. L., A. M. Long & S. C. Schultz. 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. Structure 5.1109-22.

Harris, E., K. L. Holden, D. Edgil, C. Polacek & K. Clyde. 2006. Molecular biology of flaviviruses. Novartis Found Symp 277.23-39; discussion 40, 71-3, 251-3.

Hendrix, R. W., J. G. Lawrence, G. F. Hatfull & S. Casjens. 2000. The origins and ongoing evolution of viruses. Trends Microbiol 8.504-8.

Ho, S. Y., M. J. Phillips, A. Cooper & A. J. Drummond. 2005. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol 22.1561-8.

Holm, L. & P. Rosenström. 2010. Dali server: conservation mapping in 3D. Nucleic Acids Res 38.W545-9.

Holm, L. & C. Sander. 1996. Mapping the protein universe. Science 273.595-603.

Huelsenbeck, J. P. & F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17.754-5.

Huiskonen, J. T. & S. J. Butcher. 2007. Membrane-containing viruses with icosahedrally symmetric capsids. Curr Opin Struct Biol 17.229-36.

Illergård, K., D. H. Ardell & A. Elofsson. 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77.499-508.

Iyer, L. M., E. V. Koonin, D. D. Leipe & L. Aravind. 2005. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain

proteins: structural insights and new members. Nucleic Acids Res 33.3875-96.

Junjhon, J., T. J. Edwards, U. Utaipat, V. D. Bowman, H. A. Holdaway, W. Zhang, P. Keelapang, C. Puttikhunt, R. Perera, P. R. Chipman, W. Kasinrerk, P. Malasit, R. J. Kuhn & N. Sittisombut. 2010. Influence of pr-M cleavage on the heterogeneity of extracellular dengue virus particles. J Virol 84.8353-8.

Junjhon, J., J. G. Pennington, T. J. Edwards, R. Perera, J. Lanman & R. J. Kuhn. 2014. Ultrastructural characterization and three-dimensional architecture of replication sites in dengue virus-infected mosquito cells. J Virol 88.4687-97.

Kapoor, M., L. Zhang, M. Ramachandra, J. Kusukawa, K. E. Ebner & R. Padmanabhan. 1995. Association between NS3 and NS5 proteins of dengue virus type 2 in the putative RNA replicase is linked to differential phosphorylation of NS5. J Biol Chem 270.19100-6.

Khunchai, S., M. Junking, A. Suttitheptumrong, U. Yasamut, N. Sawasdee, J. Netsawang, A. Morchang, P. Chaowalit, S. Noisakran, P. T. Yenchitsomanus & T. Limjindaporn. 2012. Interaction of dengue virus nonstructural protein 5 with Daxx modulates RANTES production. Biochem Biophys Res Commun 423.398-403.

Klenerman, P., H. Hengartner & R. M. Zinkernagel. 1997. A non-retroviral RNA virus persists in DNA form. Nature 390.298-301.

Knox, J., R. U. Cowan, J. S. Doyle, M. K. Ligtermoet, J. S. Archer, J. N. Burrow, S. Y. Tong, B. J. Currie, J. S. Mackenzie, D. W. Smith, M. Catton, R. J. Moran, C. A. Aboltins & J. S. Richards. 2012. Murray Valley encephalitis: a review of clinical features, diagnosis and treatment. Med J Aust 196.322-6.

Kohlstaedt, L. A., J. Wang, J. M. Friedman, P. A. Rice & T. A. Steitz. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science 256.1783-90.

Konagurthu, A. S., J. C. Whisstock, P. J. Stuckey & A. M. Lesk. 2006. MUSTANG: a multiple structural alignment algorithm. Proteins 64.559-74.

Koonin, E. V. 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J Gen Virol 72 (Pt 9).2197-206.

—. 2006. Temporal order of evolution of DNA replication systems inferred by comparison of cellular and viral DNA polymerases. Biol Direct 1.39.

Koonin, E. V. & V. V. Dolja. 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit Rev Biochem Mol Biol 28.375-430.

—. 2012. Expanding networks of RNA virus evolution. BMC Biol 10.54.

Koonin, E. V., T. G. Senkevich & V. V. Dolja. 2006. The ancient Virus World and evolution of cells. Biol Direct 1.29.

Korneeva, V. S. & C. E. Cameron. 2007. Structure-function relationships of the viral RNA-dependent RNA polymerase: fidelity, replication speed, and initiation mechanism determined by a residue in the ribose-binding pocket. J Biol Chem 282.16135-45.

Kozak, M. 1984. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature 308.241-6.

—. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44.283-92.

—. 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 299.1-34.

Krupovič, M. & D. H. Bamford. 2010. Order to the viral universe. J Virol 84.12476-9.

Lang, D. M., A. T. Zemla & C. L. Zhou. 2013. Highly similar structural frames link the template tunnel and NTP entry tunnel to the exterior surface in RNA-dependent RNA polymerases. Nucleic Acids Res 41.1464-82.

Le Flohic, G., V. Porphyre, P. Barbazan & J. P. Gonzalez. 2013. Review of climate, landscape, and viral genetics as drivers of the Japanese encephalitis virus ecology. PLoS Negl Trop Dis 7.e2208.

Leipe, D. D., L. Aravind & E. V. Koonin. 1999. Did DNA replication evolve twice independently? Nucleic Acids Res 27.3389-401.

Li, C. Y., Y. Zhang, Z. Wang, C. Cao, P. W. Zhang, S. J. Lu, X. M. Li, Q. Yu, X. Zheng, Q. Du, G. R. Uhl, Q. R. Liu & L. Wei. 2010. A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol 6.e1000734.

Lin, H. H., H. C. Lee, X. F. Li, M. J. Tsai, H. J. Hsiao, J. G. Peng, S. C. Sue, C. F. Qin & S. C. Wu. 2014. Dengue type four viruses with E-Glu345Lys adaptive mutation from MRC-5 cells induce low viremia but elicit potent neutralizing antibodies in rhesus monkeys. PLoS One 9.e100130.

Maeda, N., H. Fan & Y. Yoshikai. 2008. Oncogenesis by retroviruses: old and new paradigms. Rev Med Virol 18.387-405.

Maori, E., E. Tanne & I. Sela. 2007. Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. Virology 362.342-9.

Martcheva, M. 2012. An evolutionary model of influenza A with drift and shift. J Biol Dyn 6.299-332.

Matsuoka, M. & J. Yasunaga. 2013. Human T-cell leukemia virus type 1: replication, proliferation and propagation by Tax and HTLV-1 bZIP factor. Curr Opin Virol 3.684-91.

Melian, E. B., E. Hinzman, T. Nagasaki, A. E. Firth, N. M. Wills, A. S. Nouwens, B. J. Blitvich, J. Leung, A. Funk, J. F. Atkins, R. Hall & A. A. Khromykh. 2010. NS1' of flaviviruses in the Japanese encephalitis virus serogroup is a

product of ribosomal frameshifting and plays a role in viral neuroinvasiveness. J Virol 84.1641-7.
Meng, E. C., E. F. Pettersen, G. S. Couch, C. C. Huang & T. E. Ferrin. 2006. Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics 7.339.
Messina, J. P., O. J. Brady, T. W. Scott, C. Zou, D. M. Pigott, K. A. Duda, S. Bhatt, L. Katzelnick, R. E. Howes, K. E. Battle, C. P. Simmons & S. I. Hay. 2014. Global spread of dengue virus types: mapping the 70 year history. Trends Microbiol 22.138-46.
Modis, Y., S. Ogata, D. Clements & S. C. Harrison. 2004. Structure of the dengue virus envelope protein after membrane fusion. Nature 427.313-9.
Monath, T. P. 2001. Yellow fever: an update. Lancet Infect Dis 1.11-20.
Muller, D. A. & P. R. Young. 2013. The flavivirus NS1 protein: molecular and structural biology, immunology, role in pathogenesis and application as a diagnostic biomarker. Antiviral Res 98.192-208.
Murray, K. O., E. Mertens & P. Despres. 2010. West Nile virus and its emergence in the United States of America. Vet Res 41.67.
Mönttinen, H. A., J. J. Ravantti, D. I. Stuart & M. M. Poranen. 2014. Automated structural comparisons clarify the phylogeny of the right-hand-shaped polymerases. Mol Biol Evol 31.2741-52.
Nayak, A., N. Pattabiraman, N. Fadra, R. Goldman, S. L. Kosakovsky Pond & R. Mazumder. 2014. Structure-function analysis of hepatitis C virus envelope glycoproteins E1 and E2. J Biomol Struct Dyn.1-13.
Ng, K. K., J. J. Arnold & C. E. Cameron. 2008. Structure-function relationships among RNA-dependent RNA polymerases. Curr Top Microbiol Immunol 320.137-56.
Notredame, C. 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3.e123.
Nowak, M. 1990. HIV mutation rate. Nature 347.522.
Ogata, N., H. J. Alter, R. H. Miller & R. H. Purcell. 1991. Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. Proc Natl Acad Sci U S A 88.3392-6.
Ollis, D. L., P. Brick, R. Hamlin, N. G. Xuong & T. A. Steitz. 1985. Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. Nature 313.762-6.
Ortiz, A. R., C. E. Strauss & O. Olmea. 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11.2606-21.
Pandey, L. K., J. K. Mohapatra, S. Subramaniam, A. Sanyal, V. Pande & B. Pattnaik. 2014. Evolution of serotype A foot-and-mouth disease virus capsid under neutralizing antibody pressure in vitro. Virus Res 181.72-6.

Paul, D. & R. Bartenschlager. 2013. Architecture and biogenesis of plus-strand RNA virus replication factories. World J Virol 2.32-48.

Perera-Lecoin, M., L. Meertens, X. Carnec & A. Amara. 2014. Flavivirus entry receptors: an update. Viruses 6.69-88.

Petersen, L. R., A. C. Brault & R. S. Nasci. 2013. West Nile virus: review of the literature. JAMA 310.308-15.

Pettersson, J. H. & O. Fiz-Palacios. 2014. Dating the origin of the genus Flavivirus in the light of Beringian biogeography. J Gen Virol 95.1969-82.

Pickett, B. E., R. Striker & E. J. Lefkowitz. 2011. Evidence for separation of HCV subtype 1a into two distinct clades. J Viral Hepat 18.608-18.

Plazzi, F., R. R. Ferrucci & M. Passamonti. 2010. Phylogenetic representativeness: a new method for evaluating taxon sampling in evolutionary studies. BMC Bioinformatics 11.209.

Poch, O., I. Sauvaget, M. Delarue & N. Tordo. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J 8.3867-74.

Pond, S. L., B. Murrell & A. F. Poon. 2012. Evolution of viral genomes: interplay between selection, recombination, and other forces. Methods Mol Biol 856.239-72.

Pong, W. L., Z. S. Huang, P. G. Teoh, C. C. Wang & H. N. Wu. 2011. RNA binding property and RNA chaperone activity of dengue virus core protein and other viral RNA-interacting proteins. FEBS Lett 585.2575-81.

Poranen, M. M. & D. H. Bamford. 2012. Assembly of large icosahedral double-stranded RNA viruses. Adv Exp Med Biol 726.379-402.

Prangishvili, D., K. Stedman & W. Zillig. 2001. Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol 9.39-43.

Prlic, A., S. Bliven, P. W. Rose, W. F. Bluhm, C. Bizon, A. Godzik & P. E. Bourne. 2010. Pre-calculated protein structure alignments at the RCSB PDB website. Bioinformatics 26.2983-5.

Randolph, S. E. & EDEN-TBD sub-project team. 2010. Human activities predominate in determining changing incidence of tick-borne encephalitis in Europe. Euro Surveill 15.24-31.

Ravantti, J., D. Bamford & D. I. Stuart. 2013. Automatic comparison and classification of protein structures. J Struct Biol 183.47-56.

Remmert, M., A. Biegert, A. Hauser & J. Söding. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9.173-5.

Rey, F. A., F. X. Heinz, C. Mandl, C. Kunz & S. C. Harrison. 1995. The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. Nature 375.291-8.

Roiz, D., A. Vazquez, M. P. Seco, A. Tenorio & A. Rizzoli. 2009. Detection of novel insect flavivirus sequences integrated in Aedes albopictus (Diptera: Culicidae) in Northern Italy. Virol J 6.93.

Rossmann, M. G. & J. E. Johnson. 1989. Icosahedral RNA virus structure. Annu Rev Biochem 58.533-73.

Routh, A., T. Domitrovic & J. E. Johnson. 2012a. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. Proc Natl Acad Sci U S A 109.1907-12.

—. 2012b. Packaging host RNAs in small RNA viruses: an inevitable consequence of an error-prone polymerase? Cell Cycle 11.3713-4.

Routh, A., P. Ordoukhanian & J. E. Johnson. 2012c. Nucleotide-resolution profiling of RNA recombination in the encapsidated genome of a eukaryotic RNA virus by next-generation sequencing. J Mol Biol 424.257-69.

Ryabova, L. A., M. M. Pooggin & T. Hohn. 2006. Translation reinitiation and leaky scanning in plant viruses. Virus Res 119.52-62.

Růzek, D., T. S. Gritsun, N. L. Forrester, E. A. Gould, J. Kopecký, M. Golovchenko, N. Rudenko & L. Grubhoffer. 2008. Mutations in the NS2B and NS3 genes affect mouse neuroinvasiveness of a Western European field strain of tick-borne encephalitis virus. Virology 374.249-55.

Růžek, D., V. V. Yakimenko, L. S. Karan & S. E. Tkachev. 2010. Omsk haemorrhagic fever. Lancet 376.2104-13.

Sabath, N., A. Wagner & D. Karlin. 2012. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol 29.3767-80.

Salgado, P. S., M. R. Koivunen, E. V. Makeyev, D. H. Bamford, D. I. Stuart & J. M. Grimes. 2006. The structure of an RNAi polymerase links RNA silencing and transcription. PLoS Biol 4.e434.

Scheeff, E. D. & P. E. Bourne. 2005. Structural evolution of the protein kinase-like superfamily. PLoS Comput Biol 1.e49.

Scheel, T. K., A. Galli, Y. P. Li, L. S. Mikkelsen, J. M. Gottwein & J. Bukh. 2013. Productive homologous and non-homologous recombination of hepatitis C virus in cell culture. PLoS Pathog 9.e1003228.

Shatskaya, G. S. & T. M. Dmitrieva. 2013. Structural organization of viral RNA-dependent RNA polymerases. Biochemistry (Mosc) 78.231-5.

Shindyalov, I. N. & P. E. Bourne. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11.739-47.

Shiu, S. Y., M. D. Ayres & E. A. Gould. 1991. Genomic sequence of the structural proteins of louping ill virus: comparative analysis with tick-borne encephalitis virus. Virology 180.411-5.

Sippl, M. J. & M. Wiederstein. 2012. Detection of spatial correlations in protein structures and molecular complexes. Structure 20.718-28.

Smith, D. B., N. McFadden, R. J. Blundell, A. Meredith & P. Simmonds. 2012. Diversity of murine norovirus in wild-rodent populations: species-specific associations suggest an ancient divergence. J Gen Virol 93.259-66.

Smith, L. M., A. R. McWhorter, G. R. Shellam & A. J. Redwood. 2013. The genome of murine cytomegalovirus is shaped by purifying selection and extensive recombination. Virology 435.258-68.

Sorek, R. 2007. The birth of new exons: mechanisms and evolutionary consequences. RNA 13.1603-8.

Sousa, R., Y. J. Chung, J. P. Rose & B. C. Wang. 1993. Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. Nature 364.593-9.

Stadler, K., S. L. Allison, J. Schalich & F. X. Heinz. 1997. Proteolytic activation of tick-borne encephalitis virus by furin. J Virol 71.8475-81.

Steimer, L. & D. Klostermeier. 2012. RNA helicases in infection and disease. RNA Biol 9.751-71.

Stiasny, K., S. Bressanelli, J. Lepault, F. A. Rey & F. X. Heinz. 2004. Characterization of a membrane-associated trimeric low-pH-induced Form of the class II viral fusion protein E from tick-borne encephalitis virus and its crystallization. J Virol 78.3178-83.

Söding, J., A. Biegert & A. N. Lupas. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33.W244-8.

Tajima, S., R. Nerome, Y. Nukui, F. Kato, T. Takasaki & I. Kurane. 2010. A single mutation in the Japanese encephalitis virus E protein (S123R) increases its growth rate in mouse neuroblastoma cells and its pathogenicity in mice. Virology 396.298-304.

Takeshita, D. & K. Tomita. 2010. Assembly of Q{beta} viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. Proc Natl Acad Sci U S A 107.15733-8.

Tan, B. H., J. Fu, R. J. Sugrue, E. H. Yap, Y. C. Chan & Y. H. Tan. 1996. Recombinant dengue type 1 virus NS5 protein expressed in Escherichia coli exhibits RNA-dependent RNA polymerase activity. Virology 216.317-25.

Tanne, E. & I. Sela. 2005. Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. Virology 332.614-22.

Tay, M. Y., J. E. Fraser, W. K. Chan, N. J. Moreland, A. P. Rathore, C. Wang, S. G. Vasudevan & D. A. Jans. 2013. Nuclear localization of dengue virus (DENV) 1-4 non-structural protein 5; protection against all 4 DENV serotypes by the inhibitor Ivermectin. Antiviral Res 99.301-6.

Tay, M. Y., W. G. Saw, Y. Zhao, K. W. Chan, D. Singh, Y. Chong, J. K. Forwood, E. E. Ooi, G. Grüber, J. Lescar, D. Luo & S. G. Vasudevan. 2014. The C-terminal 50 amino acid residues of Dengue NS3 protein are important for NS3-NS5 interaction and viral replication. J Biol Chem.

Unni, S. K., D. Růžek, C. Chhatbar, R. Mishra, M. K. Johri & S. K. Singh. 2011. Japanese encephalitis virus: from genome to infectome. Microbes Infect 13.312-21.

Utama, A., H. Shimizu, S. Morikawa, F. Hasebe, K. Morita, A. Igarashi, M. Hatsu, K. Takamizawa & T. Miyamura. 2000. Identification and characterization of the RNA helicase activity of Japanese encephalitis virus NS3 protein. FEBS Lett 465.74-8.

Van de Peer, Y., T. Frickey, J. Taylor & A. Meyer. 2002. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. Gene 295.205-11.

Venugopal, K., T. Gritsun, V. A. Lashkevich & E. A. Gould. 1994. Analysis of the structural protein gene sequence shows Kyasanur Forest disease virus as a distinct member in the tick-borne encephalitis virus serocomplex. J Gen Virol 75 (Pt 1).227-32.

Wang, H. J., X. F. Li, Q. Ye, S. H. Li, Y. Q. Deng, H. Zhao, Y. P. Xu, J. Ma, E. D. Qin & C. F. Qin. 2014. Recombinant chimeric Japanese encephalitis virus/tick-borne encephalitis virus is attenuated and protective in mice. Vaccine 32.949-56.

Wang, R. Y., Y. R. Huang, K. M. Chong, C. Y. Hung, Z. L. Ke & R. Y. Chang. 2011. DnaJ homolog Hdj2 facilitates Japanese encephalitis virus replication. Virol J 8.471.

Ward, C. W. 1993. Progress towards a higher taxonomy of viruses. Res Virol 144.419-53.

Welsch, S., S. Miller, I. Romero-Brey, A. Merz, C. K. Bleck, P. Walther, S. D. Fuller, C. Antony, J. Krijnse-Locker & R. Bartenschlager. 2009. Composition and three-dimensional architecture of the dengue virus replication and assembly sites. Cell Host Microbe 5.365-75.

Westaway, E. G., J. M. Mackenzie, M. T. Kenney, M. K. Jones & A. A. Khromykh. 1997. Ultrastructure of Kunjin virus-infected cells: colocalization of NS1 and NS3 with double-stranded RNA, and of NS2B with NS3, in virus-induced membrane structures. J Virol 71.6650-61.

Woese, C. R., O. Kandler & M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87.4576-9.

Yamaguchi, Y., Y. Nukui, S. Tajima, R. Nerome, F. Kato, H. Watanabe, T. Takasaki & I. Kurane. 2011. An amino acid substitution (V3I) in the Japanese encephalitis virus NS4A protein increases its virulence in mice, but not its growth rate in vitro. J Gen Virol 92.1601-6.

Yon, C., T. Teramoto, N. Mueller, J. Phelan, V. K. Ganesh, K. H. Murthy & R. Padmanabhan. 2005. Modulation of the nucleoside triphosphatase/RNA helicase and 5'-RNA triphosphatase activities of Dengue virus type 2 nonstructural protein 3 (NS3) by interaction with NS5, the RNA-dependent RNA polymerase. J Biol Chem 280.27412-9.

Yu, I. M., W. Zhang, H. A. Holdaway, L. Li, V. A. Kostyuchenko, P. R. Chipman, R. J. Kuhn, M. G. Rossmann & J. Chen. 2008. Structure of the immature dengue virus at low pH primes proteolytic maturation. Science 319.1834-7.

Yu, L., K. Takeda & L. Markoff. 2013. Protein-protein interactions among West Nile non-structural proteins and transmembrane complex formation in mammalian cells. Virology 446.365-77.

Zanotto, P. M., G. F. Gao, T. Gritsun, M. S. Marin, W. R. Jiang, K. Venugopal, H. W. Reid & E. A. Gould. 1995. An arbovirus cline across the northern hemisphere. Virology 210.152-9.

Zanotto, P. M., M. J. Gibbs, E. A. Gould & E. C. Holmes. 1996a. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J Virol 70.6083-96.

Zanotto, P. M., E. A. Gould, G. F. Gao, P. H. Harvey & E. C. Holmes. 1996b. Population dynamics of flaviviruses revealed by molecular phylogenies. Proc Natl Acad Sci U S A 93.548-53.

Černý, J., B. Černá Bolfíková, P. M. de A Zanotto, L. Grubhoffer & D. Růžek. 2015. A deep phylogeny of viral and cellular right-hand polymerases. Infect Genet Evol 36.275-86.

Černý, J., B. Černá Bolfíková, J. J. Valdés, L. Grubhoffer & D. Růžek. 2014. Evolution of tertiary structure of viral RNA dependent polymerases. PLoS One 9.e96070.

Černý, Jiří, Martin Selinger, Martin Palus, Zuzana Vavrušková, Hana Tykalová, Lesley Bell-Sakyi, Libor Grubhoffer & Daniel Růžek. Expression of a second open reading frame present in the genome of tick-borne encephalitis virus strain Neudoerfl is not detectable in infected cells.

Černý, Jiří, Barbora Černá Bolfíková, Libor Grubhoffer & Daniel Růžek. Genomes of viruses classified in genus Flavivirus (family Flaviviridae) evolved via multiple recombination events.

6. PUBLICATIONS

6.1 Evolution of tertiary structure of viral RNA dependent polymerases

The article has been published in PLoS One and should be cited as: **Jiří Černý**; Barbora Černá Bolfíková; James J. Valdés; Libor Grubhoffer; Daniel Růžek: Evolution of tertiary structure of viral RNA dependent polymerases, PLoS ONE, 2014, doi: 10.1371/journal.pone.0096070

Evolution of tertiary structure of viral RNA dependent polymerases

AUTHORS:

Jiří Černý[1, 2, #], Barbora Černá Bolfíková[3], James J. Valdés[1], Libor Grubhoffer[1, 2], Daniel Růžek[1, 4]

AUTHORS' AFFILIATIONS:

[1] Institute of Parasitology, Biology Centre of the Academy of Sciences of the Czech Republic,Branišovská 31, CZ-37005 České Budějovice, Czech Republic

[2] Faculty of Science, University of South Bohemia in České Budějovice, Branišovská 31, CZ-37005 České Budějovice, Czech Republic

[3] Faculty of Tropical AgriSciences, Czech University of Life Sciences Prague, Kamýcká 129
CZ-16521 Praha 6 – Suchdol, Czech Republic

[4] Veterinary Research Institute, Hudcova 296/70, CZ-62100 Brno, Czech Republic

[#] corresponding author: e-mail: cerny@paru.cas.cz, tel: +420 387 775 451; fax: +420 385 310 388;

ABSTRACT

Viral RNA dependent polymerases (vRdPs) are present in all RNA viruses; unfortunately, their sequence similarity is too low for phylogenetic studies. Nevertheless, vRdP protein structures are remarkably conserved.

In this study, we used the structural similarity of vRdPs to reconstruct their evolutionary history. The major strength of this work is in unifying sequence and structural data into a single quantitative phylogenetic analysis, using powerful Bayesian approach.

The resulting phylogram of vRdPs demonstrates that RNA-dependent DNA polymerases (RdDPs) of viruses within *Retroviridae* family cluster in a clearly separated group of vRdPs, while RNA-dependent RNA polymerases (RdRPs) of dsRNA and +ssRNA viruses are mixed together. This evidence supports the hypothesis that RdRPs replicating +ssRNA viruses evolved multiple times from RdRPs replicating +dsRNA viruses, and *vice versa*. Moreover, our phylogram may be presented as a scheme for RNA virus evolution. The results are in concordance with the actual concept of RNA virus evolution. Finally, the methods used in our work provide a new direction for studying ancient virus evolution.

KEY WORDS

Virus evolution; viral polymerase; MrBayes; structural evolution; protein structure; HCV; HIV; Poliovirus

INTRODUCTION

RNA viruses evolve rapidly. Since viral RNA-dependent polymerases (vRdP) miss the proofreading activity they produce a high percentage of mutated variants [1]. These variants face a strong evolutionary pressure by the host immune system and a highly competitive environment between relative viruses [2]. These factors lead to a rapid diversification in the primary structure of all viral genes and proteins, and a swift establishment of new virus strains [3-5].

Despite these fast changes in the sequences of viral proteins, functions that are crucial for efficient virus reproduction must be preserved [6]. Therefore, proteins involved in important steps of the virus life cycle accumulate mutations slower and preserve a higher degree of conservation [6]. The most

conserved proteins among RNA viruses are polymerases, helicases, proteases and methyltransferases [7].

Contrary to the primary structure, the tertiary structure of most proteins sharing a common evolutionary origin remains conserved [8,9]. The most conserved part of the protein is usually the core structure essential for protein function. The core is often surrounded by less conserved structures modifying the protein function. Changes in these additional structures often lead to minor changes in protein character (e. g., different substrate specificity), but the major protein function remains unchanged.

Morphological description of protein structure can help in reconstructing protein evolutionary history. In this approach, protein structural features are encoded in a character matrix where the rows describe the individual proteins and the columns describe the individual features. This is similar to the approach used for reconstructing the evolutionary relations among fossil species [10]. Morphological data can also be coupled with sequence data to enforce the incoming information [11,12]. This approach may also be applied to proteins. For example, mixed morphological and sequence data were used to reconstruct the evolution of aminoacyl tRNA synthetases class I [13] and the protein kinase-like superfamily [14].

Among all viral proteins, vRdPs display the highest degree of conservation. Genes coding for vRdPs were found in all non-satellite RNA viruses and RNA viruses reproducing via a DNA intermediate [15]. All vRdPs contain seven typical sequence motifs (G, F, A, B, C, D and E) [16,17] that incorporate conserved amino acid residues crucial for polymerase function [18,19].

Moreover, vRdPs share remarkable structural homology. The protein structural fold resembles a right hand with subdomains termed fingers, palm and thumb [20-23]. The palm subdomain is structurally well conserved among all vRdPs. Finger and thumb subdomains are more variable, but they can be fully aligned only among RNA-dependent RNA polymerases (RdRPs) of +ssRNA viruses [21]. For most vRdPs, the finger, palm and thumb subdomains accommodate seven conserved structural motifs (homomorphs), each bearing one of the conserved sequence motif described before [24].

All vRdPs evolved from one common ancestral protein [16,20]. In the past, sequence similarity among vRdPs was used in attempts to reconstruct RNA virus evolutionary history [7,16,25-31]. Unfortunately, this sequence similarity was shown to be too low to produce an accurate sequence alignment for further phylogenetic analysis [32].

In our current work, we used the structural similarity of vRdPs to reconstruct their evolutionary history. We used the similarities of vRdPs protein structures to produce a highly accurate structure based sequence alignment for our subsequent studies. Moreover, we picked 21 biochemical and structural features of each polymerase and encoded them into the matrix that was used in a phylogenetic analysis to particularize results obtained from structure based sequence alignment analysis. In our phylogenetic analysis, we used Bayesian clustering algorithms, which are ideal for reconstruction of complicated phylogenetic relationships. The resulting phylogenetic tree describing the evolution of vRdPs has high statistical support for most branches. As vRdPs are the only universal gene in all RNA viruses, our phylogenetic tree can be understood as a scheme of RNA virus evolution.

MATERIAL AND METHODS

Selection of vRdPs for further phylogenetic studies

To find structurally homologous vRdPs, we employed the DALI server [33] using the structure of Dengue virus type 3 (DENV3) RdRP as a query (PDB number 2J7W-A). The program was run under the default conditions. DALI server automatically screens the PDB database to select structurally homologous proteins and lists them according to a decreasing Z-score, a quantitative expression of protein structure similarity [33]. Only protein structures having similarity Z score higher than 2 were taken in account since hits with lower Z-score are most likely incidental hits. The vRdPs were selected among the listed protein structures. They were assigned to the individual virus species classified into genera and families according to the actual ICTV virus taxonomy [34]. Representative structures were selected using the following criteria: (1) Maximally two polymerases from two different viruses were selected from one genus (the exception was four viruses from genus *Enterovirus*). (2) Structures with bound substrate, substrate analogue and/or template nucleic acid were favored. (3) High resolution structures were preferred. (4) Structures without

any mutation were favored. As polymerases are very active enzymes changing their topology in response to many external stimuli (bound template/nucleotide/product, actual step of polymerization cycle, etc.), the criteria for structure selection was set up to select polymerase structures under identical conditions.

The same process described above was done using three structures with the lowest structure homology to 2J7W-A as queries using the DALI sever: 3V81-C (human immunodeficiency virus 1 - HIV1), 2R7W-A (simian rotavirus - SRV) and 2PUS-A (infectious bursal disease virus - IBDV). Sets of structures selected in these three runs were compared with the first set to insure no adequate structures were missed.

Construction of structure superposition and structure based sequence alignment

Structures of selected vRdPs were superimposed using the DALI server multiple structural alignment tool [33]. DALI created structure based sequence alignment was validated and improved using the default settings in T-Coffee Expresso [35]. The resulting alignment was verified by comparison with previously published vRdP alignments [17,24,31,36,37].

The structure based sequence alignment was analyzed using the JOY server under the default conditions [38]. JOY is a program used for annotation of protein sequence alignments with 3D structural features. It is necessary in understanding the conservation of specific amino acid residues in a specific environment. JOY contains various algorithms such as DSSP [39] used for secondary structure classification. Sequence consensus and sequence conservation were calculated in Chimera implemented algorithms [40,41].

Analysis of the vRdPs structural similarities between vRdPs

Analysis of conserved amino acid residues and sequence motifs in the structural based sequence alignment as well as presence/absence of conserved structural features was done manually according to criteria previously used in describing vRdPs [20,24,42]. Comparative results were encoded into a 21-column character matrix where each column represents a single selected character typical of some but not all vRdPs. The matrix row represents each evaluated

polymerase. Structural characters were coded to MrBayes as standard data (0-9). These characters were set as unordered allowing them to move from one state to another (character designated “0” can change to “2” without passing “1”).

Construction of phylogenetic tree

Best fitting model of amino acid substitutions was tested in PROTTEST 2.4 [43] under the Akaike information criterion [44] and the Bayesian information criterion [45]. As results of the two tests were not consistent, we decided to use the most complex model, the general time reversible (GTR) model with a proportion of invariable sites and a gamma-shaped distribution of rates across sites [46,47]. Bayesian phylogenetic analysis was performed using MrBayes v3.1.2 [48]. Bayesian analysis consisted of two runs with four chains (one cold and three heated), and was run for 10 million generations sampled every 100 generations. The first 25% of samples were discarded as a burning period. Although the average standard deviation of split frequencies was much lower than 0.01, convergence of runs and chains was verified using the AWTY [49]. Analysis was run for sequence data alone and for mixed data (sequence alignment and structural character matrix) with equal settings for analysis.

RESULTS

Formation of representative set of vRdPs

The DALI server queried using the Dengue virus RdRP (2J7W-A) found 745 hits with structure similarity Z-score 2 or higher. Using the criteria described in the Material and methods section, we selected 21 vRdPs protein structures among these hits. In our subsequent query, no additional protein structures were selected from 844, 743 and 575 hits identified using 3V81-C (HIV1), 2R7W-A (SRV), and 2PUS-A (IBDV).

To ensure we did not miss any relevant structure, we browsed the PDB [50] using names of all RNA virus genera listed in the ICTV database. No additional structures were found. A preliminary notice was found about the successful crystallization of *Thosea asigna* virus RdRP (genus *Permutotetravirus*, family *Permutotetraviridae*), but the structure has not yet been published [51].

The final list included 22 vRdPs from 22 virus species in 17 virus genera and 8 virus families (see Table 1 for details). All viral families were classified in the Baltimore classes III (double stranded RNA viruses), IV (positive sense single stranded RNA viruses), and VI (Positive-sense single-stranded RNA viruses that replicate through a DNA intermediate). No polymerases of any virus classified in Baltimore class V (negative sense single stranded RNA viruses) were identified, since there was no known protein structure of any RNA dependent RNA polymerase for these viruses.

Structure superposition of vRdPs

The vRdPs from our collection represents a wide range of proteins that are different in protein size and other parameters (see Table 1). Many of them bear additional domains with non-polymerase activities that are conserved only among closely related proteins. These domains were not taken into account for subsequent analysis.

Primary and tertiary structures of domains bearing polymerase activity are similar in all selected proteins. Subdomains finger (F), palm (P), and thumb (T) are collinearly arranged in all vRdPs succeeding always as F1-P1-F2-P2-T from N- to C-terminus (see Figure S1 for details) [20-23]. Polymerase domains of selected vRdPs were superpositioned and structures typical for each of the selected viral families are highlighted in Figure 1 (for schematic structure of all vRdPs see Figure S2). Structural superposition shows a conserved architecture of vRdP subdomains and the seven conserved structural homomorphs previously described [24] are clearly visible.

An additional eighth structural helix-turn-helix motif was observed in the thumb subdomain, we call homomorph H (hmH). Despite the poorly conserved sequence of homomorph H, the structural motif is well conserved in all vRdPs (see Figure 1). To characterize its conservativeness, we calculated its RMSD among all vRdPs and compared it with the RMSD of homomorph D (hmD) that is similar in size. Results showed that hmH is as conserved as the well-established hmD (see Table S1 for further details).

Structural similarities among vRdPs

The structure similarity Z-score was calculated for all polymerase couples (see Table 2) showing extremely high protein structure similarities among vRdPs from viruses classified into one viral genus (see genus *Enterovirus* as the best example). The similarities among the vRdPs of viruses classified in the same family are slightly lower, but still very high (see family *Picornaviridae* as the best example). RdRPs of all +ssRNA viruses (except enterobacteriophage Qβ - Qβ) form a cluster of relatively highly similar structures, while structures of pseudomonas phage Φ6 (Φ6), Qβ and *Birnaviridae* RdRPs are moderately similar, and structures of reoviral RdRPs and retroviral RdDPs are similar only distantly to RdRPs of +ssRNA virus (see Table 2 for details).

We also quantified 21 attributes previously used for vRdPs description and encoded them into a 21-column character matrix (see Table 3). Features were selected and quantified manually according to criteria previously used for describing vRdPs [20,24,42] and are included in the Text S1.

Automatically created structure based alignment of selected vRdPs including annotated structural features is depicted in Figures 2, 3, and 4.

Phylogenetic characterization of vRdPs

The evolutionary history of vRdPs was reconstructed using the Bayesian clustering analysis. Sequence (structure based sequence alignment) and structural (character matrix) information were used simultaneously in a unified analysis. Combination of these datasets was used to produce a phylogenetic tree with high Bayesian posterior probabilities for most branches (see Figure 5). Despite the high Bayesian support, one polytomy appeared concerning the position of *Birnaviridae* family.

Our phylogenetic analysis classified all vRdPs into groups that correspond to the viral genera and families proposed by ICTV. RdDPs of RNA viruses replicating via DNA intermediate (Baltimore class VI) formed a clearly separated group of vRdPs. The RdRPs of +ssRNA and dsRNA viruses clustered together and did not form any separate groups. This suggests that dsRNA viruses evolved from +ssRNA viruses multiple times, and vice versa. The possible evolutionary

scenarios of vRdP evolution and its impact on the reconstruction of RNA virus evolution will be discussed further.

Usage of each data set alone was less statistically powerful than the combined analysis (see Figure S3). Despite, our results rely mostly on sequence information incoming from a structure based sequence alignment. The 21-column character matrix served as a stabilizing element that properly placed ambiguous branches and prevent against long branch artifacts (compare Figure S3 panels A and B and Figure 5).

DISCUSSION

Similarities among vRdPs

The vRdPs are an ancient and diversified enzyme group. They share only limited conservation in primary structure, however their protein structure [21,24] and the mechanism of function [19,23,42] are very similar. The vRdPs adopt a conserved right hand conformation with three subdomains termed fingers, palm and thumb. Seven conserved sequence motifs were previously described in vRdPs [16,17,37]. Moreover, amino acid residues in these motifs adopt extremely conserved position in vRdPs' [24]. Herein, we described a novel conserved structural motif named homomorph H (hmH) formed by a conserved helix-turn-helix structure in the thumb subdomain of all vRdPs. Despite its high structure conservation, and hmH primary structure is slightly conserved. Function of hmH remains elusive and further biochemical studies will be needed to elucidate it.

Presence of vRdPs in all RNA virus species allows their use in phylogenetic analysis [7,16,25-31]. This approach was disputed by an extensive study showing the sequence conservation of vRdPs is too low to be successfully and meaningfully used for phylogenetic analysis employing classical methods [32]. The similarities among vRdPs may have evolved by convergent evolution [32], however these conclusions may be challenged by several arguments. 1) The vRdPs share seven conserved sequential collinearly arranged motifs; a phenomenon highly improbable via convergence [16]. 2) The right hand conformation is not the only fold that can be adapted by RNA-dependent polymerases. Cellular RdRPs participating in RNA interference accommodate totally different double barrel conformations [52]. 3) Modern bioinformatics

approaches based on Bayesian analyses are more suitable for reconstruction of distant evolutionary relationships [53] than previously described statistical methods [32]. 4) Conserved protein tertiary structure of all vRdPs can supplement missing information in highly diverged protein sequences and allowing us to study the evolution of extremely distantly related proteins [13,14].

Nevertheless, polymerases can adopt various conformations, changing their topology in response to bound template/incoming nucleotides, steps in polymerization cycle and artificially depending on crystallization conditions. We overcome this by selecting vRdPs' representatives crystallized under similar conditions (see Material and methods).

How did the vRdPs evolve?

Our phylogram shows the RdDP of *Retroviridae* forms a clearly separate group of RNA viruses replicating via the dsDNA intermediate (Baltimore class VI). This is caused by a series of specific interactions that occurs between template, product and protein, and differs significantly between RdDPs and RdRPs [54]. For example, RdDPs accommodates a conservative aromatic amino acid residue in motif B (alignment position 525 - Figure 3). This position is occupied by aspartate or asparagine interacting with aspartate in motif A (alignment position 416 - Figure 3) in RdRPs discriminating incorporation of dNTPs instead of NTPs [20]. Moreover, the structure of RdDPs is much simpler, many structural motifs are absent, and others are highly reduced [24].

RdRP of the +ssRNA bacteriophage Qβ is the closest relative of retroviral RdDPs. The Qβ polymerase already contains all motifs typical for RdRPs, but is still simpler having no additional structural motifs [55,56]. As Qβ represents an ancient virus group [57], it is probable that the phylogram may be rooted between Qβ RdRP and retroviral RdRPs.

Rooting the evolutionary tree of vRdPs using cellular right handed polymerases as an outgroup shows, the root is positioned between bacteriophage Qβ RdRP and retroviral RdDPs (Černý et al, under submission). This is in concordance with RNA world theories and theories implicating viruses in the shift from RNA world to DNA world [58].

RdRPs of all RNA viruses are mixed together in our phylogram and they do not follow the Baltimore classification. For example RdRP of +ssRNA Qβ is closely related to the RdRPs of dsRNA viruses than to the RdRPs of other +ssRNA viruses and RdRP of dsRNA birnaviruses tends towards RdRPs of mammalian +ssRNA viruses. The RdRPs can easily replicate both ssRNA and dsRNA without any critical rearrangements in their structure. This is not surprising since picornaviral RdRP were shown to replicate dsRNA even without the aid of a helicase [59].

Primer dependence/independence also apparently evolved multiple times. RdRPs of viruses, which in our phylogram are closer to the expected root (*Leviviridae*, *Reoviridae*, *Cystoviridae*), do not require RNA or protein primer for reaction initialization [60]. This suggests that the original vRdPs were probably primer independent. *De novo* initiation is also typical for many cellular RdRPs [61].

Primer independent RdRPs of viruses from families *Flaviviridae* and *Cystoviridae* share remarkably large thumb subdomains of their RdRPs, allowing accurate positioning of the first incoming nucleotide and RNA polymerization initiation [62]. Despite that both proteins share similar interactions between enzyme, template and incoming nucleotide, the position of the priming motif is different [62].

Viruses from the family *Birnaviridae* and several other families encode cyclic permuted RdRP [31,37]. It was suggested that birnaviral RdRPs represents an ancient group of polymerases that split from other polymerases before DdDPs, DdRPs, RdDPs and RdRPs were established as four distinct groups [31]. Our results indicate RdRPs with cyclic permutation are younger and they share a common evolutionary ancestor with RdRPs of +ssRNA virus RdRPs.

What does our model of vRdPs evolution tell us about the evolution of RNA viruses?

Virus evolution is an extremely complicated story. Viral genes and proteins evolve rapidly and relative proteins share only a low degree of homology [3-5], making virus phylogenetic reconstruction difficult. It is complicated to generate a proper alignment of selected proteins and the resulting phylograms usually do not have sufficient statistical support [32]. Therefore, a qualitative

description of a set of virus features is used for reconstruction of distant phylogenetic virus relationships (capsid architecture, genome replication strategies, etc. [63,64]). Nevertheless, this approach is sensitive to recombination events between virus and host, or between different viruses, and occurs quite often resulting in a mixture of different genes[65-68]. That is why, virus evolution nowadays is not considered as a linear process, but rather as a network [69].

Absence of any universal gene shared by all viruses makes reconstruction of virus evolution even more difficult, despite that some genes are shared among many viruses. An example of such a gene is a jelly-roll capsid protein that is typical for picorna-like viruses (+ssRNA genome), *Microviridae*, *Parvoviridae* (both ssDNA), *Papylomaviridea*, *Polyomaviridae* (both dsDNA), etc. [70,71]. Jelly-roll capsid protein, however is an inappropriate candidate for a virus phylogenetic marker, since viruses sharing a jelly-roll capsid protein are only distantly related and protein is missing among closely related virus families.

Presence of the vRdPs in all RNA viruses [15] allowed to use the vRdPs as a marker for RNA virus evolution [28]. Nevertheless, their sequence similarity is too low to be used by classical phylogenetic approaches [32]. We overcome this using structure based homology of vRdPs. Our phylogram describing the evolutionary history of vRdPs may be understood as an evolutive phylogram of RNA viruses. Our results are in concordance with the actual concepts of virus evolution [63,69] and depict the polyphyletic origin of dsRNA viruses. The first group is represented by *Cystoviridae* and *Reoviridae* families, while the second group is represented by the *Birnaviridae* family. *Reoviridae* and *Cystoviridae* share many common features. Both viral groups have similar multilayer capsid organization [72]. They replicate their genome by a conservative manner inside the inner virus capsid [73]. Viruses in *Birnaviridae* family are more similar to +ssRNA viruses. Their cyclically permuted RdRPs are similar to cyclically permuted RdRPs of +ssRNA viruses from *Permutotetraviridae* [31]. Moreover, birnaviruses replicate their genome in a semiconservative manner outside the virus capsid [74] using their guanylylated RdRP as a primer [75] that is similar to protein primed replication of picornavirus-like viruses [76,77].

Mammalian +ssRNA viruses cluster together forming two monophyletic clades. The first is represented by viruses from the family *Flaviviridae*, while the second

by viruses from families *Caliciviridae* and *Picornaviridae*. Regardless that the differences between them are smaller than in the case of dsRNA viruses, both these clades differ in the same biological aspect. Flaviviruses replicates their RNA by a primer independent manner [78,79]. Their genome is either uncapped [80,81] or capped by 7-methylguanosine cap [82]. *Caliciviridae* and *Picornaviridae* use vPg protein primer that also caps their genomes [83]. These similarities between mammalian +ssRNA viruses and *Birnaviridae* show they evolved from a common ancestor [31,70,84].

The last two groups of RNA viruses, families *Leviviridae* and *Retroviridae*, are distinctly separated. These two groups seem to be extremely ancient and they probably evolved from the last universal common ancestor of all life forms – even before the cell evolution [64,85,86]. This is in concordance with recent theories about evolution of ancient life forms, the transition from the RNA into the DNA word and cell evolution [58].

Only a limited number of vRdP protein structures are known now. Nevertheless, they come out from very diverse viral groups that can serve as representatives of other virus groups (*Togaviridae* and *Coronaviridae* would most probably follow *Flaviviridae* etc.). ThevRdPs with known protein structure come from viruses that are usually important as human or veterinary pathogens or represent important biological models. There is no known vRdP protein structure of any plant, protozoan or fungal virus. Moreover, no protein structure of any –ssRNA virus RdRP is known. Since RdRPs of –ssRNA viruses share many sequence motifs with other vRdPs [87-89], their structure will most probably be similar to the structure of other RNA viruses. Likewise, vRdPs structures of plant, protozoan and fungal viruses that are often closely related to animal viruses [68] will probably be similar.

SUPPLEMENTARY DATA

Supplementary Data are available at PLoS One online: Text S1, Table S1 and Figures S1, S2, and S3. All data are available on request from the corresponding author.

ACKNOWLEDGEMENT

We would like to express our thanks to Filip Husník and Martin Pospíšek for constructive criticism of our work and for interesting suggestions.

REFERENCES

1. Steinhauer DA, Domingo E, Holland JJ (1992) Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. Gene 122: 281-288.
2. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439: 344-348.
3. Cabanillas L, Arribas M, Lázaro E (2013) Evolution at increased error rate leads to the coexistence of multiple adaptive pathways in an RNA virus. BMC Evol Biol 13: 11.
4. Smith DB, McFadden N, Blundell RJ, Meredith A, Simmonds P (2012) Diversity of murine norovirus in wild-rodent populations: species-specific associations suggest an ancient divergence. J Gen Virol 93: 259-266.
5. Pickett BE, Striker R, Lefkowitz EJ (2011) Evidence for separation of HCV subtype 1a into two distinct clades. J Viral Hepat 18: 608-618.
6. Krupovič M, Bamford DH (2010) Order to the viral universe. J Virol 84: 12476-12479.
7. Koonin EV, Dolja VV (1993) Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit Rev Biochem Mol Biol 28: 375-430.
8. Illergård K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77: 499-508.
9. Holm L, Sander C (1996) Mapping the protein universe. Science 273: 595-603.
10. Wiens J (2004) The role of morphological data in phylogeny reconstruction. Syst Biol 53: 653-661.
11. Nylander JA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL (2004) Bayesian phylogenetic analysis of combined data. Syst Biol 53: 47-67.
12. McGowen MR, Spaulding M, Gatesy J (2009) Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol Phylogenet Evol 53: 891-906.
13. Aravind L, Anantharaman V, Koonin EV (2002) Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase

nucleotide-binding domains: implications for protein evolution in the RNA. Proteins 48: 1-14.
14. Scheeff ED, Bourne PE (2005) Structural evolution of the protein kinase-like superfamily. PLoS Comput Biol 1: e49.
15. Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35: 235-241.
16. Poch O, Sauvaget I, Delarue M, Tordo N (1989) Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J 8: 3867-3874.
17. Bruenn JA (2003) A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. Nucleic Acids Res 31: 1821-1829.
18. Gohara DW, Crotty S, Arnold JJ, Yoder JD, Andino R, et al. (2000) Poliovirus RNA-dependent RNA polymerase (3Dpol): structural, biochemical, and biological analysis of conserved structural motifs A and B. J Biol Chem 275: 25523-25532.
19. Korneeva VS, Cameron CE (2007) Structure-function relationships of the viral RNA-dependent RNA polymerase: fidelity, replication speed, and initiation mechanism determined by a residue in the ribose-binding pocket. J Biol Chem 282: 16135-16145.
20. Hansen JL, Long AM, Schultz SC (1997) Structure of the RNA-dependent RNA polymerase of poliovirus. Structure 5: 1109-1122.
21. Ferrer-Orta C, Arias A, Escarmís C, Verdaguer N (2006) A comparison of viral RNA-dependent RNA polymerases. Curr Opin Struct Biol 16: 27-34.
22. Shatskaya GS, Dmitrieva TM (2013) Structural organization of viral RNA-dependent RNA polymerases. Biochemistry (Mosc) 78: 231-235.
23. Ng KK, Arnold JJ, Cameron CE (2008) Structure-function relationships among RNA-dependent RNA polymerases. Curr Top Microbiol Immunol 320: 137-156.
24. Lang DM, Zemla AT, Zhou CL (2013) Highly similar structural frames link the template tunnel and NTP entry tunnel to the exterior surface in RNA-dependent RNA polymerases. Nucleic Acids Res 41: 1464-1482.
25. Dolja VV, Carrington JC (1992) Evolution of positive-strand RNA viruses. Seminars in Virology. pp. 315-326.
26. Eickbush TH (1994) Origin and evolutionary relationships of retroelements. In: Morse SS, editor. The evolutionary biology of viruses: Raven Press,

1185 Avenue of the Americas, New York, New York 10036-2806, USA. pp. 121-157.
27. Goldbach R, Wellink J, Verver J, van Kammen A, Kasteel D, et al. (1994) Adaptation of positive-strand RNA viruses to plants. Arch Virol Suppl 9: 87-97.
28. Ward CW (1993) Progress towards a higher taxonomy of viruses. Res Virol 144: 419-453.
29. Koonin EV (1991) The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J Gen Virol 72 (Pt 9): 2197-2206.
30. Bruenn JA (1991) Relationships among the positive strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases. Nucleic Acids Res 19: 217-226.
31. Gorbalenya AE, Pringle FM, Zeddam JL, Luke BT, Cameron CE, et al. (2002) The palm subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. J Mol Biol 324: 47-62.
32. Zanotto PM, Gibbs MJ, Gould EA, Holmes EC (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J Virol 70: 6083-6096.
33. Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38: W545-549.
34. King AMQ, Adams, M.J., Carstens, E.B. and Lefkowitz, E.J. (2012) Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses.: Elsevier Academic Press, San Diego, USA.
35. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, et al. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34: W604-608.
36. Ferrer-Orta C, Arias A, Perez-Luque R, Escarmís C, Domingo E, et al. (2004) Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. J Biol Chem 279: 47212-47221.
37. Pan J, Vakharia VN, Tao YJ (2007) The structure of a birnavirus polymerase reveals a distinct active site topology. Proc Natl Acad Sci U S A 104: 7385-7390.

38. Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics 14: 617-623.
39. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.
40. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25: 1605-1612.
41. Meng EC, Pettersen EF, Couch GS, Huang CC, Ferrin TE (2006) Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics 7: 339.
42. Gong P, Peersen OB (2010) Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. Proc Natl Acad Sci U S A 107: 22505-22510.
43. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21: 2104-2105.
44. Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.
45. Schwarz G (1978) Estimating the Dimension of a Model. Annals of Statistics 6: 461-464.
46. Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86-93.
47. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39: 306-314.
48. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.
49. Wilgenbusch JC, Warren DL, Swofford DL (2004) AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.
50. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
51. Ferrero D, Buxaderas M, Rodriguez JF, Verdaguer N (2012) Purification, crystallization and preliminary X-ray diffraction analysis of the RNA-

dependent RNA polymerase from Thosea asigna virus. Acta Crystallogr Sect F Struct Biol Cryst Commun 68: 1263-1266.
52. Salgado PS, Koivunen MR, Makeyev EV, Bamford DH, Stuart DI, et al. (2006) The structure of an RNAi polymerase links RNA silencing and transcription. PLoS Biol 4: e434.
53. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.
54. Kohlstaedt LA, Wang J, Friedman JM, Rice PA, Steitz TA (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science 256: 1783-1790.
55. Kidmose RT, Vasiliev NN, Chetverin AB, Andersen GR, Knudsen CR (2010) Structure of the Qbeta replicase, an RNA-dependent RNA polymerase consisting of viral and host proteins. Proc Natl Acad Sci U S A 107: 10884-10889.
56. Takeshita D, Tomita K (2010) Assembly of Q(Takeshita & Tomita) viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. Proc Natl Acad Sci U S A 107: 15733-15738.
57. van Duin J, Tsareva N (2006) Single-stranded RNA phages. In: Calendar RL, editor. The Bacteriophages (Second ed): Oxford University Press.
58. Forterre P (2002) The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5: 525-532.
59. Cho MW, Richards OC, Dmitrieva TM, Agol V, Ehrenfeld E (1993) RNA duplex unwinding activity of poliovirus RNA-dependent RNA polymerase 3Dpol. J Virol 67: 3010-3018.
60. van Dijk AA, Makeyev EV, Bamford DH (2004) Initiation of viral RNA-dependent RNA polymerization. J Gen Virol 85: 1077-1093.
61. Makeyev EV, Bamford DH (2002) Cellular RNA-dependent RNA polymerase involved in posttranscriptional gene silencing has two distinct activity modes. Mol Cell 10: 1417-1427.
62. Butcher SJ, Grimes JM, Makeyev EV, Bamford DH, Stuart DI (2001) A mechanism for initiating RNA-dependent RNA polymerization. Nature 410: 235-240.
63. Ahlquist P (2006) Parallels among positive-strand RNA viruses, reverse-transcribing viruses and double-stranded RNA viruses. Nat Rev Microbiol 4: 371-382.

64. Bamford DH, Grimes JM, Stuart DI (2005) What does structure tell us about virus evolution? Curr Opin Struct Biol 15: 655-663.
65. Scheel TK, Galli A, Li YP, Mikkelsen LS, Gottwein JM, et al. (2013) Productive homologous and non-homologous recombination of hepatitis C virus in cell culture. PLoS Pathog 9: e1003228.
66. Smith LM, McWhorter AR, Shellam GR, Redwood AJ (2013) The genome of murine cytomegalovirus is shaped by purifying selection and extensive recombination. Virology 435: 258-268.
67. Pond SL, Murrell B, Poon AF (2012) Evolution of viral genomes: interplay between selection, recombination, and other forces. Methods Mol Biol 856: 239-272.
68. Dolja VV, Koonin EV (2011) Common origins and host-dependent diversity of plant and animal viromes. Curr Opin Virol 1: 322-331.
69. Koonin EV, Dolja VV (2012) Expanding networks of RNA virus evolution. BMC Biol 10: 54.
70. Koonin EV, Wolf YI, Nagasaki K, Dolja VV (2008) The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat Rev Microbiol 6: 925-939.
71. Ravantti J, Bamford D, Stuart DI (2013) Automatic comparison and classification of protein structures. J Struct Biol 183: 47-56.
72. Poranen MM, Bamford DH (2012) Assembly of large icosahedral double-stranded RNA viruses. Adv Exp Med Biol 726: 379-402.
73. Lawton JA, Estes MK, Prasad BV (2000) Mechanism of genome transcription in segmented dsRNA viruses. Adv Virus Res 55: 185-229.
74. Cortez-San Martín M, Villanueva RA, Jashés M, Sandino AM (2009) Molecular characterization of IPNV RNA replication intermediates during the viral infective cycle. Virus Res 144: 344-349.
75. Dobos P (1995) Protein-primed RNA synthesis in vitro by the virion-associated RNA polymerase of infectious pancreatic necrosis virus. Virology 208: 19-25.
76. Wimmer E, Hellen CU, Cao X (1993) Genetics of poliovirus. Annu Rev Genet 27: 353-436.
77. Buck KW (1996) Comparison of the replication of positive-stranded RNA viruses of plants and animals. Adv Virus Res 47: 159-251.

78. Urcuqui-Inchima S, Patiño C, Torres S, Haenni AL, Díaz FJ (2010) Recent developments in understanding dengue virus replication. Adv Virus Res 77: 1-39.
79. Rice CM (2011) New insights into HCV replication: potential antiviral targets. Top Antivir Med 19: 117-120.
80. Wang C, Sarnow P, Siddiqui A (1993) Translation of human hepatitis C virus RNA in cultured cells is mediated by an internal ribosome-binding mechanism. J Virol 67: 3338-3344.
81. Pérard J, Leyrat C, Baudin F, Drouet E, Jamin M (2013) Structure of the full-length HCV IRES in solution. Nat Commun 4: 1612.
82. Cleaves GR, Dubin DT (1979) Methylation status of intracellular dengue type 2 40 S RNA. Virology 96: 159-165.
83. Goodfellow I (2011) The genome-linked protein VPg of vertebrate viruses - a multifaceted protein. Curr Opin Virol 1: 355-362.
84. Gorbalenya AE, Koonin EV (1988) Birnavirus RNA polymerase is related to polymerases of positive strand RNA viruses. Nucleic Acids Res 16: 7735.
85. Koonin EV, Dolja VV (2006) Evolution of complexity in the viral world: the dawn of a new vision. Virus Res 117: 1-4.
86. Holmes EC (2011) What does virus evolution tell us about virus origins? J Virol 85: 5247-5251.
87. Poch O, Blumberg BM, Bougueleret L, Tordo N (1990) Sequence comparison of five polymerases (L proteins) of unsegmented negative-strand RNA viruses: theoretical assignment of functional domains. J Gen Virol 71 (Pt 5): 1153-1162.
88. Müller R, Poch O, Delarue M, Bishop DH, Bouloy M (1994) Rift Valley fever virus L segment: correction of the sequence and possible functional role of newly identified regions conserved in RNA-dependent polymerases. J Gen Virol 75 (Pt 6): 1345-1352.
89. Lukashevich IS, Djavani M, Shapiro K, Sanchez A, Ravkov E, et al. (1997) The Lassa fever virus L gene: nucleotide sequence, comparison, and precipitation of a predicted 250 kDa protein with monospecific antiserum. J Gen Virol 78 (Pt 3): 547-551.
90. Ng KK, Cherney MM, Vazquez AL, Machin A, Alonso JM, et al. (2002) Crystal structures of active and inactive conformations of a caliciviral RNA-dependent RNA polymerase. J Biol Chem 277: 1381-1387.

91. Mastrangelo E, Pezzullo M, Tarantino D, Petazzi R, Germani F, et al. (2012) Structure-based inhibition of Norovirus RNA-dependent RNA polymerases. J Mol Biol 419: 198-210.
92. Zamyatkin DF, Parra F, Alonso JM, Harki DA, Peterson BR, et al. (2008) Structural insights into mechanisms of catalysis and inhibition in Norwalk virus polymerase. J Biol Chem 283: 7705-7712.
93. Fullerton SW, Blaschke M, Coutard B, Gebhardt J, Gorbalenya A, et al. (2007) Structural and functional characterization of sapovirus RNA-dependent RNA polymerase. J Virol 81: 1858-1871.
94. Yap TL, Xu T, Chen YL, Malet H, Egloff MP, et al. (2007) Crystal structure of the dengue virus RNA-dependent RNA polymerase catalytic domain at 1.85-angstrom resolution. J Virol 81: 4753-4765.
95. Lu G, Gong P (2013) Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. PLoS Pathog 9: e1003549.
96. O'Farrell D, Trowbridge R, Rowlands D, Jäger J (2003) Substrate complexes of hepatitis C virus RNA polymerase (HC-J4): structural evidence for nucleotide import and de-novo initiation. J Mol Biol 326: 1025-1035.
97. Choi KH, Groarke JM, Young DC, Rossmann MG, Pevear DC, et al. (2004) Design, expression, and purification of a Flaviviridae polymerase using a high-throughput approach to facilitate crystal structure determination. Protein Sci 13: 2685-2692.
98. Takeshita D, Tomita K (2012) Molecular basis for RNA polymerization by Qβ replicase. Nat Struct Mol Biol 19: 229-237.
99. Ferrer-Orta C, Arias A, Pérez-Luque R, Escarmís C, Domingo E, et al. (2007) Sequential structures provide insights into the fidelity of RNA replication. Proc Natl Acad Sci U S A 104: 9463-9468.
100. Love RA, Maegley KA, Yu X, Ferre RA, Lingardo LK, et al. (2004) The crystal structure of the RNA-dependent RNA polymerase from human rhinovirus: a dual function target for common cold antiviral therapy. Structure 12: 1533-1544.
101. Gruez A, Selisko B, Roberts M, Bricogne G, Bussetta C, et al. (2008) The crystal structure of coxsackievirus B3 RNA-dependent RNA polymerase in complex with its protein primer VPg confirms the existence of a second VPg binding site on Picornaviridae polymerases. J Virol 82: 9577-9590.

102. Graham SC, Sarin LP, Bahar MW, Myers RA, Stuart DI, et al. (2011) The N-terminus of the RNA polymerase from infectious pancreatic necrosis virus is the determinant of genome attachment. PLoS Pathog 7: e1002085.
103. Garriga D, Navarro A, Querol-Audí J, Abaitua F, Rodríguez JF, et al. (2007) Activation mechanism of a noncanonical RNA-dependent RNA polymerase. Proc Natl Acad Sci U S A 104: 20540-20545.
104. Tao Y, Farsetta DL, Nibert ML, Harrison SC (2002) RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. Cell 111: 733-745.
105. Lu X, McDonald SM, Tortorici MA, Tao YJ, Vasquez-Del Carpio R, et al. (2008) Mechanism for coordinated RNA packaging and genome replication by rotavirus polymerase VP1. Structure 16: 1678-1688.
106. Das D, Georgiadis MM (2004) The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. Structure 12: 819-829.
107. Ren J, Bird LE, Chamberlain PP, Stewart-Jones GB, Stuart DI, et al. (2002) Structure of HIV-2 reverse transcriptase at 2.35-A resolution and the mechanism of resistance to non-nucleoside inhibitors. Proc Natl Acad Sci U S A 99: 14410-14415.
108. Das K, Martinez SE, Bauman JD, Arnold E (2012) HIV-1 reverse transcriptase complex with DNA and nevirapine reveals non-nucleoside inhibition mechanism. Nat Struct Mol Biol 19: 253-259.

TABLE LEGENDS

Table 1: The list of selected vRdPs

The vRdPs selected as described in Material and methods were assigned to individual viral species, genera, families and Baltimore groups. For each individual vRdP its PDB code (PDB), used protein strand (column str.), resolution (column res.) and cofactor, substrate, template, product molecules (column co-crystallized molecules) are listed.

Table 2: Comparison of structure similarity Z-score of all vRdPs

Individual vRdP structures are introduced by a PBD code-strain and they are assigned to a virus species. Note that structure similarity Z-score is high among vRdPs originating from viruses classified in the same genus (see genus

Enterovirus (written in bold) as the best example). Structural similarity is somewhat lower but still high among vRdPs from viruses classified in the same family (see family *Picornaviridae* (written in italic) as the best example). Structural similarity of vRdPs from viruses classified in different families is significantly lower and is decreasing with excepted phylogenetic relationship. Compare all other families to family *Picornaviridae*.

Table 3: Matrix describing individual features used in phylogenetic analysis of vRdPs

Individual vRdP structures are introduced by PBD code-strain and they are assigned to a virus species. Rows in the matrix represent vRdPs, while the compared features are listed as 21 columns. Compared features are: (A) polymerase product - 0 RNA, 1 DNA; (B) polymerase template - 0 RNA, 1 both DNA and RNA; (C) NA synthesis initiation - 0 *de novo*, 1 protein primer, 2 RNA primer; (D) overall polymerase domain architecture as described in [23] - 0 active site is encircled by finger tips, 1 active site is open (fingers subdomain do not touch thumb subdomain); (E) polymerase core organization - 0 ABC, 1 CAB; (F) motif F length - 0 normal (motif is F2 is present), 1 short (motif F2 is absent), 2 long (insertion is present in motif F); (G) motif F structure - 0 ββα(3_{10})β, 1 βββ, 2 ββ; (H) F - A (C) motif connection - 0 short (≤35 amino acid residues), 1 long structured (>35 amino acid residues); (I) motif A structure - 0 -3_{10}, 1 βα, 2 β3_{10}; (J) A - B motif connection - 0 ααββ, 1 αββαββ, 2 ββ; (K) length of helix in motif B - 0 normal (≤21 amino acid residues), 1 long (>22 amino acid residues); (L) kink in motif B - 0 absent, 1 present; (M) B - C (D) motifs connection - 0 very short (≤5 amino acid residues), 1 loop (6-14 amino acid residues), 2 long helical (≥15 amino acid residues, at least 8 amino acid residues long helix); (N) motif C length - 0 short (10 amino acid residues), 1 long (>10 amino acid residues); (O) C (B) - D motifs connection - 0 short loop (≤5 amino acid residues), 1 long loop (>5 amino acid residues); (P) motif D structure - 3_{10}α-, 1 α-, 2αβ; (Q) position of helix in motif D - 0 normal position, 1 shifted position; (R) D - E motif connection - 0 short (<20 amino acid residues), 1 long structured (<20 amino acid residues); (S) motif E structure - 0 wide, 1 narrow; (T) thumb domain size - 0 large (>180 amino acid residues), 1 small (<180 amino acid residues); (U) priming motif - 0 none, 1 priming loop in thumb subdomain, 2 priming loop in palm subdomain, 3 polymerase C terminal part. Symbols α, β, 3_{10}, and L mean α helix, β strand, 3_{10} helix, and loop, respectively.

FIGURE LEGENDS

Figure 1: Protein structures of selected vRdPs representatives

Nine representatives of the selected vRdPs were chosen. Their structures are shown as a ribbon diagram. All molecules are oriented in the same orientation with finger subdomain on the left, the palm on the bottom and the thumb on the right. The catalytic site is positioned in the centre of each molecule and in some protein structures it is enclosed by the finger tips located at the top of each protein structure. Conserved protein structures typical of vRdPs (homomorphs) are highlighted by colours: violet (hmG), dark blue (hmF), dark green (hmA), light green (hmB), yellow (hmC), orange (hmD) red (hmE), and pink (hmH). Molecular rendering in this figure were created with Swiss PDB Viewer.

Figure 2: Structure based sequence alignment of vRdPs finger subdomain

vRdPs are listed at the beginning of each row by the name of the virus encoding the appropriate vRdP followed by vRdP PBD code. The number at the beginning and at the end of each row indicates the position of the first and last amino acid residue on the appropriate row in the full-length protein bearing polymerase activity (including all additional protein domains). The numbering above the alignment describes position of individual amino acid residues in the alignment. Amino acid residues forming α helices, 3_{10} helices, and β strands are written by red, green, and blue, respectively. Solvent accessible amino acid residues are written in lower case letters; solvent inaccessible by upper case letters. Amino acid residues with positive phi torsion angle, amino acid residues hydrogen bound to main-chain amide, or amino acid residues hydrogen bound to main-chain carbonyl are underlined, written in bold, or in italic, respectively. Most frequent amino acid residues at each alignment position are listed in a row called consensus. Highly conserved positions (more than 80%) are indicated by uppercase violet letters. The 100% conserved amino acid residues are shown by uppercase red letters. Most upper row shows Clustal calculated consensus. Amino acid residues in conserved sequence motifs G and F typical for all vRdPs are highlighted by violet and dark blue colour frames. Amino acid residues it the conserved structural homomorhps hmG and hmF are highlighted the same but lighter colours.

Figure 3: Structure based sequence alignment of vRdPs palm subdomain

Alignment of vRdPs is as in Figure 2. Amino acid residues in conserved sequence motifs F, A, B, and C are highlighted by dark blue, dark green, light green, and yellow frames. Amino acid residues it the conserved structural homomorhps are highlighted the same but lighter colours. The only three 100% conserved amino acid residues in the entire alignment (an arginine residue at position 327 in motif F, an aspartate residue at position 411 in motif, and a glycine residue at position 517 in motif B). The fourth 100% conserved amino acid residue is an aspartate residue in motif C. Despite this aspartate residue is superpostionable in protein structures, it is placed on different position in structure based sequence alignment of protein primary structures thanks to cyclic permutation in IBDV and IPNV RdRPs (see position 397 for birnaviral RdRPs and position 580 for remaining vRdPs).

Figure 4: Structure based sequence alignment of vRdPs thumb subdomain

Alignment of vRdPs is as in Figure 2 and 3. Amino acid residues in conserved sequence motifs D and E are highlighted by orange and red frames. Amino acid residues in the conserved structural homomorhps are highlighted the same but lighter colours. hmH homomorph is highlighted in pink.

Figure 5: Phylogenetic tree of vRdPs evolution

Phylogenetic tree was calculated by an analysis unifying sequence and structure information. Only names of virus species coding vRdPs are listed in the tree. Individual virus species are grouped in genera (blue) and families (red) according actual ICTV virus taxonomy.

SUPPLEMENTARY DATA LEGENDS

Figure S1: Linear organization of protein domains of vRdPs

The vRdP polymerase finger, palm and thumb subdomains are highlighted by blue, green and red. Remaining protein domains are colored by yellow. Conserved sequential and structural features are not shown. Diagram is in scale.

Figure S2: Protein structures of all vRdPs involved in analysis

Molecule positioning is the same as in Figures 1. Polymerase subdomains are highlighted as in the Figure S1: finger subdomain by blue, palm subdomain by green, thumb subdomain by red. Other protein domains are not visible. Molecular rendering in this figure were created with Swiss PDB Viewer.

Figure S3: Phylogenetic tree of vRdPs evolution based only on sequence or structure data

Phylogenetic trees were calculated using only sequence (A) or structure (B) borne information. Only names used for virus species coding vRdPs are listed in the tree.

Table S1: Comparison of hmH and hmE

The RMSD of hmH and hmE were calculated for all individual couples of vRdPs and compared in table. Individual vRdP structures introduced by PBD code-strain are assigned to virus species. Row E shows RMSD values for hmE. Row H shows adequate values for hmH. It is apparent that RMSD values for hmH are comparable with values for hmE and they are often even lower.

TABLES

Table 1: The list of selected vRdPs

Baltimore class	family	genus	virus	abbreviation	viral RNA dependent polymerase				
					PDB	str.	res. [Å]	cocrystallized molecules	citation
+ssRNA viruses	*Caliciviridae*	*Lagovirus*	Rabbit hemorrhagic disease virus	RHEV	1KHV	B	2,5	Lu^{2+}	
		Norovirus	Murine norovirus	MuNORV1	3UQS	A	2	SO_4^{2-}	
			Norovirus	NORV	3BSO	A	1,74	Mg^{2+}, CTP, RNA	
		Sapovirus	Sapporo virus	SappV	2CKW	A	2,3		
	Flaviviridae	*Flavivirus*	Dengue virus 3	DENV3	2J7W	A	2,6	Zn^{2+}, GTP	
			Japanese encephalitis virus	JEV	4K6M	A	2,6	SAH, SO_4^{2-}, Zn^{2+}	
		Hepacivirus	Hepatitis C virus 1	HCV1	1NB6	A	2,6	Mn^{2+}, UTP	
		Pestivirus	Bovine viral diarrhea virus	BVDV1	1S49	A	3	GTP	
	Leviviridae	*Allolevivirus*	Enterobacterio phage Qβ	Qβ	3AVX	A	2,41	Ca2+, 3´dGTP, RNA	
	Picornaviridae	*Aphthovirus*	Foot and mouth disease virus	FMDV	2E9Z	A	3	Mg2+, UTP, PP_i, RNA	
		Enterovirus	Humane rhinovirus 16 A	HuRV16A	1XR7	A	2,3		
			Coxsackie virus B3	CoxVB3	3CDW	A	2,5	PP_i	
			Humane rhinovirus 1B	HuRV1B	1XR6	A	2,5	K^+	
			Poliovirus 1	PolV	3OLB	A	2,41	Zn2+, ddCTP, RNA	
ds RNA viruses	*Birnaviridae*	*Aquabirnavirus*	Infectious pancreatic necrosis virus	IPNV	2YI9	A	2,2	Mg^{2+}	
		Avibirnavirus	Infectious bursal disease virus	IBDV	2PUS	A	2,4		
	Cystoviridae	*Cystovirus*	Pseudomonas phage phi6	Φ6	1HI0	P	3	Mn^{2+}, Mg^{2+}, GTP, DNA	
	Reoviridae	*Orthoreovirus*	Mammalian orthoreovirus 3	MORV3	1N35	A	2,5	Mn2+, 3´dCTP, RNA	
		Rotavirus	Simian rotavirus Sa11	SRV	2R7W	A	2,6	GTP, RNA	
Reverse trans-cribing viruses	*Retroviridae*	*Gammaretrovirus*	Moloney murine leukemia virus	MoMLV	1RW3	A	3		
		Lentivirus	Human immunodeficiency virus 2	HIV2	1MU2	A	2,35	SO_4^{2-}	
			Human immunodeficiency virus 1	HIV1	3V81	C	2,85	nepavirine, DNA	

Table 2: Comparison of structure similarity Z-score of all vRdPs

		DENV	JEV	BVDV1	HCV1	PolV1	HuRV16	HuRV1B	CoxVB3	FMDV	NORV	MuNORV1	RHEV	SappV	Φ6	Qβ	IBDV	IPNV	SRV	MORV3	HIV1	HIV2
		2J7W-A	4K6M-A	1S49-A	1NB6-A	3OLB-A	1XR7-A	1XR6-A	3CDW-A	2E9Z-A	3BSO-A	3UQS-A	1KHV-B	2CKW-A	1HI0-P	3AVX-A	2PUS-A	2YI9-A	2R7W-A	1N35-A	3V81-C	1MU2-A
JEV	4K6M-A	42,9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BVDV1	1S49-A	22,8	21,7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HCV1	1NB6-A	20,5	17,4	27,4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PolV1	3OLB-A	18,1	16,8	25,3	21,5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HuRV16	1XR7-A	18,2	16,6	25,1	20,9	52,4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HuRV1B	1XR6-A	18	16,5	24,8	20,7	52,2	56,7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CoxVB3	3CDW-A	18	16,3	25,2	21	53,1	52,4	53,1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FMDV	2E9Z-A	19,2	17,2	26,5	21,6	41,5	41,3	41	41,6	-	-	-	-	-	-	-	-	-	-	-	-	-
NORV	3BSO-A	20,5	17,5	27,1	23,8	32	32,3	38,1	31,8	32,4	-	-	-	-	-	-	-	-	-	-	-	-
MuNORV1	3UQS-A	20,9	17,7	28	25,2	31,1	31,5	31,2	31,4	32,2	51	-	-	-	-	-	-	-	-	-	-	-
RHEV	1KHV-B	18,7	17,9	27,4	24,3	32,4	33	32,9	33	32,4	39,3	42,7	-	-	-	-	-	-	-	-	-	-
SappV	2CKW-A	17,5	15	24,7	20,6	30,4	30,8	30,8	30,9	30,8	39,1	39,4	43,9	-	-	-	-	-	-	-	-	-
Φ6	1HI0-P	14,8	10,6	4,1	16,4	17,2	17	16,9	17,7	15,7	18,5	19,1	17,7	14,1	-	-	-	-	-	-	-	-
Qβ	3AVX-A	11,1	7,7	14,8	14,1	14	13,5	13,6	14,5	13,8	13,2	14,4	14,9	12,6	12,3	-	-	-	-	-	-	-
IBDV	2PUS-A	8,4	6,6	10,7	9,5	12,1	12,1	11,9	12,6	12,9	13,4	13,3	12,6	12,9	9,5	6	-	-	-	-	-	-
IPNV	2YI9-A	9,8	6,7	13,9	12,9	12,4	12,3	12,1	13	13,5	15,5	14,2	14	13,2	10,7	7,7	42,5	-	-	-	-	-
SRV	2R7W-A	8,9	9	10,2	10,5	9,7	9,4	8,3	8,4	9,3	9,4	9,1	10,4	8,5	9,9	7,8	4,6	4,6	-	-	-	-
MORV3	1N35-A	6,5	4	10,3	7,6	7,8	7,3	7,1	7,8	8,1	7,9	7,9	8,1	8	8,4	8	6,5	6,6	15,4	-	-	-
HIV1	3V81-C	4,7	1,6	6,3	6,5	5,4	5,5	4,9	4,8	5,3	5,5	5,7	5,7	4,9	3,8	5,8	2,8	2,3	4	5,9	-	-
HIV2	1MU2-A	5,4	4	7,9	7,4	6,2	6,6	6,8	6,9	6,1	7,6	7,9	6,5	7,4	5,5	7,7	3,6	4,3	4,6	5,1	28,5	-
MoMLV	1RW3-A	4,7	3,4	7,9	6,2	7,2	7,4	7	6,8	6	7,6	6,8	7,5	7,4	4,9	6,2	2,6	3	4	3,9	18,2	20,7

Table 3: Matrix describing individual features used in phylogenetic analysis of vRdPs

Virus	Family	Genus	PDB ID	Ch.	Features																			
					A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
DENV3	*Flaviviridae*	*Flavivirus*	2J7W	A	0	0	0	0	0	0	N	1	0	0	0	0	2	0	0	0	0	0	0	0	1
JEV	*Flaviviridae*	*Flavivirus*	4K6M	A	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	1
BVDV1	*Flaviviridae*	*Pestivirus*	1S49	A	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	1
HCV1	*Flaviviridae*	*Hepacivirus*	1NB6	A	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	1	0	0	0	0	1
PolV1	*Picornaviridae*	*Enterovirus*	3OLB	A	0	0	1	0	0	0	0	1	2	0	0	0	1	1	0	2	0	0	0	1	0
HuRV16	*Picornaviridae*	*Enterovirus*	1XR7	A	0	0	1	0	0	0	0	1	2	0	0	0	1	1	0	2	0	0	0	1	0
HuRV1B	*Picornaviridae*	*Enterovirus*	1XR6	A	0	0	1	0	0	0	0	1	2	0	0	0	1	1	0	2	0	0	0	1	0
CoxVB3	*Picornaviridae*	*Enterovirus*	3CDW	A	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	2	0	0	0	1	0
FMDV	*Picornaviridae*	*Aphthovirus*	2E9Z	A	0	0	1	0	0	0	0	1	2	0	0	0	1	1	0	2	0	0	0	1	0
NORV	*Caliciviriade*	*Norovirus*	3BSO	A	0	0	1	0	0	0	0	1	2	0	0	0	1	1	0	2	0	0	0	1	0
MuNORV1	*Caliciviriade*	*Norovirus*	3UQS	A	0	0	1	0	0	0	0	1	2	0	0	0	1	1	0	1	0	0	0	1	0
RHEV	*Caliciviriade*	*Lagovirus*	1KHV	B	0	0	1	0	0	0	0	1	1	0	1	0	1	1	0	2	0	0	0	1	0
SappV	*Caliciviriade*	*Sapovirus*	2CKW	A	0	0	1	0	0	0	0	1	2	0	1	0	1	1	0	1	0	0	0	1	0
Φ6	*Cystoviridae*	*Cystovirus*	1HI0	P	0	0	0	0	0	2	1	1	1	0	0	0	2	1	0	2	1	0	1	1	2
Qβ	*Leviviridae*	*Allolevivirus*	3AVX	A	0	0	0	1	0	1	1	1	2	0	0	0	1	0	0	1	0	0	1	1	0
IBDV	*Birnaviridae*	*Avibirnavirus*	2PUS	A	0	0	1	1	1	0	0	1	1	0	0	0	0	1	0	2	0	1	0	1	0
IPNV	*Birnaviridae*	*Aquabirnavirus*	2YI9	A	0	0	1	1	1	0	0	1	1	0	0	0	0	1	0	2	0	1	0	1	0
SRV	*Reoviridae*	*Rotavirus*	2R7W	A	0	0	0	0	0	1	2	1	1	0	0	0	0	1	1	2	0	0	1	1	3
MORV3	*Reoviridae*	*Orthoreovirus*	1N35	A	0	0	0	0	0	1	2	1	1	1	1	1	2	1	1	2	0	0	1	1	3
HIV1	*Retroviridae*	*Lentivirus*	3V81	C	1	1	2	1	0	1	2	0	2	2	0	1	0	1	0	1	0	0	1	1	0
HIV2	*Retroviridae*	*Lentivirus*	1MU2	A	1	1	2	1	0	1	2	0	2	2	0	1	0	1	0	1	0	0	1	1	0
MoMLV	*Retroviridae*	*Gammaretroviru*	1RW3	A	1	1	2	1	0	1	2	0	2	2	0	1	0	1	0	1	0	0	1	1	0

FIGURES

Figure 1: Protein structures of selected vRdPs representatives

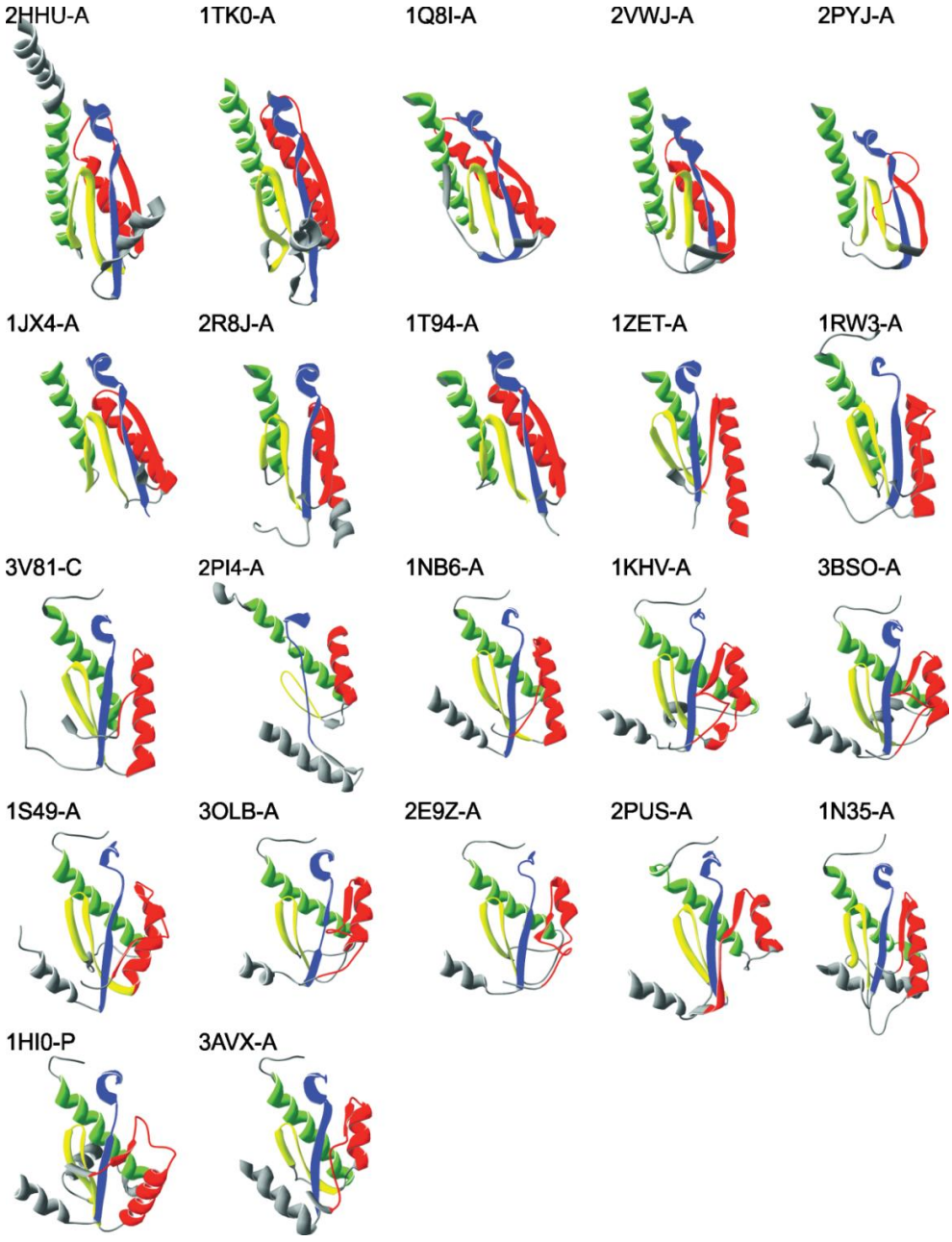


Figure 2: Structure based sequence alignment of vRdPs finger subdomain

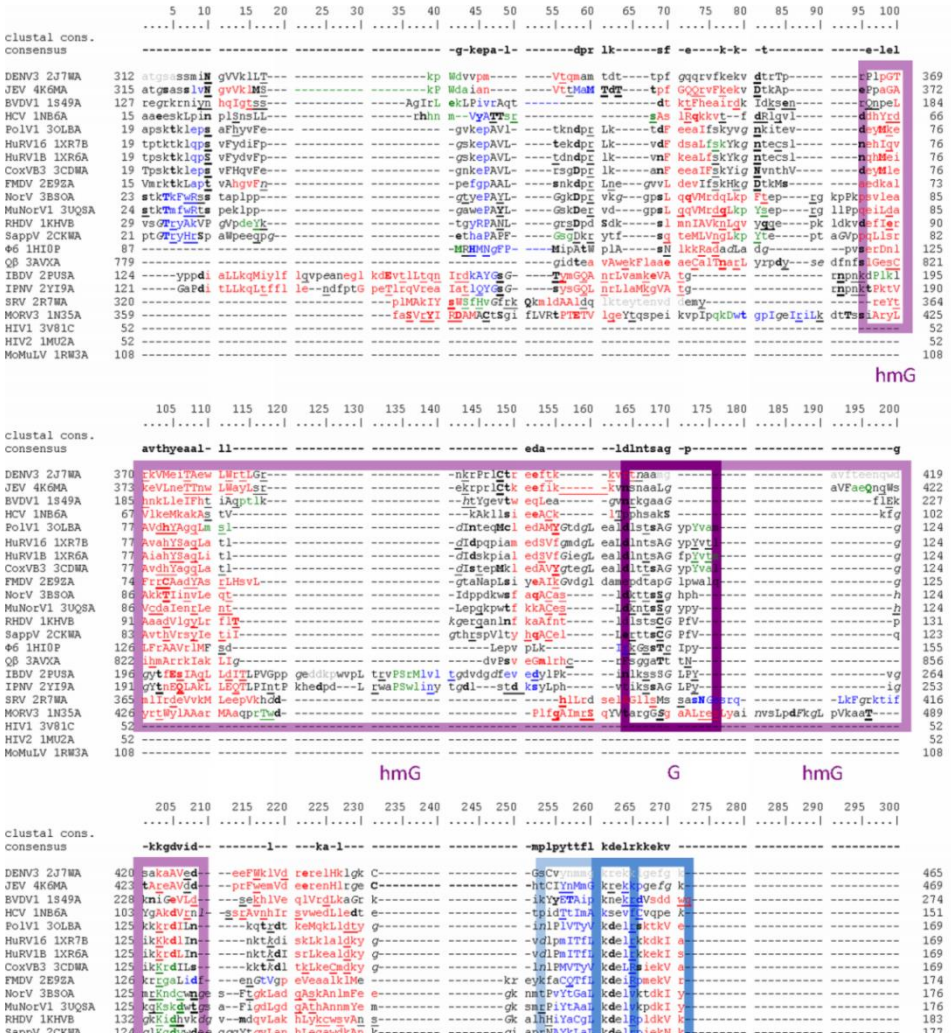


Figure 3: Structure based sequence alignment of vRdPs palm subdomain

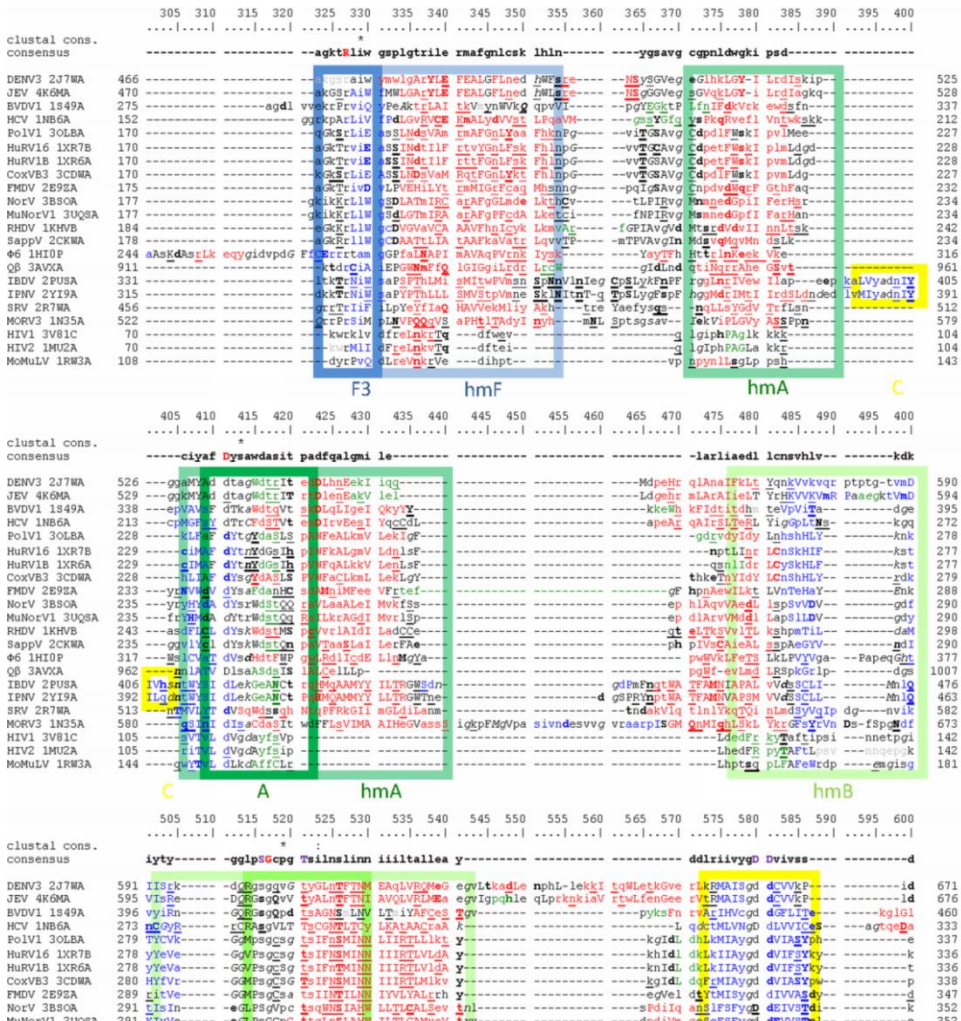
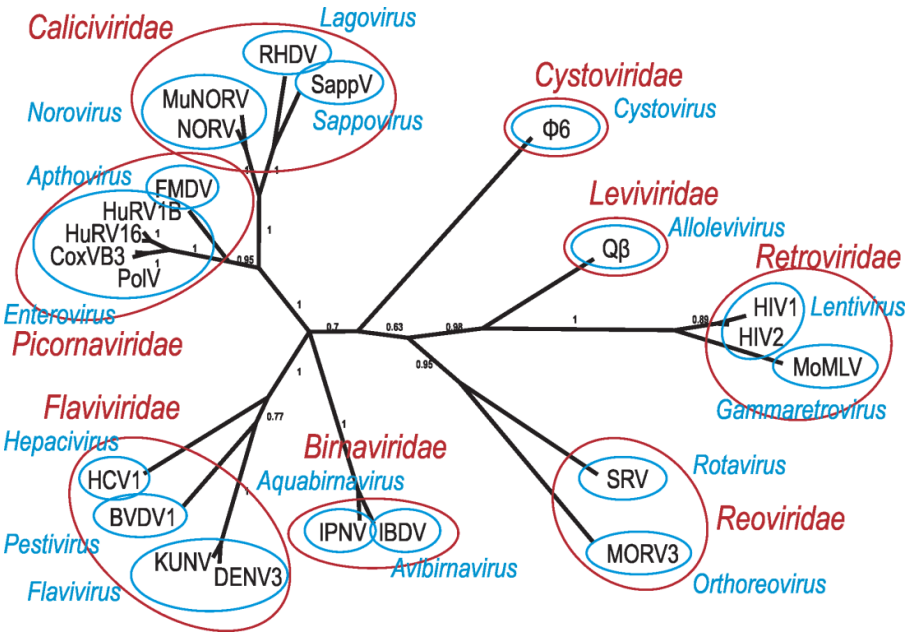


Figure 4: Structure based sequence alignment of vRdPs thumb subdomain

Figure 5: Phylogenetic tree of vRdPs evolution



SUPPLEMENTARY DATA

Figure S1: Linear organization of protein domains of vRdPs

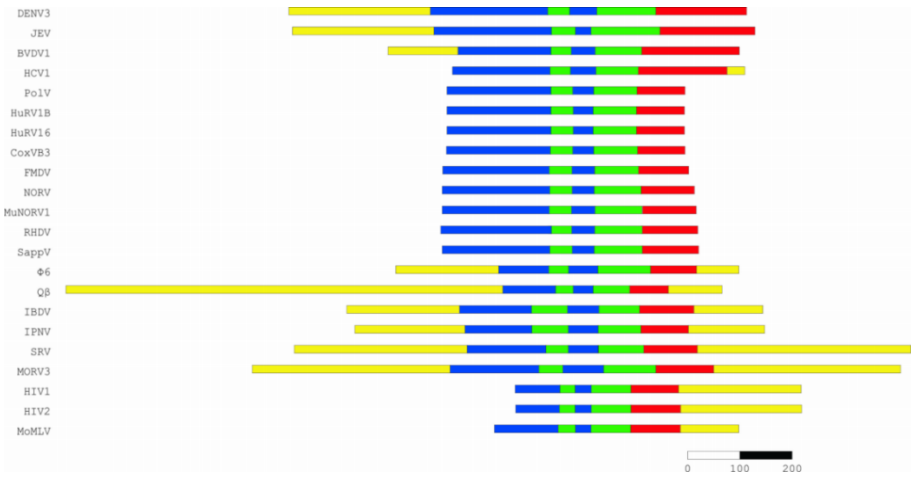
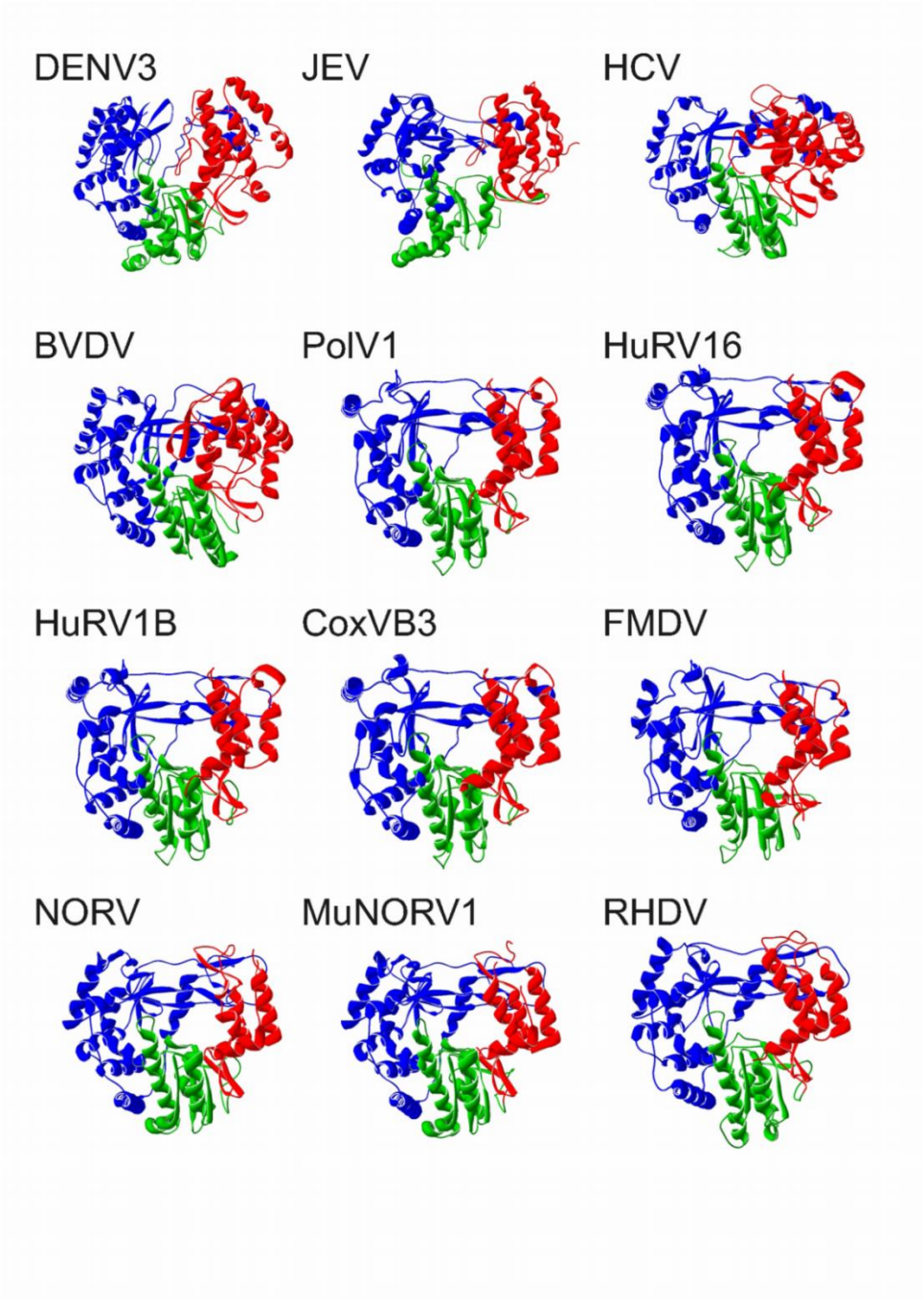


Figure S2: Protein structures of all vRdPs involved in analysis



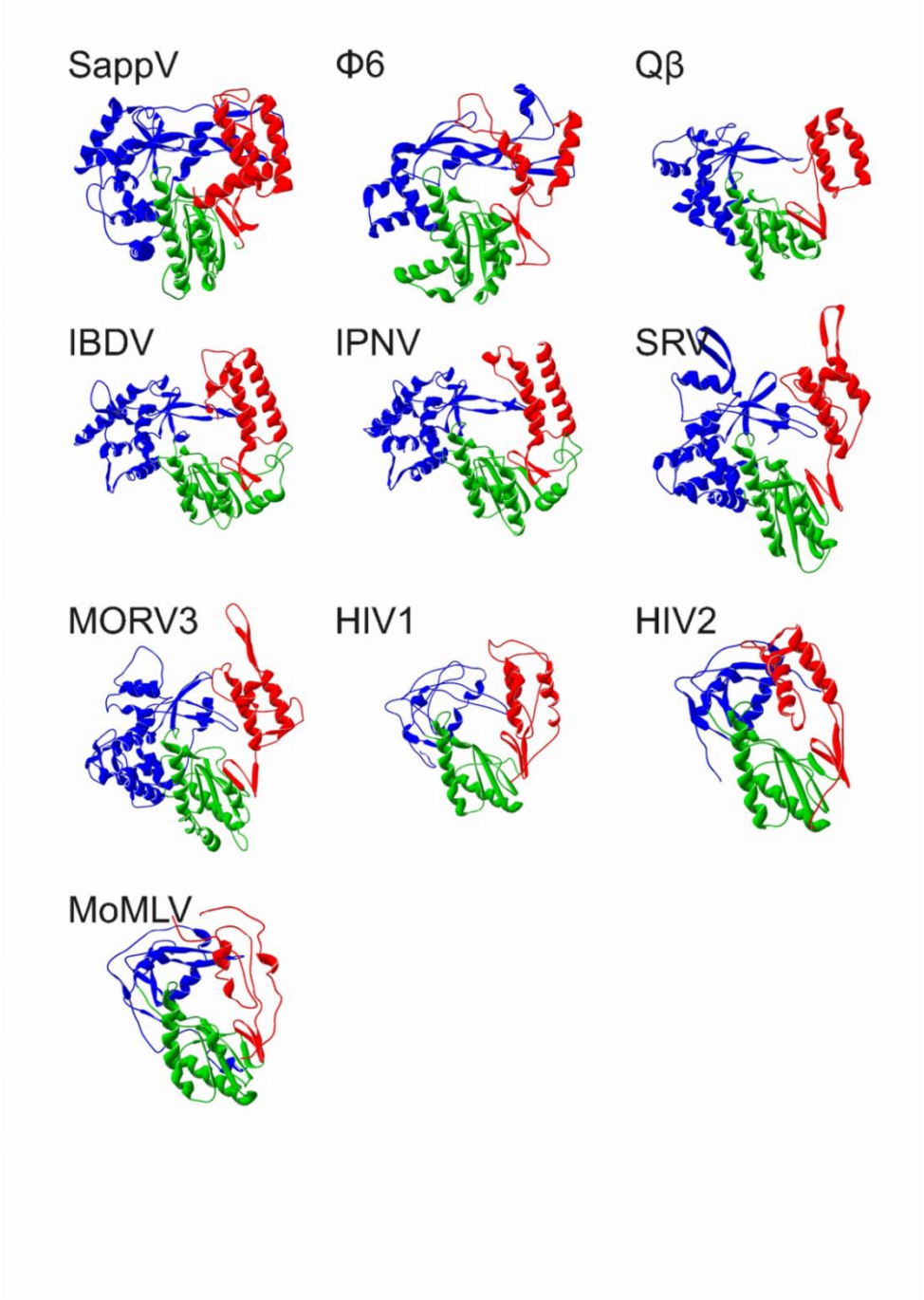

SappV
Φ6
Qβ
IBDV
IPNV
SRV
MORV3
HIV1
HIV2
MoMLV

Figure S3: Phylogenetic tree of vRdPs evolution based only on sequence or structure data

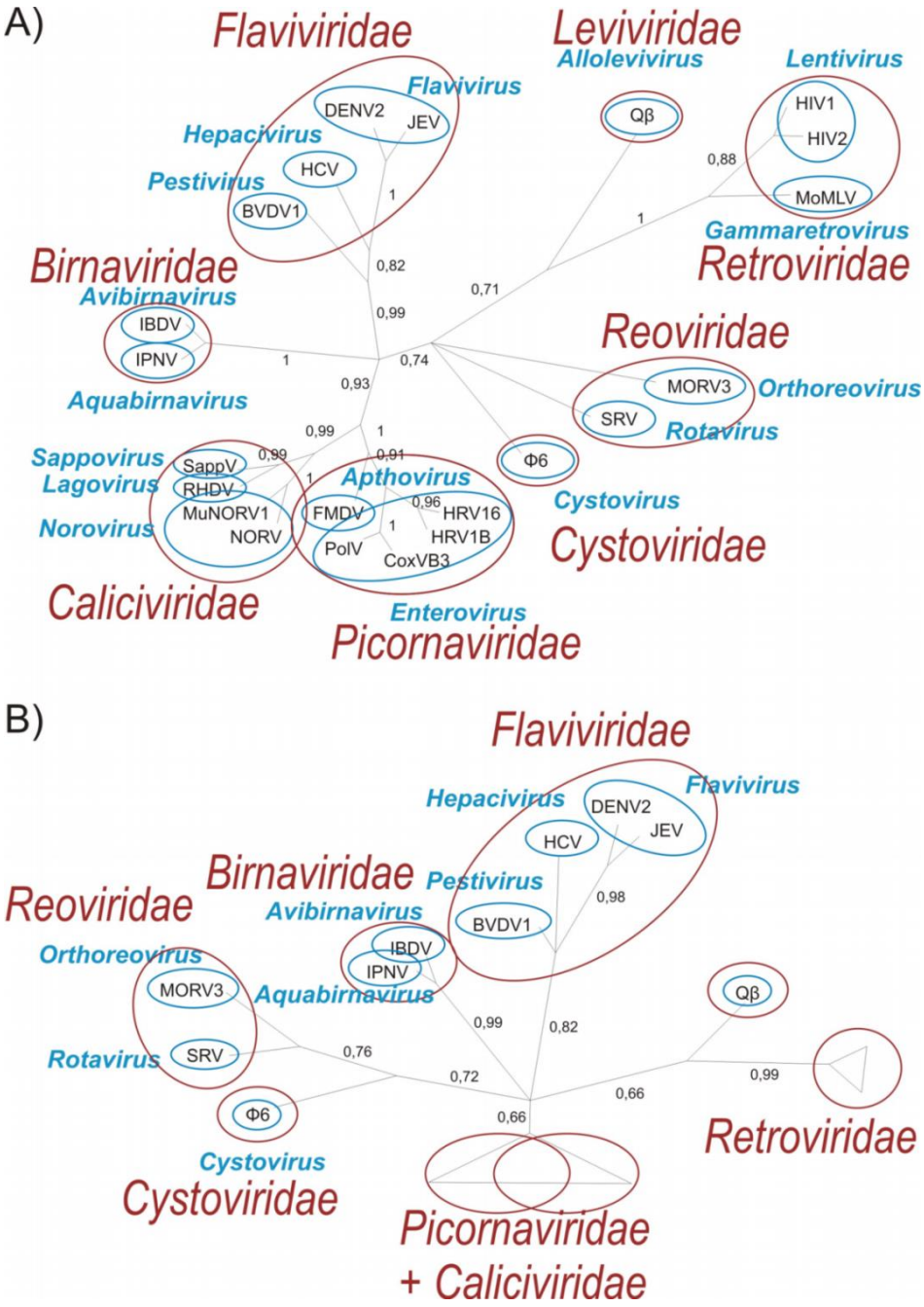


Table S1: Comparison of hmH and hmE

			DENV	JEV	BVDV1	HCV1	PoIV1	HuRV16	HuRV1B	CoxVB3	FMDV	NORV	MuNORV1	RHEV	SappV	Φ6	Qβ	IBDV	IPNV	SRV	MORV3	HIV1	HIV2
			2I7W-A	4K6M-A	1S49-A	1NB6-A	3OLB-A	1XR7-A	1XR6-A	3CDW-A	2E9Z-A	3BSO-A	3UQS-A	1HKV-B	2CKW-A	1HI0-P	3AVX-A	2PUS-A	2YI9-A	2R7W-A	1N35-A	3V81-C	1MU2-A
JEV	4K6M-A	E	0.4 (36)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	0.2 (32)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BVDV1	1S49-A	E	1.1 (36)	1.2 (36)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	0.9 (32)	0.9 (32)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HCV1	1NB6-A	E	2.6 (36)	1.6 (36)	1.6 (36)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	2.6 (31)	2.7 (31)	2.8 (31)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PoIV1	3OLB-A	E	2.0 (34)	3.0 (34)	1.9 (34)	1.7 (34)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	3.2 (33)	3.2 (33)	3.8 (33)	2.5 (33)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HuRV16	1XR7-A	E	1.6 (34)	1.5 (34)	1.6 (34)	1.6 (34)	0.6 (34)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	3.1 (33)	3.2 (33)	3.8 (33)	2.5 (33)	0.4 (33)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HuRV1B	1XR6-A	E	1.8 (34)	1.9 (34)	2.0 (34)	1.7 (34)	0.5 (34)	0.4 (34)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	3.3 (33)	3.1 (33)	3.8 (33)	2.5 (33)	0.4 (33)	0.3 (33)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CoxVB3	3CDW-A	E	2.3 (34)	2.8 (34)	1.9 (34)	1.8 (34)	0.32 (34)	0.7 (34)	0.7 (34)	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	3.1 (33)	3.1 (33)	3.7 (33)	2.4 (33)	0.3 (33)	0.3 (33)	0.4 (33)	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FMDV	2E9Z-A	E	1.4 (34)	1.3 (34)	1.5 (34)	1.7 (34)	1.2 (34)	1.3 (34)	1.4 (34)	1.2 (34)	-	-	-	-	-	-	-	-	-	-	-	-	-
		H	3.1 (27)	3.0 (27)	4.0 (27)	2.9 (27)	2.1 (27)	2.1 (27)	2.2 (27)	2.1 (27)	-	-	-	-	-	-	-	-	-	-	-	-	-
NORV	3BSO-A	E	3.0 (31)	3.6 (31)	4.4 (31)	2.0 (31)	2.9 (31)	1.8 (31)	2.3 (31)	3.2 (31)	1.8 (31)	-	-	-	-	-	-	-	-	-	-	-	-
		H	3.5 (32)	3.5 (32)	3.2 (32)	2.8 (32)	1.5 (32)	1.5 (32)	1.5 (32)	1.5 (32)	2.5 (32)	-	-	-	-	-	-	-	-	-	-	-	-
MuNORV	3UQS-A	E	2.2 (31)	1.9 (31)	2.8 (31)	2.1 (31)	2.9 (31)	2.8 (31)	1.6 (31)	2.8 (31)	1.3 (31)	1.2 (31)	-	-	-	-	-	-	-	-	-	-	-
		H	2.8 (32)	2.8 (32)	2.4 (32)	2.8 (32)	1.5 (32)	1.4 (32)	1.5 (32)	1.4 (32)	2.4 (32)	1.1 (32)	-	-	-	-	-	-	-	-	-	-	-
RHEV	1HKV-B	E	2.4 (31)	3.5 (31)	2.9 (31)	3.1 (31)	3.0 (31)	3.3 (31)	3.4 (31)	2.4 (31)	1.5 (31)	1.5 (31)	1.6 (31)	-	-	-	-	-	-	-	-	-	-
		H	2.7 (32)	2.7 (32)	2.7 (32)	2.4 (32)	1.3 (32)	1.4 (32)	1.5 (32)	1.3 (32)	2.3 (32)	1.7 (32)	1.1 (32)	-	-	-	-	-	-	-	-	-	-
SappV	2CKW-A	E	2.1 (30)	3.8 (30)	3.4 (30)	2.3 (30)	2.9 (30)	1.9 (30)	3.4 (30)	3.2 (30)	3.4 (30)	1.9 (30)	2.5 (30)	1.2 (30)	-	-	-	-	-	-	-	-	-
		H	2.9 (32)	2.8 (32)	3.4 (32)	2.6 (32)	1.5 (32)	1.5 (32)	1.6 (32)	1.5 (32)	2.3 (32)	2.7 (32)	0.7 (32)	1.67 (32)	-	-	-	-	-	-	-	-	-
Φ6	1HI0-P	E	1.8 (33)	3.3 (33)	3.8 (33)	3.9 (33)	2.1 (33)	3.6 (33)	2.9 (33)	2.0 (33)	3.3 (33)	2.6 (33)	2.7 (33)	3.1 (33)	2.9 (33)	-	-	-	-	-	-	-	-
		H	3.7 (32)	3.7 (32)	3.9 (32)	3.9 (32)	3.4 (32)	3.1 (32)	3.1 (32)	3.2 (32)	2.6 (32)	3.7 (32)	3.9 (32)	4.0 (32)	4.1 (32)	-	-	-	-	-	-	-	-
Qβ	3AVX-A	E	2.1 (26)	2.4 (26)	2.6 (26)	3.0 (26)	3.0 (26)	1.8 (26)	2.1 (26)	3.0 (26)	2.4 (26)	2.1 (26)	2.9 (26)	2.2 (26)	2.4 (26)	3.0 (26)	-	-	-	-	-	-	-	-
		H	0.5 (35)	0.5 (35)	0.9 (35)	0.9 (35)	0.6 (35)	0.7 (35)	0.6 (35)	0.6 (35)	0.4 (35)	2.7 (35)	0.6 (35)	0.7 (35)	0.6 (35)	0.4 (35)	-	-	-	-	-	-	-	-
IBDV	2PUS-A	E	2.8 (32)	2.9 (32)	3.2 (32)	2.7 (32)	2.0 (32)	2.2 (32)	2.2 (32)	2.1 (32)	2.1 (32)	2.3 (32)	2.1 (32)	3.0 (32)	2.9 (32)	3.1 (32)	2.6 (32)	-	-	-	-	-	-	-
		H	1.3 (32)	1.3 (32)	1.1 (32)	2.6 (32)	3.3 (32)	3.1 (32)	3.2 (32)	3.1 (32)	2.5 (32)	2.8 (32)	2.3 (32)	2.3 (32)	2.0 (32)	4.6 (32)	0.5 (32)	-	-	-	-	-	-	-
IPNV	2YI9-A	E	3.2 (33)	3.1 (33)	3.0 (33)	2.7 (33)	2.1 (33)	2.1 (33)	2.2 (33)	2.1 (33)	2.0 (33)	2.7 (33)	2.0 (33)	2.7 (33)	2.9 (33)	3.0 (33)	3.0 (33)	0.5 (33)	-	-	-	-	-	-
		H	1.4 (32)	1.3 (32)	1.2 (32)	2.6 (32)	2.9 (32)	3.0 (32)	3.0 (32)	2.8 (32)	2.6 (32)	2.8 (32)	2.4 (32)	2.3 (32)	2.0 (32)	4.5 (32)	0.6 (32)	0.5 (32)	-	-	-	-	-	-
SRV	2R7W-A	E	3.6 (25)	3.5 (25)	3.3 (25)	3.8 (25)	2.9 (25)	2.6 (25)	2.9 (25)	3.0 (25)	3.8 (25)	2.5 (25)	2.7 (25)	2.6 (25)	2.9 (25)	3.2 (25)	2.1 (25)	3.4 (25)	2.8 (25)	-	-	-	-	-
		H	2.3 (38)	2.4 (38)	2.8 (38)	2.8 (38)	2.6 (38)	3.0 (38)	2.6 (38)	2.6 (38)	2.3 (38)	2.6 (38)	2.5 (38)	2.7 (38)	2.4 (38)	3.8 (38)	0.5 (38)	2.7 (38)	2.7 (38)	-	-	-	-	-
MORV3	1N35-A	E	1.9 (26)	2.5 (26)	3.3 (26)	3.1 (26)	2.2 (26)	3.4 (26)	2.2 (26)	2.0 (26)	2.4 (26)	2.1 (26)	2.0 (26)	2.0 (26)	1.9 (26)	2.5 (26)	2.2 (26)	2.5 (26)	2.6 (26)	1.8 (26)	-	-	-	-
		H	2.5 (36)	2.5 (36)	2.7 (36)	2.7 (36)	2.0 (36)	2.1 (36)	2.0 (36)	2.0 (36)	2.6 (36)	2.5 (36)	2.7 (36)	2.2 (36)	2.2 (36)	3.9 (36)	0.5 (36)	2.3 (36)	2.2 (36)	2.2 (36)	-	-	-	-
HIV1	3V81-C	E	3.1 (23)	2.1 (23)	2.6 (23)	2.4 (23)	3.2 (23)	3.8 (23)	3.4 (23)	3.0 (23)	3.1 (23)	2.0 (23)	1.7 (23)	1.9 (23)	2.6 (23)	1.9 (23)	1.8 (23)	2.4 (23)	2.3 (23)	2.1 (23)	1.4 (23)	-	-	-
		H	1.8 (29)	1.7 (29)	2.0 (29)	2.7 (29)	3.0 (29)	2.7 (29)	2.8 (29)	2.6 (29)	3.6 (29)	2.6 (29)	2.0 (29)	2.3 (29)	2.8 (29)	3.4 (29)	0.3 (29)	2.1 (29)	2.1 (29)	1.9 (29)	2.5 (29)	-	-	-
HIV2	1MU2-A	E	2.0 (23)	3.1 (23)	2.6 (23)	2.6 (23)	3.0 (23)	3.5 (23)	2.7 (23)	3.1 (23)	2.8 (23)	2.2 (23)	2.1 (23)	2.4 (23)	3.0 (23)	2.3 (23)	2.1 (23)	2.2 (23)	2.1 (23)	2.6 (23)	1.6 (23)	2.0 (23)	-	-
		H	1.9 (29)	1.7 (29)	2.2 (29)	2.8 (29)	2.9 (29)	2.7 (29)	2.7 (29)	2.9 (29)	3.8 (29)	2.6 (29)	2.0 (29)	2.7 (29)	2.8 (29)	3.2 (29)	0.2 (29)	2.1 (29)	2.2 (29)	2.1 (29)	2.6 (29)	0.6 (29)	-	-
MoMLV	1RW3-A	E	3.5 (23)	2.1 (23)	1.6 (23)	1.8 (23)	2.3 (23)	2.8 (23)	3.1 (23)	2.2 (23)	2.8 (23)	1.7 (23)	1.4 (23)	1.8 (23)	2.4 (23)	2.2 (23)	2.0 (23)	1.4 (23)	1.3 (23)	1.8 (23)	2.1 (23)	2.2 (23)	2.4 (23)	-
		H	3.1 (29)	2.0 (29)	2.8 (29)	2.7 (29)	2.0 (29)	2.9 (29)	2.9 (29)	2.0 (29)	1.7 (29)	1.8 (29)	2.1 (29)	3.1 (29)	2.1 (29)	3.0 (29)	1.7 (29)	2.5 (29)	2.4 (29)	2.7 (29)	1.3 (29)	1.0 (29)	1.8 (29)	-

6.2 Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients

The article was published in Tick and Tick-borne diseases and should be cited as Petra Formanová; **Jiří Černý**; Barbora Černá Bolfíková; James J. Valdés; Irina Kozlova; Yuri Dzhioev; Daniel Růžek: Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients. Ticks and Tick-Borne Diseases, 2015, 6(1):38-46. doi:10.1016/j.ttbdis.2014.09.002

Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients

Petra Formanová,[1,2] Jiří Černý,[3,4] Barbora Černá Bolfíková,[5] James J. Valdés,[3] Irina Kozlova,[6,7] Yuri Dzhioev,[6,7] and Daniel Růžek[1,2,3,4]*

(1) Department of Virology, Veterinary Research Institute, Hudcova 70, CZ-62100 Brno, Czech Republic

(2) Faculty of Science, Masaryk University, Kotlářská 267/2, CZ-61137 Brno, Czech Republic

(3) Institute of Parasitology, Biology Centre of the Academy of Sciences of the Czech Republic, Branišovská 31, CZ-37005 České Budějovice, Czech Republic

(4) Faculty of Science, University of South Bohemia, Branišovská 31, CZ-37005 České Budějovice, Czech Republic

(5) Faculty of Tropical AgriSciences, Czech University of Life Sciences Prague; Kamýcká 126, CZ-16521 Prague, Czech Republic

(6) Institute of Biomedical Technology, Irkutsk State Medical University of Russian Ministry of Health, Krasnogo Vosstanija 1, Irkutsk, 664003, Russia

(7) FSSFE Scientific Centre of Family Health and Human Reproduction Problems, Siberian Branch of the Russian Academy of Medical Sciences, Timirjazeva Street 16, 664003 Irkutsk, Russia

*Author for Correspondence
Daniel Růžek

Department of Virology
Veterinary Research Institute
Hudcova 70
CZ-62100 Brno
Czech Republic
e-mail: ruzekd@paru.cas.cz
phone: +420-5-3333-1101
fax: +420-5-4121-1229

Abstract:

Tick-borne encephalitis virus (TBEV) causes tick-borne encephalitis (TBE), one of the most important human neuroinfections across Eurasia. Up to date, only three full genome sequences of human European TBEV isolates are available, mostly due to difficulties with isolation of the virus from human patients. Here we present full genome characterization of an additional five low-passage TBEV strains isolated from human patients with severe forms of TBE. These strains were isolated in 1953 within Central Bohemia in the former Czechoslovakia, and belong to the historically oldest human TBEV isolates in Europe. We demonstrate here that all analyzed isolates are distantly phylogenetically related, indicating that the emergence of TBE in Central Europe was not caused by one predominant strain but rather a pool of distantly related TBEV strains. Nucleotide identity between individual sequenced TBEV strains ranged from 97.5 to 99.6% and all strains shared large deletions in the 3´ non-coding region, which has been recently suggested to be an important determinant of virulence. The number of unique amino acid substitutions varied from 3 to 9 in individual isolates, but no characteristic amino acid substitution typical exclusively for all human TBEV isolates was identified when compared to the isolates from ticks. We did, however, correlate that the exploration of the TBEV envelope glycoprotein by specific antibodies were in close proximity to these unique amino acid substitutions. Taken together, we report here the largest number of patient-derived European TBEV full genome sequences to date and provide a platform for further studies on evolution of TBEV since the first emergence of human TBE in Europe.

Key words: tick-borne encephalitis virus; tick-borne encephalitis; genome analysis; human patients

Introduction

Tick-borne encephalitis (TBE) is the most important arboviral infection in Europe and Central and Eastern Asia. More than 13,000 human TBE cases are reported annually (Mansfield et al., 2009). The disease is caused by tick-borne encephalitis virus (TBEV), a member of the genus *Flavivirus*, family *Flaviviridae* (Mansfield et al., 2009).

TBEV is an enveloped virus with approximately 11kb long single-stranded RNA genome of positive polarity. The genomic RNA contains one open reading frame (ORF) encoding single polyprotein. It is co- and post-translationally cleaved by viral and host proteases into three structural (capsid (C), membrane (M; derived from its precursor, prM) and envelope (E)) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B and NS5) (Monath and Heinz, 1996; Rice, 1996). Structural proteins are responsible for packaging of virus genome and budding of viral capsids through cellular membranes. Non-structural proteins catalyze replication of viral genome and regulate host-antiviral response.

The main ORF is flanked with 5' and 3' non-coding regions (NCRs). The 5' NCR has a length of approximately 100 bp and is relatively homogenous on both size and sequence. The 3' NCR is extremely heterogeneous in length (751 bp in TBEV strain Neudoerfl, 445nt in TBEV strain Hypr) (Wallner et al., 1996). Rarely, the 3' NCR of some TBEV strains contains a shorter poly(A) tail (Wallner et al., 1996; Frey et al., 2014). Both NCRs contain conserved secondary structures that are supposed to be involved in TBEV genome amplification, translation and packaging (Gritsun et al., 1997).

Based on phylogenetic analysis, TBEV can be divided onto three subtypes: the European subtype (Eu-TBEV), the Siberian subtype (S-TBEV), and the Far Eastern subtype (FE-TBEV) (Ecker et al., 1999). Members of these three subtypes differ in their geographical distribution, virulence, and clinical severity of caused diseases (Mansfield et al., 2009).

Although the medical and economic impact of TBE is high, the TBEV strains isolated from patients remain largely unstudied and only a few complete genome sequences of human Eu-TBEV strains have been reported until now. This paucity is caused by the difficulty in obtaining TBEV isolates from humans –

the virus can be isolated from blood during the first (nonspecific) phase of the infection or from *post mortem* brain tissue. During the neurological phase of the infection, the virus is rarely present in the blood or the cerebrospinal fluid of the patients (Růžek et al., 2010) and most isolation attempts are usually unsuccessful.

Almost all Eu-TBEV strains with known genome sequence were isolated from ticks or rodents. However, analysis of complete nucleotide sequences of strains isolated from patients with variable disease severities is crucial for detection of mutations in the TBEV genome that determine the pathogenicity for humans (Belikov et al., 2014). Currently, only three complete Eu-TBEV genome sequences are available, which were isolated from human patients. Strain "Hypr" was isolated in 1953 from the blood of a diseased young boy in Czechoslovakia (Pospíšil 90 et al., 1954; Wallner et al., 1996). Strain "Est3476" was obtained from a serum sample of patient from Estonia (Golovljova et al., 2004). Finally, strain "Ljubljana 1" was isolated in 1992 from blood of a TBE patient from Slovenia (Fajs et al., 2012). The largest set of European patient-derived TBEV sequences was provided by analysis of E gene sequences of 15 strains and NS5 gene sequences of 17 strains (Fajs et al., 2012).

Recently, a comparison of 34 genomes of FE-TBEV strains isolated from patients with different disease severities identified specific mutations responsible for differences in pathogenicity of FE97 TBEV strains (Leonova et al., 2014; Belikov et al., 2014). However, there are large differences in sequence of FE-TBEV and Eu-TBEV that also underlines a need of analysing patient-derived Eu-TBEV complete genomes.

The TBEV of Central Europe was first isolated in 1948 in the former Czechoslovakia (Krejčí, 1949; Gallia et al., 1949). The TBEV strains analyzed in this study belong, therefore, together with other strains from the late 1940s and early 1950s, are the oldest human TBEV isolates in Europe. Here, we report a total of five full genome sequences from patient-derived European TBEV strains to date. We also provide a platform to further analyse TBEV evolution and its antigenic properties since the first TBEV emergence in Europe.

Material and Methods

Five archival low-passage TBEV strains were selected for the full genome sequence analysis. These strains were isolated from the blood of patients hospitalized with TBEV infection during the TBEV outbreak in 1953 in Central Bohemia (Czechoslovakia). All patients had severe course of the TBE. RNA was isolated from 20% suckling mouse brain suspension using QIAamp Viral RNA Mini Kit (Qiagen). Reverse transcription was performed using ProtoScript® First Strand cDNA Synthesis Kit (New England Biolabs). The 35 overlapping DNA fragments were produced by PPP Master Mix (Top-Bio, sequence of primers is available on request) as described previously (Růžek et al., 2008). The PCR products were then sequenced directly by commercial service (SEQme, Czech Republic). The deduced whole genome sequences were deposited in the GenBank database under accession numbers: KJ922512-KJ922516. Both nucleotide and deduced amino acid sequences were analysed using BioEdit Sequence Alignment Editor, version 7.2.0 (Hall, 1999) and MultAlin (Corpet, 1988), aligned by Muscle in MEGA version 5 (Tamura et al., 2007). For complete sequence comparisons we used 60 complete genomes of TBEV together with Turkish sheep encephalitis virus (TSEV; GenBanki Accession Number: DQ235151.1), Spanish sheep encephalitis virus (SSEV; DQ235152.1) and Louping ill virus (LIV; Y07863.1) deposited in GenBank database. For detection of selection pressure acting on individual genes we calculated the ratios of non-synonymous and synonymous nucleotide substitutions per site (dN/dS) of the available TBEV sequences using MEGA 124 version 5 (Tamura et al., 2007).

The predicted secondary structure of the NCRs were produced using Mfold server (http://mfold.rna.albany.edu) under default conditions.

Best fitting model of nucleotide substitutions was tested in jModelTest (Darriba et al 2012). The general time reversible (GTR) model was selected as the best fitting model. Bayesian phylogenetic analysis was performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003). Bayesian analysis consisted of two runs with four chains (one cold and three heated), and was run for 10 million generations sampled every 500 generations. The first 25% of samples were discarded as a burning period. The average standard deviation of split frequencies was 0.001 showing convergence of all chains.

We used 1SVB to depict structure of TBEV protein E (Rey et al 1995). Structures of proteins NS1, NS3, and NS5 were modelled by homology modeling on Phyre2

server (Kelley and Sternberg 2009) and proteins were modelled according 4O6C (Akey et al 2014), 2VBC (Luo et al 2008), and 4K6M (Lu and Gong 2013) templates. Molecular rendering was done using PDB Swiss Viewer (Guex and Peitsch 1997). The TBEV protein E crystal structure and predicted models were prepared and refined by adding hydrogen atoms, optimization of the hydroden-bond network, followed by a full minimization of the system to remove steric clashes (i.e., overlapping atoms) using the Schrodinger's Maestro software (Li et al. 2007). The prepared structures were then submitted to the ElliPro server to predict epitope(s) position(s). The ElliPro server uses the tertiary structures to predict epitope regions based on their particular scoring function (Ponomarenko et al., 2008). For the antibody-antigen docking we used the SwarmDock server (Torchala et al. 2013a, 2013b; Torchala and Bates 2014) that incorporates flexible protein-protein docking by exploring around the Cartesian center of mass of the receptor (the antigen) and including minimization steps for the whole system. Once energy favorable poses are generated they minimized again sent to the user.

Results and Discussion

We have sequenced and analyzed the complete genomes of 5 Eu-TBEV strains Skrivanek, Petracova, Vlasaty, Tobrman and Kubinova isolated from patients with severe TBE in 1953. Nucleotide identity between individual sequenced TBEV strains ranged 97.5% – 99.6%. All isolates, therefore, represent unique strains, although they were isolated during the same season and in the same geographic region. The length of the nucleotide sequence of the genomes ranged from 10,777 to 10,979 nucleotides. The differences in genome length were due to the variable length of the 3' NCR. The ORF of all isolates were of standard length (10,245 nt). Nucleotide identities were between 97.5% – 97.7% for TBEV strain Neudoerfl and 97.3% – 97.5% for TBEV strain 159 Hypr. Amino acid identities were around 99.1% with TBEV strain Neudoerfl and 98.8 – 98.9% with the strain Hypr (Table 1).

Phylogenetic relationship was established on the basis of full genome sequences including both NCRs. The results showed that all newly sequenced TBEV strains are representatives of Eu-TBEV subtype but they do not form a monophyletic group despite that they were isolated during one season and in the same region (Fig. 1). The TBEV strain Vlasaty was most closely related to

TBEV strains K23, a tick-derived TBEV strain originating near Karlsruhe (Baden-Württemberg, Germany). The strains Tobrman and Petracova formed a monophyletic group related to the strain Neudoerfl isolated from ticks close to same name village in Burgenland (Austria). TBEV strain Skrivanek cladded with TBEV strain Hypr that was isolated from a human patient from Moravia (Czech Republic). TBEV strain Kubinova clustered together with TBEV strain AS33 isolated from a tick in Bavaria (Germany). The only fact that from our TBEV phylogeny based on full genome sequences is a slight tendency to group on the base of geographical location. TBEV strains originating from central Europe form a basal group, from which the strains isolated in northeastern Europe, southeastern Europe, and South Korea diverged. This is in concordance with recent theory formulated on phylogenetic comparison of large E gene dataset (Weidmann et al., 2011, 2013).

The genomic 5‘ NCRs of all five isolates were conserved in length and had 132 nucleotides, the same length as in the majority of Eu-TBEV strains. The heterogeneity in 5’ NCR was between 0 – 4.9%, among newly sequenced TBEV strains. The sequence identity in 5’ NCR among the newly sequenced TBEV strains and TBEV strain Neudoerfl varyied between 96.8 – 99.2%. The sequence identity to strain Hypr was much lower varying between 95.1 – 96.8%. The 5‘ NCR positions 79-132 were completely conserved in all of the analyzed strains, with no nucleotide substitutions. Prediction of the 5‘ NCR of the analyzed strains revealed only some minor insignificant differences in the conformations of the 2D structure and we could not identify any substitutions attributed to the higher pathogenicity of the strains used in this study in comparison to strains isolated from ticks (not shown).

The genomic 3’ NCR is heterogeneous in length, ranging from 403 to 620 nucleotides in different strains, depending on the length of deletions (Figure 2). All of the newly sequenced TBEV isolates lacked poly(A) region (Fig. 2). The largest deletion was observed in the 3’ UTR of the strain Kubinova and this deletion encompassed virtually the whole variable part of the 3’ UTR (Fig. 2). The deletions in 3’ NCR represent the major difference between TBEV strains isolated from humans or other vertebrates and ticks. However, the observed deletions had no significant effect on the 2D topology of the conserved loops formed by the conserved terminal part of 3’ UTR (not shown). The origin of heterogeneity in the 3´ NCR was discussed to be associated with virus

propagation *in vitro*, as well as polymerase stumbling across extensive secondary structures of the viral RNA (Frey et al., 2014; Mandl et al., 1998). Some studies, in which the poly(A) or the whole 194 variable part of 3' UTR was abridged or removed, came to the conclusions that these variations do not have significant effects on virus properties (Mandl et al., 1998). It was then demonstrated more recently that deletions in the variable 3' NCR can represent a critical virulence factor enhancing virus multiplication and pathogenicity in the mouse brain (Sakai et al., 2014). Large deletions in the 3' UTR, including extensive deletions covering almost the entire 3' NCR were reported in TBEV strains isolated from patients in the Far East (Belikov et al., 2014). In our previous study, an attenuated TBEV strain (263), isolated from field ticks, was either serially subcultured, 5 times in mice, or at 40 °C in PS cells, producing 2 independent strains, 263-m5 and 263-TR with identical genomes; both strains exhibited increased plaque size, neuroinvasiveness and temperature-resistance. Sequencing revealed two unique amino acid substitutions located in NS2B and NS3 genes, but also large deletion in the 3' NCR in comparison to the parental attenuated strain (Růžek et al., 2008). With respect to recent observations, we hypothesize that in addition to the mutations in the NS2B and NS3 also the deletion in 3' NCR contributed to increased neuroinvasiveness of the 263-m5 and 263-TR strains (Růžek et al., 2008). Based on all data available, the presence of extended deletions in the 3' NCR seems to be a

common feature of highly virulent TBEV strains and that the full-length 3' NCR is significant for the survival of TBEV in tick cells (Wallner et al., 1995; Růžek et al., 2008; Belikov et al., 2014). But the mechanism of the occurrence of these deletions, their role and their importance to the evolution of the viral population remain uncertain (Belikov et al., 2014) and requires additional studies.

Many single amino acid substitutions observed in our strains were randomly distributed along the polyprotein. The number of unique amino acid substitutions varied from 3 to 9 in individual isolates, but no characteristic amino acid substitution typical for all human TBEV isolates was identified when compared to the isolates from ticks. In total, 25 unique amino acid substitutions were found in the genomes of the analyzed strains. Table 2 shows a summary of the identified substitutions with comparison to the prototypic TBEV strain Neudoerfl. No unique amino acid substitutions were found in NS2B and NS4B

genes. Mutations were most often located in the third codon position, but in case of Met192/968→Tyr (NS1, strain Skrivanek), all three codon positions were changed. Strains Kubinova and Skrivanek contained substitutions typical for Turkish sheep encephalitis virus (GenBank Access. No. DQ235151.1); i.e., Asp74/186→Glu (prM, strain Kubinova) and Gln146/1635→His (NS3, strain Skrivanek). The substitution Gln256/1745→His in NS3 protein is specific just for the strains Petracova and Tobrman and then for a single FE-TBEV strain 886 (GenBank Access. No. EF469662.1).

The most interesting specific mutations are Ile692/3203→Ser (Skrivanek) and Ile692/3203→Thr (Petracova and Tobrman) in NS5 protein. The second substitution can be found only in European human pathogenic TBEV strain Ljubljana I (GenBank Access. No. JQ654701.1). The substitution is localized in hmD region forming a template entry channel of TBEV polymerase. We speculate that it may be responsible for better interaction with host replication 229 trans-acting factors. However, most amino acid substitutions found in our strains may be incidental or represent a result of adaptation of the virus to various environments. As shown in Fig. 3, the unique amino acid substitutions are mostly distributed “randomly” in the 3D model of the proteins. The exact effect of each of the identified amino acid substitutions independently or in combination with other substitution(s) on biological properties of the virus strains needs to be investigated using reverse genetics approach.

Using these tertiary predicted models (NS1, N3 and NS5) and the available crystal structure of TBEV envelope glycoprotein (E; PDB: 1SVB) we were able to hypothesize about these “random” substitutions. As predicted by the ElliPro server (Ponomarenko et al., 2008) all substitutions for each respective structure depicted in Figure 4 occur within and near regions with a high probability of being recognized by an antibody (Fig. 4A). To further explore any antigenic properties these substitutions may possess, we used the crystal structures of the antibodies from the envelope glycoprotein complexes of the West Nile virus (PDB: 3I50) and the Dengue virus (PDB: 3UAJ) for protein-protein docking (i.e., antibody-antigen). Both envelope glycoproteins are ~40% identical to TBEV E and are recognized by antibodies at polar ends of their conserved tertiary structures (Fig. 4C), thus serving as positive controls for antibody-antigen docking of the TBEV E protein. Tertiary predicted structures NS1, N3 and NS5 were not used for docking since no homologous crystal structures were found

in complex with an antibody, as predicted by the DALI server (Holm and Rosenström 2010). Docking results show that the West Nile antibody explores within the first 200 residues of the predicted epitope regions for TBEV E protein, while the Dengue antibody explores the entire envelope glycoprotein (Fig. 4B). This suggests that the Dengue antibody may target the TBEV E protein more efficiently since it explores (and may bind to) regions with limited amino acid substitutions (Fig. 4B). Figure 5C also depicts that the exploration of both antibodies comes in close proximity to their respective native positions. These data may be extremely informative since Spurrier et al. (2014) discovered that immunogenic regions with high variability (i.e., substitutions) showed reduced response to antibodies against the gp120 of HIV. Therefore, understanding the all255

atom exploration of specific antibodies may provide better preventative measures.

For analysis of dN/dS ratios we used all available TBEV genome sequences and the strains were analyzed according to the three TBEV subtypes (Fig. 5A). The ratio dN/dS reveals that all three datasets have undergone a purifying (negative) selection throughout their evolution (dN/dS < 0.05). This purifying selection (i.e., deleterious mutations) may be caused by specific host-pathogen interactions (or other environmental factors) that TBEV is subjected to. This observation is in accordance with previously published data (Holmes, 2003; Belikov et al., 2014). In order to understand how selective constraints differ between different regions of the TBEV genome we estimated the dN/dS for individual genes. All genes were under a purifying selection, although slight differences were found in the dN/dS ratios between individual 264 genes in the TBEV genome. However, differences in the dN/dS were found in some genes between the individual subtypes. In particular, Eu-TBEV has the dN/dS of 0.073 in NS4B gene, while FE-TBEV and S-TBEV have dN/dS of 0.025 and 0.027, respectively. When we compare the dN/dS ratios in Eu-TBEV strains isolated from human patients and ticks, most genes have similar dN/dS ratios, but we can see again a difference in NS4B: the dN/dS of 0.128 in tick-derived Eu-TBEV, but 0.038 for patient-derived Eu-TBEV. This indicates that NS4B is in different TBEV subtypes under different selection pressures and that differences can also be found between strains isolated from ticks and human patients (Fig. 5B). The NS4B is known to interact with the helicase domain of NS3 and may serve as an

interferon antagonist (Munos-Jordan et al., 2005). However, the importance of our finding is unclear and requires further study.

The severity of TBE may depend on various factors that include the inoculation dose, exposure time and virulence of the virus (Belikov et al., 2014; Leonova et al., 2014; Růžek et al., 2008), the age, sex and immune status of the host (Růžek et al., 2009), and also susceptibility based on the host's genetic background (Palus et al., 2013; Kindberg et al., 2008, 2011; Barkhash et al., 2010, 2012). Field TBEV strains are very heterogeneous with respect to their pathogenicity for humans (Belikov et al., 2014; Růžek et al., 2008). Therefore, analysis of TBEV strains isolated from human patients with severe forms of TBE is crucial for identification of molecular determinants that make these strains pathogenic for humans.

Here we present the largest number of patient-derived European TBEV full genome sequences to date and their molecular characterization. However, more human TBEV strains need to be analyzed to better understand what determines some TBEV strains to cause dangerous life-threating encephalitis in humans, while other do not give rise to any clinical manifestations. Our data can also represent a platform for further studies on evolution of TBEV since the first emergence of human TBE in Europe.

Acknowledgements

The authors are greatly indebted to Dr. Vlasta Danielová for providing us with the TBEV strains and for general support of our work. This study was supported by by the Russian Scientific Foundation (project No. 14-15-00615), the Academy of Sciences of the Czech Republic (Z60220518), Czech Science Foundation projects Nos. P502/11/2116, P302/12/2490 and GA14-29256S, and the project LO1218 with a financial support from the MEYS of the Czech Republic under the NPU I program. The founders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

Akey, D.L., Brown, W.C., Dutta, S., Konwerski, J., Jose, J., Jurkiw, T.J., DelProposto, J., Ogata, C.M., Skiniotis, G., Kuhn, R.J., Smith, J.L. 2014.

Flavivirus NS1 structures reveal surfaces for associations with membranes and the immune system. Science 343(6173), 881-885.

Barkhash, A.V., Perelygin, A.A., Babenko, V.N., Myasnikova, N.G., Pilipenko, P.I., Romaschenko, A.G., Voevoda, M.I., Brinton, M.A. 2010. Variability in the 2'-5'-oligoadenylate synthetase gene cluster is associated with human predisposition to tick-borne encephalitis virus-induced disease. J. Infect. Dis. 202(12), 1813-1818.

Barkhash, A.V., Perelygin, A.A., Babenko, V.N., Brinton, M.A., Voevoda, M.I. 2012. Single nucleotide polymorphism in the promoter region of the CD209 gene is associated with human predisposition to severe forms of tick-borne encephalitis. Antiviral Res. 93(1), 64-68.

Belikov, S.I., Kondratov, I.G., Potapova, U.V., Leonova, G.N. 2014. The Relationship between the Structure of the Tick-Borne Encephalitis Virus Strains and Their Pathogenic Properties. PLoS One. 9(4), e94946.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 16(22):10881-10890.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods. 9(8), 772.

Ecker, M., Allison, S.L., Meixner, T., Heinz, F.X. 1999. Sequence analysis and genetic classification of tick-borne encephalitis viruses from Europe and Asia. J. Gen. Virol. 80 (Pt 1), 179-185.

Ponomarenko, J., Bui, H.-H., Li, W., Fusseder, N., Bourne, P.E., Sette, A., Peters, B. 2008. ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 9, 514.

Fajs, L., Durmiši, E., Knap, N., Strle, F., Avšič-Županc, T. 2012. Phylogeographic characterization of tick-borne encephalitis virus from patients, rodents and ticks in Slovenia. PLoS One. 7(11), e48420.

Frey, S., Essbauer, S., Zöller, G., Klempa, B., Dobler, G., Pfeffer, M. 2014. Full genome sequences and preliminary molecular characterization of three tick-borne encephalitis virus strains isolated from ticks and a bank vole in Slovak Republic. Virus Genes 48(1), 184-188.

Gallia, F., Rampas, J., Hollender, L. 1949. [Laboratory infection caused by tick-borne encephalitis virus] (In Czech) Čas. Lék. čes. 88, 224-229.

Golovljova, I., Vene, S., Sjölander, K.B., Vasilenko, V., Plyusnin, A., Lundkvist, A. 2004. Characterization of tick-borne encephalitis virus from Estonia. J. Med. Virol. 74(4), 580-588.

Gritsun, T.S., Venugopal, K., Zanotto, P.M., Mikhailov, M.V., Sall, A.A., Holmes, E.C., Polkinghorne, I., Frolova, T.V., Pogodina, V.V., Lashkevich, V.A., Gould, E.A. 1997. Complete sequence of two tick-borne flaviviruses isolated from Siberia and the UK: analysis and 332 significance of the 5' and 3'-UTRs. Virus Res. 49(1), 27-39.

Guex, N., Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18(15), 2714-2723.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95-98.

Holm, L., Rosenström, P. 2010. Dali server: conservation mapping in 3D. Nucleic Acids Res. 38, W545- 549.

Holmes, E.C. 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J. Virol. 77(20), 11296-11298.

Kelley, L.A., Sternberg, M.J. 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat. Protoc. 4(3), 363-371.

Kindberg, E., Mickiene, A., Ax, C., Akerlind, B., Vene, S., Lindquist, L., Lundkvist, A., Svensson, L. 2008. A deletion in the chemokine receptor 5 (CCR5) gene is associated with tickborne encephalitis. J. Infect. Dis. 197(2), 266-269.

Kindberg, E., Vene, S., Mickiene, A., Lundkvist, Å., Lindquist, L., Svensson, L. 2011. A functional Toll-like receptor 3 gene (TLR3) may be a risk factor for tick-borne encephalitis virus (TBEV) infection. J. Infect. Dis. 203(4), 523-528.

Krejčí, J. 1949. Isolement d´un virus noveau en course d´un epidémie de meningoencephalite dans la region de Vyškov (Moraviae). Presse Méd. (Paris) 74, 1084, 1949.

Leonova, G.N., Maystrovskaya, O.S., Kondratov, I.G., Takashima, I., Belikov, S.I. 2014. The nature of replication of tick-borne encephalitis virus strains isolated from residents of the Russian Far East with inapparent and clinical forms of infection. Virus Res. 189C:34-42. doi: 10.1016/j.virusres.2014.04.004. [Epub ahead of print]

Li, X., Jacobson, M.P., Zhu, K., Zhao, S., Friesner, R.A. 2007. Assignment of polar states for protein amino acid residues using an interaction cluster

decomposition algorithm and its application to high resolution protein structure modeling. Proteins: Struct., Funct., Bioinf. 66(4), 824-837.

Lu, G., Gong, P. 2013. Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. PLoS Pathog. 9(8), e1003549.

Luo, D., Xu, T., Hunke, C., Grüber, G., Vasudevan, S.G., Lescar, J. 2008. Crystal structure of the NS3 protease-helicase from dengue virus. J. Virol. 82(1), 173-183.

Mandl, C.W., Holzmann, H., Meixner, T., Rauscher, S., Stadler, P.F., Allison, S.L., Heinz, F.X. 1998. Spontaneous and engineered deletions in the 3' noncoding region of tick-borne encephalitis virus: construction of highly attenuated mutants of a flavivirus. J. Virol. 72(3), 2132-2140.

Mansfield, K.L., Johnson, N., Phipps, L.P., Stephenson, J.R., Fooks, A.R., Solomon, 366 T. 2009. Tick-borne encephalitis virus - a review of an emerging zoonosis. J. Gen. Virol. 90(Pt 8), 1781-1794.

Muñoz-Jordán, J.L., Laurent-Rolle, M., Ashour, J., Martínez-Sobrido, L., Ashok, M., Lipkin, W.I., García-Sastre, A. 2005. Inhibition of alpha/beta interferon signaling by the NS4B protein of flaviviruses. J. Virol. 79(13), 8004-8013.

Palus, M., Vojtíšková, J., Salát, J., Kopecký, J., Grubhoffer, L., Lipoldová, M., Demant, P., Růžek, D. 2013. Mice with different susceptibility to tick-borne encephalitis virus infection show selective neutralizing antibody response and inflammatory reaction in the central nervous system. J. Neuroinflammation 10, 77.

Pospíšil, L., Jandásek, L., Pešek, J. 1954. *Isolation of new strains of tick-borne encephalitis virus, Brno region, summer 1953] (In Czech), Lék. listy 9, 3-5.

Rey, F.A., Heinz, F.X., Mandl, C., Kunz, C., Harrison, S.C. 1995. The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. Nature 375(6529), 291-298.

Ronquist, F., Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Růzek, D., Gritsun, T.S., Forrester, N.L., Gould, E.A., Kopecký, J., Golovchenko, M., Rudenko, N., Grubhoffer, L. 2008. Mutations in the NS2B and NS3 genes affect mouse neuroinvasiveness of a Western European field strain of tick-borne encephalitis virus. Virology 374(2), 249-255.

Růzek, D., Salát, J., Palus, M., Gritsun, T.S., Gould, E.A., Dyková, I., Skallová, A., Jelínek, J., Kopecký, J., Grubhoffer, L. 2009. CD8+ T-cells mediate immunopathology in tick-borne encephalitis. Virology 384(1), 1-6.

Růžek, D., Dobler, G., Donoso Mantke, O. 2010. Tick-borne encephalitis: pathogenesis and clinical implications. Travel Med. Infect. Dis. 8(4), 223-232.

Sakai, M., Yoshii, K., Sunden, Y., Yokozawa, K., Hirano, M., Kariwa, H. 2014. Variable region of the 3' UTR is a critical virulence factor in the Far-Eastern subtype of tick-borne encephalitis virus in a mouse model. J. Gen. Virol. 95(Pt 4), 823-835.

Spurrier, B., Sampson, J., Gorny, M.K., Zolla-Pazner, S., Kong, X.P. 2014. Functional implications of the binding mode of a human conformation-dependent V2 monoclonal antibody against HIV. J. Virol. 88(8), 4100-4112.

Torchala, M., Moal, I.H., Chaleil, R.A.G., Fernandez-Recio, J., Bates, P.A. 2013a. SwarmDock: a server for flexible protein-protein docking. Bioinformatics 29, 807-809.

Torchala, M., Moal, I.H., Chaleil, R.A.G., Agius, R.., Bates, P.A. 2013b. A Markov-chain model description of binding funnels to enhance the ranking of docked solutions. Proteins 81, 2143-2149.

Torchala, M., Bates, P.A. 2014. Predicting the Structure of Protein-400 Protein Complexes Using the SwarmDock Web Server, in: Kihara D. (Ed.) *Protein Structure Prediction 3rd Edition (Methods in Molecular Biology, Vol. 1137)*, Springer, New York, pp. 181-197.

Tamura, K., Dudley, J., Nei, M., Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24(8), 1596-1599.

Wallner, G., Mandl, C.W., Kunz, C., Heinz, F.X. 1995. The flavivirus 3'-noncoding region: extensive size heterogeneity independent of evolutionary relationships among strains of tick-borne encephalitis virus. Virology 213(1), 169-178.

Wallner, G., Mandl, C.W., Ecker, M., Holzmann, H., Stiasny, K., Kunz, C., Heinz, F.X. 1996. Characterization and complete genome sequences of high- and low- virulence variants of tick-borne encephalitis virus. J. Gen. Virol. 77(Pt 5), 1035-1042.

Weidmann, M., Ruzek, D., Krivanec, K., Zöller, G., Essbauer, S., Pfeffer, M., Zanotto, P.M., Hufert, F.T., Dobler, G. 2011. Relation of genetic phylogeny

and geographical distance of tick-borne encephalitis virus in central Europe. J. Gen. Virol. 92(Pt 8), 1906-1916.

Weidmann, M., Frey, S., Freire, C.C., Essbauer, S., Růžek, D., Klempa, B., Zubrikova, D., Vögerl, M., Pfeffer, M., Hufert, F.T., Zanotto, P.M., Dobler, G. 2013. Molecular phylogeography of tick-borne encephalitis virus in central Europe. J. Gen. Virol. 94(Pt 9), 2129-2139.

Figure and Table Legends

Figure 1. Phylogenetic relationships among European TBEV strains with fully sequenced genome. Phylogenetic analysis was done based on full-genome nucleotide sequences. The TBE patient derived TBEV strains are highlighted by bold. What is clearly visible is that human derived TBEV strains do not form a single monophyletic group, but they are randomly dispersed in the cladogram. Also the newly sequenced TBEV strain all isolated from Czech patients do not tend to form a monophyletic group but they are phylogenetically mixed among other European TBEV strains with mild tendency to form a central European cluster.

Figure 2. Alignment of 3'NCR of the analyzed strains and compared with 3'NCR from the strains Neudoerfl and Hypr.

Figure 3. Placement of amino acid substitutions on TBEV proteins: Placement of amino acid is shown on the protein structure for the crystallized flaviviral protein E and the modeled tertiary structures NS1, NS3 and NS5 (as there is no specific substitution in methyltransferase domain of NS5 protein only polymerase domain is visualized here). Only substitutions that were specifically exclusive for newly sequenced patient isolates of TBEV or for maximally two other TBEV strains are shown. Substitutions specific for strains Skrivanek (Sk), Vlasaty (Vl), Petracova (Pet),Tobrman (To), and Kubinova (Ku) are shown in red, yellow, green, blue, and violet respectively.

Figure 4. Epitope predictions and antibody docking of the TBEV proteins: Panel A depicts the scoring function for epitope prediction (y-axis) based on the methods employed by the ElliPro server (Ponomarenko et al., 2008) and the residue position (x-axis) of each epitope for TBEV strains E (blue), NS1 (red), NS3 (green) and NS5 (magenta). The points on top of the scatter plot indicate the position of the amino acid substitutions for each strain (as indicated in

Figure 3). The correlation shown (B) is between the epitope regions predicted by the ElliPro server (x-axis) and the contact residues for the top 10 docked poses predicted by SwarmDock server (Torchala et al. 2013a, 2013b; Torchala and Bates 2014) using the E strain (PDB: 1SVB) and the antibodies from PDBs 3I50 (blue) and 3UAJ (red). Panel C shows the superposition for the homologous flaviviruses of the TBEV E strain (PDB: 1SVB; green), West Nile in complex with the E53 antibody Fab (PDB: 3I50; blue) and Dengue in complex with the Fab fragment of the chimpanzee monoclonal antibody 5H2 (PDB: 3UAJ; red) in a 180° turn. The native position of the respective antibodies for West Nile (blue) and Dengue (red) viruses are shown in cartoon with the center of mass of the top 10 docked poses from SwarmDock (color coded spheres that match the respective antibody type).

Figure 5. For detection of selection pressure acting on individual genes we calculated the ratios of non-synonymous and synonymous nucleotide substitutions per site (dN/dS) of the available TBEV sequences and compared the dN/dS ratios in individual genes of TBEV strains from European, Siberian and Far Eastern subtype (A). Within the European subtype, we compared the dN/dS ratios of individual genes from TBEV strains isolated from ticks and human patients (B).

Table 1. Comparison (similarity in percentage) among nucleotide sequence (below the diagonal) and deduced amino acid sequence (above the diagonal) of the analyzed strains and the strains Neudoerfl and Hypr.

Table 2. List of the amino acid substitutions of the analyzed TBEV strains in comparison to the strain Neudoerfl.

Supplementary Figure 1. Polyprotein alignment of the analyzed strains and compared with selected representatives of each TBEV subtype.

Tables:

Table 1:

	Vlasaty	Tobrman	Skrivanek	Petracova	Kubinova	Hypr	Neudoerfl
Vlasaty	x	99.4	99.2	99.2	99.3	98.8	99.1
Tobrman	97.50	x	99.4	99.8	99.3	98.9	99.1
Skrivanek	97.63	97.55	x	99.3	99.3	98.9	99.1
Petracova	97.45	99.79	97.52	x	99.2	98.9	99.1
Kubinova	97.71	97.59	97.74	97.53	x	98.9	99.1
Hypr	97.26	97.32	97.45	97.31	97.42	x	98.8
Neudoerf	97.45	97.75	97.50	97.74	97.47	97.24	X

Comparison of nucleotide sequences is based on the complete genome sequences including the noncoding regions.

Table 2:

region + number of unique aa changes		Sk	Vl	Pet	To	Ku
C	3		Arg80→Lys	Thr24→Met*, Arg30→Gly*	Thr24→Met*, Arg30→Gly*	Val107→Ile*
prM	1			Thr15/127→Ile *	Thr15/127→Ile *	Asp74/186→Glu**(DQ235151)
glyc. M	3		Leu42/247→Pro*, Thr48/253→Ile *	Ala62/267→Val, Tyr74/279→His *	Ala62/267→Val	**Lys40/245→Arg**
M	*1*	*Asn52/332→Ser*, Thr81/361→Ile **(GQ266392), ***Ile167/447→Val***, Ser349/629→Phe**(AF069066)	Arg20/300→Lys**(HM051171), ***Ile167/447→Val***	***Ile167/447→Val***, Ser169/449→Thr*	***Ile167/447→Val***	***Ile167/447→Val***
NS1	4	Glu51/827→Asp*, Met192/968→Tyr***, ***Tyr271/1047→His***		Ser71/847→Leu*, ***Tyr271/1047→His***	***Tyr271/1047→His***	Val2/778→Ile**(GU121642),Glu52/828→Gly*, His177/953→Gln*, Ile194/970→Val, ***Tyr271/1047→His***, Asn289/1065→Asp**(GQ266392)
NS2A	4	**Thr33/1161→Ser, Ile53/1181→Met, *Glu127/1255→Asp, Val201/1329→Ile, Gly206/1334→Arg***	Val41/1169→Ile*, Thr109/1237→Ile, ***Glu127/1255→Asp, Val179/1307→Ile, Val201/1329→Ile, Gly206/1334→Arg***	**Ile53/1181→Met, *Arg99/1227→Gly*, Glu127/1255→Asp, Val201/1329→Ile, Gly206/1334→Arg***	**Ile53/1181→Met, *Arg99/1227→Gly*, Glu127/1255→Asp, Val201/1329→Ile, Gly206/1334→Arg***	Val103/1231→Ile*, ***Glu127/1255→Asp, Val201/1329→Ile, Gly206/1334→Arg***
NS2B 1492	*0*	***Tyr61/1419→His, Ile100/1458→Leu***	***Tyr61/1419→His, Ile100/1458→Leu***	***Tyr61/1419→His, Ile100/1458→Leu***	***Tyr61/1419→His, Ile100/1458→Leu***	***Tyr61/1419→His, Ile100/1458→Leu***
NS3	*2*	***Arg106/1595→Lys***, Gln146/1635→His**(DQ235151), Ser558/2047→Asn	Asn37/1526→Ser*, ***Arg106/1595→Lys***, Arg124/1613→Lys	***Arg106/1595→Lys*** Gln256/1745→His** Lys599/2089→Arg	***Arg106/1595→Lys*** Arg124/1613→Lys Gln256/1745→His**(EF469662)	***Arg106/1595→Lys*** Lys174/1663→Arg** (AM600965), Asn242/1733→Ile*(j.a.)
NS4A	1			Val55/2165→Ala*(j.a.)	Val55/2165→Ala*	
NS4B	*0*	***Ala16/2275→Val***, Arg21/2280→Gln, ***Thr178/2437→Ile***	***Ala16/2275→Val***, Arg21/2280→Gln, ***Thr178/2437→Ile***, Gly227/2486→Ser** (DQ486861, AB062063)	***Ala16/2275→Val***, , ***Thr178/2437→Ile***	***Ala16/2275→Val***, Arg21/2280→Gln, ***Thr178/2437→Ile***	***Ala16/2275→Val***, Arg21/2280→Gln, ***Thr178/2437→Ile***
NS5	*5*	***Val51/2562→Met***, Arg101/2612→Lys, ***Lys108/2619→Arg, Lys253/2764→Arg***, *Arg397/2908→Lys*, ***Arg434/2945→His***, Ile692/3203→Ser*, Thr724/3235→Ser, ***Ala786/3297→Val, Arg830/3341→Lys, Arg855/3366→Lys***, Asp8913402→Gly*	***Val51/2562→Met, Lys108/2619→Arg, Lys253/2764→Arg***, Met394/2905→Arg*, *Arg397/2908→Lys*, ***Arg434/2945→His***, Ala634/3145→Thr, ***Ala786/3297→Val, Arg830/3341→Lys, Arg855/3366→Lys***	***Val51/2562→Met, Lys108/2619→Arg, Lys253/2764→Arg***, *Arg397/2908→Lys*, ***Arg434/2945→His***, Val661/3172→Ile*, Ile692/3203→Thr**(JQ654701), ***Ala786/3297→Val, Arg830/3341→Lys, Arg855/3366→Lys***	Arg14→2525Gln, ***Val51/2562→Met, Lys108/2619→Arg, Lys253/2764→Arg***, *Arg397/2908→Lys*, ***Arg434/2945→His***, Ile692/3203→Thr**(JQ654701), ***Ala786/3297→Val, Arg830/3341→Lys, Arg855/3366→Lys***	***Val51/2562→Met, Lys108/2619→Arg, Lys253/2764→Arg***, *Arg397/2908→Lys*, ***Arg434/2945→His***, Gln571/3082→His*, ***Ala786/3297→Val, Arg830/3341→Lys, Arg855/3366→Lys***

* unique substitution; ** substitution found in one or two other strains, (GenBank accession number
in parentheses); *** substitution of the whole nucleotide triplet. bold: identical substitution to the
strain Hypr; italics: identical substitution to the strain 263

Figures:

Figure 1:

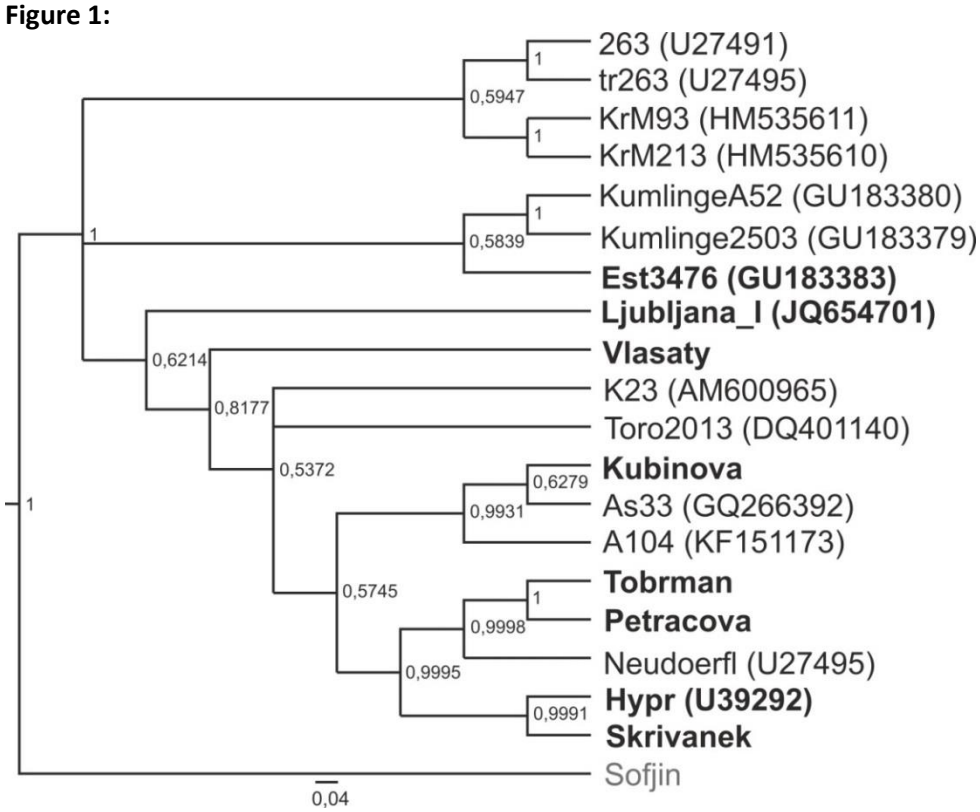


Figure 2:

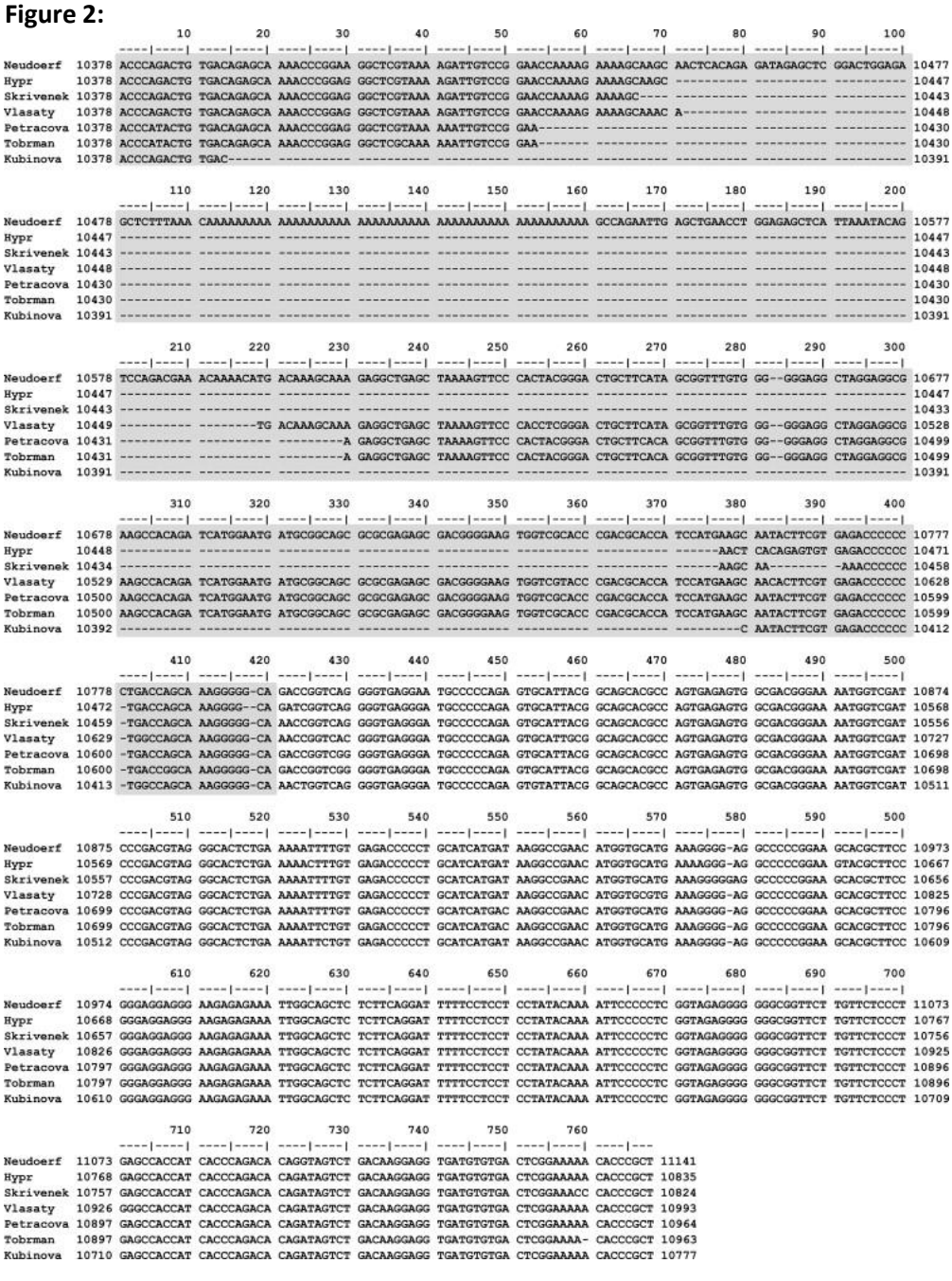

10 20 30 40 50 60 70 80 90 100
Neudoerf 10378 ACCCAGACTG TGACAGAGCA AAACCCGGAA GGCTCGTAAA AGATTGTCCG GAACCAAAAG AAAAGCAAGC AACTCACAGA GATAGAGCTC GGACTGGAGA 10477
Hypr 10378 ACCCAGACTG TGACAGAGCA AAACCCGGAG GGCTCGTAAA AGATTGTCCG GAACCAAAAG AAAAGCAAGC ---------- ---------- ---------- 10447
Skrivenek 10378 ACCCAGACTG TGACAGAGCA AAACCCGGAG GGCTCGTAAA AGATTGTCCG GAACCAAAAG AAAAGC---- ---------- ---------- ---------- 10443
Vlasaty 10378 ACCCAGACTG TGACAGAGCA AAACCCGGAG GGCTCGTAAA AGATTGTCCG GAACCAAAAG AAAAGCAAAC A--------- ---------- ---------- 10448
Petracova 10378 ACCCATACTG TGACAGAGCA AAACCCGGAG GGCTCGTAAA AAATTGTCCG GAA------- ---------- ---------- ---------- 10430
Tobrman 10378 ACCCATACTG TGACAGAGCA AAACCCGGAG GGCTCGCAAA AAATTGTCCG GAA------- ---------- ---------- ---------- 10430
Kubinova 10378 ACCCAGACTG TGAC------ ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10391
110 120 130 140 150 160 170 180 190 200
Neudoerf 10478 GCTCTTTAAA CAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA GCCAGAATTG AGCTGAACCT GGAGAGCTCA TTAAATACAG 10577
Hypr 10447 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10447
Skrivenek 10443 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10443
Vlasaty 10448 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10448
Petracova 10430 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10430
Tobrman 10430 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10430
Kubinova 10391 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10391
210 220 230 240 250 260 270 280 290 300
Neudoerf 10578 TCCAGACGAA ACAAAACATG ACAAAGCAAA GAGGCTGAGC TAAAAGTTCC CACTACGGGA CTGCTTCATA GCGGTTTGTG GG--GGGAGG CTAGGAGGCG 10677
Hypr 10447 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10447
Skrivenek 10443 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10433
Vlasaty 10449 ---------- --------TG ACAAAGCAAA GAGGCTGAGC TAAAAGTTCC CACCTCGGGA CTGCTTCATA GCGGTTTGTG GG--GGGAGG CTAGGAGGCG 10528
Petracova 10431 ---------- ---------- ---------A GAGGCTGAGC TAAAAGTTCC CACTACGGGA CTGCTTCACA GCGGTTTGTG GG--GGGAGG CTAGGAGGCG 10499
Tobrman 10431 ---------- ---------- ---------A GAGGCTGAGC TAAAAGTTCC CACTACGGGA CTGCTTCACA GCGGTTTGTG GG--GGGAGG CTAGGAGGCG 10499
Kubinova 10391 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- 10391
310 320 330 340 350 360 370 380 390 400
Neudoerf 10678 AAGCCACAGA TCATGGAATG ATGCGGCAGC GCGCGAGAGC GACGGGGAAG TGGTCGCACC CGACGCACCA TCCATGAAGC AATACTTCGT GAGACCCCCC 10777
Hypr 10448 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ------AACT CACAGAGTGT GAGACCCCCC 10471
Skrivenek 10434 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ------AAGC AA-------- -AAACCCCCC 10458
Vlasaty 10529 AAGCCACAGA TCATGGAATG ATGCGGCAGC GCGCGAGAGC GACGGGGAAG TGGTCGTACC CGACGCACCA TCCATGAAGC AACACTTCGT GAGACCCCCC 10628
Petracova 10500 AAGCCACAGA TCATGGAATG ATGCGGCAGC GCGCGAGAGC GACGGGGAAG TGGTCGCACC CGACGCACCA TCCATGAAGC AATACTTCGT GAGACCCCCC 10599
Tobrman 10500 AAGCCACAGA TCATGGAATG ATGCGGCAGC GCGCGAGAGC GACGGGGAAG TGGTCGCACC CGACGCACCA TCCATGAAGC AATACTTCGT GAGACCCCCC 10599
Kubinova 10392 ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------C AATACTTCGT GAGACCCCCC 10412
410 420 430 440 450 460 470 480 490 500
Neudoerf 10778 CTGACCAGCA AAGGGGG-CA GACCGGTCAG GGGTGAGGAA TGCCCCCAGA GTGCATTACG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10874
Hypr 10472 -TGACCAGCA AAGGGG--CA GATCGGTCAG GGGTGAGGGA TGCCCCCAGA GTGCATTACG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10568
Skrivenek 10459 -TGACCAGCA AAGGGGG-CA AACCGGTCAG GGGTGAGGGA TGCCCCCAGA GTGCATTACG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10556
Vlasaty 10629 -TGGCCAGCA AAGGGGG-CA AACCGGTCAC GGGTGAGGGA TGCCCCCAGA GTGCATTGCG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10727
Petracova 10600 -TGACCAGCA AAGGGGG-CA GACCGGTCGG GGGTGAGGGA TGCCCCCAGA GTGCATTACG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10698
Tobrman 10600 -TGACCGGCA AAGGGGG-CA GACCGGTCGG GGGTGAGGGA TGCCCCCAGA GTGCATTACG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10698
Kubinova 10413 -TGGCCAGCA AAGGGGG-CA AACTGGTCAG GGGTGAGGGA TGCCCCCAGA GTGTATTACG GCAGCACGCC AGTGAGAGTG GCGACGGGAA AATGGTCGAT 10511
510 520 530 540 550 560 570 580 590 500
Neudoerf 10875 CCCGACGTAG GGCACTCTGA AAAATTTTGT GAGACCCCCT GCATCATGAT AAGGCCGAAC ATGGTGCATG AAAGGGG-AG GCCCCCGGAA GCACGCTTCC 10973
Hypr 10569 CCCGACGTAG GGCACTCTGA AAAACTTTGT GAGACCCCCT GCATCATGAT AAGGCCGAAC ATGGTGCATG AAAAGGG-AG GCCCCCGGAA GTACGCTTCC 10667
Skrivenek 10557 CCCGACGTAG GGCACTCTGA AAAATTTTGT GAGACCCCCT GCATCATGAT AAGGCCGAAC ATGGTGCATG AAAGGGGGAG GCCCCCGGAA GCACGCTTCC 10656
Vlasaty 10728 CCCGACGTAG GGCACTCTGA AAAATTTTGT GAGACCCCCT GCATCATGAT AAGGCCGAAC ATGGTGCGTG AAAGGGG-AG GCCCCCGGAA GCACGCTTCC 10825
Petracova 10699 CCCGACGTAG GGCACTCTGA AAAATTTTGT GAGACCCCCT GCATCATGAC AAGGCCGAAC ATGGTGCATG AAAGGGG-AG GCCCCCGGAA GCACGCTTCC 10796
Tobrman 10699 CCCGACGTAG GGCACTCTGA AAAATTCTGT GAGACCCCCT GCATCATGAC AAGGCCGAAC ATGGTGCATG AAAGGGG-AG GCCCCCGGAA GCACGCTTCC 10796
Kubinova 10512 CCCGACGTAG GGCACTCTGA AAAATTCTGT GAGACCCCCT GCATCATGAT AAGGCCGAAC ATGGTGCATG AAAGGGG-AG GCCCCCGGAA GCACGCTTCC 10609
610 620 630 640 650 660 670 680 690 700
Neudoerf 10974 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 11073
Hypr 10668 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 10767
Skrivenek 10657 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 10756
Vlasaty 10826 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 10925
Petracova 10797 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 10896
Tobrman 10797 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 10896
Kubinova 10610 GGGAGGAGGG AAGAGAGAAA TTGGCAGCTC TCTTCAGGAT TTTTCCTCCT CCTATACAAA ATTCCCCCTC GGTAGAGGGG GGGCGGTTCT TGTTCTCCCT 10709
710 720 730 740 750 760
Neudoerf 11073 GAGCCACCAT CACCCAGACA CAGGTAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAAAA CACCCGCT 11141
Hypr 10768 GAGCCACCAT CACCCAGACA CAGATAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAAAA CACCCGCT 10835
Skrivenek 10757 GAGCCACCAT CACCCAGACA CAGATAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAACC CACCCGCT 10824
Vlasaty 10926 GGGCCACCAT CACCCAGACA CAGATAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAAAA CACCCGCT 10993
Petracova 10897 GAGCCACCAT CACCCAGACA CAGATAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAAAA CACCCGCT 10964
Tobrman 10897 GAGCCACCAT CACCCAGACA CAGATAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAAAA- CACCCGCT 10963
Kubinova 10710 GAGCCACCAT CACCCAGACA CAGATAGTCT GACAAGGAGG TGATGTGTGA CTCGGAAAAA CACCCGCT 10777

Figure 3:

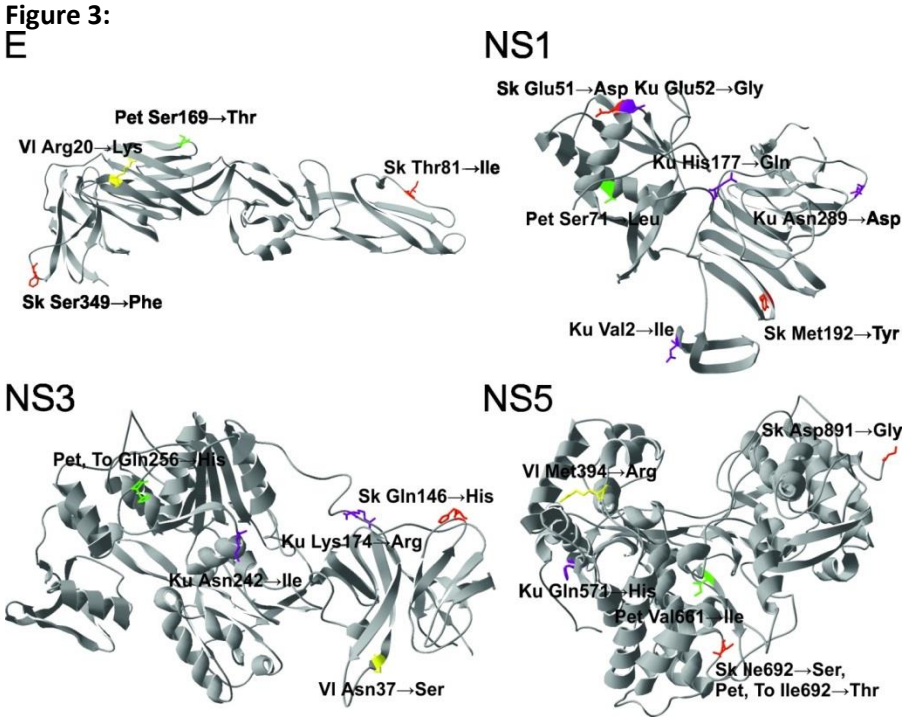


Figure 4:

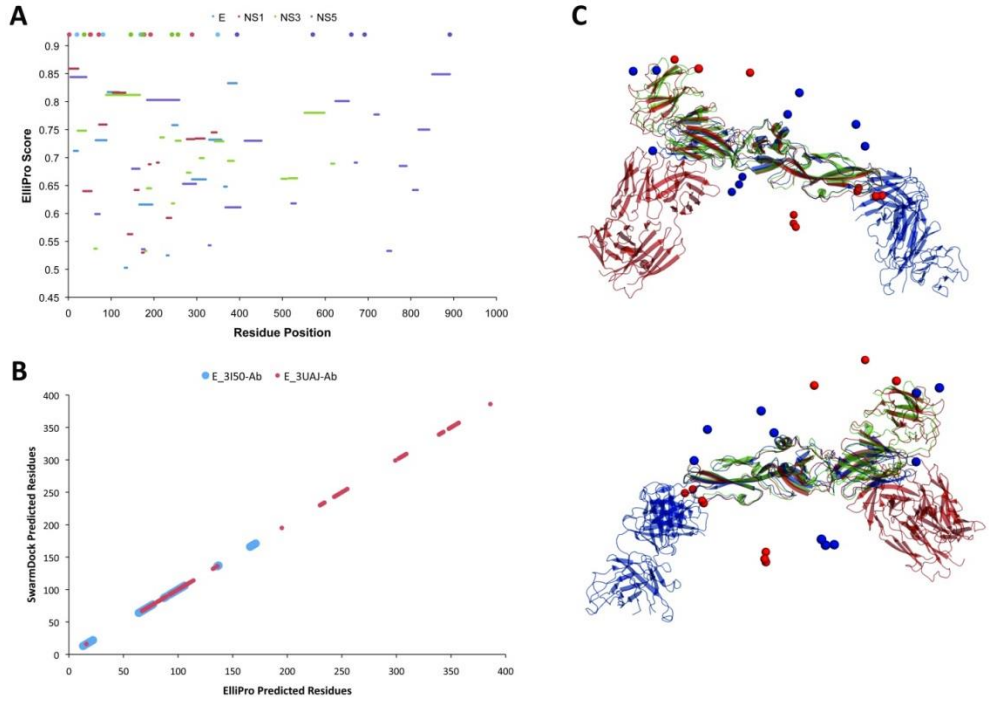
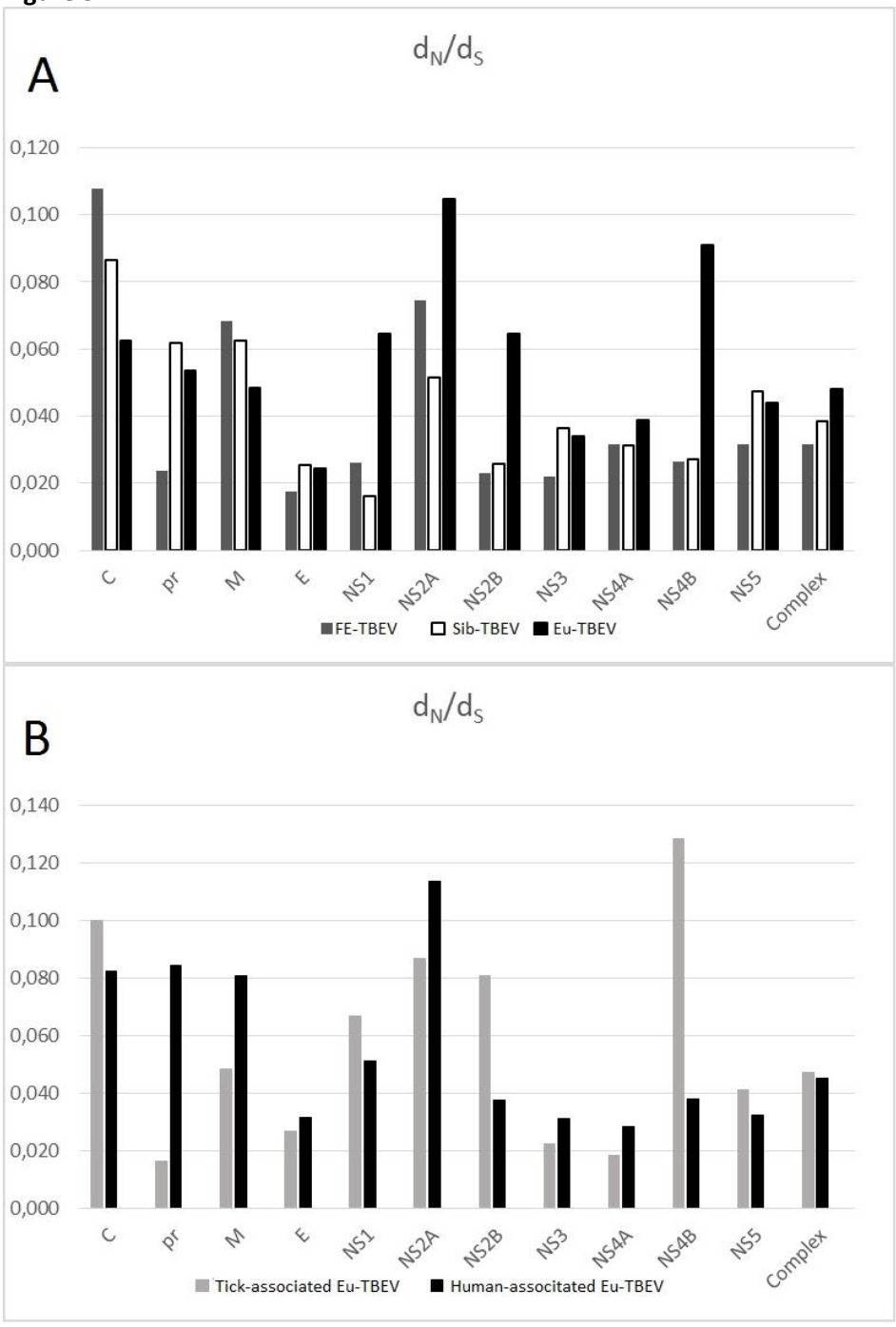
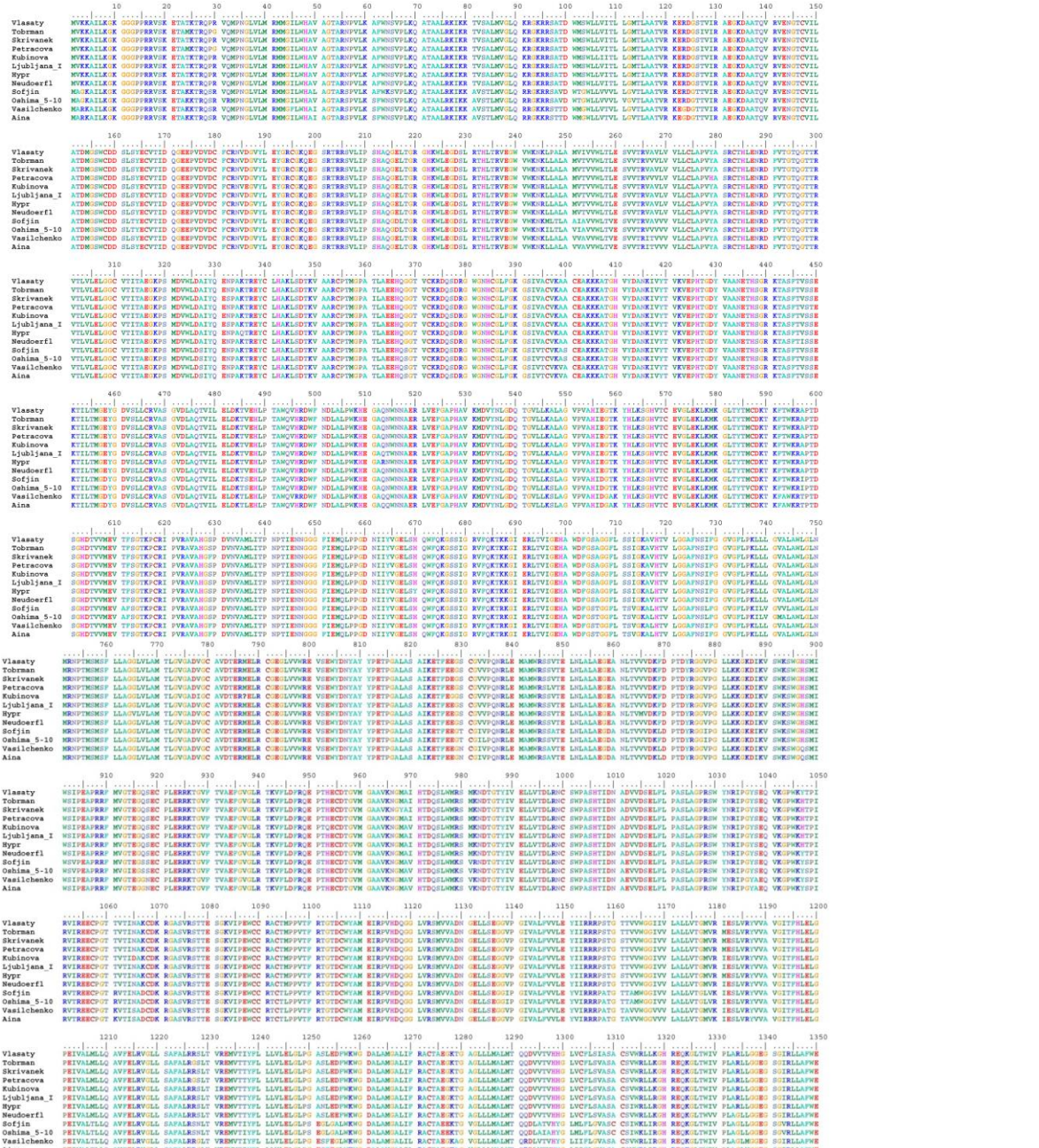
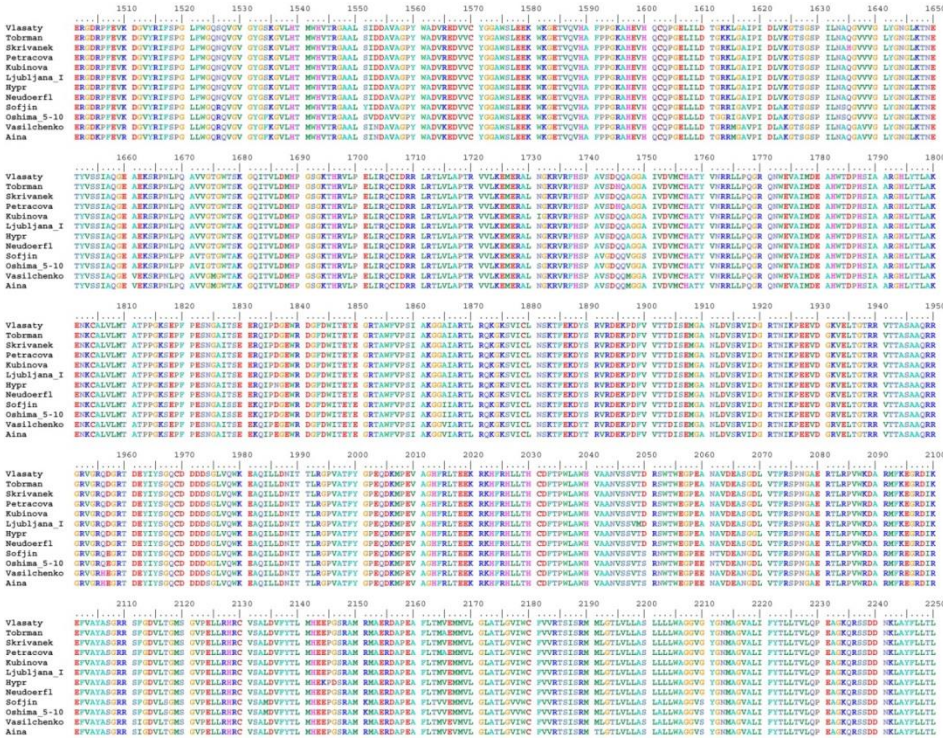


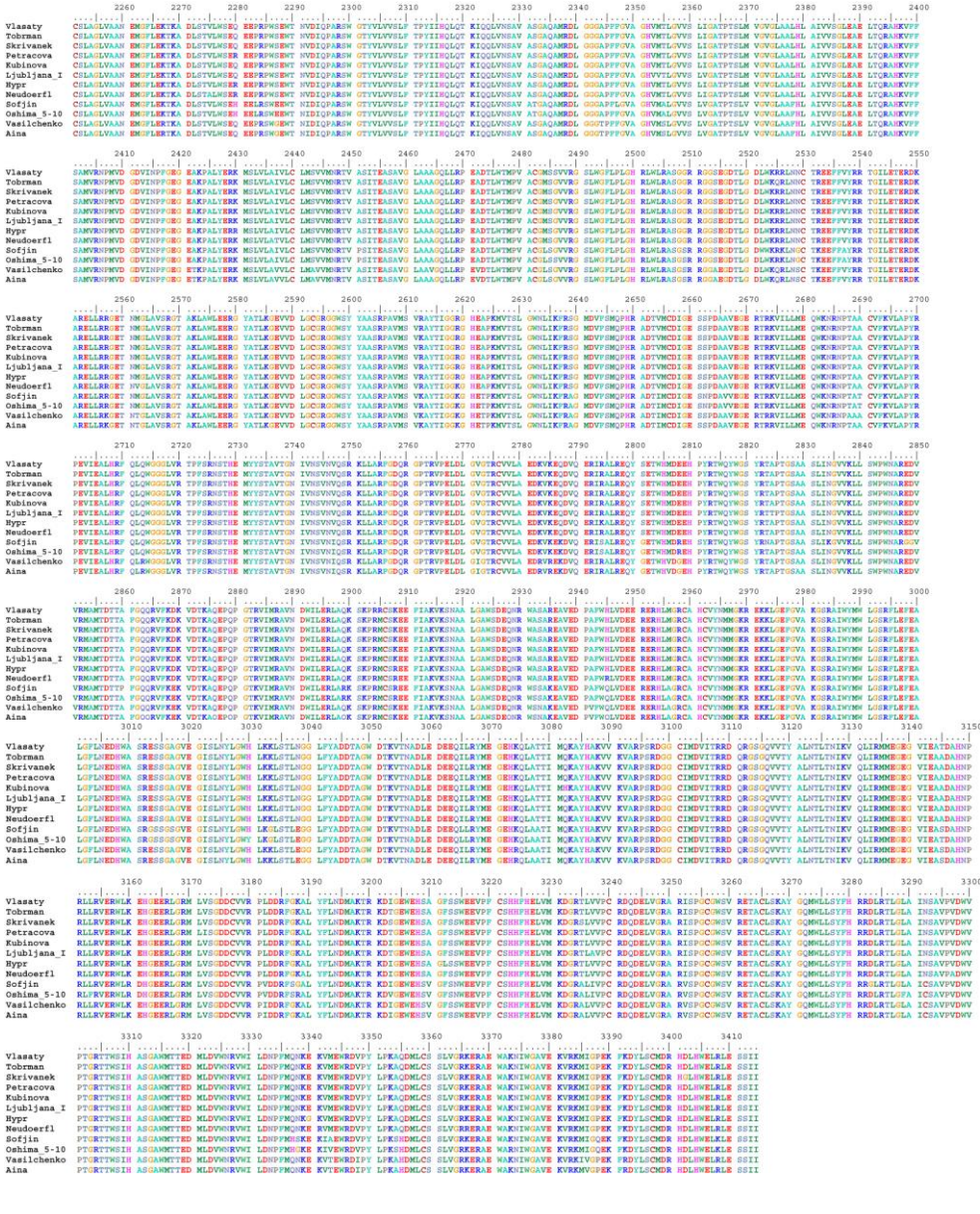
Figure 5:


A
d_N/d_S
0,120
0,100
0,080
0,060
0,040
0,020
0,000
C
pr
M
E
NS1
NS2A
NS2B
NS3
NS4A
NS4B
NS5
Complex
FE-TBEV
Sib-TBEV
Eu-TBEV
B
d_N/d_S
0,140
0,120
0,100
0,080
0,060
0,040
0,020
0,000
C
pr
M
E
NS1
NS2A
NS2B
NS3
NS4A
NS4B
NS5
Complex
Tick-associated Eu-TBEV
Human-associtated Eu-TBEV

Supplementary figure 1:







6.3 A deep phylogeny of viral and cellular right-hand polymerases

The article was published in Infections, Genetids and Evolution and should by cited as **Jiří Černý**; Barbora Černá Bolfíková; Paolo M. de A. Zanotto, Libor Grubhoffer, Daniel Růžek: A deep phylogeny of viral and cellular right-hand polymerases. Infection, Genetics and Evolution, 2015 Sep 30. pii: S1567-1348(15)00402-5. doi: 10.1016/j.meegid.2015.09.026.

TITLE

A deep phylogeny of viral and cellular right-hand polymerases

AUTHORS

Jiří ČERNÝ (a, b), Barbora ČERNÁ BOLFÍKOVÁ (c), Paolo M. de A. ZANOTTO (d), Libor GRUBHOFFER (a, b), Daniel RŮŽEK (a, e)

AFFILIATION

(a) Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, České Budějovice, Branišovská 31, 370 05, Czech Republic

(b) Faculty of Science, University of South Bohemia in České Budějovice, České Budějovice, Branišovská 31, 370 05, Czech Republic

(c) Faculty of Tropical AgriSciences, Czech University of Life Sciences Prague, Prague 6 - Suchdol, Kamýcká 126, 165 21, Czech Republic

(d) Department of Microbiology, Biomedical Sciences Institute – ICB II University of Sao Paulo, 05508-000 Sao Paulo, Brazil

(e) Veterinary Research Institute, Brno, Hudcova 296/70, 621 00, Czech Republic

Corresponding author: Jiří Černý, e-mail: cerny@paru.cas.cz, tel: +420 387 775 451, fax: +420 385 310 388, address: Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 31, 370 05 České Budějovice, Czech Republic

ABSTRACT

Right-hand polymerases are important players in genome replication and repair in cellular organisms as well as in viruses. All right-hand polymerases are grouped into seven related protein families: viral RNA-dependent RNA polymerases, reverse transcriptases, single-subunit RNA polymerases, and DNA polymerase families A, B, D, and Y. Although the evolutionary relationships of right-hand polymerases *within* each family have been proposed, evolutionary relationships *between* families remain elusive because their sequence similarity is too low to allow classical phylogenetic analyses. The structure of viral RNA-dependent RNA polymerases recently was shown to be useful in inferring their evolution. Here, we address evolutionary relationships between right-hand polymerase families by combining sequence and structure information. We used a set of 22 viral and cellular polymerases representing all right-hand polymerase families with known protein structure. In contrast to previous studies, which focused only on the evolution of particular families, the current approach allowed us to present the first robust phylogenetic analysis unifying evolution of all right-hand polymerase families. All polymerase families branched into discrete lineages, following a fairly robust adjacency pattern. Only single-subunit RNA polymerases formed an inner group within DNA polymerase family A. RNA-dependent RNA polymerases of RNA viruses and reverse transcriptases of retroviruses formed two sister groups and were distinguishable from all other polymerases. DNA polymerases of DNA bacteriophages did not form a monophyletic group and are phylogenetically mixed with cellular DNA polymerase families A and B. Based on the highest genetic variability and structural simplicity, we assume that RNA-dependent RNA polymerases are the most ancient group of right-hand polymerases, in agreement with the RNA World hypothesis, because RNA-dependent RNA polymerases are enzymes that could serve in replication of RNA genomes. Moreover, our results show that protein structure can be used in phylogenetic analyses of distantly related proteins that share only limited sequence similarity.

HIGHLIGHTS

- Usage of both sequence and structure of right-hand polymerase can reveal their evolution. Analyzing both structure and sequence yields higher-resolution phylogenetic trees than when only one type of characters is used.

- Compared to trees based on sequence data only, these trees have fewer polytomies.
- viral RdRPs and reverse transcriptases polymerases form 2 groups distinct from DNA polymerases.
- High variability implies viral RNA polymerases are original right-hand polymerases.

KEYWORDS

Right-hand polymerase, polymerase evolution, virus evolution, structural evolution, protein tertiary structure

INTRODUCTION

Right-hand polymerases are important players in genome replication and repair in *Eubacteria, Archaea, Eukarya*, and viruses. Genes coding for right-hand polymerases are present in genomes of all cellular life forms and in the vast majority of viruses (Koonin, 2006). Right-hand polymerases are a monophyletic group that evolved from one common ancestor in the very early stages of life evolution (Delarue et al., 1990). Nevertheless, it is not known whether the common ancestor was a processive polymerase or a non-processive nucleotidyl transferase. According to the Structural Classification of Proteins (SCOP) database (Murzin et al., 1995), the superfamily of right-hand polymerases consists of six families: i) viral RNA-dependent RNA polymerases, which are responsible for replication and transcription of viral genomes (Ferrer-Orta et al., 2006); ii) reverse transcriptases, involved in replication of reverse-transcribing viruses (Miller and Robinson, 1986); iii) single-subunit RNA polymerases, important for transcription in T-odd phages, α-Proteobacteria, and mitochondria (Cermakian et al., 1997; Shutt and Gray, 2006); iv) DNA polymerase family A, involved in replication of T-odd phages or in repair of cellular DNA (Shutt and Gray, 2006); v) DNA polymerase family B, important for replication in the vast majority of DNA viruses as well as eukaryotes (Zhu and Ito, 1994); and vi) DNA polymerase family Y, involved in repair of eukaryotic DNA (Sale et al., 2012).

Apart from the right-hand polymerases, many life forms also use evolutionarily unrelated polymerases, such as i) multi-subunit RNA polymerases, which are involved in RNA transcription; ii) barrel-shaped cellular RNA-dependent RNA

polymerases, involved in RNA interference (Cramer, 2002; Salgado et al., 2006); iii) bacterial DNA polymerase family C, major players in bacterial genome replication (Timinskas et al., 2014); and iv) the DNA polymerase family X, such as DNA polymerase β, which are important for DNA repair (Pelletier et al., 1994; Sawaya et al., 1994).

All right-hand polymerases fold into a right hand–resembling structure containing three subdomains called fingers, palm, and thumb (Hansen et al., 1997; Kohlstaedt et al., 1992; Ollis et al., 1985; Sousa et al., 1993). The conserved protein core, responsible for nucleotide polymerization, is formed by the palm subdomain. It folds into an RNA recognition motif (RRM) containing four conserved sequence motifs (A, B, C, and D) (Lang et al., 2013). The thumb and fingers subdomains are variable, and they can be aligned only among closely related polymerases (Lang et al., 2013).

Evolutionary relationships within each of the seven families of right-hand polymerases have been extensively studied, and partial phylogenies for some of them have been obtained (Cerný et al., 2014; Filée et al., 2002; Koonin, 1991; Villarreal and DeFilippis, 2000). Nevertheless, evolutionary relationships between the individual polymerase families within the right-hand polymerase superfamily are not fully understood, primarily because sequence differences between homologous but highly diverged polymerases are too high to allow for classical distance-based phylogenetic studies (Zanotto et al., 1996). Recently, Mönttinen and colleagues (Mönttinen et al., 2014) inferred the evolutionary relationships between right-hand polymerase families using the HSF program, which performs comparison and classification of protein structures (Ravantti et al., 2013). This approach allowed proposing evolutionary relationships among polymerases with known structure, giving particularly reliable phylogenies for polymerases within each family. Nevertheless, the statistical support for inter-family associations was still quite low (Mönttinen et al., 2014).

In contrast to protein sequence, which may diverge considerably over time, protein structure changes much more slowly (Holm and Sander, 1996). It is maintained by the high plasticity of interactions among several amino acid residues. Particular intra- and inter-chain interactions are achieved in a variety of ways (hydrogen bonding, stacking interactions of aromatic residues, hydrophobic interactions, etc.) without substantial changes in the protein fold,

despite extensive sequence divergence (Illergård et al., 2009). The protein core is the most conserved part of all proteins. Amino acid residues involved in important contacts are usually not only well conserved but also are located at the same positions of the conserved folds (Illergård et al., 2009). The protein core is surrounded by less conserved region, which show higher sequence similarity only among closely related proteins. Changes in these domains lead to changes in enzyme specificity or to changes in protein interacting partners (Lu et al., 2013). Nevertheless, conserved residues present in highly divergent proteins may not convey sufficient phylogenetic signal to unveil deeper ancestral relationships among organisms (Zanotto et al., 1996). For this reason, the evolutionary stability of protein tertiary structures can be used to reconstruct the evolutionary relationships of distantly related proteins.

One of the approaches to increasing phylogenetic evidence is to create a character matrix quantifying the morphological features of the studied proteins. Such a matrix can then be combined with protein sequence alignment during phylogenetic inference to increase the amount of available useful information (Scheeff and Bourne, 2005).

In this study, we present the first robust phylogenetic tree to describe evolutionary relationships among right-hand polymerases based on comparison of both their structure and sequence. The resulting tree allowed us to speculate about the evolutionary history of right-hand polymerases and their role in the evolution of life.

MATERIALS AND METHODS

Selection of right-hand polymerase representatives

The polymerases were selected from the SCOP database (Murzin et al., 1995) superfamily of RNA/DNA polymerases (e.8.1). This condition leads to quite a narrow definition of right-hand polymerases because it includes only polymerases with known tertiary protein structure while excluding, for example, all eukaryote-infecting DNA virus polymerases for which structural information is missing. Some polymerases are not listed in the SCOP superfamily e.8.1, despite apparently being members of it, as is the case with Qβ phage polymerase (PDB ID 3AVX) (Takeshita and Tomita, 2010), which was arbitrarily added to our list despite not being listed in the e.8.1 superfamily.

Selected polymerases were clustered via BLASTCLUST (Altschul et al., 1997) to allow grouping using an identity cut-off of 40%. Proteins with higher sequence identity can be easily aligned using only sequence information (Elofsson, 2002; Illergård et al., 2009). The representatives of polymerase groups created by BLASTCLUST were selected manually. Structures with a bound template, substrate, and/or primer, structures of non-mutated proteins, high-resolution structures, and structures with maximal solved protein chain length were preferred to minimize differences arising from conformational changes in polymerases at different steps of the enzymatic cycle.

Comparison of right-hand polymerase structures and sequences

Structural superposition of selected right-hand polymerases was calculated using the DALI server (Holm and Rosenström, 2010). The structure-based sequence alignment of the polymerase palm subdomain sequences was generated using an automatic algorithm implemented in T-Coffee Expresso (Armougom et al., 2006). The known tertiary structure of selected polymerases was used to improve the final alignment (Armougom et al., 2006).

A character matrix describing structural features of selected right-hand polymerases was constructed manually. Individual quantified protein features were selected on an empirical basis by comparing the structural and functional features used previously for the description of these enzymes (Gong and Peersen, 2010; Hansen et al., 1997; Lang et al., 2013; Sousa et al., 1993; Steitz, 1999; Černý et al., 2014). Each of the matrix columns represents a single selected character typical for at least one but not all viral RNA-dependent RNA polymerases (RdRPs) while the matrix rows represent each evaluated polymerase. The structural characters were coded for subsequent analysis in MrBayes as standard data (0–9). Their character was set as unordered, allowing them to move freely from one state to another (*e.g.*, a character designated as “0” can change to “2” without passing “1”).

Phylogenetic analyses

The best-fitting model of amino acid residue substitutions was tested in PROTTEST 2.4 (Abascal et al., 2005). The BLOSUM matrix, with a proportion of

invariable sites and a gamma-shaped distribution of rates across sites (Yang, 1994), was chosen. Phylogenetic analysis was performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003). MrBayes was selected for analysis because it is the best currently available method for reconstruction of distant evolutionary relationships that is less prone to attracting long branches using proper model and appropriate taxon sampling (Glenner et al., 2004; Huelsenbeck and Ronquist, 2001). The analysis was run using a mixed dataset including both sequence and structural features (datatype=mixed). The analysis consisted of two runs with four chains (one cold and three heated) and was run for 10 million generations and sampled every 100 generations. The first 25% of the samples were discarded as a burn-in period. The average standard deviation of the split frequencies was significantly below 0.01. Chain convergence was verified with the AWTY system (Wilgenbusch et al., 2004). The equal settings were used in analyses of phylogenetic tree stability. Moreover, datasets with (i) excluded individual conserved motifs or (ii) excluded individual representatives of all polymerase families were used to verify the robustness of the phylogenetic tree topology. This verification allowed us to detect possible systematic sources of error during the inferential process. The first approach is intended to evaluate the variation in the contribution of phylogenetic signal along the alignment during the phylogenetic inference. The second is a kind of jackknifing, which we performed to reveal artificial results originating from long-branch attraction between individual polymerase families (Husmeier and Mantzaris, 2008; Lyons-Weiler and Hoelzer, 1997).

Testing of congruence between structure- and sequence-borne phylogenetic information

We also performed a series of experiments to test the level of agreement between sequence- and structure-based phylogenetic trees. There are several well-tested state transition probability matrices in use for amino acid–based phylogenetic inference (Abascal et al., 2005; Posada and Buckley, 2004). That is not the case, however, for structural information–based character-state matrices, such as the one we constructed, because there is no probabilistic basis for structural change and stasis. Therefore, it is paramount to evaluate if the signals obtained from sequence and structure are congruent (*i.e.*, support the same tree).

To test the congruence, we created a set of 87 alignments using a sliding window of size 5, 10, 20, or 50 amino acid residues moving along the polymerase protein alignment at five amino acid residues per step. Only alignments in which at least one amino acid residue was present in each sequence were used. Sequences in the sliding window were multiplied to the length of 200 amino acid residues. The original alignment and all of these random alignments were used to produce a phylogenetic tree using neighbor-Joining with the p-distance method in Mega 6 (Tamura et al., 2013). The resulting trees were compared with a phylogenetic tree generated by MrBayes based only on structure information using Robinson–Foulds distance (Makarenkov and Leclerc, 2000; Robinson and Foulds, 1981).

RESULTS

Selection of right-hand polymerase representatives

The final set of polymerases included 22 enzymes representing six polymerase families: viral RdRPs; viral RNA-dependent DNA polymerases (RdDPs); DNA-dependent DNA polymerase (DdDP) families A, B, and Y; and single-subunit DNA-dependent RNA polymerases (DdRPs) (Table 1).

Comparison of right-hand polymerase structure and sequence

The overall protein architecture of these proteins was compared (Fig. 1), and only the palm subdomain was included in further studies. The protein structures of all selected right-hand polymerase palm subdomains were superimposed, and conserved sequence motifs were mapped onto them (Fig. 2). Finally, a structure-based sequence alignment was generated covering the entire palm subdomain of all selected right-hand polymerases (Fig. 3). The only two 100% conserved amino acid residues are two aspartate residues in motifs A and C (Fig. 3), which are responsible for the binding of divalent metal ions crucial for the terminal nucleotidyl transfer reaction (Hansen et al., 1997). These aspartate residues are structurally superimposable for all right-hand polymerases, being positioned at the end of the first RRM β-strand in motif A (i), and in the turn between the second and third RRM β-strands in motif C (ii) (Fig. 3).

Quantification of structural similarities

To avoid circularity, any systematics procedures relies on the choice and definition of characters before the inferential procedure starts. Therefore, we established a criterion to build a set of binary-state structural characters, by means of which we selected and quantified 4 functional and 22 structural features for subsequent phylogenetic analysis. Characters describing these features were encoded into a character-state matrix (Table 2). The individual two-state characters were defined as follows.

1) Polymerase template: In native systems, these are right-hand polymerases that use DNA only (DdDPs and DdRPs), RNA only (viral RdRPs), or both (viral RdDPs) as a template during replication *in vivo* (Johansson and Dixon, 2013; Ng et al., 2008; Sale et al., 2012). In artificial systems, some RNA-dependent RNA polymerases also may use DNA as the template and vice versa (Arnold et al., 1999). This potential was not taken into account because it is not a native characteristic of these enzymes.

2) Polymerase product: During genome replication, the right-hand polymerases produce either DNA or RNA daughter molecules *in vivo* (Johansson and Dixon, 2013; Ng et al., 2008; Sale et al., 2012). Under artificial conditions, some polymerases can produce both (Arnold et al., 1999), but this possibility was not taken in account.

3) Polymerization initiation: Right-hand polymerases can start nucleic acid polymerization either *de novo* or using RNA or protein primers (Ferrer-Orta et al., 2006; Johansson and Dixon, 2013; Ng et al., 2008; Sale et al., 2012).

4) Additional protein domains: Additional protein domains can be attached to right-hand polymerases and provide higher fidelity in removing improperly incorporated nucleotides (Wu and Beese, 2011), degrade the template molecule (Schneider et al., 2014), or interact with polymerase partners (Tao et al., 2002).

5) Overall polymerase architecture: The succession of fingers, palm, and thumb subdomain modules varies in different right-hand polymerases. A part of the finger subdomain is always embedded in the middle of a palm subdomain. The remaining part of the finger subdomain can be positioned at the N-terminal

part of the polymerase or it does not have to be developed (Fig. 1). The thumb subdomain is located at the C-terminal end of most right-hand polymerases, but in the case of single-subunit RNA polymerases and DNA polymerases I, it can be located at the N terminus (Ollis et al., 1985; Sousa et al., 1993).

6) Overall polymerase conformation: The finger subdomain of some viral RdRPs contains protrusions called "fingertips." These fingertips interact directly with the thumb subdomain, encircling whole polymerase active sites. Polymerases with a whole active site encircled by fingertips were marked as closed (Ferrer-Orta et al., 2006); the other polymerases were marked as open.

7) Size of the F1 subdomain: The part of the finger subdomain located at the N-terminal end (F1) is missing in some polymerases (D polymerases I, single-subunit RNA polymerases). Other polymerases contain F subdomains of various lengths (Fig. 1).

8) Total size of the finger subdomain: The finger subdomain is quite variable in length, from only a few amino acid residues to long domains containing a few hundred residues (Fig. 1).

9) Size of the palm subdomain: The palm subdomain is very conservative in length with some differences mainly due to the length of helix-bearing conserved sequence motif B (Figs. 2 and 3).

10) Palm domain organization: The succession of conserved sequence motifs is highly conserved among right-hand polymerases. They are arranged in alphabetical order: A, B, C, and D. In RdRPs of viruses within the family *Birnaviridae*, the conserved sequence motifs are reordered, succeeding in order C, A, B, and D (Gorbalenya et al., 2002; Pan et al., 2007).

11) Structure preceding motif A: Conserved sequence motif A is located at the N terminus of the palm subdomain. In some polymerases, the motif is located at the very N-terminal end of the palm subdomain. In other polymerases, this motif can be preceded by a helix or β strand (Figs. 2 and 3).

12) Helix in motif A: The structure of motif A is extremely conserved. It forms a conserved β strand followed by an α or 3_{10} helix (Figs. 2 and 3).

13) Amino acid residue at alignment position 40: The amino acid residue at alignment position 40 is important for selection of incoming nucleotides. Viral RNA polymerases accommodate an acidic amino acid residue in this position while DNA polymerases contain an aromatic residue in position 40 (Hansen et al., 1997).

14) Amino acid residue at alignment position 56: A glycine residue before motif B (alignment position 56) is one of the classical markers of viral right-hand polymerases (Bruenn, 2003).

15) Length of helix in motif B: The helix accommodating conserved sequence motif B is extremely long in DNA polymerases I and single-subunit RNA polymerases. In other right-hand polymerases, the helix is much shorter (Figs. 2 and 3).

16) Amino acid residue at alignment position 64: The amino acid residue at position 64 is crucial for distinguishing between NTP and dNTP. In RNA polymerases, this position is occupied by an asparagine, aspartate, or glutamate residue, which allows interaction with the 2' hydroxide of an incoming nucleotide ribose. DNA polymerases accommodate an aromatic or short aliphatic amino acid residue, which does not allow such interactions (Hansen et al., 1997).

17) Interaction between amino acid residues at alignment positions 40 and 64: In RNA polymerases, there is a hydrogen bond between amino acid residues in positions 40 and 64. In nonreplicative DNA polymerases, the contact is provided by hydrophobic interaction (Hansen et al., 1997).

18) Kink in the helix in motif B: The helix accommodating conserved sequence motif B is usually straight. In single-subunit RNA polymerases and DNA polymerases I, the helix accommodates a kink in its N-terminal part while in viral RdDPs, the kink is positioned in the C-terminal part of the helix (Figs. 2 and 3).

19) B (A) - C motif connection: A loop preceding two antiparallel β strands accommodating conserved motif B is usually very short, being formed only by a few amino acid residues. In some polymerases, it can accommodate a short

helix. In RdRPs of viruses from the families *Reoviridae* and *Cystoviridae*, the loop can be formed by a long helix (Figs. 2 and 3).

20) Two antiparallel β strands in motif B: Two antiparallel β strands accommodating a conserved sequence motif B are a key marker of right-hand polymerases. Nevertheless, the β strands are not formed in some polymerases, and the position is occupied only by β-like stretches (Figs. 2 and 3).

21) Amino acid residue at alignment position 116: The amino acid residue at alignment position 116 is involved in coordination of the divalent ions necessary for the terminal nucleotide transfer reaction (Gong and Peersen, 2010). In viral RdRPs and RdDPs, the position is occupied by a glutamate or glutamine residue while in DdDPs and single-subunit RNA polymerases, the same position is occupied by aspartate or serine residue (Figs. 2 and 3).

22) C - D motif connection: The conserved sequence motifs C and D are usually directly connected. Nevertheless, in eukaryotic DdDP family Y, a whole protein domain is inserted between these conserved motifs (Pata, 2010).

23) Helix structure in motif D: The helix accommodating conserved sequence motif D is a right-hand polymerase marker. In Φ29 DNA polymerase, the helix is not fully formed (Figs. 2 and 3).

24) Helix position in motif D: The helix accommodating the conserved sequence motif D is usually a part of RRM. In Φ6 RdRP, the position of the helix is shifted (Fig. 2) (Butcher et al., 2001).

25) Length of the helix in motif D: The length of the helix accommodating the conserved sequence motif D is variable. Some helices are very short while some helices are extended at the N or C terminus (Figs. 2 and 3).

26) β strand in motif D: The β strand accommodating the conserved sequence motif D is quite variable. The strand can be fully absent, formed only as a β-like stretch, or fully formed (Figs. 2 and 3).

Although several other sets of characters and options of states could have been chosen, we encoded a set of characters and states on the basis of well-defined polymerase features (Černý et al., 2014; Gong and Peersen, 2010; Hansen et al.,

1997; Lang et al., 2013; Sousa et al., 1993; Steitz, 1999). Moreover, given that no structural character state coding system was available at the time of our analysis and that several multistate encodings could be possible, we chose to use binary encoding, which facilitated the inclusion of both sequence and structure data within the same Bayesian inferential framework under MrBayes.

Evolution of right-hand polymerases

In the resulting tree unifying the structure- and sequence-borne information, shown in Fig. 4, all polymerases were classified into appropriate protein families (Filée et al., 2002). All polymerase families were clearly separated. Moreover, in the resulting tree unifying the structure- and sequence-borne information all internal splits in the phylogeny of the right-hand polymerase families had a high support (Fig. 4). Single-subunit DdRPs, represented by T7 RNA polymerase, formed an inner clade in DdDP family A, which is in concordance with previously published results (Doublié et al., 1998). All viral RdRPs and the viral RdDPs included in this study formed two clearly separated sister groups. In contrast, DdDP replicating genomes of dsDNA phages were phylogenetically mixed among the DdDPs of families A and B.

The branching pattern of the polymerase families in the tree was stable. The mutual position of polymerase families in the tree was not influenced by deletion of any individual conserved sequence motif (Fig. S1) or any single polymerase family (Fig. S2). Thus, our results are not affected by artifacts coming from extremely strong phylogenetic signals present in small parts of our alignments or by long-branch attraction.

The only polymerase with a position that was not conserved was T7 RNA polymerase, the only representative of single-subunit RNA polymerases in our study. These polymerases are sometimes listed together with DdDP family A (Filée et al., 2002). Deletion of the whole DdDP family A but not T7 RNA polymerase could lead to an observed unstable resolution of the T7 RNA polymerase branching order during phylogenetic inference.

We also reconstructed the evolution of the right-hand polymerases based exclusively on the structure-based sequence alignment or on the character matrix. Both sequence and structure information–based trees had topology similar to the tree based on mixed data. The sole important difference between

the trees based only on structure or sequence information and the tree based on mixed data was the lower statistical support for individual branches and presence of more polytomies than in the case of the unifying method (Fig. S3).

Structure- and sequence-borne phylogenetic information is correlated

Finally, we checked whether the sequence- and structure-derived phylogenetic signals correlate with each other. It is well known that protein sequence and structure are tightly bound. Nevertheless, they are in principle different levels of description, each with significant synonymia and redundancy (*i.e.*, interchangeability among distinct but equivalent amino acids and structural features). Therefore, it is not necessary that the phylogenetic signals provided by these two distinct levels of description of proteins agree with each other. Moreover, no correlation would be observable if a structure-based phylogenetic tree were to be calculated on the basis of uninformative or incongruous structural features. Therefore, we estimated the Robinson–Foulds distance between the phylogenetic trees created on the basis of sequence and structure information and compared it with the Robinson–Foulds distance between the original structure-based phylogenetic tree and 87 phylogenetic trees based on randomized sequence alignment.

The Robinson–Foulds distance between the original structure- and sequence-based phylogenetic trees was estimated to be 12, lower than the estimated Robinson–Foulds distance between the original structure-based phylogenetic tree and 86 of the 87 phylogenetic trees based on randomized sequence alignment. Only in one case was the distance the same (Table S1). This result clearly shows that we chose appropriate structural markers that appear to agree with the information they provide about the evolution of the polymerases. This last finding is very important because it validates the use of two independent information sources, at different levels of description (*i.e.*, structural and primary sequence data) in inferences of deep phylogenetic associations.

DISCUSSION

Does combining sequence and structural data allow a longer-distance view of the phylogenetic horizon?

Evolutionary relationships between distantly related proteins are extremely difficult to study because insufficient sequence similarity does not allow for the precise sequence alignments required for further phylogenetic studies (Elofsson, 2002). False-negatives can arise, as happened, for example, in the case of viral RNA-dependent RNA polymerases (Zanotto et al., 1996). Additional evolutionary information is therefore necessary to overcome the lack of information in the protein sequence.

It was proved that inclusion of the protein structure could bring the lacking information (Mönttinen et al., 2014; Ravantti et al., 2013; Scheeff and Bourne, 2005; Černý et al., 2014) because of the high stasis of protein structures (Holm and Sander, 1996; Illergård et al., 2009). When applying protein structure as a trait useful for evolutionary inference, two different approaches may be employed: i) similarities among protein structures may be searched by an automatic alignment program (Mönttinen et al., 2014; Ravantti et al., 2013), or ii) they can be found, compared, and evaluated manually (Aravind et al., 2002a; Scheeff and Bourne, 2005; Černý et al., 2014). The second approach is a variation of classical evolutionary studies that used morphological similarities to reconstruct phylogenetic relationships between animal or plant species (Willi, 1947). This approach is still used, for example, in paleontology, where no molecular data usually are available (Tschopp et al., 2015). The manual approach to protein structure comparison has several positives and negatives. On the downside, we can include only proteins with 3D structures available (which prevented us from using polymerases of DNA viruses, for example). On the plus side, manual structure-based phylogenetic analysis is very flexible, and many different features, which could be very difficult to quantify automatically, can be included and used to characterize well-studied proteins. The choice of markers must rely on the empirical information available in the literature (here we used mostly Černý et al., 2014; (Gong and Peersen, 2010; Hansen et al., 1997; Lang et al., 2013; Sousa et al., 1993; Steitz, 1999; Černý et al., 2014), which may introduce unknown and unpredictable sources of biases and shortcomings. Therefore, we would argue that, as we show here, it is very

important to see if phylogenetic reconstructions based on structure comparison agree with sequence-based phylogenetic trees, which stands as a validation for deep phylogenetic associations not available from sequence information alone.

A brief history of the replicases

In this work, we unraveled the phylogenetic relationships among 22 right-hand polymerases representing all right-hand polymerases with known protein structure listed in the SCOP database (Murzin et al., 1995). All polymerase families included in our study branched into discrete, fairly well-supported lineages. Nevertheless, the position of some proteins within these families differed from previously published studies on evolution of individual polymerase families, possibly because of the large scale of this study and low number of taxa included in each polymerase family (Cerný et al., 2014; Filée et al., 2002; Villarreal and DeFilippis, 2000). Nevertheless, the main goal of this study was to elucidate relationships among polymerase families and not between individual polymerases within the families.

The branching pattern between polymerase families in our study is very similar to the branching pattern between polymerase families that was recently published by Mönttinen and colleagues (Mönttinen et al., 2014). Compared to their work, our approach led to higher statistical support for inter-familiar branches (Fig. 4). The concordance between the results coming from these two alternative studies shows that right-hand polymerase families really evolved according to the inferred pattern.

The evolution of polymerases is intertwined with that of their encoding genomes. There is no reason to advocate that DdDPs replicated primitive RNA genomes, and it is more reasonable to argue that they have evolved among organisms using DNA genomes. Therefore, no strong explanation is readily available for the presence of DNA polymerases in the RNA world. If we try to reconstruct the evolutionary history of right-hand polymerases from the perspective of the currently widely accepted model of genome evolution, the RdRPs thus appear to be the most ancient group of polymerases (Forterre, 2006b; Koonin et al., 2006). It is plausible to assume that the extant viral RdRPs represent an ancient group of enzymes, related to polymerases used to replicate RNA genomes in the RNA World stage of evolution (Koonin, 1991).

Later in evolution, some ancient viruses could have begun using DNA instead of RNA to encode their genomes (Forterre, 2002, 2006b), leading to a switch from the RNA to DNA world, the appearance of reverse transcriptases, and finally to DNA-dependent DNA and RNA polymerases (Forterre, 2013; Lazcano et al., 1988). This scenario is in concordance with the RNA World hypothesis and highlights viral RNA-dependent RNA polymerases as living "fossils" that share the most common features with polymerases used for replication of RNA genomes in the RNA world (Prangishvili et al., 2006).

We can only speculate that viral RdRPs may be a *bona fide* outgroup because they were the most variable and divergent family in terms of both sequence and structure included in our study. This could be explained also by the rapid evolution of viral RdRPs and by the sampling biases because viral RdRP structures are the most studied among right-hand polymerases. Nevertheless, the extreme diversity between viral RdRPs in numerous aspects (primer independence/RNA primers/protein primers, extreme size difference, presence of polymerases with reordered active site topology etc.) indicates that viral RdRPs are probably the most ancient group of right-hand polymerases, which is in concordance with the theory of the RNA World.

Several other protein families, which are not included in our study, could be included within right-hand polymerases. Typical example are archaeal genome replicating DNA polymerase family D (Cann and Ishino, 1999) and retrotransposon reverse transcriptases (Inouye and Inouye, 1995), which share unifying sequence features with the right-hand polymerases but their palm subdomain structure remains unsolved and therefore, they could not be included in this study. Right-hand resembling structure is present also in telomerase (Mitchell et al., 2010; Nakamura et al., 1997) and PRP8 (Dlakić and Mushegian, 2011; Galej et al., 2013), which were excluded from this study as they are not included in the SCOP superfamily of right-hand polymerases. Previous studies showed that the reverse transcriptases including retrotransposon reverse transcriptases, telomerase, PRP8 as well as viral RdDPs form a monophyletic group (Belfort et al., 2011; Makarova et al., 2002; Mönttinen et al., 2014; Nakamura et al., 1997). We have no reason to doubt that inclusion of these proteins in our study would lead to similar results resulting in the same model of polymerase evolution from RdRPs via RdDPs to DdDPs.

Finally, it has to be mentioned that right-hand polymerase are distantly related to other RRM motif containing proteins (Aravind et al., 2002b). It would be very interesting to prepare similar but widen study including ever these proteins but it is behind the scope of this article.

Considerations of deep phylogenetic inferences about polymerases

The dataset used in this study, including 22 polymerases, is rather small. The number was limited for several reasons, as follows: i) protein structures of polymerases from many different life domains (for example, eukaryotic DNA viruses – mostly DdDP family B or *Archaea* – DdDP family D) are not available, and ii) many well-resolved polymerase structures come from closely related species (for example, RNA viruses within family *Picornaviridae*). It would have made no sense to include closely related enzymes in our study because doing so would not have brought any additional information about deeper phylogenetic relationships among the right-hand polymerase families (Elofsson, 2002; Illergård et al., 2009). Therefore, we filtered these polymerases out. The third reason is that SCOP classifies proteins based on regularities in their secondary and tertiary structures (Chothia et al., 1977; Levitt and Chothia, 1976; Richardson, 1976). This approach allows effective classification of relatively simple single-domain proteins. Nevertheless, the classification of large, multi-domain proteins is problematic. Therefore, some multi-domain proteins are not listed in the SCOP superfamily of right-hand polymerases despite containing the polymerase fold. Good examples are the flavivirus polymerases from the genus Flavivirus, family *Flaviviridae*, which are not listed in SCOP despite being related to Hepatitis C virus polymerase. Furthermore, other proteins that are related to right-hand polymerases, such as telomerase and PRP8 (Dlakić and Mushegian, 2011; Galej et al., 2013), are not listed in the SCOP superfamily of right-hand polymerases, so we did not include them in our study.

Nevertheless, we believe that our dataset was sufficient to provide meaningful support for our main result, which is a description of evolutionary relationships between right-hand polymerase families. We certainly have proposed an approach that can be used to expand the right-hand polymerase phylogenetic tree when more structures are made available.

Our claims would be seriously challenged if structural and sequence similarities among right-hand polymerases were to have evolved by convergence. Such an event cannot be ruled out, but it seems to be less likely for several reasons. First, all right-hand polymerases share a number of collinearities. Their palm subdomains always fold in the RRM motif, with particular secondary structures occurring in the same order. The same is true for the conserved sequence motifs, which are accommodated on these conserved secondary structures (Steitz, 1999) (the only exceptions being the birnaviral RdRPs, which evolved from the classic fold by cyclical permutation; (Gorbalenya et al., 2002; Pan et al., 2007). Second, the palm subdomain of all right-hand polymerases is always divided into two parts by a portion of the finger subdomain that always occurs after motif A (Fig. 1). Third, even though right-hand polymerases are the most common enzymes with polymerase activity, they do not represent the only possible fold. Mammalian DNA polymerase β (Sawaya et al., 1994), bacterial DdDP family C (Lamers and O'Donnell, 2008), and cellular RdRPs (Salgado et al., 2006) can also catalyze nucleic acid polymerization by employing an entirely different protein fold, which shows that the right-hand–resembling structure is not the only functional polymerase fold.

Differences between virus polymerase- and virus capsid-based evolutionary studies?

Basically viruses can be characterized by two key features: i) a virus genome replicated (usually but not necessarily) by a virus polymerase, and ii) a virus capsid, which consists of one or more capsid proteins. The importance of these two features in defining viruses is open to discussion. From the outset of molecular evolution studies based on nucleotide sequences, viral genomes were assumed to be the most important aspect for comparative studies, eventually almost replacing viral morphology and serology in viral systematics. Given the fact that polymerases shared sequence similarity among distantly related virus families, they became widely used as marker genes to study phylogenetic relationships between distantly related viruses (Bruenn, 1991; Dolja and Carrington, 1992; Eickbush, 1994; Goldbach et al., 1994; Gorbalenya et al., 2002; Koonin, 1991; Koonin and Dolja, 1993; Poch et al., 1989; Ward, 1993).

This approach was seriously challenged by further studies. First, it was shown that the polymerase sequence by itself does not offer sufficient phylogenetic information (Zanotto et al., 1996). Second, the horizontal gene transfer of polymerase genes was described, as for example in the cases of the related phi29 and T7 phages (both order *Caudovirales*), which encode for totally unrelated DdDPs (family B in phi29 and family A in T7) (Filée et al., 2002). Third, the DdDPs exhibit a profound dichotomy; replicases of *Archaea*, *Eukarya*, and the vast majority of DNA viruses are right-hand DNA polymerases from families A, B, and D while replicases of *Eubacteria* but also a very narrow group of viruses are DdDP family C, which are totally unrelated to right-hand polymerases (Filée et al., 2002). Finally, some viruses do not encode their own polymerase at all, and they are fully dependent on the host replication apparatus.

Most of these arguments against polymerases as suitable evolutionary markers can be dealt with. i) Polymerase structure can be used to overcome low sequence similarity, as was done in this and previously published research (Mönttinen et al., 2014; Černý et al., 2014), allowing for deeper phylogenetic reconstructions that can be statistically validated. ii) If polymerases are not used as a standalone marker but together with other phylogenetic markers such as virus capsid, however, they can help with filtering out inferential systematic errors. iii) The vast majority of bacteria-infecting viruses use right-hand polymerases to replicate their genomes, and most bacteria use right-hand polymerases, at least in some processes, while the DdDP family C is missing in *Archaea*, *Eukarya*, and most viruses (Filée et al., 2002).

The biggest advantage of polymerases is that as they are present also in cellular organisms, they may help us in reconstruction of virus-cell evolutionary relationships. The overall picture of right hand-polymerase evolution as well as their presence in all life forms indicate that they may reflect the original polymerase fold and all other polymerase types (barrel-shaped cellular RNA-dependent RNA polymerases, bacterial DNA polymerase family C etc.) may evolved later. Wider discussions about the relationship between right-hand polymerases and bacterial replicases and about the evolutionary mechanisms underpinning their distribution in the biota are beyond the scope of this work but have been previously addressed in numerous excellent reviews (Forterre, 2002, 2005, 2006a, 2013; Koonin and Dolja, 2006; Koonin et al., 2006; Koonin et

al., 2008; Leipe et al., 1999). We hope that our findings show that the use of polymerases as marker genes to study the evolutionary relationships among distantly related viruses is meaningful and may be informative about the evolution of virus genomes (de Andrade Zanotto and Krakauer, 2008).

CONCLUSIONS

We reconstructed deep evolutionary relationships among right-hand polymerases by using not only the sequence but also the structural and functional features of these enzymes. Both of these sources of data share a phylogenetic signal. All polymerase families branched into discrete lineages, following a fairly robust adjacency pattern. Only single-subunit RNA polymerases formed an inner group within DNA polymerase family A. RNA-dependent RNA polymerases of RNA viruses and reverse transcriptases of retroviruses form two sister monophyletic groups and are distinguishable from all other polymerases. Based on the highest genetic variability and structural simplicity, we assume that RNA-dependent RNA polymerases are the most ancient group of right-hand polymerases. This inference is in concordance with the RNA World hypothesis, in which enzymes similar to current RNA-dependent RNA polymerases could have been used for replication of RNA genomes of ancient life entities. Our methodological approach can be of immediate use because it proposes a useful topological constraint for heuristic searches using a higher number of replicase sequences or could be extended to incorporate polymerases whose structures become available in the future.

ACKNOWLEDGMENTS

This work was supported by the Czech Science Foundation [P502/11/2116 and 14-29256S to D. R. and 15-03044S and P302/12/2490 to L. G.]; Grant Agency of University of South Bohemia [155/2013/P to L. G.]; Internal Grant Agency of the University of Life Sciences in Prague [CIGA 20134311 to B. C. B.]; the Ministry of Education, Youth and Sports of the Czech Republic [Z60220518 to D. R.]; ANTIGONE [278976 to L. G.]; and the Ministry of Education, Youth and Sports of the Czech Republic under the NPU I program (LO1218). P.M.A.Z. holds a CNPq and FAPESP grants. J.C. is a postdoctoral fellow supported by the project Postdok_BIOGLOBE (CZ. 1.07/2.3.00/30.0032), co-financed by the European Social Fund and state budget of the Czech Republic. The funders had no role in

the study design, data collection, and analysis or the decision to publish or preparation of the manuscript.

REFERENCES

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-2105.

Abrescia, N.G., Bamford, D.H., Grimes, J.M., Stuart, D.I., 2012. Structure unifies the viral universe. Annu Rev Biochem 81, 795-822.

Akita, F., Chong, K.T., Tanaka, H., Yamashita, E., Miyazaki, N., Nakaishi, Y., Suzuki, M., Namba, K., Ono, Y., Tsukihara, T., Nakagawa, A., 2007. The crystal structure of a virus-like particle from the hyperthermophilic archaeon Pyrococcus furiosus provides insight into the evolution of viruses. J Mol Biol 368, 1469-1483.

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

Aravind, L., Anantharaman, V., Koonin, E.V., 2002a. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins 48, 1-14.

Aravind, L., Mazumder, R., Vasudevan, S., Koonin, E.V., 2002b. Trends in protein evolution inferred from sequence and structure analysis. Curr Opin Struct Biol 12, 392-399.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., Notredame, C., 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res 34, W604-608.

Arnold, J.J., Ghosh, S.K., Cameron, C.E., 1999. Poliovirus RNA-dependent RNA polymerase (3D(pol)). Divalent cation modulation of primer, template, and nucleotide selection. J Biol Chem 274, 37060-37069.

Bamford, D.H., 2003. Do viruses form lineages across different domains of life? Res Microbiol 154, 231-236.

Bamford, D.H., Grimes, J.M., Stuart, D.I., 2005. What does structure tell us about virus evolution? Curr Opin Struct Biol 15, 655-663.

Belfort, M., Curcio, M.J., Lue, N.F., 2011. Telomerase and retrotransposons: reverse transcriptases that shaped genomes. Proc Natl Acad Sci U S A 108, 20304-20310.

Bruenn, J.A., 1991. Relationships among the positive strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases. Nucleic Acids Res 19, 217-226.

Bruenn, J.A., 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. Nucleic Acids Res 31, 1821-1829.

Butcher, S.J., Grimes, J.M., Makeyev, E.V., Bamford, D.H., Stuart, D.I., 2001. A mechanism for initiating RNA-dependent RNA polymerization. Nature 410, 235-240.

Cann, I.K., Ishino, Y., 1999. Archaeal DNA replication: identifying the pieces to solve a puzzle. Genetics 152, 1249-1267.

Cermakian, N., Ikeda, T.M., Miramontes, P., Lang, B.F., Gray, M.W., Cedergren, R., 1997. On the evolution of the single-subunit RNA polymerases. J Mol Evol 45, 671-681.

Cerný, J., Cerná Bolfíková, B., Valdés, J.J., Grubhoffer, L., Růžek, D., 2014. Evolution of tertiary structure of viral RNA dependent polymerases. PLoS One 9, e96070.

Chothia, C., Levitt, M., Richardson, D., 1977. Structure of proteins: packing of alpha-helices and pleated sheets. Proc Natl Acad Sci U S A 74, 4130-4134.

Cramer, P., 2002. Multisubunit RNA polymerases. Curr Opin Struct Biol 12, 89-97.

de Andrade Zanotto, P.M., Krakauer, D.C., 2008. Complete genome viral phylogenies suggests the concerted evolution of regulatory cores and accessory satellites. PLoS One 3, e3500.

Delarue, M., Poch, O., Tordo, N., Moras, D., Argos, P., 1990. An attempt to unify the structure of polymerases. Protein Eng 3, 461-467.

Dlakić, M., Mushegian, A., 2011. Prp8, the pivotal protein of the spliceosomal catalytic center, evolved from a retroelement-encoded reverse transcriptase. RNA 17, 799-808.

Dolja, V.V., Carrington, J.C., 1992. Evolution of positive-strand RNA viruses, Seminars in Virology, pp. 315-326.

Doublié, S., Tabor, S., Long, A.M., Richardson, C.C., Ellenberger, T., 1998. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. Nature 391, 251-258.

Eickbush, T.H., 1994. Origin and evolutionary relationships of retroelements, in: Morse, S.S. (Ed.), The evolutionary biology of viruses. Raven Press, 1185 Avenue of the Americas, New York, New York 10036-2806, USA, pp. 121-157.

Elofsson, A., 2002. A study on protein sequence alignment quality. Proteins 46, 330-339.

Ferrer-Orta, C., Arias, A., Escarmís, C., Verdaguer, N., 2006. A comparison of viral RNA-dependent RNA polymerases. Curr Opin Struct Biol 16, 27-34.

Filée, J., Forterre, P., Sen-Lin, T., Laurent, J., 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol 54, 763-773.

Forterre, P., 2002. The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5, 525-532.

Forterre, P., 2005. The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie 87, 793-803.

Forterre, P., 2006a. The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117, 5-16.

Forterre, P., 2006b. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. Proc Natl Acad Sci U S A 103, 3669-3674.

Forterre, P., 2013. Why are there so many diverse replication machineries? J Mol Biol 425, 4714-4726.

Galej, W.P., Oubridge, C., Newman, A.J., Nagai, K., 2013. Crystal structure of Prp8 reveals active site cavity of the spliceosome. Nature 493, 638-643.

Glenner, H., Hansen, A.J., Sørensen, M.V., Ronquist, F., Huelsenbeck, J.P., Willerslev, E., 2004. Bayesian inference of the metazoan phylogeny; a combined molecular and morphological approach. Curr Biol 14, 1644-1649.

Goldbach, R., Wellink, J., Verver, J., van Kammen, A., Kasteel, D., van Lent, J., 1994. Adaptation of positive-strand RNA viruses to plants. Arch Virol Suppl 9, 87-97.

Gong, P., Peersen, O.B., 2010. Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. Proc Natl Acad Sci U S A 107, 22505-22510.

Gorbalenya, A.E., Pringle, F.M., Zeddam, J.L., Luke, B.T., Cameron, C.E., Kalmakoff, J., Hanzlik, T.N., Gordon, K.H., Ward, V.K., 2002. The palm

subdomain-based active site is internally permuted in viral RNA-dependent RNA polymerases of an ancient lineage. J Mol Biol 324, 47-62.

Hansen, J.L., Long, A.M., Schultz, S.C., 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. Structure 5, 1109-1122.

Holm, L., Rosenström, P., 2010. Dali server: conservation mapping in 3D. Nucleic Acids Res 38, W545-549.

Holm, L., Sander, C., 1996. Mapping the protein universe. Science 273, 595-603.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Husmeier, D., Mantzaris, A.V., 2008. Addressing the shortcomings of three recent Bayesian methods for detecting interspecific recombination in DNA sequence alignments. Stat Appl Genet Mol Biol 7, Article 34.

Illergård, K., Ardell, D.H., Elofsson, A., 2009. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77, 499-508.

Inouye, S., Inouye, M., 1995. Structure, function, and evolution of bacterial reverse transcriptase. Virus Genes 11, 81-94.

Johansson, E., Dixon, N., 2013. Replicative DNA polymerases. Cold Spring Harb Perspect Biol 5.

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., Steitz, T.A., 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science 256, 1783-1790.

Koonin, E.V., 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J Gen Virol 72 (Pt 9), 2197-2206.

Koonin, E.V., 2006. Temporal order of evolution of DNA replication systems inferred by comparison of cellular and viral DNA polymerases. Biol Direct 1, 39.

Koonin, E.V., Dolja, V.V., 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit Rev Biochem Mol Biol 28, 375-430.

Koonin, E.V., Dolja, V.V., 2006. Evolution of complexity in the viral world: the dawn of a new vision. Virus Res 117, 1-4.

Koonin, E.V., Senkevich, T.G., Dolja, V.V., 2006. The ancient Virus World and evolution of cells. Biol Direct 1, 29.

Koonin, E.V., Wolf, Y.I., Nagasaki, K., Dolja, V.V., 2008. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat Rev Microbiol 6, 925-939.

Krupovič, M., Bamford, D.H., 2010. Order to the viral universe. J Virol 84, 12476-12479.

Lamers, M.H., O'Donnell, M., 2008. A consensus view of DNA binding by the C family of replicative DNA polymerases. Proc Natl Acad Sci U S A 105, 20565-20566.

Lang, D.M., Zemla, A.T., Zhou, C.L., 2013. Highly similar structural frames link the template tunnel and NTP entry tunnel to the exterior surface in RNA-dependent RNA polymerases. Nucleic Acids Res 41, 1464-1482.

Lazcano, A., Guerrero, R., Margulis, L., Oró, J., 1988. The evolutionary transition from RNA to DNA in early cells. J Mol Evol 27, 283-290.

Leipe, D.D., Aravind, L., Koonin, E.V., 1999. Did DNA replication evolve twice independently? Nucleic Acids Res 27, 3389-3401.

Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552-558.

Lu, G., Hu, Y., Wang, Q., Qi, J., Gao, F., Li, Y., Zhang, Y., Zhang, W., Yuan, Y., Bao, J., Zhang, B., Shi, Y., Yan, J., Gao, G.F., 2013. Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. Nature 500, 227-231.

Lyons-Weiler, J., Hoelzer, G.A., 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. Mol Phylogenet Evol 8, 375-384.

Makarenkov, V., Leclerc, B., 2000. Comparison of additive trees using circular orders. J Comput Biol 7, 731-744.

Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B., Koonin, E.V., 2002. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. Nucleic Acids Res 30, 482-496.

Miller, R.H., Robinson, W.S., 1986. Common evolutionary origin of hepatitis B virus and retroviruses. Proc Natl Acad Sci U S A 83, 2531-2535.

Mitchell, M., Gillis, A., Futahashi, M., Fujiwara, H., Skordalakes, E., 2010. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. Nat Struct Mol Biol 17, 513-518.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247, 536-540.

Mönttinen, H.A., Ravantti, J.J., Stuart, D.I., Poranen, M.M., 2014. Automated structural comparisons clarify the phylogeny of the right-hand-shaped polymerases. Mol Biol Evol 31, 2741-2752.

Nakamura, T.M., Morin, G.B., Chapman, K.B., Weinrich, S.L., Andrews, W.H., Lingner, J., Harley, C.B., Cech, T.R., 1997. Telomerase catalytic subunit homologs from fission yeast and human. Science 277, 955-959.

Ng, K.K., Arnold, J.J., Cameron, C.E., 2008. Structure-function relationships among RNA-dependent RNA polymerases. Curr Top Microbiol Immunol 320, 137-156.

Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G., Steitz, T.A., 1985. Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. Nature 313, 762-766.

Pan, J., Vakharia, V.N., Tao, Y.J., 2007. The structure of a birnavirus polymerase reveals a distinct active site topology. Proc Natl Acad Sci U S A 104, 7385-7390.

Pata, J.D., 2010. Structural diversity of the Y-family DNA polymerases. Biochim Biophys Acta 1804, 1124-1135.

Pelletier, H., Sawaya, M.R., Kumar, A., Wilson, S.H., Kraut, J., 1994. Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. Science 264, 1891-1903.

Poch, O., Sauvaget, I., Delarue, M., Tordo, N., 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. EMBO J 8, 3867-3874.

Posada, D., Buckley, T.R., 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53, 793-808.

Prangishvili, D., Forterre, P., Garrett, R.A., 2006. Viruses of the Archaea: a unifying view. Nat Rev Microbiol 4, 837-848.

Ravantti, J., Bamford, D., Stuart, D.I., 2013. Automatic comparison and classification of protein structures. J Struct Biol 183, 47-56.

Richardson, J.S., 1976. Handedness of crossover connections in beta sheets. Proc Natl Acad Sci U S A 73, 2619-2623.

Robinson, D.R., Foulds, L.R., 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53, 131-147.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Sale, J.E., Lehmann, A.R., Woodgate, R., 2012. Y-family DNA polymerases and their role in tolerance of cellular DNA damage. Nat Rev Mol Cell Biol 13, 141-152.

Salgado, P.S., Koivunen, M.R., Makeyev, E.V., Bamford, D.H., Stuart, D.I., Grimes, J.M., 2006. The structure of an RNAi polymerase links RNA silencing and transcription. PLoS Biol 4, e434.

Sawaya, M.R., Pelletier, H., Kumar, A., Wilson, S.H., Kraut, J., 1994. Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism. Science 264, 1930-1935.

Scheeff, E.D., Bourne, P.E., 2005. Structural evolution of the protein kinase-like superfamily. PLoS Comput Biol 1, e49.

Schneider, A., Peter, D., Schmitt, J., Leo, B., Richter, F., Rösch, P., Wöhrl, B.M., Hartl, M.J., 2014. Structural requirements for enzymatic activities of foamy virus protease-reverse transcriptase. Proteins 82, 375-385.

Shutt, T.E., Gray, M.W., 2006. Bacteriophage origins of mitochondrial replication and transcription proteins. Trends Genet 22, 90-95.

Sousa, R., Chung, Y.J., Rose, J.P., Wang, B.C., 1993. Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. Nature 364, 593-599.

Steitz, T.A., 1999. DNA polymerases: structural diversity and common mechanisms. J Biol Chem 274, 17395-17398.

Takeshita, D., Tomita, K., 2010. Assembly of Q(Takeshita & Tomita) viral RNA polymerase with host translational elongation factors EF-Tu and -Ts. Proc Natl Acad Sci U S A 107, 15733-15738.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30, 2725-2729.

Tao, Y., Farsetta, D.L., Nibert, M.L., Harrison, S.C., 2002. RNA synthesis in a cage--structural studies of reovirus polymerase lambda3. Cell 111, 733-745.

Timinskas, K., Balvočiūtė, M., Timinskas, A., Venclovas, Č., 2014. Comprehensive analysis of DNA polymerase III α subunits and their homologs in bacterial genomes. Nucleic Acids Res 42, 1393-1413.

Tschopp, E., Mateus, O., Benson, R.B., 2015. A specimen-level phylogenetic analysis and taxonomic revision of Diplodocidae (Dinosauria, Sauropoda). PeerJ 3, e857.

Villarreal, L.P., DeFilippis, V.R., 2000. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. J Virol 74, 7079-7084.

Ward, C.W., 1993. Progress towards a higher taxonomy of viruses. Res Virol 144, 419-453.

Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2004. AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.

Willi, H., 1947. Problemen deg biologische Systematik, Forschugen und Fortschritte, pp. 276-279.

Wu, E.Y., Beese, L.S., 2011. The structure of a high fidelity DNA polymerase bound to a mismatched nucleotide reveals an "ajar" intermediate conformation in the nucleotide selection mechanism. J Biol Chem 286, 19758-19767.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39, 306-314.

Zanotto, P.M., Gibbs, M.J., Gould, E.A., Holmes, E.C., 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J Virol 70, 6083-6096.

Zhu, W., Ito, J., 1994. Family A and family B DNA polymerases are structurally related: evolutionary implications. Nucleic Acids Res 22, 5177-5183.

Černý, J., Černá Bolfíková, B., Valdés, J.J., Grubhoffer, L., Růžek, D., 2014. Evolution of tertiary structure of viral RNA dependent polymerases. PLoS One 9, e96070.

FIGURE LEGENDS

Figure 1: Schematic structure and domain organization of right-hand polymerases

A) The structure of many polymerases resembles a right hand. The three domains, termed fingers, palm, and thumb (depicted in orange, purple, and cyan, respectively) can be clearly distinguished (additional domains, presented in many polymerases, are depicted in grey). Although the structure of the fingers and thumb subdomains is variable and conserved only among closely related polymerases, the palm subdomain always contains the so-called RNA recognition motif (RRM), formed by four antiparallel β-strands packed beneath two α-helices. This conserved structural motif is formed by sequence motifs called A, B, C, and D (depicted in blue, dark green, yellow, and red, respectively). B) Despite the fact that the domains in right-hand polymerases are arranged in various ways, two important collinearities can be described: (i) the palm subdomain is always divided by the finger subdomain into two parts, and (ii) the N-terminal part of the palm subdomain always contains conserved motif A while the C-terminal portion bears motifs B, C, and D. The only exception is the RdRPs of viruses within the family Birnaviridae where motif C is included in the N-terminal portion of the palm subdomain. The rearrangement in the linear sequence of the conserved motif was produced by a circular permutation. Despite this rearrangement, the position of the conserved motifs within the protein structure is almost identical (Gorbalenya et al., 2002; Pan et al., 2007). Virus names are as follows: MMLV – Moloney murine leukemia virus; HIV1 – Human immunodeficiency virus 1; HCV – Hepatitis C virus; BVDV – Bovine viral diarrhea virus; NORV – Norwalk virus; RHDV – Rabbit hemorrhagic disease virus; POLV – Poliovirus; FMDV – Foot and mouth disease virus; IBDV – Infectious bursal disease virus; MORV – Mammalian orthoreovirus.

Figure 2: Palm domain structure of selected polymerases

The structures of all 22 selected polymerase palm subdomains are depicted in the same orientation. Conserved motifs A, B, C, and D are shown in blue, green, yellow, and red, respectively. The molecular rendering in this figure was created in Swiss PDB Viewer.

Figure 3: Structure-based sequence alignment of right-hand polymerases

The PDB ID of each individual polymerase is listed at the beginning of each row. The numbers at the beginning and the end of each row respectively indicate the positions of the first and last amino acid residues on the appropriate row in the full-length protein with polymerase activity (including all additional protein

domains). The numbering above the alignment describes the position of individual amino acid residues in the alignment. The amino acid residues located in conserved sequence motifs A, B, C, and D are highlighted by color, as in Figure 1: blue (A), green (B), yellow (C), and red (D). Amino acid residues forming α helices, 3_{10} helices, and β strands are in red, green, and blue, respectively. Solvent-accessible amino acid residues are in lower-case letters, and solvent-inaccessible residues are in upper-case letters. Amino acid residues with a positive phi torsion angle, amino acid residues hydrogen bonded to a main-chain amide, or amino acid residues hydrogen bonded to a main-chain carbonyl are underlined, in bold, or in italics, respectively. The bottom row shows the Clustal consensus. Note that there are only two 100% conserved amino acid residues in the entire alignment: aspartate residues at positions 35 and 115 in motifs A and C, respectively.

Figure 4: Phylogenetic tree of right-hand polymerases

The phylogenetic tree was calculated by a Bayesian analysis unifying sequence and structural information. Individual polymerases are listed in the tree using the appropriate PDB IDs. Polymerase families are highlighted by colored ellipses. The phylogenetic relationships among viral RdRPs could not be solved with meaningful statistical significance at this scale.

TABLE LEGENDS

Table 1: Selected representatives of right-hand polymerases

Twenty-two representatives of different polymerases with a known protein structure were selected from the SCOP superfamily of DNA/RNA polymerases (e.8.1) as described in 2.1. The selected polymerases were classified into six protein families. Furthermore, the proteins were assigned to corresponding protein types and to organisms coding these proteins. For all protein groups, SCOP right-hand polymerase nomenclature was used. The structure of each protein is characterized by a PDB ID and the corresponding chain ID (c). The resolution of protein structure (res.) and co-crystalized molecules are depicted.

Table 2: Character matrix

Individual polymerase structures are introduced with a PBD ID code and assigned to appropriate organisms and polymerase families. The 26 selected characteristic features of individual polymerases are listed in the matrix as follows: (1) polymerase template: 0 only DNA, 1 both RNA and DNA, and 2 only RNA; (2) polymerase product: 0 DNA, 1 RNA; (3) polymerization initiation: 0 RNA primer, 1 *de novo*, and 2 protein primer; (4) additional protein domains: 0 present, 1 absent; (5) overall polymerase architecture: 0 T-P1-F-P2, 1 F1-P1-F2-P2-T, and 2 P1-F-P2-T; (6) overall polymerase conformation: 0 open, 1 closed; (7) size of F1 subdomain: 0 absent, 1 <70, 2 70–150, and 3 >150; (8) size of total finger subdomain: 0 very short (<35), 1 short (36–59), 2 normal (60–79), 3 long (80–149), and 4 very long (>150); (9) size of palm subdomain: 0 short (<150), 1 long (>150); (10) palm domain organization: 0 ABCD, 1 CABD; (11) structure before motif A: 0 none, 1 helix, and 2 strand; (12) helix in motif A: 0 α helix, 1 3_{10} helix; (13) amino acid residue at alignment position 40: 0 acidic residue, 1 aromatic residue, and 2 other; (14) amino acid residue at alignment position 56: 0 glycine, 1 other; (15) length of helix in motif B: 0 long, 1 normal; (16) amino acid residue at alignment position 64: 0 aromatic amino acid residue, 1 asparagine, aspartate, or glutamate residue, and 2 short aliphatic amino acid residue; (17) interaction between amino acid residues at alignment positions 40 and 64: 0 none, 1 hydrophobic interaction, 2 hydrogen bond; (18) kink in helix in motif B: 0 in N-terminal part of helix, 1 no kink, and 2 in C-terminal part of helix; (19) B (A) - C motif connection: 0 very short loop, 1 structured, and 2 very long and structured; (20) two antiparallel β strands in motif B: 0 present, 1 not formed and β-like stretches only; (21) amino acid residue at alignment position 116: 0 glutamate residue, 1 aspartate or asparagine residue, and 2 other; (22) C - D motif connection: 0 normal, 1 inserted protein domain; (23) helix in motif D: 0 α helix, 1 helix-like structure; (24) helix position in motif D: 0 normal, 1 shifted; (25) length of helix in motif D: 0 normal, 1 extended at N terminus, 2 extended at C terminus, and 3 very short; (26) β strand in motif D: 0 absent, 1 long β strand, 2 no formed β strand and β-like stretches only, and 3 short β strand.

SUPPLEMENTARY FIGURE LEGENDS

Figure S1: Phylogenetic tree of right-hand polymerases without the removed motifs

We removed sequences and structural features corresponding to motifs A (A), B (B), C (C), and D (D) to test the stability of the phylogenetic tree and the distribution of the phylogenetic data in the structure-based sequence alignment and character matrix. The deletion of any of these motifs did not lead to substantial changes in the topology of the resulting tree. This outcome showed that the phylogenetic signal was regularly distributed among the whole alignment and matrix. Nevertheless, the deletion of substantial conserved motifs led to a decrease in the statistical significance of individual branches and to the appearance of new polytomies.

Figure S2: Phylogenetic tree of right-hand polymerases without the removed polymerase families

We removed sequences and structural features corresponding to individual polymerase families [A) DdDP family A, B) DdDP family B, C) DdDP family Y, D) single-subunit RNA polymerases, E) viral reverse transcriptases, and F) viral RdRPs] to test the stability of the phylogenetic tree and the impact of individual polymerase families on deep branching. The deletion of a polymerase family did not lead to substantial changes in the resulting tree. This outcome showed that the absence/presence of individual polymerase families did not have an impact on the tree arrangement.

Figure S3: Phylogenetic tree of right-hand polymerases calculated using only the structure-based sequence alignment or character matrix

The deletion of either the character matrix (A) or sequence alignment (B) led to a decrease in the statistical significance of most branches, and new polytomies appeared. Nevertheless, the overall structure of the phylogenetic tree remained similar.

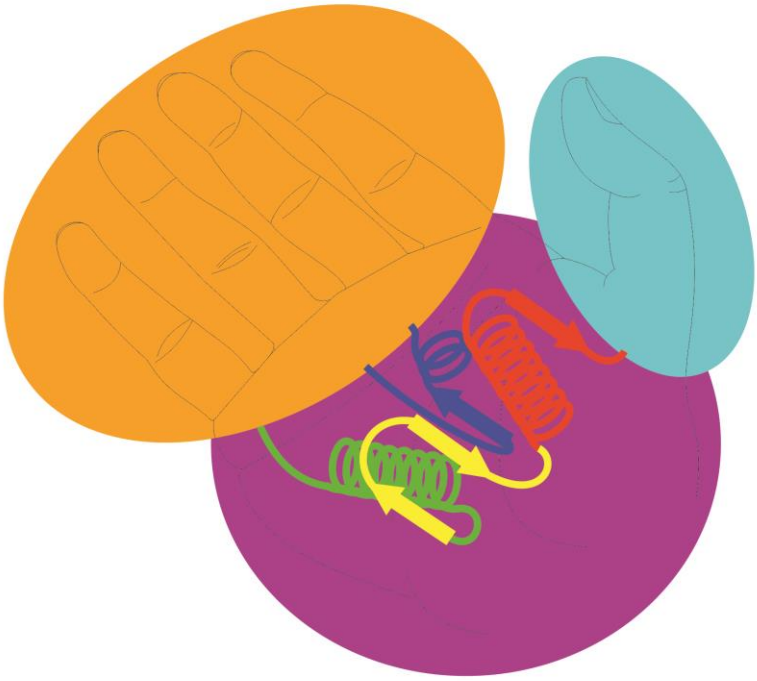
SUPPLEMENTARY TABLE

Table S1: Robinson–Foulds distances between the structure- and sequence-only–based phylogenetic trees

FIGURES

Figure 1:

A)



B)

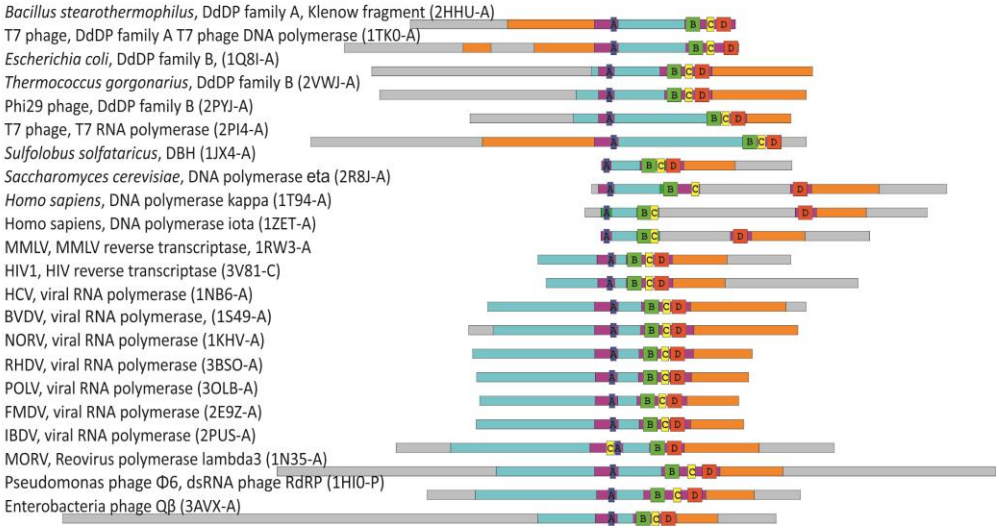


Figure 2:

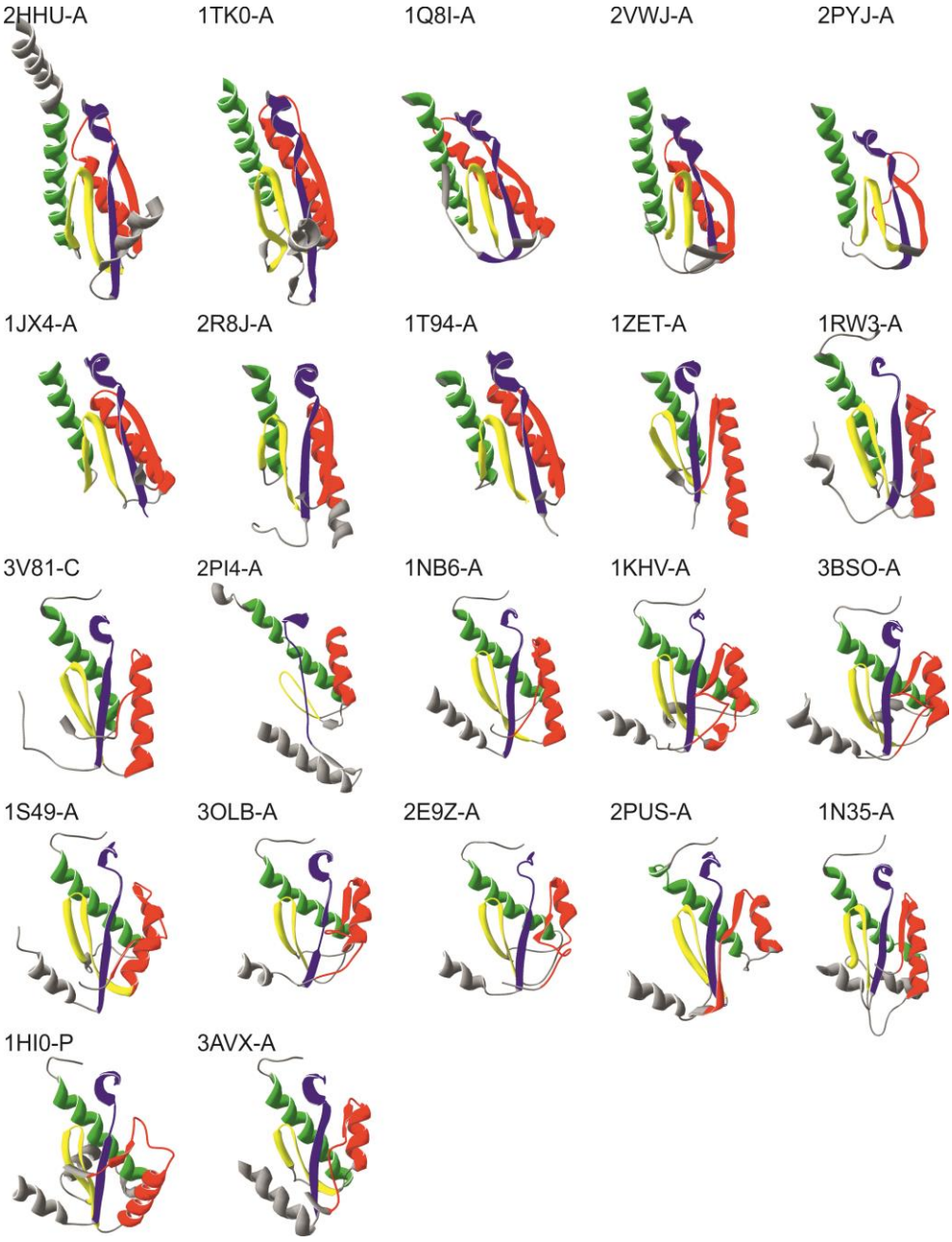

2HHU-A
1TK0-A
1Q8I-A
2VWJ-A
2PYJ-A
1JX4-A
2R8J-A
1T94-A
1ZET-A
1RW3-A
3V81-C
2PI4-A
1NB6-A
1KHV-A
3BSO-A
1S49-A
3OLB-A
2E9Z-A
2PUS-A
1N35-A
1HI0-P
3AVX-A

Figure 3:

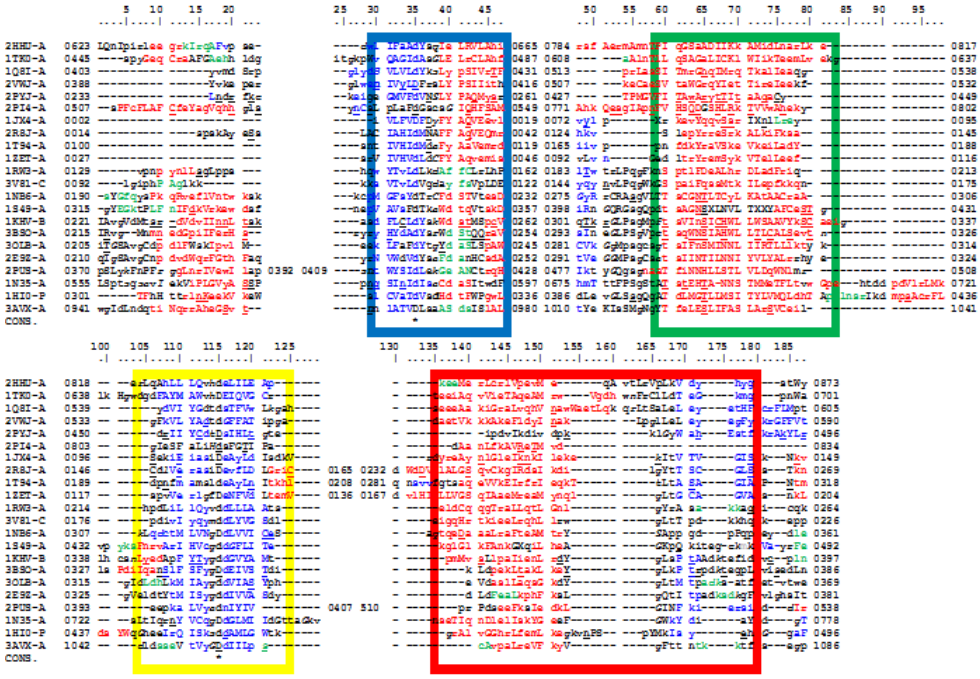


Figure 4:

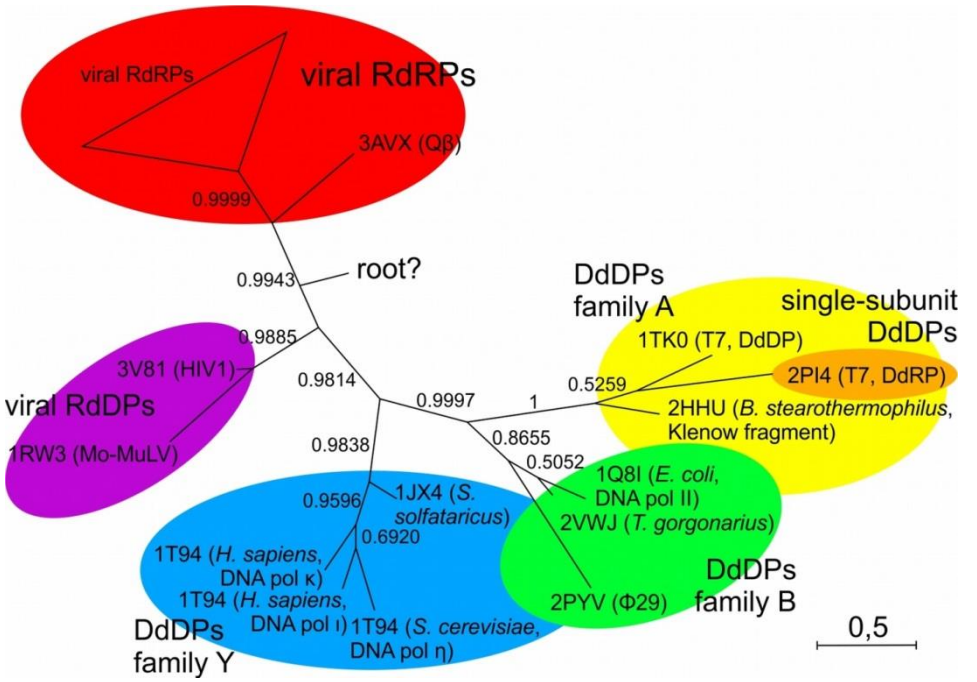


Table 1:

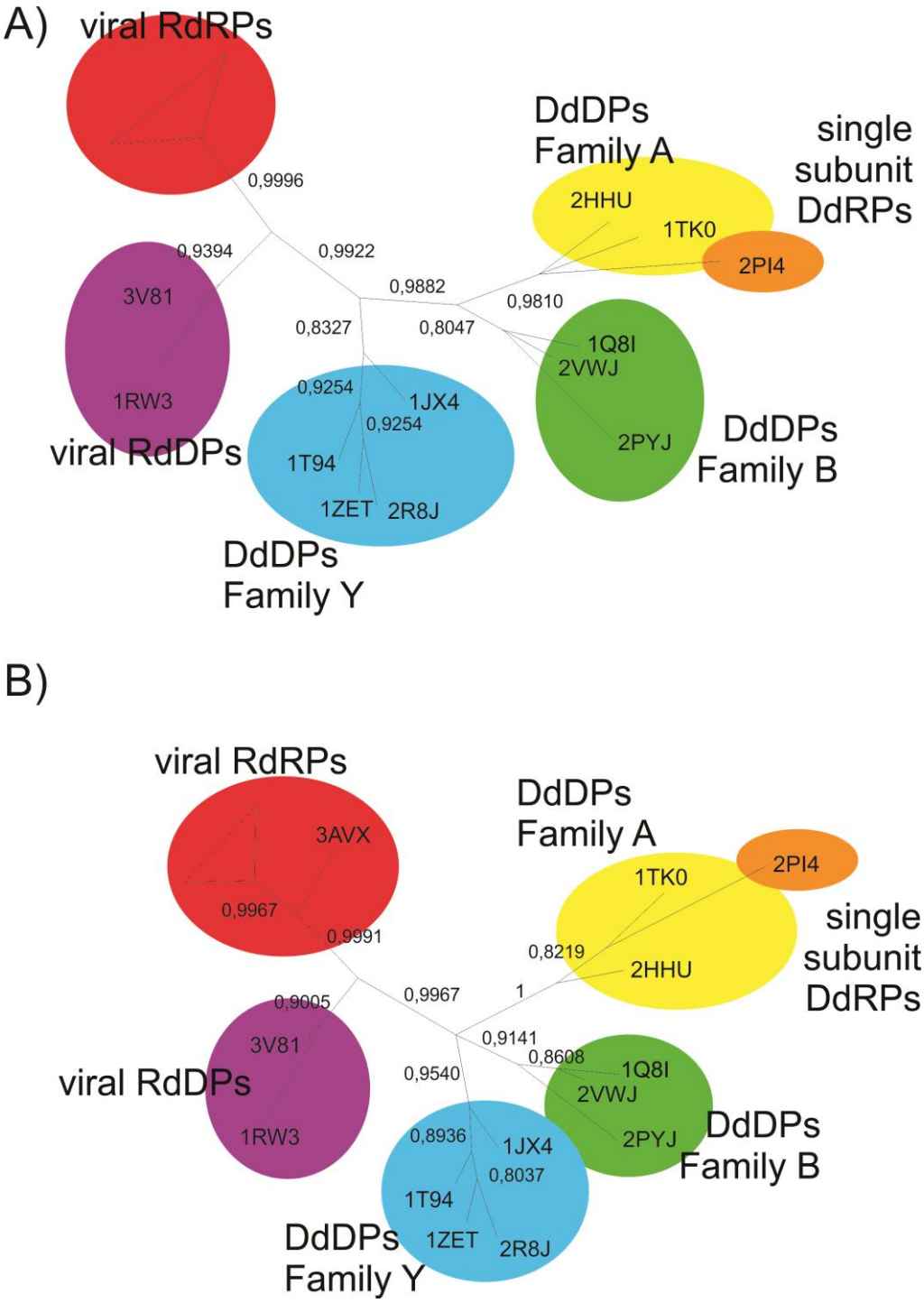
Protein family	Protein type	organism	PDB ID	ch.	res. [A]	cocrystalized molecules
Family A DNA polymerases	DNA polymerase I (Klenow fragment)	*Bacillus stearothermophilus*	2HHU	A	1,8	template, primer, dCTP
	T7 phage DNA polymerase	T7 phage	1TK0	A	2,3	template, primer, ddCTP
Family B DNA polymerases	Family B DNA polymerases	*Escherichia coli*	1Q8I	A	2	-
		Thermococcus gorgonarius	2VWJ	A	2,78	template
		Phi29 phage	2PYJ	A	2,03	template, primer, dTTP
DNA polymerase family Y	DinB homolog (DBH)	*Sulfolobus solfataricus*	1JX4	A	1,7	template, primer, ddADP
	DNA polymerase eta	*Saccharomyces cerevisiae*	2R8J	A	3,1	template, primer, dCTP
	DNA polymerase kappa	*Homo sapiens*	1T94	A	2,4	-
	DNA polymerase iota	*Homo sapiens*	1ZET	A	2,3	template, primer, dTTP
Reverse transcriptases	MMLV reverse transcriptase	Moloney murine leukemia virus	1RW3	A	3	-
	HIV reverse transcriptase	Human immunodeficiency virus 1	3V81	C	2,85	template, primer, nepavirine
Single-subunit RNA polymerases	T7 RNA polymerase	T7 phage	2PI4	A	2,5	template, product, 3'dGTP
Viral RNA dependent RNA polymerases	Viral RNA polymerases	Hepatitis C virus	1NB6	A	2,6	-
		Bovine viral diarrhea virus	1S49	A	2,5	-
		Rabbit hemorrhagic disease virus	1KHV	A	1,74	template, primer, CTP
		Norwalk virus	3BSO	A	3	GTP
		Poliovirus	3OLB	A	2,41	template, product, ddCTP
		Foot and mouth disease virus	2E9Z	A	3	template, product, UTP
		Infectious bursal disease virus	2PUS	A	2,4	-
	Reovirus polymerase lambda3	Mammalian orthoreovirus	1N35	A	2,5	template, product, 3'dCTP
	dsRNA phage RdRP	Pseudomonas phage Φ6	1HI0	P	3	template, GTP
		Enterobacteria phage Qβ	3AVX	A	2,41	template, product, 3'dGTP

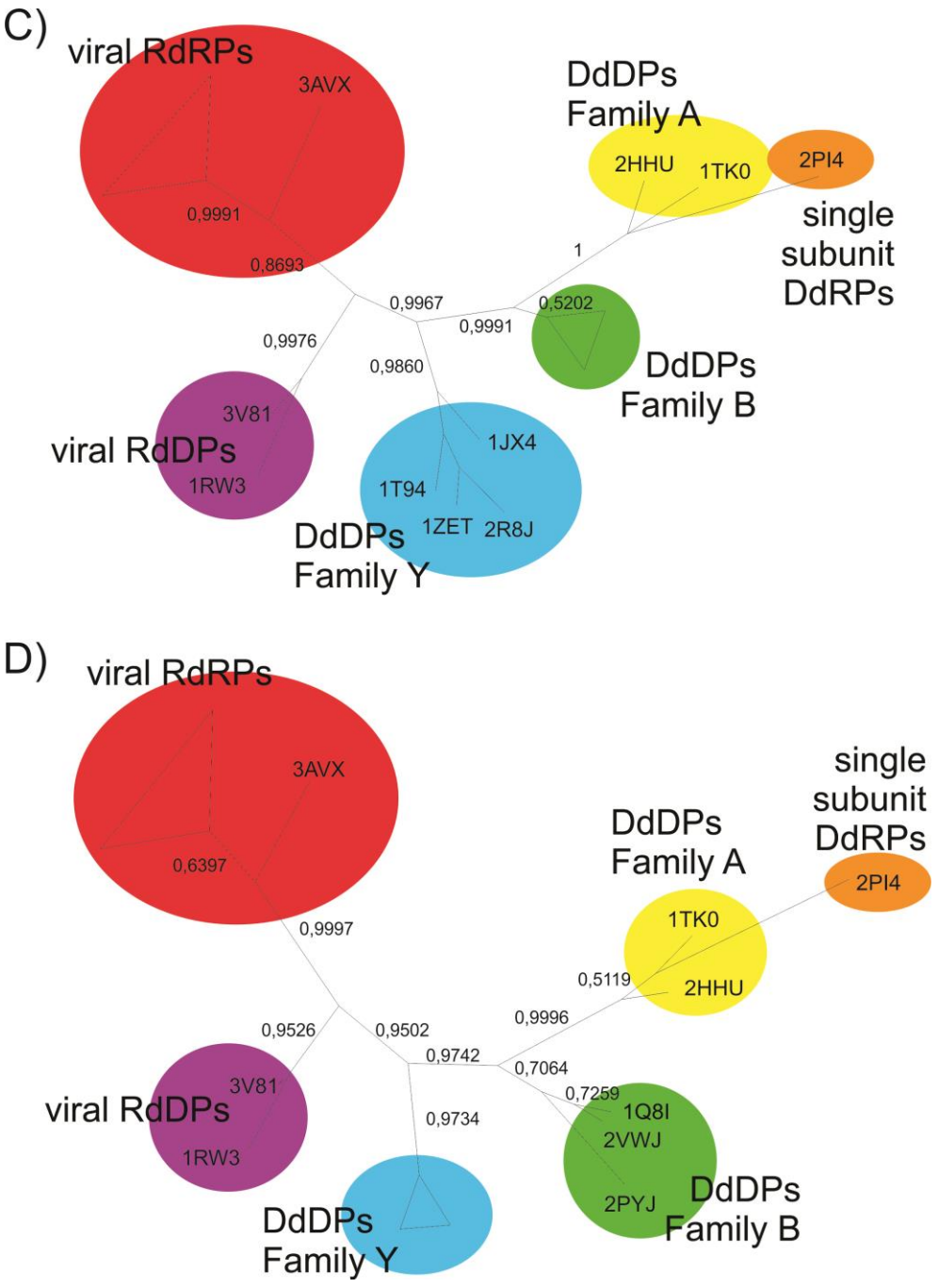
Table 2:

protein family	organism	pdb ID	features												
			1	2	3	4	5	6	7	8	9	10	11	12	13
Family A DNA polymerases	*Bacillus stearothermophilus*	2hhu	0	0	0	0	0	0	0	3	1	0	1	0	0
	T7 phage	1tk0	0	0	0	0	0	0	0	3	1	0	1	0	0
Family B DNA polymerases	*Escherichia coli*	1q8i	0	0	0	0	1	0	1	3	1	0	2	0	1
	Thermococcus gorgonarius	2vwj	0	0	0	0	1	0	1	3	1	0	2	0	1
	Phi29 phage	2pyj	0	0	0	0	1	0	1	4	0	0	1	0	1
DNA polymerase family Y	*Sulfolobus solfataricus*	1jx4	0	0	0	0	2	0	0	2	0	0	0	0	1
	Saccharomyces cerevisiae	2r8j	0	0	0	0	2	0	0	3	0	0	0	0	1
	Homo sapiens	1t94	0	0	0	0	2	0	0	2	0	0	0	0	1
	Homo sapiens	1zet	0	0	0	0	2	0	0	2	0	0	0	0	1
Reverse transcriptases	Moloney murine leukemia virus	1rw3	1	0	0	0	1	0	2	0	0	0	1	1	1
	Human immunodeficiency virus 1	3v81	1	0	0	0	1	0	2	0	0	0	1	1	1
One-subunit RNA pol.	T7 phage	2pi4	0	1	1	0	0	0	0	4	0	0	1	0	2
Viral RNA dependent RNA polymerases	Hepatitis C virus	1nb6	2	1	1	1	1	1	3	1	1	0	1	0	0
	Bovine viral diarrhea virus	1s49	2	1	1	0	1	1	3	1	1	0	1	0	0
	Rabbit hemorrhagic disease virus	1khv	2	1	2	1	1	1	3	1	0	0	1	0	0
	Norwalk virus	3bso	2	1	2	1	1	1	3	1	0	0	1	1	0
	Poliovirus	3olb	2	1	2	1	1	1	3	1	0	0	1	1	0
	Foot and mouth disease virus	2e9z	2	1	2	1	1	1	3	1	0	0	1	1	0
	Infectious bursal disease virus	2pus	2	1	2	0	1	1	3	2	0	1	1	1	0
	Mammalian orthoreovirus	1n35	2	1	2	0	1	1	3	2	0	0	1	0	0
	Pseudomonas phage Φ6	1hi0	2	1	1	0	1	1	3	1	0	0	1	0	0
	Enterobacteria phage Qβ	3avx	2	1	1	0	1	0	2	1	0	0	1	1	2

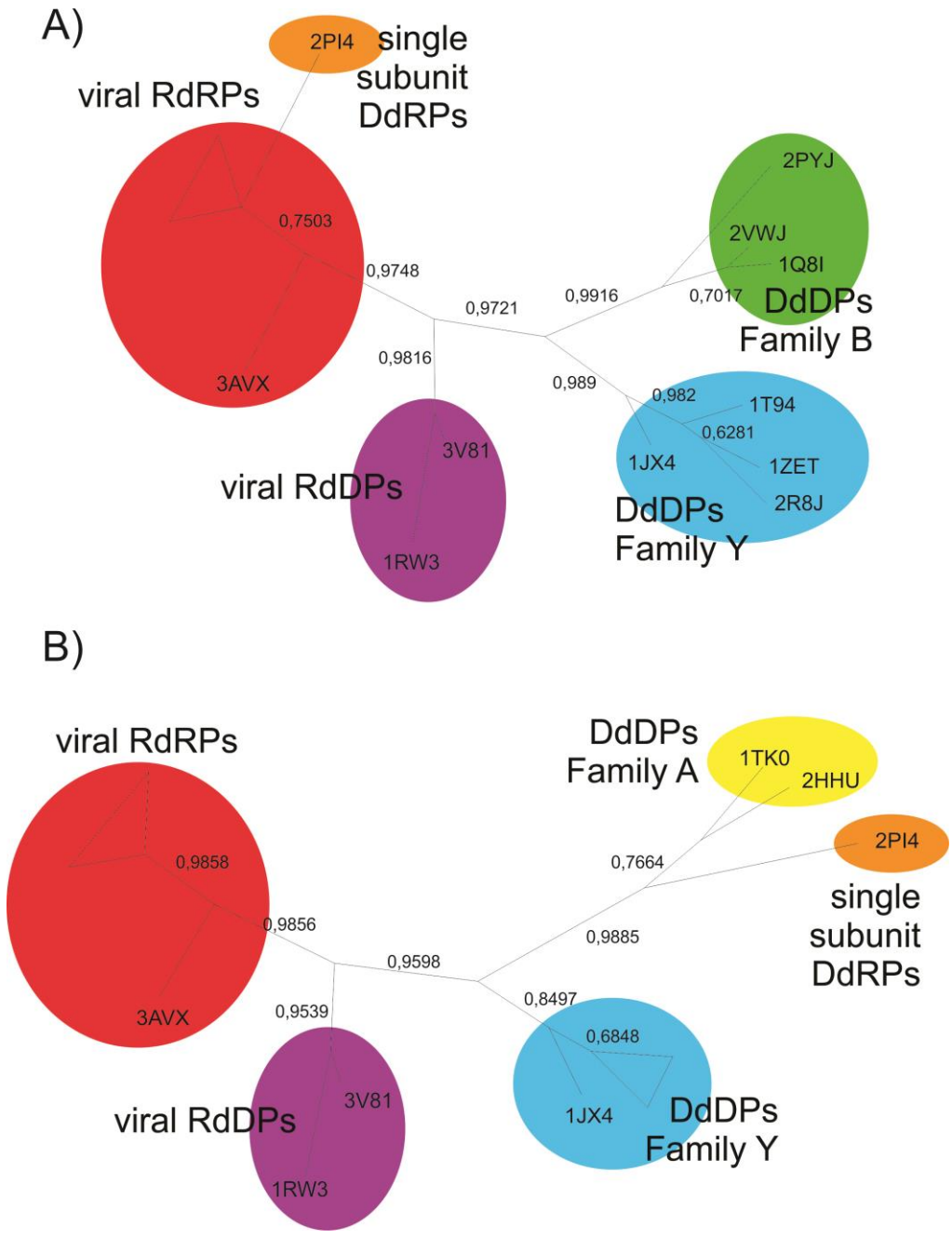
protein family	organism	pdb ID	features												
			14	15	16	17	18	19	20	21	22	23	24	25	26
Family A DNA polymerases	*Bacillus stearothermophilus*	2hhu	1	0	2	0	0	0	0	0	0	0	0	0	1
	T7 phage	1tk0	1	1	2	0	1	1	0	0	0	0	0	0	1
Family B DNA polymerases	*Escherichia coli*	1q8i	1	1	2	0	1	0	0	2	0	0	0	1	1
	Thermococcus gorgonarius	2vwj	1	1	2	0	1	0	0	2	0	0	0	0	1
	Phi29 phage	2pyj	1	1	2	0	1	0	0	2	0	1	0	3	1
DNA polymerase family Y	*Sulfolobus solfataricus*	1jx4	1	1	0	1	1	0	0	2	0	0	0	0	1
	Saccharomyces cerevisiae	2r8j	1	1	0	1	1	0	0	0	1	0	0	2	1
	Homo sapiens	1t94	1	1	0	1	1	0	0	0	1	0	0	2	1
	Homo sapiens	1zet	1	1	0	1	1	0	0	0	1	0	0	2	1
Reverse transcriptases	Moloney murine leukemia virus	1rw3	0	1	0	1	2	0	0	1	0	0	0	0	2
	Human immunodeficiency virus 1	3v81	0	1	0	1	2	0	0	1	0	0	0	0	2
One-subunit RNA pol.	T7 phage	2pi4	1	0	1	2	0	0	1	1	0	0	0	0	0
Viral RNA dependent RNA polymerases	Hepatitis C virus	1nb6	0	1	1	2	1	0	0	1	0	0	0	0	2
	Bovine viral diarrhea virus	1s49	0	1	1	2	1	1	0	1	0	0	0	0	2
	Rabbit hemorrhagic disease virus	1khv	0	1	1	2	1	1	0	1	0	0	0	3	3
	Norwalk virus	3bso	0	1	1	2	1	1	0	1	0	0	0	3	3
	Poliovirus	3olb	0	1	1	2	1	1	0	1	0	0	0	3	3
	Foot and mouth disease virus	2e9z	0	1	1	2	1	1	0	1	0	0	0	3	3
	Infectious bursal disease virus	2pus	0	1	1	2	1	0	0	1	0	0	0	3	1
	Mammalian orthoreovirus	1n35	0	1	0	2	1	2	0	1	0	0	0	0	3
	Pseudomonas phage Φ6	1hi0	0	1	2	0	1	2	0	1	0	0	1	0	3
	Enterobacteria phage Qβ	3avx	0	1	1	0	1	0	0	1	0	0	0	3	2

Supplementary figure 1:

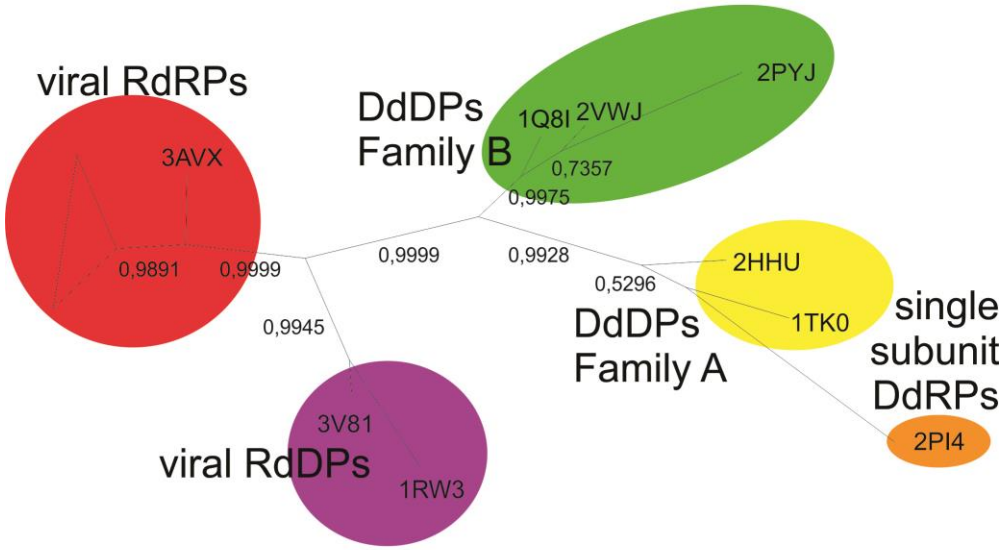



C)
viral RdRPs
3AVX
0,9991
0,8693
DdDPs
Family A
2HHU
1TK0
2PI4
single
subunit
DdRPs
1
0,9967
0,5202
0,9991
0,9976
0,9860
DdDPs
Family B
3V81
viral RdDPs
1RW3
1JX4
1T94
1ZET
2R8J
DdDPs
Family Y
D)
viral RdRPs
3AVX
0,6397
single
subunit
DdRPs
DdDPs
Family A
2PI4
1TK0
0,9997
0,5119
2HHU
0,9996
0,9526
0,9502
0,9742
0,7064
0,7259
3V81
viral RdDPs
1RW3
0,9734
1Q8I
2VWJ
2PYJ
DdDPs
Family B
DdDPs
Family Y

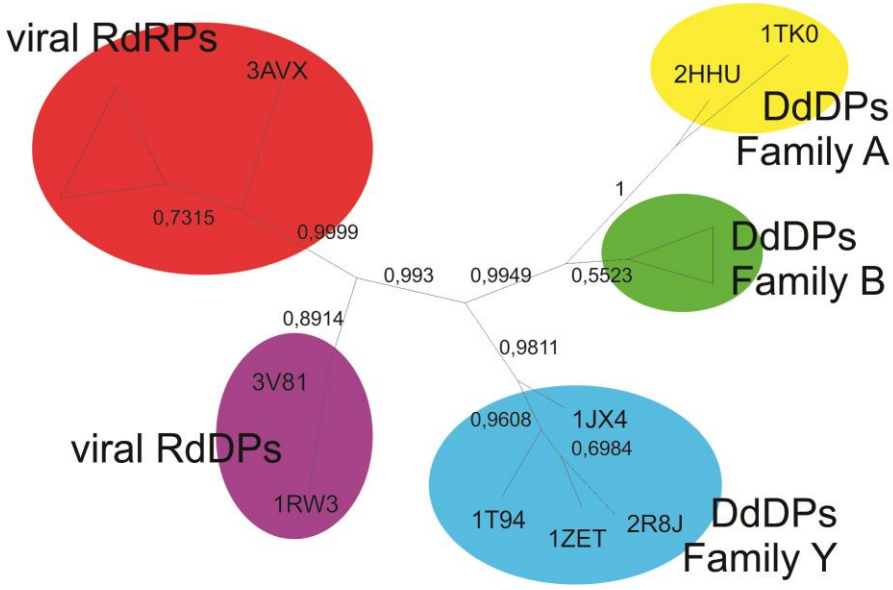
Supplementary figure 2:

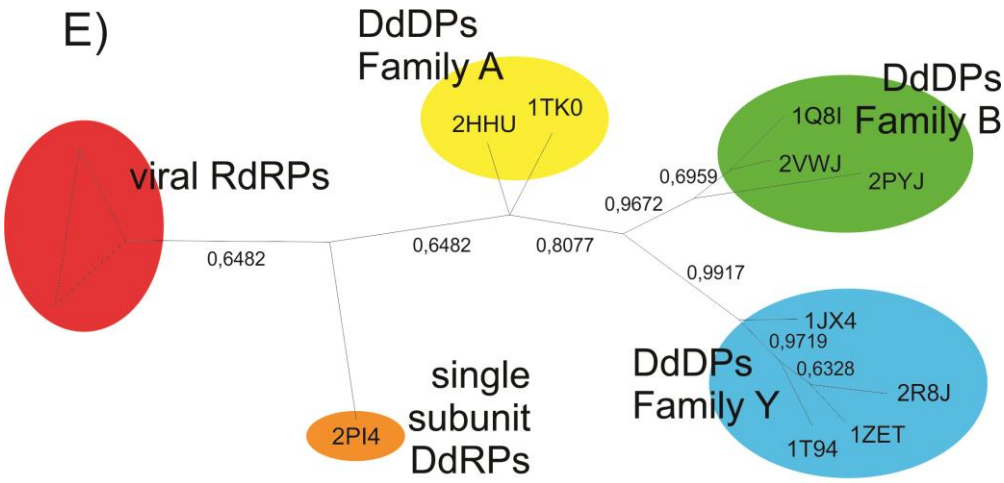


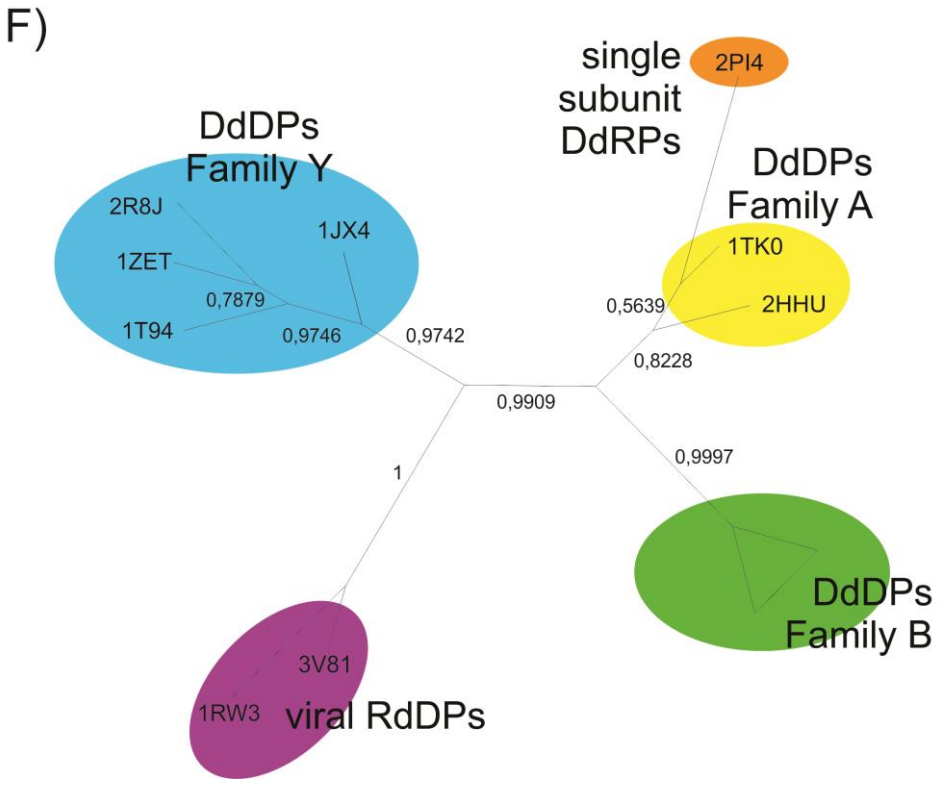
C)


viral RdRPs
3AVX
0,9891
0,9999
DdDPs
Family B
1Q8I
2VWJ
2PYJ
0,7357
0,9975
0,9999
0,9928
2HHU
0,5296
1TK0
DdDPs
Family A
single
subunit
DdRPs
2PI4
0,9945
3V81
viral RdDPs
1RW3

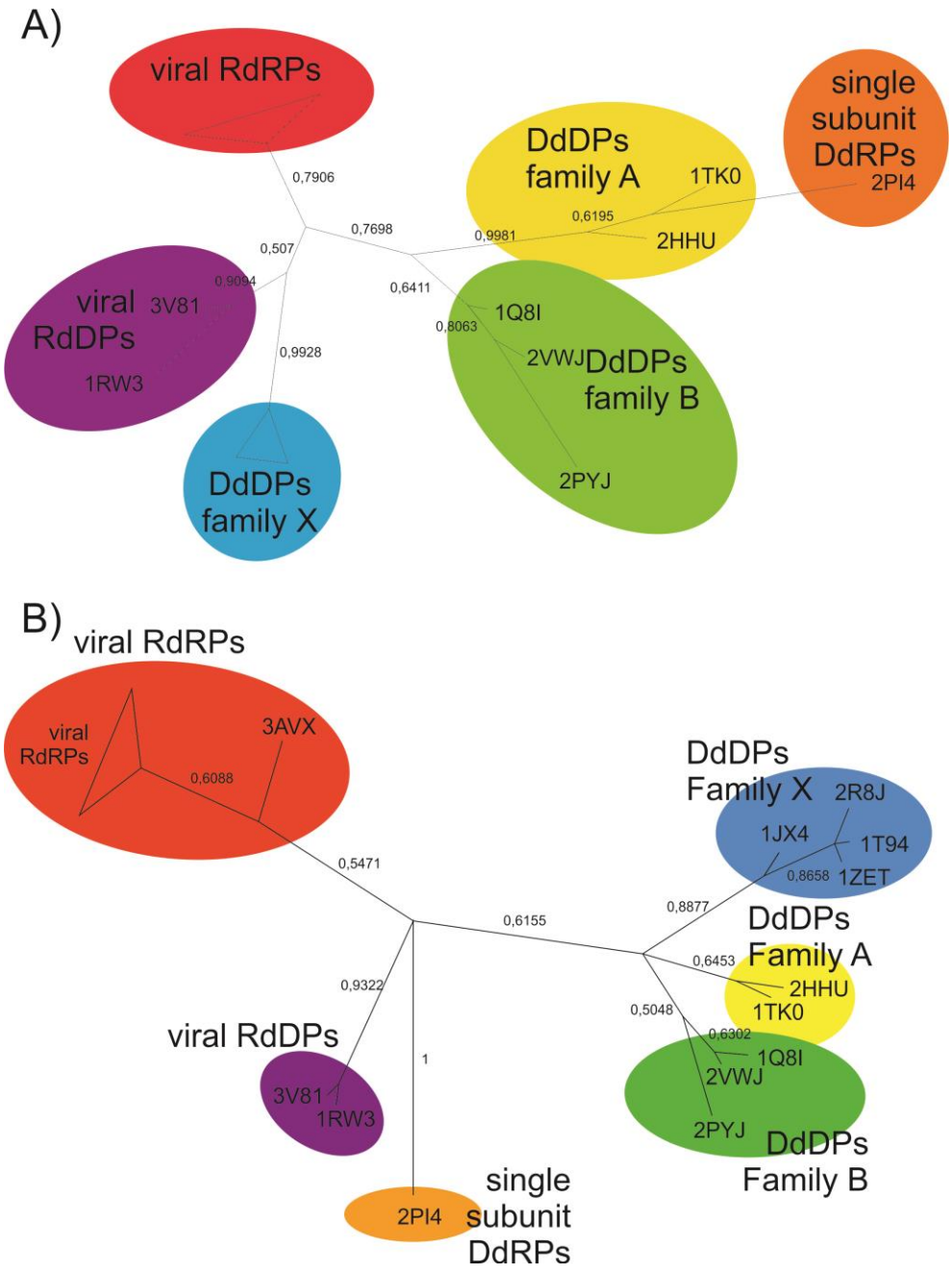
D)


viral RdRPs
3AVX
0,7315
0,9999
1TK0
2HHU
DdDPs
Family A
1
DdDPs
Family B
0,993
0,9949
0,5523
0,8914
0,9811
3V81
viral RdDPs
1RW3
0,9608
1JX4
0,6984
1T94
1ZET
2R8J
DdDPs
Family Y


E)
DdDPs
Family A
2HHU
1TK0
DdDPs
Family B
1Q8I
2VWJ
2PYJ
viral RdRPs
0,6959
0,9672
0,6482
0,6482
0,8077
0,9917
1JX4
0,9719
0,6328
2R8J
DdDPs
Family Y
1ZET
1T94
single
subunit
DdRPs
2PI4


F)
single
subunit
DdRPs
2PI4
DdDPs
Family Y
DdDPs
Family A
2R8J
1JX4
1TK0
1ZET
0,7879
0,5639
2HHU
1T94
0,9746
0,9742
0,8228
0,9909
0,9997
1
DdDPs
Family B
3V81
1RW3
viral RdDPs

Supplementary figure 3:



Supplementary table 1:

sliding window size	position	Robinson-Foulds distance	sliding window size	position	Robinson-Foulds distance	sliding window size	position	Robinson-Foulds distance
50	1-50	16	50	171-35	28	10	26-35	28
50	6-55	16	50	176-40	18	10	31-40	20
50	11-60	14	50	181-45	16	10	36-45	18
50	16-65	14	20	16-35	28	10	41-50	18
50	21-70	14	20	21-40	22	10	46-55	25
50	26-75	14	20	26-45	16	10	56-65	16
50	31-80	12	20	31-50	16	10	61-70	18
50	36-85	14	20	36-55	18	10	66-75	26
50	41-90	14	20	41-60	14	10	71-80	26
50	46-95	16	20	46-65	18	10	76-85	28
50	51-100	16	20	51-70	18	10	101-110	30
50	56-105	18	20	56-75	16	10	106-115	20
50	61-110	20	20	61-80	18	10	116-125	30
50	66-115	20	20	66-85	26	10	136-145	30
50	71-120	20	20	71-90	26	10	141-150	30
50	76-125	22	20	76-95	28	10	146-155	30
50	81-130	24	20	91-110	30	5	26-30	24
50	86-135	24	20	96-115	20	5	31-35	26
50	91-140	24	20	101-120	24	5	36-40	20
50	96-145	26	20	106-125	24	5	41-45	22
50	101-150	24	20	111-130	22	5	61-65	20
50	106-155	22	20	116-135	30	5	66-70	30
50	116-165	30	20	126-145	30	5	71-75	28
50	121-170	26	20	131-150	30	5	76-80	28
50	126-175	26	20	136-155	30	5	106-110	30
50	131-180	24	20	141-160	30	5	111-115	20
50	136-185	22	20	146-165	30	5	116-120	28
50	141-5	22	20	151-170	28	5	141-145	30
50	151-15	28	10	21-30	28	5	145-150	30

6.4 Expression of a second open reading frame present in the genome of tick-borne encephalitis virus strain Neudoerfl is not detectable in infected cells.

Manusript is under revision process in Virus Genes.

AUTHORS:

Jiří Černý (1, 2, 3)*, Martin Selinger (1, 2), Martin Palus (1, 2, 3), Zuzana Vavrušková (1), Hana Tykalová (1, 2), Lesley Bell-Sakyi (4), Libor Grubhoffer (1, 2), Daniel Růžek (1, 3)

TITLE:

Expression of a second open reading frame present in the genome of tick-borne encephalitis virus strain Neudoerfl is not detectable in infected cells.

AUTHORS AFFILIATIONS:

(1) Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 31, 370 05 České Budějovice, Czech Republic

(2) Faculty of Science, University of South Bohemia in České Budějovice, Branišovská 31, 370 05 České Budějovice, Czech Republic

(3) Veterinary Research Institute, Hudcova 296/70, 621 00 Brno, Czech Republic

(4) The Pirbright Institute, Ash Road, Pirbright, Woking, Surrey GU24 0NF, UK

CORRESPONDING AUTHOR:

Jiří Černý, Tel: +420 387 775 451; Fax: +420 385 310 388; e-mail: cerny@paru.cas.cz

ABSTRACT:

A short upstream open reading frame (uORF) was recently identified in the 5' untranslated region of some tick borne encephalitis virus (TBEV) strains. However, it is not known if this TBEV uORF (TuORF) codes for a peptide. Here we show that TuORF forms two phylogenetically separated clades which are typical of European and Siberian TBEV subtypes. Both these clades are under

positive evolutionary selection pressure. Theoretically, TuORF may code for a short hydrophobic peptide embedded in a biological membrane. However, expression of TuORF was not detectable by immunoblotting and immunofluorescence in mammalian or tick cell lines infected with TBEV strain Neudoerfl. As the TuORF sequence is evolutionarily very stable, we may speculate that it has a different biological role in the TBEV life cycle such as regulation of TBEV polyprotein expression.

KEY WORDS:

TBEV, uORF, TuORF, immunoblotting, immunofluorescence

INTRODUCTION:

Tick-borne encephalitis virus (TBEV), the causative agent of tick-borne encephalitis (TBE), is a typical representative of the genus *Flavivirus*, family *Flaviviridae* (1, 2). It is endemic in most of Central and Eastern Europe and North Asia (3) where it is the most medically important flavivirus (4). Despite the availability of effective vaccination in endemic regions, TBEV infects thousands of people annually. Many of them develop clinical manifestations of TBE, often followed by permanent decrease in their life quality. TBEV mortality varies according to the TBEV subtype (4).

The TBEV genomic RNA, which is approximately 11,000 nt long, serves also as viral mRNA. It contains a single open reading frame (ORF) encoding one polyprotein. Translation of this ORF is initiated by a classical cap-dependent scanning mechanism (5). The polyprotein is co- and post-translationally cleaved into three structural (C, M and E) and seven nonstructural (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) proteins (6). Apart from the major proteins, some flaviviruses such as Japanese encephalitis virus (JEV) and West Nile virus (WNV) produce minor proteins and peptides. Each minor protein is usually specific only for a narrow group of closely-related flaviviruses. NS1' produced by JEV (7, 8) and WARF4 produced by WNV (9, 10) are typical examples of such flaviviral minor proteins. Both these minor proteins are encoded by alternative open reading frames and produced via a ribosome frame-shifting process (11, 12). While the role of WARF4 is unknown, JEV NS1' plays an important role in virus-

host interaction, especially in virus neuroinvasiveness (8, 13) and JEV genomic RNA replication (14).

The presence of a short upstream open reading frame (uORF) in the 5' untranslated region (UTR) of some TBEV strains has been reported (15). Expression and functional importance of this second ORF (here called TuORF) remain unknown. In the present study, we investigated the expression of the hypothetical TuORF-encoded peptide in mammalian and tick cells by Western blotting and indirect immunofluorescence.

METHODS:

TBEV strains, cell lines, synthetic TuORF peptide and anti-TuORF antibodies

Low passage TBEV strain Neudoerfl (4th passage) (kindly provided by F. X. Heinz) and the strain Hypr (unknown passage history) were used in this study. TuORF presence and absence in TBEV strains Neudoerfl and Hypr respectively were verified by sequencing. Human cell lines of neural origin comprising neuroblastoma (UKF-NB-4), medulloblastoma (DAOY) and glioblastoma cells (16) and the *Ixodes ricinus* tick cell line IRE/CTVM19 (17) were used. A synthetic version of the TuORF peptide (sequence MRLLRTALAAVGLKKKC) and anti-TuORF protein A-purified mouse and rabbit polyclonal antibodies were produced by GenScript (USA). Because of high hydrophobicity, the most hydrophilic part of the peptide was synthesized together with an additional hydrophilic tail in order to obtain sufficient yields of the artificial peptide.

Bioinformatics characterization of TBEV 5' UTR and TuORF peptide

One hundred closest homologues of the TBEV strain Neudoerfl 5'UTR were identified in GenBank using the blastn algorithm (18). TBEV strains containing uORF were manually selected and classified into appropriate TBEV subtypes. Alignment of selected 5'UTRs was constructed using ClustalX (19). Protein sequences of hypothetical TuORF peptides were deduced from nucleotide sequences using the ExPASy – Translate tool (20).

Distant homologues of the TBEV TuORF peptide were sought using HHPred (21), HHblits (22), and Psi-blast algorithms (23). Basic biophysical characteristics of the TuORF peptide from TBEV strain Neudoerfl were predicted using ProtParam (24). TuORF peptide secondary structure was predicted using Jpred

(25). TuORF peptide position in the cell membrane was predicted by TMpred (26).

Phylogenetic analysis and selective constraint calculation

Phylogenetic analysis of TuORF evolution was carried out using MEGA6 (27). Protein and nucleic acid sequence alignments were processed by the neighbor-joining method using 1000 bootstrap replicates.

To calculate selective constraint, codon based sequence alignment of TuORF was constructed on the GUIDANCE server (28, 29), using the implemented ClustalW algorithm (19). The dN and dS difference was calculated in MEGA6 (27). Analyses were conducted using the Nei-Gojobori method (31). The analysis involved 17 nucleotide sequences. The variance of the difference was computed using 1000 bootstrap replicates. All ambiguous positions were removed for each sequence pair. There were a total of 29 positions in the final dataset. Wilcoxon tests were used to assess the significance of linked and unlinked synonymous and nonsynonymous scores, respectively.

Western blot assay

Mammalian and tick cell lines were infected with TBEV strain Neudoerfl at a multiplicity of infection (MOI) of 10. Virus adsorption was carried out for 1 hour. At several time points post infection (3, 6, 12, 18, 24, and 48 h in the case of mammalian cell lines, and 24, 92, 168, and 336 h in case of the tick cell line), the cells were harvested and lysed. Equal amounts of whole cell protein were separated by SDS-PAGE and transferred to nitrocellulose membranes. Transferred proteins were labeled with primary mouse or rabbit polyclonal anti-TuORF antibodies (GenScript, USA). All primary antibodies were diluted 1:200 in a 5% solution of dried milk in PBS (5% milk). Subsequently, primary antibodies were detected by horse secondary antibody conjugated with alkaline phosphatase (Vector Laboratories, USA) diluted 1:2000 in 5% milk. Labeled proteins were visualized by chemiluminescence assay using CPD Star Reagent (NEB, USA).

Immunofluorescence staining

Neuroblastoma cells were infected with TBEV strains Neudoerfl and Hypr at a MOI of either 1 or 10. Virus adsorption was carried out for 1 h. At several time

points post infection (12, 24, 48, and 72 h), cells were fixed in 4% paraformaldehyde for 15 min, rinsed in PBS and permeabilized with 0.1% Triton X-100 for 5 min. Fixed cells were treated with 50 mM NH_4Cl in a 1% solution of bovine serum albumin (BSA) in PBS to block formaldehyde autofluorescence. Further, cells were blocked with 3% BSA dissolved in PBS and labeled with either mouse or rabbit polyclonal anti-TuORF antibody (GenScript) and with chicken polyclonal anti-NS3 antibody (reactive with TBEV NS3 protein) (32). After washing in PBS, the cells were labeled with goat anti-rabbit and goat anti-chicken secondary antibodies conjugated with DyLight 594 and DyLight488, respectively (Vector Laboratories). Subsequently, the cells were mounted in Vectashield mounting medium (Vector Laboratories). Examination was done on an Olympus BX-51 fluorescence microscope equipped with an Olympus DP-70 CCD camera.

RESULTS:

An upstream ORF is present in the 5'UTR of numerous (but not all) TBEV strains as well as in the 5'UTR of some other flaviviruses

A TuORF was identified in 43 of 100 tested TBEV strains. TuORF was present in strains representative of all TBEV subtypes (European, Siberian, and Far Eastern - Supplementary Table 1). The length of the TuORF varied between 36 and 93 nt; correspondingly, the length of coded peptides varied between 13 and 31 amino acid residues (Figure 1). The modal length of the hypothetical TuORF peptide in European subtype TBEV strains was 23 amino acid residues. The most frequently-seen length of the TuORF peptide in Siberian subtype TBEV strains was 21 amino acid residues. The longest TuORF peptide was in Far Eastern TBEV strains where it could be up to 31 amino acid residues in length. The N terminal part of the TuORF peptide is conserved while its C terminal part accommodates many substitutions typical for either European or Asian TBEV subtypes (Figure 1B).

Among other tick-borne flaviviruses, uORFs were found in all 5'UTR sequences of Langat virus (LGTV) (AF253419.1, AF253420.1, EU790644.1), Kama virus (KAMV) (NC_023439.1, KF815940.1), and Karshi virus (KARV) (DQ462443.1) available in GenBank (Supplementary Figure 1). LANV and KAMV uORFs are, respectively, 339nt and 51nt long and they exceed the 5'UTR continuing also into the main ORF. In KARV, the initiating AUG codon is immediately followed

by a UAG amber stop codon. Among mosquito-borne flaviviruses, the uORF was detected only in St. Louis encephalitis virus (DQ525916.1) (Supplementary Figure 1). Sequences of these uORFs as well as the sequences of the possibly-encoded peptides are unrelated to TuORF. Sequences of other screened tick- and mosquito-borne flaviviruses did not contain any uORF (a complete list of flaviviruses that do or do not contain a uORF in their 5' UTR is shown in Supplementary Table 2).

Evolutionary history of TuORF

Reconstruction of its evolutionary history and determination of any selection pressure would indicate if the TuORF peptide has a molecular function or whether it is only a free rider in the TBEV genome.

First we reconstructed phylogenetic relationships among the TuORFs of the different TBEV strains. Nucleic acid- and protein-based analysis revealed existence of three TuORF groups corresponding to the European, Siberian and Far Eastern TBEV strains (Supplementary Figure 2). Only the position of the European strain Ek-328 in the phylogenetic tree is uncertain, possibly due to its origin. It was created by multiple passaging of TBEV in mice, which may have led to accumulation of multiple mutations (33).

To see if the uORF coding for the TuORF peptide is under selection pressure, we compared the proportion of nonsynonymous (dN) and synonymous (dS)substitutions appearing in the TuORF of different TBEV strains. A dN higher than dS 1 implies positive selection, while a dN lower than dS 1 indicates negative (purifying) selection. In the case of TuORF the overall average of dN and dS shows that number of nonsynonymous mutations is significantly higher than the number of synonymous mutations which shows that TuORF is under positive selection pressure (Table 1). Nevertheless, this trend is only poorly or not at all visible in pairwise analyses or in overall analyses done on data subsets containing only individual TBEV subtypes (Supplementary Table 3).

Bioinformatics characterization of the putative TuORF peptide

The TuORF peptide is a highly hydrophobic peptide. According to *in silico* prediction, TuORF should form a single helix embedded into a membrane with its N terminus protruding outside (Supplementary Table 4) possibly into the

lumen of the endoplasmic reticulum. No TuORF peptide homologues were found among any other protein sequences in GenBank.

The TuORF peptide was not detected in TBEV-infected cells by immunoblotting

To test TuORF peptide expression in TBEV-infected cells, we infected three human neural cell lines and one tick cell line with TBEV Neudoerfl strain as described in Methods. Neither human nor tick cells were positive for TuORF peptide expression at any time point tested while the positive control (synthetic peptide loaded onto the gel) returned a strong positive signal in all cases (Supplementary Figure 3). The results indicate that the TuORF peptide either was not expressed in the cell lines tested or its expression was extremely low, below the detection limit of the immunoblotting, which was 100ng (Supplementary Figure 4).

TuORF peptide expression was not visible in TBEV-infected cells using indirect immunofluorecence

To confirm the immunoblotting experiment results, we explored TuORF peptide expression in TBEV-infected neuroblastoma cells using indirect immunofluorescence. Both Neudoerfl (encodes for TuORF) and Hypr (does not encode for TuORF) strains of TBEV were used. Anti-TuORF staining with mouse or rabbit polyclonal antibodies did not produce any visible signal from either TBEV strain (Figure 2). Control anti-NS3 immunofluorescence staining showed a very bright signal increasing in intensity with the time post TBEV infection (Figure 2). These results show that either TBEV Neudoerfl-infected cells do not express the TuORF peptide or that TuORF peptide expression was under the detection limit of the indirect immunofluorescence.

DISCUSSION:

Minor peptides occur in some flaviviruses; for example JEV NS1' protein (7, 8) and WNV WARF4 protein (9). Presence of a uORF in the TBEV 5'UTR was described previously (15). However, it has not been determined whether or not a peptide coded by TBEV uORF is expressed in TBEV-infected cells.

Here we showed that the putative peptide coded by the TBEV strain Neudoerfl uORF was not detectably expressed in the TBEV-infected human or tick cell

lines tested. As two sets of polyclonal antibodies were used for TuORF peptide detection it is very unlikely that the negative results were caused by inability of the antibodies to detect the natural TuORF peptide.

These results can be explained in at least three different ways. (i) The simplest explanation is that the TuORF peptide is not produced in TBEV infected cells and TuORF itself is just a product of random mutation. This explanation is also supported by evolutionary analyses. (ii) The TuORF peptide may be produced under different conditions from those tested in our experiments. TBEV infects various cell types during mammalian host infection and neural cells are only the final targets (34). Other target cells such as dendritic cell, macrophage, and spleen cells are infected during primary viremia; in some of these cells the TuORF peptide may be produced. (iii) TuORF peptide is expressed in TBEV-infected cells but is rapidly degraded and therefore it is impossible to detect it.

The bioinformatics analyses showed that TuORF is present in some (but not all) TBEV strains belonging to all three TBEV subtypes. Individual TuORFs specific for European, Siberian, and Far Eastern subtypes differ in both nucleotide and amino acid sequence (Figure 1) and they form three monophyletic clades which can be clearly distinguished in the phylogenetic tree (Supplementary Figure 3). TBEV is not the only Flavivirus containing a uORF in its 5'UTR. uORFs were also detected in other flaviviruses as LGTV, KAMV, KARV, and SLEV (Supplementary Table 2). Nevertheless these uORFs do not share any sequence similarity with TuORF (Supplementary Figure 1).

It is likely that TuORF evolved by mutation of the GUG codon, which is present in TBEV strains without TuORF, to an initiating AUG codon. The TBEV 5'UTR is extremely structured (35). All the structures are very conserved and they have crucial functions in TBEV genome replication (36) and polyprotein expression (37). Therefore all mutations in the TuORF peptide have to be assessed in respect of preservation of the 5'UTR structure. The GUG/AUG codon is positioned at the base of the stem loop 1 (SL1) structure (35). As the first guanosine in GUG is not a part of SL1 but is located in the preceding internal loop, GUG can mutate to AUG without affecting the 5'UTR secondary structure.

The TBEV 5'UTR has numerous sequence-variable but structurally extremely-conserved regions, which affect TBEV replication and translation (38). Mutational analyses of these regions showed that secondary structures, but not

primary sequence, in these regions are responsible for their function (39, 40). TuORF is located in SL1, which is one of the most important structures in the TBEV 5'UTR (38). Therefore it is not surprising that the proportion of nonsynonymous mutations (dN) exceeds the proportion of synonymous mutations (dS) in this region. This indicates that the putative TuORF peptide, if expressed, does not have an exact, precisely defined role in the TBEV life cycle.

It is possible that TuORF can regulate expression of the major TBEV ORF by itself. Translation regulation by uORFs is a well-known and intensively-studied process. In most cases uORF down-regulates gene expression (43). The rate of down-regulation depends on sequence context of the uORF initiation codon, uORF length, and distance between uORF and major ORF (44). In the case of TuORF, down-regulation of the major ORF would not be great. The AUG codon initiating TuORF is in a suboptimal sequence context (acgTgcAUGC) which is far from the optimal Kosak sequence (gccRccAUGG) (45, 46). Also the length of TuORF is rather short and the distance between TuORF and the major TBEV ORF is sufficient for possible translation reinitiation. This allows us to speculate that a high proportion of ribosomes would pass the TuORF initiation codon by leaky scanning and initiate translation on the major TBEV ORF initiation codon. Nevertheless, the exact effect of TuORF presence on major TBEV polyprotein production remains unknown.

SUMMARY

We showed that uORFs are present in some strains of TBEV, LGTV, KAMV, KARV, and SLEV. TuORF sequence conservation among different TBEV subtypes is low. The TuORF peptide was not detectably expressed in TBEV strain Neudoerfl-infected cells. Therefore, we can assume that uORFs play either a minor or no role in flavivirus infection.

COMPETING INTERESTS:

The authors have declared no competing interests.

AUTHORS CONSTRIBUTION:

JC did all bioinformatics predictions and phylogenetic calculations; he participated in the immunofluorescence and western blot experiments and he drafted the manuscript. JC, MS, MP, and ZV grew the cells and did TBEV

infections. MS and ZV carried out the immunofluorescence and Western blot experiments, assisted by JC, HT and MP. LBS provided the tick cell line and critically revised the manuscript. LG and DR supervised all work and participated in the manuscript revisions.

ACKNOWLEDGMENTS:

We would like to thank B. Černá Bolfíková for her help with phylogenetic analyses, F. X. Heinz (Medical University of Vienna, Austria) for TBEV strain Neudoerfl, T. Eckschlager (Chales University in Prague, Czech Republic) for human neural cell lines and M. Bloom (Rocky Mountain Laboratories, USA) for anti-NS3 antibody. The tick cell line IRE/CTVM19 was provided by the Tick Cell Biobank.

This work was supported by the Czech Science Foundation [P502/11/2116 and GA14-29256S to D. R. and 15-03044S to L. G.), Grant Agency of University of South Bohemia [155/2013/P to L. G.], the Ministry of Education, Youth and Sports of the Czech Republic [Z60220518 to D. R.], ANTIGONE [278976 to L. G.], and by project LO1218, with financial support from the MEYS of the Czech Republic under the NPU I program. J. C. is a postdoctoral fellow supported by the project Postdok BIOGLOBE (CZ. 1.07/2.3.00/30.0032) co-financed by the European Social Fund and state budget of the Czech Republic. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

LITERATURE

1. Gritsun T.S., Nuttall P.A., and Gould E.A., Adv Virus Res 61, 317-371, 2003.
2. King A.M.Q., Adams, M.J., Carstens, E.B. and Lefkowitz, E.J., Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego, USA, 2012.
3. Ecker M., Allison S.L., Meixner T., and Heinz F.X., J Gen Virol 80 (Pt 1), 179-185, 1999.

4. Gritsun T.S., Lashkevich V.A., and Gould E.A., Antiviral Res 57, 129-146, 2003.
5. Hoenninger V.M., Rouha H., Orlinger K.K., Miorin L., Marcello A., Kofler R.M., and Mandl C.W., Virology 377, 419-430, 2008.
6. Harris E., Holden K.L., Edgil D., Polacek C., and Clyde K., Novartis Found Symp 277, 23-39; discussion 40, 71-23, 251-253, 2006.
7. Blitvich B.J., Scanlon D., Shiell B.J., Mackenzie J.S., and Hall R.A., Virus Res 60, 67-79, 1999.
8. Melian E.B., Hinzman E., Nagasaki T., Firth A.E., Wills N.M., Nouwens A.S., Blitvich B.J., Leung J., Funk A., Atkins J.F., Hall R., and Khromykh A.A., J Virol 84, 1641-1647, 2010.
9. Faggioni G., Pomponi A., De Santis R., Masuelli L., Ciammaruconi A., Monaco F., Di Gennaro A., Marzocchella L., Sambri V., Lelli R., Rezza G., Bei R., and Lista F., Virol J 9, 283, 2012.
10. Faggioni G., Ciammaruconi A., De Santis R., Pomponi A., Scicluna M.T., Barbaro K., Masuelli L., Autorino G., Bei R., and Lista F., Int J Mol Med 23, 509-512, 2009.
11. Firth A.E., and Atkins J.F., Virol J 6, 14, 2009.
12. Sun J., Yu Y., and Deubel V., Microbes Infect 14, 930-940, 2012.
13. Ye Q., Li X.F., Zhao H., Li S.H., Deng Y.Q., Cao R.Y., Song K.Y., Wang H.J., Hua R.H., Yu Y.X., Zhou X., Qin E.D., and Qin C.F., J Gen Virol 93, 1959-1964, 2012.
14. Satchidanandam V., Uchil P.D., and Kumar P., Novartis Found Symp 277, 136-145; discussion 145-138, 251-133, 2006.
15. Chausov E.V., Ternovoi V.A., Protopopova E.V., Kononova J.V., Konovalova S.N., Pershikova N.L., Romanenko V.N., Ivanova N.V., Bolshakova N.P., Moskvitina N.S., and Loktev V.B., Vector Borne Zoonotic Dis 10, 365-375, 2010.
16. Ruzek D., Vancova M., Tesarova M., Ahantarig A., Kopecky J., and Grubhoffer L., J Gen Virol 90, 1649-1658, 2009.
17. Bell-Sakyi L., Zweygarth E., Blouin E.F., Gould E.A., and Jongejan F., Trends Parasitol 23, 450-457, 2007.
18. Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J., J Mol Biol 215, 403-410, 1990.

19. Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., and Higgins D.G., Bioinformatics 23, 2947-2948, 2007.
20. Artimo P., Jonnalagedda M., Arnold K., Baratin D., Csardi G., de Castro E., Duvaud S., Flegel V., Fortier A., Gasteiger E., Grosdidier A., Hernandez C., Ioannidis V., Kuznetsov D., Liechti R., Moretti S., Mostaguir K., Redaschi N., Rossier G., Xenarios I., and Stockinger H., Nucleic Acids Res 40, W597-603, 2012.
21. Söding J., Biegert A., and Lupas A.N., Nucleic Acids Res 33, W244-248, 2005.
22. Remmert M., Biegert A., Hauser A., and Söding J., Nat Methods 9, 173-175, 2012.
23. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., and Lipman D.J., Nucleic Acids Res 25, 3389-3402, 1997.
24. Gasteiger E. H.C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. in M. W.J. (ed). *Protein Identification and Analysis Tools on the ExPASy Server*. Humana Pres, 2005, pp. 571-607.
25. Cuff J.A., Clamp M.E., Siddiqui A.S., Finlay M., and Barton G.J., Bioinformatics 14, 892-893, 1998.
26. Hofmann K., and Stoffel W. TMbase - A database of membrane spanning proteins segments. Biol Chem Hoppe-Seyler, 1993.
27. Tamura K., Stecher G., Peterson D., Filipski A., and Kumar S., Mol Biol Evol 30, 2725-2729, 2013.
28. Penn O., Privman E., Ashkenazy H., Landan G., Graur D., and Pupko T., Nucleic Acids Res 38, W23-28, 2010.
29. Penn O., Privman E., Landan G., Graur D., and Pupko T., Mol Biol Evol 27, 1759-1767, 2010.
30. Korber B. in Rodrigo A. G. L.G.H. (ed). Computational Analysis of HIV Molecular Sequences. Kluwer Academic Publishers, Dordrecht, Netherlands, 2000, pp. 55-72.
31. Nei M., and Gojobori T., Mol Biol Evol 3, 418-426, 1986.
32. Mitzel D.N., Best S.M., Masnick M.F., Porcella S.F., Wolfinbarger J.B., and Bloom M.E., Virology 381, 268-276, 2008.
33. Romanova L.I.u., Gmyl A.P., Dzhivanian T.I., Bakhmutov D.V., Lukashev A.N., Gmyl L.V., Rumyantsev A.A., Burenkova L.A., Lashkevich V.A., and Karganova G.G., Virology 362, 75-84, 2007.

34. Růžek D., Dobler G., and Donoso Mantke O., Travel Med Infect Dis 8, 223-232, 2010.
35. Tuplin A., Evans D.J., Buckley A., Jones I.M., Gould E.A., and Gritsun T.S., Nucleic Acids Res 39, 7034-7048, 2011.
36. Li X.F., Jiang T., Yu X.D., Deng Y.Q., Zhao H., Zhu Q.Y., Qin E.D., and Qin C.F., J Gen Virol 91, 1218-1223, 2010.
37. Paranjape S.M., and Harris E., Curr Top Microbiol Immunol 338, 15-34, 2010.
38. Gritsun T.S., and Gould E.A., Virology 366, 8-15, 2007.
39. Gebhard L.G., Filomatori C.V., and Gamarnik A.V., Viruses 3, 1739-1756, 2011.
40. Lodeiro M.F., Filomatori C.V., and Gamarnik A.V., J Virol 83, 993-1008, 2009.
41. Nooij F.J., Van der Sluijs-Gelling A.J., Jol-Van der Zijde C.M., Van Tol M.J., Haas H., and Radl J., J Immunol Methods 134, 273-281, 1990.
42. Walker M.J., Montemagno C., Bryant J.C., and Ghiorse W.C., Appl Environ Microbiol 64, 2281-2283, 1998.
43. Firth A.E., and Brierley I., J Gen Virol 93, 1385-1409, 2012.
44. Ryabova L.A., Pooggin M.M., and Hohn T., Virus Res 119, 52-62, 2006.
45. Kozak M., Nature 308, 241-246, 1984.
46. Kozak M., Cell 44, 283-292, 1986.

TABLES:

Table 1 - Determination of selection pressure on the TuORF peptide:

Overall analysis revealed significant positive selection acting on the complete set of TuORF peptides. This evolutionary trend was not confirmed at the level of TuORFs encoded by individual TBEV subtypes. The probability of rejecting the null hypothesis of strict-neutrality (dN = dS) in favor of the alternative hypothesis (Negative selection: dN < dS, any selection pressure: dN≠dS, or positive selection: dN<dS) is shown. P values lower than 0.05 are considered significant at the 5% level and are shown in bold type. The values were calculated as described in Methods.

	Negative selection		Any selection pressure		Positive selection	
	dS-dN	P	dN-dS	p	dN-dS	p
all TuORFs	-2.4251	1	**2.2685**	**0.0251**	**2.3365**	**0.0106**
TuORFs of European TBEV strains	0.364	0.3583	-0.3907	0.6967	-0.3971	1
TuORFs of Siberian TBEV strains	-0.501	1	0.4858	0.628	0.4996	0.3091
TuORFs of Far Eastern TBEV strains	0.3579	0.3605	-0.34	0.7392	-0.3348	1

FIGURE LEGENDS:

Figure 1 - Comparison of TuORF nucleotide and protein sequences:

Full length sequence of TBEV 5´UTR strain Neudoerfl (A). uORF sequence is marked in color, while remaining part of the 5´UTR is in grey. uORF start and stop codons as well as major ORF start codons are underlined. Alignment of uORF nucleotide sequences (B) and TuORF protein sequences (C) of various TBEV strains. GenBank accession numbers of all nucleotide sequences used in this study are listed in Supplementary Table 1. Protein sequences of hypothetical TuORF peptides were deduced from nucleotide sequences as indicated in Methods.

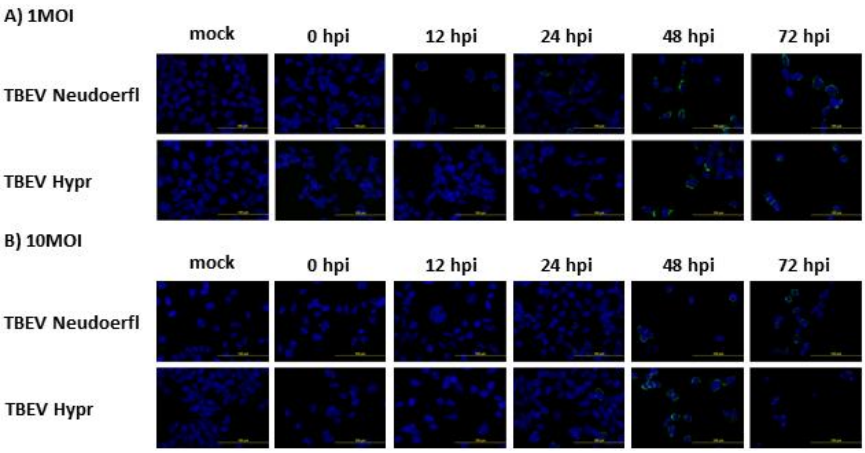
Figure 2 – Attempted detection of TuORF peptide expression by immunofluorescence:

Human neuroblastoma cells were infected by TBEV strains Neudoerfl (sample, TuORF containing TBEV strain) and Hypr (negative control, TuORF lacking TBEV strain). Mock- and TBEV-infected (MOI of 1, panel A; MOI of 10, panel B) cells were grown and fixed at various time points were stained with anti-NS3 antibody (green) and anti-TuORF antibody (red), and counterstained with DAPI (blue). No positive response for TuORF was detected at any time post infection while NS3 protein was already detectable 12 h post infection.

FIGURES:

Figure 1:

Figure 2:



SUPPLEMENTARY TABLES:

Supplementary table 1 - A list of TBEV strains with uORF in their 5´UTR:

TuORF is present in the 5´UTR of many (but not all) TBEV strains representing all TBEV subtypes (European, Siberian, and Far East). The same TuORF sequence is often found in multiple TBEV strains (such strains are grouped together in one row). In such cases only one strain (written in bold) was randomly selected as a representative and used in subsequent analysis. TuORF is absent from most TBEV strains. GenBank accession numbers are shown in brackets.

with TuORF	European TBEv strains	A104 (KF151173.1), Ljubljana I (JQ654701.1), Kumlinge A52 (GU183380.1), **Neudoerfl (U27495.1)**
		temperature-resistant variant of strain 263 (DQ153877.1), 263 (U27491.1)
		KrM 93 (HM535611.1), KrM 213 (HM535610.1)
		AS33 (GQ266392.1)
		Toro-2003 (DQ401140.2)
		EK-328 (DQ486861.1)
	Siberian TBEV strains	**L22 (EU715149.1)**
		K2 (EU715157.1), L23 (EU715151.1), K1 (EU715156.1), L32n (EU715154.1), L27 (EU715153.1), **L22-3 (EU715150.1)**, i36-6 (EU715146.1)
		i6-6 (EU715162.1), K42 (EU715148.1), i113n (EU715147.1), i32L (EU715145.1), i32-3 (EU715144.1), c34-16 (EU715142.1), c30-1 (EU715141.1), c22 (EU715140.1), c34-20 (EU715143.1) **Latvia-1-96 (GU183382.1)**
		Yar 71 (EU444077.1), Yar 114 (EU444078.1) **O-13 (EU715155.1)**
		L23-3 (EU715152.1)
		c19-5 (EU715139.1)
	Far Eastern TBEV strains	**Oshima 5-10 (AB062063.2)**,Oshima 08-As (AB753012.1)
		MDJ01 (JQ650522.1) **Prot1 (EU715174.1)**
		TE05 (EU715168.1)
without TuORF	European TBEV strains	Hypr (U39292.1), Hypr (M76660.1)
	Siberian TBEV strains	Zausaev (AF527415.1), Tms Bird-08-75 (KC602128.1), Tms Bird-10-87 (KC602125.1), Cht-653 (JN003207.1), Kolarovo-2008 (FJ968751.1), Tms Bird-08-29 (KC602127.1), Zabaikalye 11-99 (KC414090.1), Komi-10-04 (JX628793.1), Vasilchenko (AF069066.1), Tms 10-18 (KC663433.1), Tms Bird-10-54 (KC602126.1), Komi-10-01 (JX628792.1), TBE6 (EU715163.1), 11-7TBE (EU715158.1)

Supplementary table 2 – A list of other Flavivirus species in which uORFs were identified

A uORF was identified in the 5´UTR of three other tick-borne flaviviruses (LGTV, KAMV, KARV) and one mosquito-borne flavivirus (SLEV). All other flaviviruses lack any AUG in their 5´UTR. GenBank accession numbers are shown in brackets.

uORF	Flavivirus group	Flavivirus species
uORF identified	Tick-borne flaviviruses	Langat virus (AF253419.1, AF253420.1, EU790644.1), Karshi virus (DQ462443.1), Kama virus (NC_023439.1, KF815940.1)
	Mosquito-borne flaviviruses	St. Louis encephalitis virus (DQ525916.1)
No uORF identified	Tick-borne flaviviruses	Louping ill virus (Y07863.1, KJ495985.1, KJ495984.1, KJ495983.1, KF056331.1), Kyasanur forest disease virus (JF416958.1, HM055369.1, X74111.1, JF416960.1, JF416959.1), Alkhurma virus (AF331718.1, JF416957.1, JF416957.1, AF331718.1, JF416962.1, JF416961.1, JF416956.1, JF416955.1, JF416954.1, JF416953.1, JF416952.1, JF416951.1, JF416950.1, JF416949.1, JF416967.1, JF416966.1, JF416963.1, JX271893.1, JX271892.1, JF416964.1), Deer tick Virus (AF311056.1, AF357218.1), Powassan virus (KJ746872.1, HM440559.1, HM440561.1, HM440558.1, HM440560.1, HM440562.1, HM440563.1, EU670438.1, L06436.1, EU770575.1, HQ231414.1, HQ231415.1), Tyuleniy virus (NC_023424.1, F815939.1)
	Mosquito-borne flaviviruses	Dengue virus 2 (NC_001474.2), Yellow fever virus (NC_002031.1), Japanese encephalitis virus (NC_001437.1), West Nile virus (NC_001563.2), Murray Valley encephalitis virus (NC_000943.1)
5'UTR not or not fully sequenced	Tick-borne flaviviruses	Gadgets Gully virus, Royal Farm virus (DQ235149.1), Kadam virus (DQ235146.1), Meaban virus (DQ235144.1), Saumarez reef virus (DQ235150.1)

Supplementary table 3 – Pairwise codon based analyses of selection pressure affecting TuORF evolution

The probability of rejecting the null hypothesis of strict-neutrality (dN = dS) in favor of the alternative hypothesis (a) Negative evolution: dN > dS, b) any evolutionary pressure: dN≠dS, or positive selection dN<dS) (below diagonal) is shown. Values of P less than 0.05 are considered significant at the 5% level and are shown in bold type. The test statistic (dN - dS) is shown above the diagonal. Analyses were conducted as described in Methods.

a) Negative (purifying) selection):

	Neudoerfl	KrM_93	AS33	tr263	Toro-2003	EK-328	Oshima_5-10	MDJ01	Prot1	TE05	L23-3	i36-6	c19-5	O-13	L22	L22-3	Latvia-1-96
Neudoerfl		0,6797	0,6797	0,0000	-1,4039	-0,9924	-1,2123	-1,2509	-2,1530	-1,0074	-2,3785	-2,3785	-2,5536	-2,5522	-2,1837	-2,3775	-0,1840
KrM_93	0,2490		1,3718	0,6797	0,1321	-0,9924	-1,1549	-0,5932	-2,1130	-0,9430	-0,9708	-0,9708	-0,9262	-2,7106	-0,7727	-1,0361	0,6241
AS33	0,2490	0,0863		0,6797	0,1321	-0,9924	-0,4707	-1,2509	-1,2685	-0,3628	-0,9708	-0,9708	-0,9262	-1,2213	-0,7727	-1,0361	0,6241
tr263	1,0000	0,2490	0,2490		-1,4039	-0,9924	-1,2123	-1,2509	-2,1530	-1,0074	-2,3785	-2,3785	-2,5536	-2,5522	-2,1837	-2,3775	-0,1840
Toro-2003	1,0000	0,4476	0,4476	1,0000		-1,3916	-0,7050	-0,8371	-1,9261	-0,5566	-1,2372	-1,2372	-1,4525	-1,5046	-0,9962	-1,2873	0,2748
EK-328	1,0000	1,0000	1,0000	1,0000	1,0000		0,0907	-0,4209	-0,5800	0,0907	-1,6883	-1,6883	-1,6883	-1,6883	-1,3916	-1,3916	-1,3916
Oshima_5	1,0000	1,0000	1,0000	1,0000	1,0000	0,4639		-0,2555	0,3231	0,9404	-1,6658	-1,6658	-1,4149	-1,4831	-1,5051	-1,7315	-0,1448
MDJ01	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		-0,9513	-0,8825	-1,0384	-1,0384	-1,0384	-1,0399	-0,9221	-0,9562	0,3508
Prot1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,3736	1,0000		0,4250	-2,9781	-2,9781	-2,8511	-2,8486	-2,8478	-2,9749	-0,4681
TE05	1,0000	1,0000	1,0000	1,0000	1,0000	0,4639	0,1745	1,0000	0,3358		-1,2876	-1,2876	-1,0351	-1,1201	-1,1479	-1,3719	-0,2563
L23-3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		0,0000	-0,9960	-0,9959	-0,9959	-1,4022	0,6571
i36-6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		-0,9960	-0,9959	-0,9959	-1,4022	0,6571
c19-5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		-1,4024	-1,4023	-1,7093	0,6804
O-13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		-1,4022	-1,7090	0,4092
L22	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		-0,9959	0,8769
L22-3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000		0,6215
Latvia-1-96	1,0000	0,2669	0,2669	1,0000	0,3920	1,0000	1,0000	0,3632	1,0000	1,0000	0,2562	0,2562	0,2488	0,3416	0,1911	0,2677	

b) Any selection pressure

	Neudoerfl	KrM_93	AS33	tr263	Toro-2003	EK-328	Oshima_5-10	MDJ01	Prot1	TE05	L23-3	i36-6	c19-5	O-13	L22	L22-3	Latvia-1-96
Neudoerfl		-0,6797	-0,6797	0,0000	1,4039	0,9924	1,2123	1,2509	**2,1530**	1,0074	**2,3785**	**2,3785**	**2,5536**	**2,5522**	**2,1837**	**2,3775**	0,1840
KrM_93	0,4980		-1,3718	-0,6797	-0,1321	0,9924	1,1549	0,5932	**2,1130**	0,9430	0,9708	0,9708	0,9262	**2,7106**	0,7727	1,0361	-0,6241
AS33	0,4980	0,1727		-0,6797	-0,1321	0,9924	0,4707	1,2509	1,2685	0,3628	0,9708	0,9708	0,9262	1,2213	0,7727	1,0361	-0,6241
tr263	1,0000	0,4980	0,4980		1,4039	0,9924	1,2123	1,2509	**2,1530**	1,0074	**2,3785**	**2,3785**	**2,5536**	**2,5522**	**2,1837**	**2,3775**	0,1840
Toro-2003	0,1629	0,8951	0,8951	0,1629		1,3916	0,7050	0,8371	1,9261	0,5566	1,2372	1,2372	1,4525	1,5046	0,9962	1,2873	-0,2748
EK-328	0,3230	0,3230	0,3230	0,3230	0,1666		-0,0907	0,4209	0,5800	-0,0907	1,6883	1,6883	1,6883	1,6883	1,3916	1,3916	1,3916
Oshima_5	0,2278	0,2504	0,6387	0,2278	0,4822	0,9278		0,2555	-0,3231	-0,9404	1,6658	1,6658	1,4149	1,4831	1,5051	1,7315	0,1448
MDJ01	0,2134	0,5541	0,2134	0,2134	0,4042	0,6746	0,7988		0,9513	0,8825	1,0384	1,0384	1,0384	1,0399	0,9221	0,9562	-0,3508
Prot1	**0,0333**	**0,0367**	0,2071	**0,0333**	0,0565	0,5630	0,7472	0,3434		-0,4250	**2,9781**	**2,9781**	**2,8511**	**2,8486**	**2,8478**	**2,9749**	0,4681
TE05	0,3158	0,3476	0,7174	0,3158	0,5788	0,9278	0,3489	0,3793	0,6716		1,2876	1,2876	1,0351	1,1201	1,1479	1,3719	0,2563
L23-3	**0,0190**	0,3336	0,3336	**0,0190**	0,2184	0,0939	0,0984	0,3012	**0,0035**	0,2004		0,0000	0,9960	0,9959	0,9959	1,4022	-0,6571
i36-6	**0,0190**	0,3336	0,3336	**0,0190**	0,2184	0,0939	0,0984	0,3012	**0,0035**	0,2004	1,0000		0,9960	0,9959	0,9959	1,4022	-0,6571
c19-5	**0,0119**	0,3562	0,3562	**0,0119**	0,1490	0,0939	0,1597	0,3012	**0,0051**	0,3027	0,3213	0,3213		1,4024	1,4023	1,7093	-0,6804
O-13	**0,0120**	**0,0077**	0,2244	**0,0120**	0,1350	0,0939	0,1407	0,3005	**0,0052**	0,2649	0,3213	0,3213	0,1634		1,4022	1,7090	-0,4092
L22	**0,0309**	0,4412	0,4412	**0,0309**	0,3212	0,1666	0,1349	0,3583	**0,0052**	0,2533	0,3213	0,3213	0,1634	0,1634		0,9959	-0,8769
L22-3	**0,0190**	0,3023	0,3023	**0,0190**	0,2005	0,1666	0,0859	0,3409	**0,0035**	0,1726	0,1634	0,1634	0,0900	0,0900	0,3213		-0,6215
Latvia-1-96	0,8543	0,5338	0,5338	0,8543	0,7839	0,1666	0,8851	0,7263	0,6406	0,7981	0,5124	0,5124	0,4976	0,6831	0,3823	0,5354	

c) Positive (directional) selection

	Neudoerfl	KrM_93	AS33	tr263	Toro-2003	EK-328	Oshima_5-10	MDJ01	Prot1	TE05	L23-3	i36-6	c19-5	O-13	L22	L22-3	Latvia-1-96
Neudoerfl		-0,6797	-0,6797	0,0000	**1,4039**	0,9924	1,2123	1,2509	**2,1530**	1,0074	**2,3785**	**2,3785**	**2,5536**	**2,5522**	**2,1837**	**2,3775**	0,1840
KrM_93	1,0000		-1,3718	-0,6797	-0,1321	0,9924	1,1549	0,5932	**2,1130**	0,9430	0,9708	0,9708	0,9262	2,7106	0,7727	1,0361	-0,6241
AS33	1,0000	1,0000		-0,6797	-0,1321	0,9924	0,4707	1,2509	1,2685	0,3628	0,9708	0,9708	0,9262	1,2213	0,7727	1,0361	-0,6241
tr263	1,0000	1,0000	1,0000		1,4039	0,9924	1,2123	1,2509	**2,1530**	1,0074	**2,3785**	**2,3785**	**2,5536**	**2,5522**	**2,1837**	**2,3775**	0,1840
Toro-2003	**0,0815**	1,0000	1,0000	0,0815		1,3916	0,7050	0,8371	**1,9261**	0,5566	1,2372	1,2372	1,4525	1,5046	0,9962	1,2873	-0,2748
EK-328	1,1615	1,1615	1,1615	1,1615	0,0833		-0,0907	0,4209	0,5800	-0,0907	**1,6883**	**1,6883**	**1,6883**	**1,6883**	1,3916	1,3916	1,3916
Oshima_5	0,1139	0,1252	0,3193	0,1139	0,2411	1,0000		0,2555	-0,3231	-0,9404	**1,6658**	**1,6658**	1,4149	1,4831	1,5051	**1,7315**	0,1448
MDJ01	0,1067	0,2771	0,1067	0,1067	0,2021	0,3373	0,3994		0,9513	0,8825	1,0384	1,0384	1,0384	1,0399	0,9221	0,9562	-0,3508
Prot1	**0,0167**	**0,0183**	0,1035	**0,0167**	**0,0282**	0,2815	1,0000	0,1717		-0,4250	**2,9781**	**2,9781**	**2,8511**	**2,8486**	**2,8478**	**2,9749**	0,4681
TE05	0,1579	0,1738	0,3587	0,1579	0,2894	1,0000	1,0000	0,1896	1,0000		1,2876	1,2876	1,0351	1,1201	1,1479	1,3719	0,2563
L23-3	**0,0095**	0,1668	0,1668	**0,0095**	0,1092	**0,0470**	**0,0492**	0,1506	**0,0018**	0,1002		0,0000	0,9960	0,9959	0,9959	1,4022	-0,6571
i36-6	**0,0095**	0,1668	0,1668	**0,0095**	0,1092	**0,0470**	**0,0492**	0,1506	**0,0018**	0,1002	1,0000		0,9960	0,9959	0,9959	1,4022	-0,6571
c19-5	**0,0060**	0,1781	0,1781	**0,0060**	0,0745	**0,0470**	0,0798	0,1506	**0,0026**	0,1514	0,1606	0,1606		1,4024	1,4023	**1,7093**	-0,6804
O-13	**0,0060**	**0,0039**	0,1122	**0,0060**	0,0675	**0,0470**	0,0703	0,1502	**0,0026**	0,1325	0,1606	0,1606	0,0817		1,4022	**1,7090**	-0,4092
L22	**0,0155**	0,2206	0,2206	**0,0155**	0,1606	0,0833	0,0675	0,1792	**0,0026**	0,1267	0,1606	0,1606	0,0817	0,0817		0,9959	-0,8769
L22-3	**0,0095**	0,1511	0,1511	**0,0095**	0,1002	0,0833	**0,0430**	0,1705	**0,0018**	0,0863	0,0817	0,0817	**0,0450**	**0,0450**	0,1607		-0,6215
Latvia-1-96	0,4272	1,0000	1,0000	0,4272	1,0000	0,0833	0,4425	1,0000	0,3203	0,3991	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	

Supplementary table 4 - Predicted biochemical features of the TBEV Neudoerfl TuORF peptide

Characteristic	Value	Program
Number of amino acids	23	ProtParam
Molecular weight [Da]	2678.2	
Theoretical pI	9.3	
Number of negatively charged residues (D + E)	1	
Number of positively charged residues (R + K)	3	
Grand average of hydropathicity (GRAVY)	0.826	
Secondary structure	Helical	Jpred
Orientation in membrane	N outside (561)	TMpred

SUPPLEMENTARY FIGURE LEGENDS:

Supplementary Figure 1 – Sequence analysis of uORFs detected in other Flavivirus species:

Full length sequence of 5´UTRs of Flavivirus species with detected uORFs (A). uORF sequences are marked in color, while remaining part of the 5´UTR is in grey. uORF start and stop codons as well as major ORF start codons are underlined. Sequence of putative peptides encoded by detected uORFs (B). Alignment of putative peptides encoded by detected Flavivirus uORFs (C). GenBank accession numbers of all nucleotide sequences used in this study are listed in Supplementary Table 2. Protein sequences of hypothetical TuORF peptides were deduced from nucleotide sequences as indicated in Methods.

Supplementary Figure 2 - Phylogenetic analysis of TuORF relationships:

Phylogenetic analysis based on nucleotide (A) and protein (B) sequences of TuORF showed existence of three clearly separated phylogenetic clades. Only bootstrap values on which the tree separation into three clades is based are shown. The first clade unites European subtype TBEV strains (encircled in red). The second clade includes Siberian subtype TBEV strains (encircled in blue). The third clade comprises Far Eastern subtype TBEV strains (encircled in green).

Supplementary Figure 3 – Detection of the TuORF peptide by immunoblotting:

Immunoblotting analysis was done on human neuroblastoma, glioblastoma, and medulloblastoma cell lines and on the tick cell line IRE/CTVM19 infected with TBEV strain Neudoerfl as described in Methods. No positive signal was detected for TuORF peptide in the cell lysates, while the positive control (artificial TuORF – marked by asterisk) always gave a very strong response.

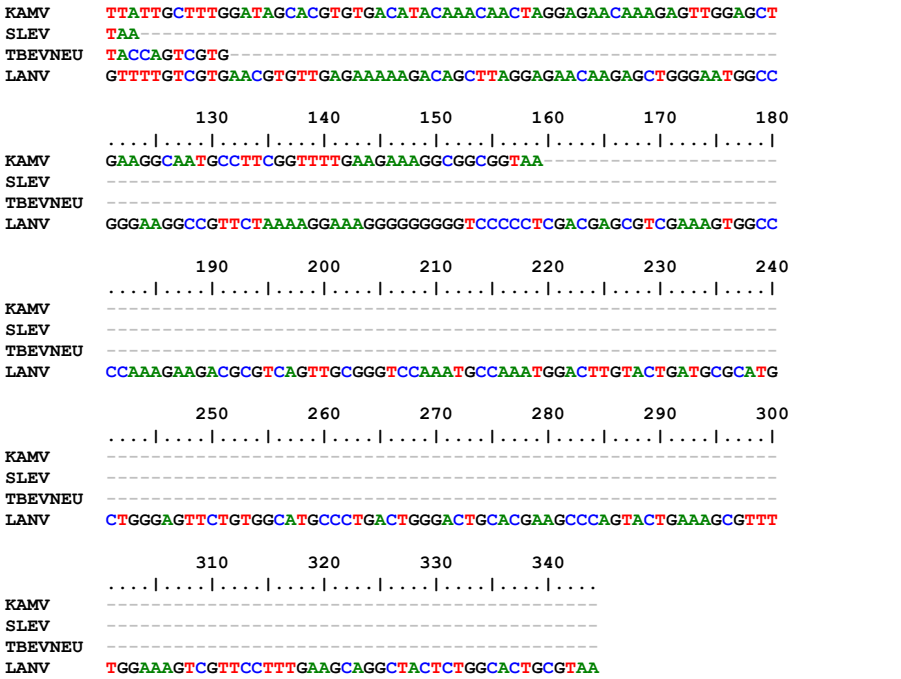
Supplementary Figure 4 – Detection limit of the synthetic TuORF peptide by immunoblotting:

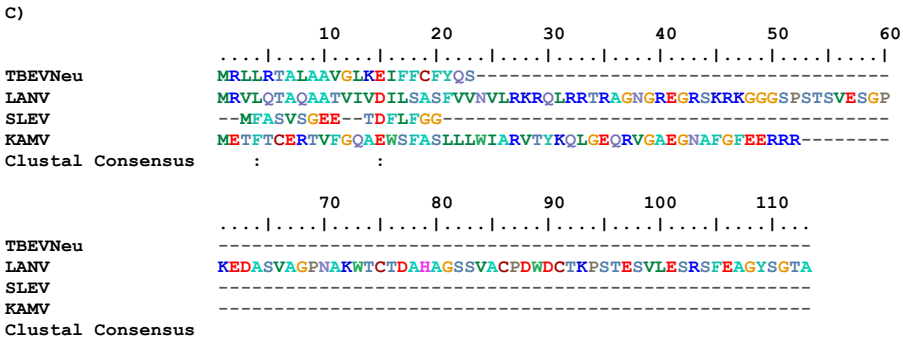
To estimate the detection limit of the TuORF peptide by immunoblotting, we tested different concentrations of the synthetic TuORF in ten-fold dilutions from 10µg to 0.1ng. The lowest detectable amount of the synthetic TuORF was 100ng.

SUPPLEMENTARY FIGURE:

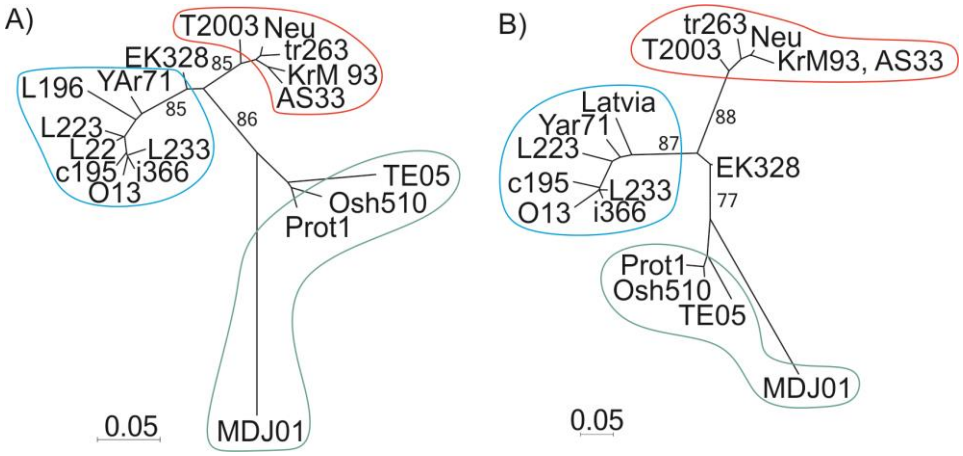
Supplementary Figure 1 – Sequence analysis of uORFs detected in other Flavivirus species:

```
A)
>TBEV Neu
AGATTTTCTT GCACGTGCAT GCGTTTGCTT CGGACAGCAT TAGCAGCGGT TGGTTTGAAA GAGATATTCT
TTTGTTTCTA CCAGTCGTGA ACGTGTTGAG AAAAAGACAG CTTAGGAGAA CAAGAGCTGG GGATG
>LANV (AF253419.1)
AGATTTTCTT GCGCGTGCAT GCGTGTGCTT CAGACAGCCC AGGCAGCGAC TGTGATTGTG GATATTCTTT
CTGCAAGTTT TGTCGTGAAC GTGTTGAGAA AAAGACAGCT TAGGAGAACA AGAGCTGGGA ATGGCCGGGA
AGGCCGTTCT AAAAGGAAAG GGGGGGGGTC CCCCTCGACG AGCGTCGAAA GTGGCCCCAA AGAAGACGCG
TCAGTTGCGG GTCCAAATGC CAAATGGACT TGTACTGATG CGCATGCTGG GAGTTCTGTG GCATGCCCTG
ACTGGGACTG CACGAAGCCC AGTACTGAAA GCGTTTTGGA AAGTCGTTCC TTTGAAGCAG GCTACTCTGG
CACTGCGTAA
>KAMV (KF815940.1)
CTCTTCCCCC CTCCTTCTTG AGTATATGTT CACGTGTGAA CGCACTGTCT TTGGTCAGGC AGAGTGGTCT
TTTGCGTCGT TATTGCTTTG GATAGCACGT GTGACATACA AACAACTAGG AGAACAAAGA GTTGGAGCTG
AAGGCAATGC CTTCGGTTTT GAAGAAAGGC GGCGGTAA
>KARV (DQ462443.1)
AGATTTTCTT GCATGTGAGT GAGTTGACTT TAGTCAGTCC GCTCAGCAAG AGTGCTTTGA TATTGTTTTT
GGAGCAAGTT TGTTAACGTG TTGAGAAAAA GACAGCTTAG GAGAACAAGA GCTGGGGATG
>SLEV
AGATGTTCGC GTCGGTGAGC GGAGAGGAAA CAGATTTCCT TTTTGGAGGA TAATAACTTA ACTTGACTGC
GAACAGTTTT TTAGCAGGGA ATTACCCAAT G

B)
                10        20        30        40        50        60
        ....|....|....|....|....|....|....|....|....|....|....|....|
KAMV    --ATGTTCACGT--GTGAACG--CACTGTCTTTGGTCAGGCAGAGTGGTCTTTTGCGTCG
SLEV    --ATGTTCGCGTCGGTGAGCG--GA--------GAGGAAACAGATTTCCTTTTTGGAGGA
TBEVNEU ATGCGTTTGCTTCGGACAGCATTAGCAGCGGTTGGTTTGAAAGAGATATTCTTTTGTTTC
LANV    ATGCGTGTGCTTCAGACAGCCCAGGCAGCGACTG--TGATTGTGGATATTCTTTCTGCAA

                70        80        90        100       110       120
        ....|....|....|....|....|....|....|....|....|....|....|....|
```

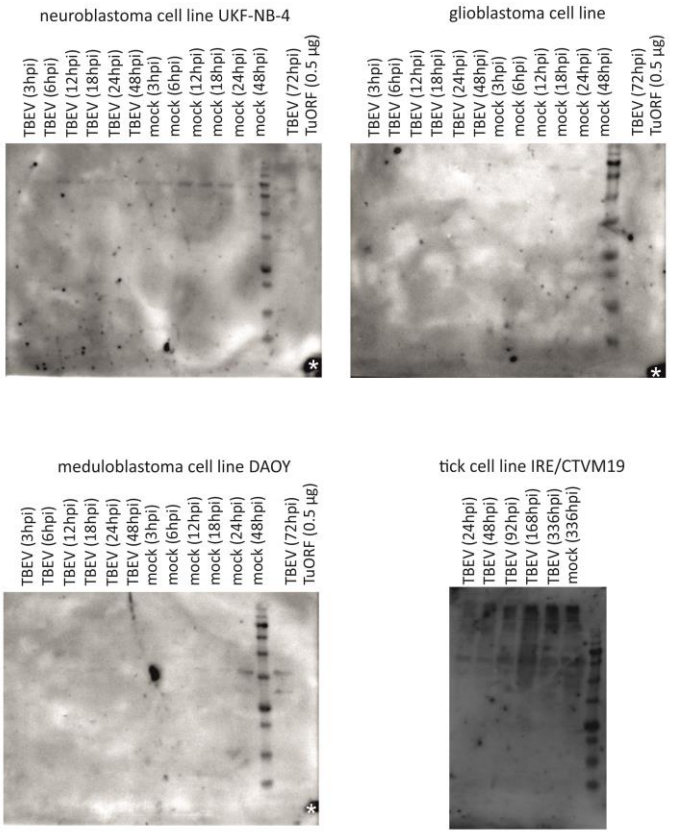




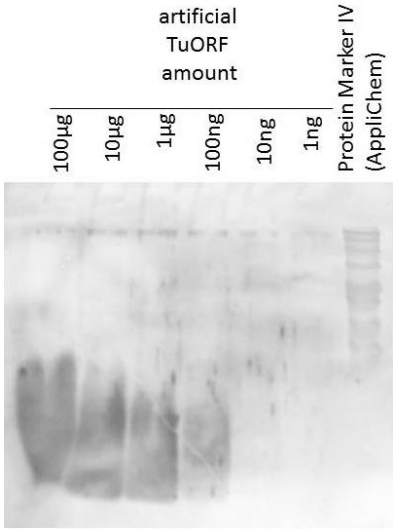
Supplementary figure 2:



Supplementary figure 3:



Supplementary figure 4:



6.5 Genomes of viruses classified in genus Flavivirus (family *Flaviviridae*) evolved via multiple recombination events

The manuscript is under revision process in BMC Evolutionary Biology

Genomes of viruses classified in genus Flavivirus (family *Flaviviridae*) evolved via multiple recombination events

Jiří Černý (1, 2, 3, #), Barbora Černá Bolfíková (4), Libor Grubhoffer (1, 2), and Daniel Růžek (1, 3)

1) Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 31, CZ-37005 České Budějovice, Czech Republic

2) Faculty of Science, University of South Bohemia in České Budějovice, Branišovská 31, CZ-37005 České Budějovice, Czech Republic

3) Veterinary Research Institute, Hudcova 296/70, CZ-62100 Brno, Czech Republic

4) Faculty of Tropical AgriSciences, Czech University of Life Sciences Prague, Kamýcká 129
CZ-16521 Praha 6 – Suchdol, Czech Republic

#) corresponding author: Jiří Černý, Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, Branišovská 31, CZ-37005 České Budějovice, Czech Republic, e-mail: cerny@paru.cas.cz, tel.: +420 387 775 451

ABSTRACT:

Background: The structure of a virus genome determines many of the virus characteristics. However, the evolutionary mechanisms behind viral genome evolution are not well understood. Here we focused on the genome evolution of viruses classified in the genus Flavivirus (family *Flaviviridae*).

Results: We performed an intensive sequence- and structure-based search to find distant viral and cellular homologues of Flavivirus proteins. Then, we aligned these sequences using advanced alignment algorithms, incorporating structural information whenever available. Finally, we reconstructed the evolution of selected proteins using Bayesian algorithms. Our analyses showed

that Flavivirus genomes are the outcome of a process of mosaic evolution, as most proteins or even protein domains evolved independently. Proteins C, M, NS1, NS2A, NS2B, NS4A, and NS4B do not have detectable homologues. NS3 is the only Flavivirus protein which shares a common evolutionary history across the whole *Flaviviridae* family. In contrast, Flavivirus protein E and the methyltransferase domain of NS5 do not have any homologues in other Flaviviridae genera; rather they have close cellular homologues. Therefore we think they were "kidnapped" by flaviviruses in early phases of their evolution. Finally, *Flaviviridae* polymerases (including the Flavivirus NS5 polymerase domain) do not form a monophyletic group in our analysis. Instead, Flavivirus polymerases are phylogenetically separated from other polymerases of *Flaviviridae* family by the polymerases of Turnip yellow mosaic virus, Hepatitis E virus and Chikungunya virus.

Conclusions: Flavivirus evolution should not be understood as a linear process but rather as a network, in which present day viruses are tangles of genes that each have their own individual evolutionary history.

KEY WORDS

Flavivirus, genome, gene, evolution, recombination,

BACKGROUND:

Genome structure is a key factor that determines the whole virus life cycle. Despite their importance, the evolutionary mechanisms behind the evolution of viral genomes are not well understood. In this study we focus on the intriguing question: Which mechanisms stand behind the genome evolution of viruses classified within the genus Flavivirus (family *Flaviviridae*)?

The genus Flavivirus includes important human pathogens. Typical examples are the four serotypes of Dengue virus (DENV1-4), Yellow fever virus (YFV), West Nile virus (WNV), Japanese encephalitis virus (JEV), Tick-borne encephalitis virus (TBEV) [1, 2]. Effective vaccination is available against some flaviviruses, but effective anti-flavivirus treatments are urgently needed [3]. Comparing Flavivirus proteins and their close relatives from host cells can help us to understand the evolutionary processes behind the evolution of Flavivirus genomes, but also to detect features common in Flavivirus proteins and absent

from host proteins. Therefore, it is a key step in the rational design of highly targeted anti-Flavivirus drugs.

The Flavivirus genome is formed by a single RNA molecule of positive polarity, which is approximately 11,000 nucleotides long [4]. The genomic RNA consists of a single open reading frame (ORF) flanked by 5' and 3' untranslated regions. The ORF is translated into a polyprotein, which is co- and posttranslationally cleaved into three structural (C, M, and E) and seven nonstructural (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) proteins. Proteins NS3 and NS5 each consist of two clearly distinguishable domains. The N-terminal domain of NS3 bears protease activity (NS3Pro), and the C-terminal domain is a helicase (NS3Hel). The N-terminal domain of NS5 catalyzes methylation of the mRNA cap (NS5Met) and the C-terminal domain is the viral polymerase (NS5Pol) [4].

According to current knowledge, seven Flavivirus proteins (M, C, NS1, NS2A, NS2B, NS4A, and NS4B) have no homologues outside of the genus Flavivirus. The remaining proteins are part of large protein families: protein E is a member of the Class II fusion proteins [5], NS3Pro belongs to the PA proteases [6], NS3Hel is a SF2 helicase [7], NS5Met is classified as a Ftsj-like methyltransferase [8], and NS5Pol is a viral RNA-dependent RNA polymerase [9].

Phylogenetic relationships of Flavivirus proteins to other viral and cellular proteins are unknown. This is due to the fast evolution of viral proteins, which rapidly leads to extreme sequence divergence, preventing the use of classical, distance-based phylogenetic methods [10]. Fortunately, it has become possible to detect homologies and to reconstruct the evolutionary relationships of very divergent proteins thanks to progress in development of very sensitive homology search algorithms [11-13], alignment algorithms that use structural information [14], together with the progress in phylogenetic methods [15, 16].

In this study we used modern powerful bioinformatics algorithms to reevaluate current knowledge about classification of Flavivirus proteins, which was established in early 1990s. To do that, we performed an extensive search for distant homologues of Flavivirus proteins. Detected homologues were used in deciphering of Flavivirus protein evolutionary history. The comparison of all *Flaviviridae* proteins evolution showed that evolutionary history of *Flaviviridae* proteins is very different, despite all viruses classified within *Flaviviridae* family

share a similar genome structure. Therefore, we conclude that Flavivirus genome evolved as a mosaic via several recombination events.

RESULTS:

Protein E

No sequence homologues of flavivirus protein E were found, but the Togaviral E1 protein and nematode EFF1 protein were detected as structural homologues (both classified in the family of Class II fusion proteins). Further, the lancelate BRAFL protein was detected as a sequence homologue of EFF1 (see Table S1 for the list of detected homologues of Flavivirus protein E and File S1A for the alignment of protein E from several members of genus Flavivirus and their detected homologues). As only a very limited number of homologues was detected, we did not perform BLASTCLUST clustering (see Methods). Phylogenetic analysis of homologues of Flavivirus protein E showed that Togavirus and Flavivirus envelope proteins form two monophyletic groups. These two groups are phylogenetically separated by EFF1 and BRAFL proteins (Fig. 1A for phylogenetic tree describing evolution of Flavivirus protein E and File S2A for a phylogenetic tree showing evolutionary relationship of Flavivirus protein E to all detected homologues).

Protease domain of protein NS3 (NS3Pro)

We did not identify any proteases or other proteins outside of the PA protease superfamily as sequence or structural homologues of Flavivirus NS3Pro. For further phylogenetic study, we selected 8 *Flaviviridae* representatives, 8 other viral proteases, and 13 cellular proteases of the PA protease superfamily (Table S1 and File S1B). Phylogenetic analysis of NS3Pro homologues showed that all *Flaviviridae* proteases form a monophyletic group within the PA protease superfamily (Fig. 1B and File S2B).

Helicase domain of protein NS3 (NS3Hel)

Only proteins classified within the helicase superfamilies SF1 and SF2 were identified as homologues of Flavivirus NS3Hel. Ten viral and 27 cellular helicases of superfamilies SF1 and SF2 were selected for the phylogenetic study (Table S1 and File S1C), together with 8 representatives of *Flaviviridae* NS3Hel. The phylogenetic analysis showed that *Flaviviridae* NS3Hels form a

monophyletic group clearly distinguishable from other SF1/SF2 helicases (Fig. 1C and File S2C).

Methyltransferase domain of protein NS5 (NS5Met)

Only proteins classified within the Ftsj-like superfamily were detected as structure or sequence homologues of the five Flavivirus NS5Mets. Ten viral and 21 cellular methyltransferases of the Ftsj-like superfamily were included in the phylogenetic study (Table S1 and File S1D). It showed that Flavivirus NS5Mets can be grouped with 23S 2'O rRNA methyltransferases from *Vibrio* genus of Gammaproteobacteria (Fig. 1D and File S2D).

Polymerase domain of protein NS5 (NS5pol)

All proteins identified under defined criteria as as homologues of Flavivirus polymerase were viral RNA-dependent RNA polymerases. Eight representatives of *Flaviviridae* polymerases and 12 other viral RNA-dependent RNA polymerases were included in the phylogenetic study (Table S1, File S1E). Surprisingly, the analysis results showed that *Flaviviridae* polymerases do not form a monophyletic group. Rather, Flavivirus NS5Pols are phylogenetically separated from Hepacivirus, Pestivirus, and Pegivirus proteins NS5A by the polymerases of Chikungunya virus (*Togaviridae*, Alphavirus), Hepatitis E virus (*Hepeviridae*, Hepevirus), and Turnip yellow mosaic virus (*Tymovirales, Tymoviridae*, Tymovirus) (Fig 1E and File S2E).

Flavivirus ORFans

No homologues of the Flavivirus proteins C, M, NS1, NS2A, NS2B, NS4A, and NS4B were found. Therefore, these proteins can be considered as Flavivirus ORFans - open reading frames with no detectable sequence similarity to any other ORF in the databases [17] (Fig. 2).

DISCUSSION:

Almost half of the Flavivirus polyprotein length occupies true ORFans

With our current knowledge, seven out of ten Flavivirus proteins, representing roughly 46% of the Flavivirus polyprotein length, can be considered as true Flavivirus ORFans, lacking any homologues even in other *Flaviviridae* genera (Fig. 2). Accordingly, there is no experimental evidence that these Flavivirus

ORFans can be functionally supplemented by proteins from other *Flaviviridae* genera, neither *in cis* nor *in trans* [18-20].In addition, the function of small *Flaviviridae* proteins differs across different genera. For example in the genus Flavivirus, the role of NS3Pro cofactor is fulfilled by NS2B [21, 22], whereas in the genus Hepacivirus, the same role is managed by NS4A [23]. Therefore we think that these proteins originated within the ancestor of the Flavivirus genus.

NS3 is the only one protein linearly evolving across whole *Flaviviridae* family

Both NS3 domains (the NS3Pro protease and the NS3Hel helicase) are members of large protein superfamilies, respectively the PA protease and the SF1/SF2 helicases [6, 7, 24]. No proteins out of these superfamilies were identified among Flavivirus NS3 homologues. We could not reconstruct the complete evolutionary history of these superfamilies, owing to the extreme sequence divergence of the PA proteases and SF1/SF2 helicases. However, we were able to reconstruct the evolution of several individual protein families. The protease and helicase domains of *Flaviviridae* NS3 each formed a monophyletic group. Also the protease and helicase of the Flavivirus genus form a monophyletic group within the group of Flaviviridae proteases and helicases. It makes NS3 protein the only one true flaviviral protein being linearly evolved across whole *Flaviviridae* family (Fig. 1).

Envelope protein and Flavivirus methyltransferase have close cellular homologues

Flavivirus E and NS5Met are cases totally opposed to that of NS3. These proteins have no homologues in other *Flaviviridae* genera, but have homologues in other viral and cellular proteins [8, 25, 26]. Here we showed that the closest homologues of Flavivirus E and NS5Met are cellular proteins, which shows that these proteins were kidnapped from cellular organisms (either from a flavivirus host or from host parasites/symbionts) and incorporated into the Flavivirus genome by recombination.

Flavivirus protein E is most closely related to a cellular Class II fusion protein from the lancelate *Branchiostoma floridae*. Recently analyses suggested that the Flavivirus protein E had originated directly from the Alphavirus E1 protein [25]. Nevertheless, our phylogenetic analysis showed that Flavivirus and Alphavirus envelope proteins do not form sister phylogenetic groups but are

phylogenetically separated by nematode and lancelate Class II fusion proteins. This finding challenges the currently widely accepted theory about a direct Alphavirus-Flavivirus envelope protein transfer.

No proteins out of the superfamily of Ftsj-like methyltransferases were identified among Flavivirus NS5Met homologues. The closest homologue of flavivirus NS5Met, which form a monophyletic group with Flavivirus NS5Met is the gammaproteobacterial rRNA methyltransferase from the *Vibrio* bacteria (Fig. 1). Bacterial rRNA methyltransferase was probably incorporated into the flavivirus genome during coinfection of a host by a pre-flavivirus and a bacteria. Bacteria-Flavivirus coinfections are quite common both in vectors and hosts of flaviviruses [27-30].

Flavivirus polymerase has closer relatives in other virus families than in *Flaviviridae*

Flavivirus NS5Pol belongs to the superfamily of right-hand polymerases that includes eukaryotic, archaeal, and viral replicases. Genes coding for RNA polymerases are present in all RNA viruses [9]. Therefore, polymerases are widely used as a RNA virus evolution marker gene [9, 31-34]. Previous phylogenetic studies showed that the Flavivirus NS5Pol forms a monophyletic group with the Hepacivirus and Pestivirus NS5A [9, 35]. In contrast, here we showed that *Flaviviridae* polymerases do not form a monophyletic group but that they are phylogenetically separated by polymerases of totally unrelated viruses (Fig. 1). Strong statistical support of our result indicates that it is not an experimental artefact. This discrepancy between our present results and previously published studies may be caused by incomplete sampling in previous studies, where only a subset of viral polymerases was chosen (i.e. polymerases with known tertiary structure). Further, more detailed phylogenetic studies are necessary to solve this discrepancy.

Flavivirus genomes are the result of a process of mosaic evolution

Our intensive database search using the most powerful modern algorithms did not reveal any novel unexpected homologues of Flavivirus proteins. On the other hand, detection of even very distant homologues allowed us for the first

time in history to reconstruct and compare the evolutionary relationships of all Flavivirus proteins.

Out of roughly 3400 amino acid residues in the Flavivirus polyprotein, only NS3, representing 13% of the total genome length (617 amino acids residues in DENV2) is linearly inherited across the whole *Flaviviridae* family. The remaining 87% of Flavivirus genome are either i) Flavivirus ORFans (C, M, NS1, NS2A, NS2B, NS4A, and NS4B – 46% of Flavivirus genome), ii) genes which have no homologues in other *Flaviviridae* genera but that have close cellular and viral homologues (E and NS5Met – 22% of the Flavivirus genome), or iii) genes that have homologues in other *Flaviviridae* genera but even closer homologues in other viruses (NS5Pol – 19% of Flavivirus genome).

Thus, the flavivirus genome is an extremely patchy structure, in which individual genes or even their domains have a very different evolutionary history (Fig. 2). This "patchiness" is most probably a result of multiple recombination events that occurred during the early history of the Flavivirus genome. This hypothesis is supported by two arguments: i) Even genomes of very distantly related members of the Flavivirus genus share the same evolutionary history; ii) No horizontal gene transfer between nowadays flaviviruses [36] or from cellular hosts to flaviviruses has been observed (even with an extremely low frequency) in nowadays flaviviruses.

Reading frame shifts may pose the major limitation for our study

Studies comparing the evolutionary history of individual flavivirus genes at the RNA level would complement our work. It is possible that differences in Flaviviridae proteins, manifesting as a totally different evolutionary history, result from insertion or deletion events that lead to reading frame shifts [37, 38]. At present, phylogenetic studies at the protein level cannot reveal such events. Nevertheless, nucleotide-based studies on the complete *Flaviviridae* family would be very complicated, owing to the low sequence similarity shared at the RNA level [39, 40].

For these reasons, our multiple recombination theory is currently the only statistically testable theory describing the formation of Flavivirus genomes. Moreover, the genome "patchiness" we observed is in concordance with previous works suggesting that multiple recombination may be the key force

behind formation of virus genomes [41]. If viral genomes are products of multiple recombination events, virus evolution cannot be understood as a linear process but rather as a network composed of the evolution of individual genes.

CONCLUSIONS:

Evolution of viral genomes is one of the most intriguing questions in modern virus evolutionary biology. In this study we focused on evolution of genes and genomes of viruses classified within genus Flavivirus, family *Flaviviridae*. We performed an extensive database search for sequence and structure homologues of individual Flavivirus proteins. Despite no unexpected proteins were detected we could use the resulting set of Flavivirus protein homologues to reconstruct their evolutionary history not only within the genus Flavivirus but also within the context of appropriate protein superfamilies for the first time in history. Resulting evolutionary trees showed that most Flavivirus proteins share very different evolutionary history. Proteins C, M, NS1, NS2A, NS2B, NS4A, and NS4B are true Flavivirus ORFans. NS3 is the only Flavivirus protein which shares a common evolutionary history across the whole *Flaviviridae* family. Protein E and the methyltransferase domain of NS5 have close cellular homologues but no homologues in other Flaviviridae genera which indicate that they were “kidnapped” by flaviviruses in early phases of their evolution. Finally, Flavivirus polymerases are phylogenetically separated from other *Flaviviridae* polymerases by the polymerases of viruses from different taxa. These results show that Flavivirus genome is very patchy structure being evolved by multiple recombination events.

METHODS:

Sample selection

Sequence homologues of individual proteins of DENV2 (GenBank Accession Number: NP_056776), WNV (YP_001527877), YEV (NP_041726), and TBEV (NP_043135) were searched using PSI-BLAST [12], HHpred [11], HHblits [13]. All search algorithms were run with default settings. The first 50 sequences with the highest E-value coming from nonflaviviral species were selected for further evaluation.

Whenever the 3D structure of a Flavivirus protein was available, it was used to search for structural homologues using DALI [42]. The search was run with default settings. If several structures of the same protein were available, the one with the highest resolution was used in the search. We selected for evaluation the sequences from the first 50 structures with the highest DALI Z-score coming from distinct species outside of the Flavivirus genus.

Sets of selected homologous protein sequences were clustered using BLASTCLUST [12] with an identity cut-off of 60%. Only one representative was chosen from each group for the phylogenetic analysis, since proteins in each group are closely related and their inclusion in the phylogenetic study would not bring additional information [43, 44]. The only exception were *Flaviviridae* proteins. Wheenver possible, we included five representatives of the genus Flavivirus (DENV, YFV, JEV/WNV/KUNV, MEAV, and TBEV), and one representative from each of the genera Hepacivirus, Pestivirus, and Pegivirus genus.

Protein multiple sequence alignment

Selected proteins were aligned using T-Coffee package aligning algorithms as Expresso and Psi-Coffee [45]. Structural information was used to improve the alignment whenever it was available. Amino acids aligned with low accuracy (alignment score lower than 10%) were trimmed out before the resulting alignments were used for the phylogenetic study.

Phylogenetic analyses

The best fitting models of amino acid substitutions were tested using PROTTEST 2.4 [46]. Phylogenetic analyses were performed using MrBayes v3.1.2 [16]. MrBayes was selected for analysis, since it is the best currently available method for the reconstruction of distant evolutionary relationships, and is less prone to long branch attraction when a proper model and appropriate taxon sampling are used [15, 47]. The analysis parameters are listed in Table S2. The final average standard deviation of the split frequencies of all analyses was always significantly below 0.01. The chain convergence was verified using AWTY [48].

AVAILABILITY OF SUPPORTING DATA:

The data sets supporting the results of this article are included within the article (and its additional files).

LIST OF ABBREVIATIONS USED

BVDV Bovine viral diarrhea virus

DENV Dengue virus

HCV Hepatitis C virus

HEV Hepatitis E virus

CHIKV Chikungunya virus

JEV Japanese encephalitis virus

KUNV Kunjin virus

MEAV Meaban virus

PEGVA Pegivirus A

SFV Semliki Forest virus

TBEV Tick-borne encephalitis virus

TYMV Turnip yellow mosaic virus

WNV West Nile virus

YFV Yellow fever virus

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

JC performed the database search, homologue alignment, interpreted the results, and wrote the manuscript. BCB calculated the evolutionary history of Flavivirus homologues. DR and LR helped with result interpretation, assisted with manuscript preparation and supervised whole process.

ACKNOWLEDGEMENTS:

We thank David Karlin from the University of Oxford for his constructive suggestions, critical reading of the manuscript, and English corrections.

This work was supported by the Czech Science Foundation [P502/11/2116 and 14-29256S to D. R. and 15-03044S to L. G.]; Grant Agency of University of South Bohemia [155/2013/P to L. G.]; Internal Grant Agency of the University of Life Sciences in Prague [CIGA 20134311 to B. C. B.]; the Ministry of Education, Youth and Sports of the Czech Republic [Z60220518 to D. R.]; ANTIGONE [278976 to L. G.]; and the Ministry of Education, Youth and Sports of the Czech Republic under the NPU I program [LO1218 to D. R.]. J.C. is a postdoctoral fellow supported by the project Postdok_BIOGLOBE (CZ. 1.07/2.3.00/30.0032) co-financed by the European Social Fund and state budget of the Czech Republic. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

LITERATURE:

1. Beck A, Guzman H, Li L, Ellis B, Tesh RB, Barrett AD: **Phylogeographic reconstruction of African yellow fever virus isolates indicates recent simultaneous dispersal into east and west Africa**. *PLoS Negl Trop Dis* 2013, **7**(3):e1910.
2. Messina JP, Brady OJ, Scott TW, Zou C, Pigott DM, Duda KA, Bhatt S, Katzelnick L, Howes RE, Battle KE *et al*: **Global spread of dengue virus types: mapping the 70 year history**. *Trends Microbiol* 2014, **22**(3):138-146.
3. Ishikawa T, Yamanaka A, Konishi E: **A review of successful flavivirus vaccines and the problems with those flaviviruses for which vaccines are not yet available**. *Vaccine* 2014, **32**(12):1326-1337.

4. Harris E, Holden KL, Edgil D, Polacek C, Clyde K: **Molecular biology of flaviviruses**. *Novartis Found Symp* 2006, **277**:23-39; discussion 40, 71-23, 251-253.
5. Stiasny K, Bressanelli S, Lepault J, Rey FA, Heinz FX: **Characterization of a membrane-associated trimeric low-pH-induced Form of the class II viral fusion protein E from tick-borne encephalitis virus and its crystallization.** *J Virol* 2004, **78**(6):3178-3183.
6. Aleshin AE, Shiryaev SA, Strongin AY, Liddington RC: **Structural evidence for regulation and specificity of flaviviral proteases and evolution of the Flaviviridae fold**. *Protein Sci* 2007, **16**(5):795-806.
7. Luo D, Xu T, Watson RP, Scherer-Becker D, Sampath A, Jahnke W, Yeong SS, Wang CH, Lim SP, Strongin A *et al*: **Insights into RNA unwinding and ATP hydrolysis by the flavivirus NS3 protein**. *Embo J* 2008, **27**(23):3209-3219.
8. Koonin EV: **Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and lambda 2 protein of reovirus**. *J Gen Virol* 1993, **74 (Pt 4)**:733-740.
9. Mönttinen HA, Ravantti JJ, Stuart DI, Poranen MM: **Automated structural comparisons clarify the phylogeny of the right-hand-shaped polymerases**. *Mol Biol Evol* 2014, **31**(10):2741-2752.
10. Zanotto PM, Gibbs MJ, Gould EA, Holmes EC: **A reevaluation of the higher taxonomy of viruses based on RNA polymerases**. *J Virol* 1996, **70**(9):6083-6096.
11. Söding J, Biegert A, Lupas AN: **The HHpred interactive server for protein homology detection and structure prediction**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W244-248.
12. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
13. Remmert M, Biegert A, Hauser A, Söding J: **HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment**. *Nat Methods* 2012, **9**(2):173-175.
14. Notredame C: **Recent evolutions of multiple sequence alignment algorithms**. *PLoS Comput Biol* 2007, **3**(8):e123.
15. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees**. *Bioinformatics* 2001, **17**(8):754-755.
16. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.
17. Yin Y, Fischer D: **Identification and investigation of ORFans in the viral world**. *BMC Genomics* 2008, **9**:24.

18. Khromykh AA, Varnavski AN, Westaway EG: **Encapsidation of the flavivirus kunjin replicon RNA by using a complementation system providing Kunjin virus structural proteins in trans**. *J Virol* 1998, **72**(7):5967-5977.
19. Khromykh AA, Sedlak PL, Westaway EG: **cis- and trans-acting elements in flavivirus RNA replication**. *J Virol* 2000, **74**(7):3253-3263.
20. Herod MR, Schregel V, Hinds C, Liu M, McLauchlan J, McCormick CJ: **Genetic complementation of hepatitis C virus nonstructural protein functions associated with replication exhibits requirements that differ from those for virion assembly**. *J Virol* 2014, **88**(5):2748-2762.
21. Chambers TJ, Weir RC, Grakoui A, McCourt DW, Bazan JF, Fletterick RJ, Rice CM: **Evidence that the N-terminal domain of nonstructural protein NS3 from yellow fever virus is a serine protease responsible for site-specific cleavages in the viral polyprotein**. *Proc Natl Acad Sci U S A* 1990, **87**(22):8898-8902.
22. Falgout B, Pethel M, Zhang YM, Lai CJ: **Both nonstructural proteins NS2B and NS3 are required for the proteolytic processing of dengue virus nonstructural proteins**. *J Virol* 1991, **65**(5):2467-2475.
23. Failla C, Tomei L, De Francesco R: **Both NS3 and NS4A are required for proteolytic processing of hepatitis C virus nonstructural proteins**. *J Virol* 1994, **68**(6):3753-3760.
24. Jankowsky A, Guenther UP, Jankowsky E: **The RNA helicase database**. *Nucleic Acids Res* 2011, **39**(Database issue):D338-341.
25. DuBois RM, Vaney MC, Tortorici MA, Kurdi RA, Barba-Spaeth G, Krey T, Rey FA: **Functional and evolutionary insight from the crystal structure of rubella virus protein E1**. *Nature* 2013, **493**(7433):552-556.
26. Pérez-Vargas J, Krey T, Valansi C, Avinoam O, Haouz A, Jamin M, Raveh-Barak H, Podbilewicz B, Rey FA: **Structural basis of eukaryotic cell-cell fusion**. *Cell* 2014, **157**(2):407-419.
27. Popov VL, Korenberg EI, Nefedova VV, Han VC, Wen JW, Kovalevskii YV, Gorelova NB, Walker DH: **Ultrastructural evidence of the ehrlichial developmental cycle in naturally infected Ixodes persulcatus ticks in the course of coinfection with Rickettsia, Borrelia, and a flavivirus**. *Vector Borne Zoonotic Dis* 2007, **7**(4):699-716.
28. Hunfeld KP, Allwinn R, Peters S, Kraiczy P, Brade V: **Serologic evidence for tick-borne pathogens other than Borrelia burgdorferi (TOBB) in Lyme borreliosis patients from midwestern Germany**. *Wien Klin Wochenschr* 1998, **110**(24):901-908.
29. Daniel M, Materna J, Honig V, Metelka L, Danielová V, Harcarik J, Kliegrová S, Grubhoffer L: **Vertical distribution of the tick Ixodes ricinus and tick-borne pathogens in the northern Moravian mountains**

correlated with climate warming (Jeseníky Mts., Czech Republic). *Cent Eur J Public Health* 2009, **17**(3):139-145.
30. Pugliese A, Beltramo T, Torre D: **Seroprevalence study of Tick-borne encephalitis, Borrelia burgdorferi, Dengue and Toscana virus in Turin Province**. *Cell Biochem Funct* 2007, **25**(2):185-188.
31. Villarreal LP, DeFilippis VR: **A hypothesis for DNA viruses as the origin of eukaryotic replication proteins**. *J Virol* 2000, **74**(15):7079-7084.
32. Filée J, Forterre P, Sen-Lin T, Laurent J: **Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins**. *J Mol Evol* 2002, **54**(6):763-773.
33. Cerný J, Cerná Bolfíková B, Valdés JJ, Grubhoffer L, Růžek D: **Evolution of tertiary structure of viral RNA dependent polymerases**. *PLoS One* 2014, **9**(5):e96070.
34. Koonin EV: **The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses**. *J Gen Virol* 1991, **72 (Pt 9)**:2197-2206.
35. Černý J, Černá Bolfíková B, Valdés JJ, Grubhoffer L, Růžek D: **Evolution of tertiary structure of viral RNA dependent polymerases**. *PLoS One* 2014, **9**(5):e96070.
36. Baillie GJ, Kolokotronis SO, Waltari E, Maffei JG, Kramer LD, Perkins SL: **Phylogenetic and evolutionary analyses of St. Louis encephalitis virus genomes**. *Mol Phylogenet Evol* 2008, **47**(2):717-728.
37. Light S, Basile W, Elofsson A: **Orphans and new gene origination, a structural and evolutionary perspective**. *Curr Opin Struct Biol* 2014, **26**:73-83.
38. Keese PK, Gibbs A: **Origins of genes: "big bang" or continuous creation?** *Proc Natl Acad Sci U S A* 1992, **89**(20):9489-9493.
39. Gritsun DJ, Jones IM, Gould EA, Gritsun TS: **Molecular archaeology of Flaviviridae untranslated regions: duplicated RNA structures in the replication enhancer of flaviviruses and pestiviruses emerged via convergent evolution**. *PLoS One* 2014, **9**(3):e92056.
40. Eddy SR: **Homology searches for structural RNAs: from proof of principle to practical use**. *RNA* 2015, **21**(4):605-607.
41. Koonin EV, Dolja VV: **Expanding networks of RNA virus evolution**. *BMC Biol* 2012, **10**:54.
42. Holm L, Rosenström P: **Dali server: conservation mapping in 3D**. *Nucleic Acids Res* 2010, **38**(Web Server issue):W545-549.
43. Elofsson A: **A study on protein sequence alignment quality**. *Proteins* 2002, **46**(3):330-339.
44. Illergård K, Ardell DH, Elofsson A: **Structure is three to ten times more conserved than sequence--a study of structural response in protein cores**. *Proteins* 2009, **77**(3):499-508.

45. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C: **Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W604-608.
46. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution**. *Bioinformatics* 2005, **21**(9):2104-2105.
47. Glenner H, Hansen AJ, Sørensen MV, Ronquist F, Huelsenbeck JP, Willerslev E: **Bayesian inference of the metazoan phylogeny; a combined molecular and morphological approach**. *Curr Biol* 2004, **14**(18):1644-1649.
48. **AWTY: A system for graphical exploration of MCMC convergence in Bayesian phylogenetic inference.** [http://ceb.csit.fsu.edu/awty]

FIGURES:

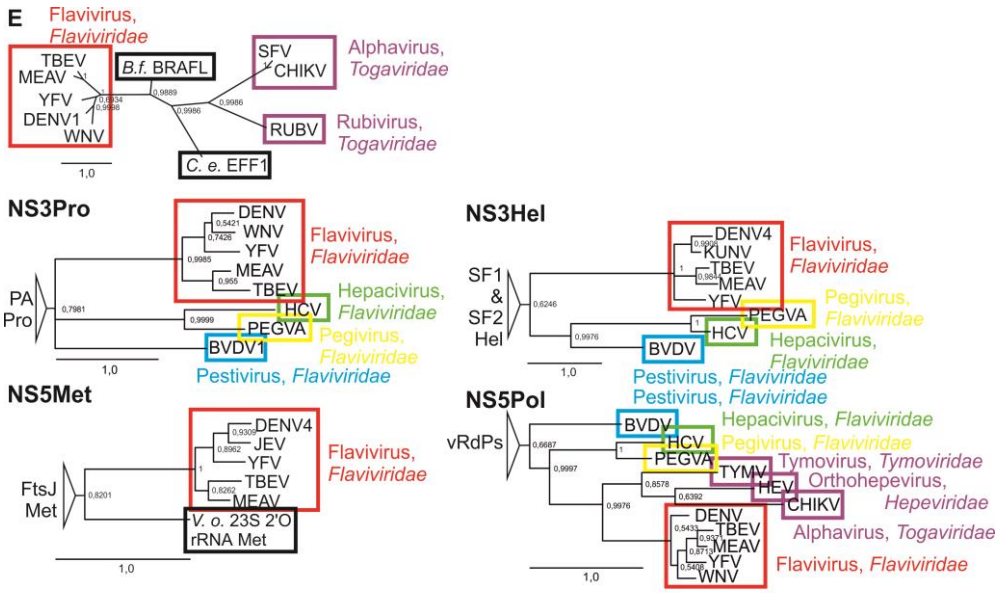


Figure 1 – Evolutionary history of Flavivirus proteins: Flavivirus proteins E, NS3Pro, NS3Hel, NS5Met, and NS5Pol are classified respectively into protein superfamilies PA, SF1/SF2, Ftsj-like, and viral RNA-dependent RNA polymerases. Both domains of *Flaviviridae* NS3 form monophyletic groups within the corresponding protein superfamilies. Protein E and the methyltransferase domain of protein NS5 do not have homologues in other Flaviviridae genera. Their closest homologues are lancelet EFF1 proteins and *Vibrio* 23S 2'O methyltransferases respectively. *Flaviviridae* polymerases do not form a monophyletic group but are phylogenetically separated by polymerases of unrelated viruses.

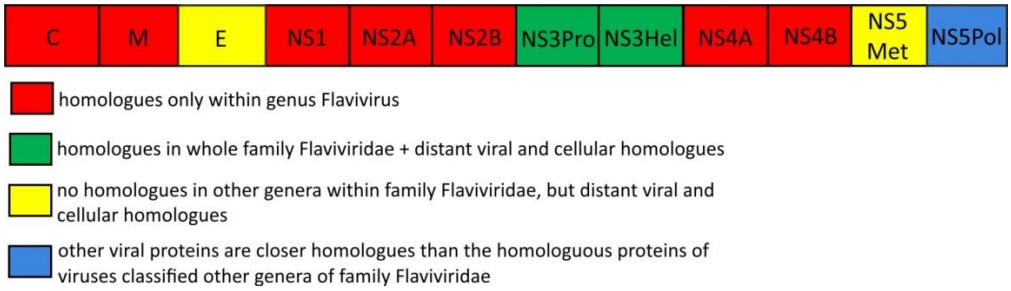


Figure 2 – Structure of the Flavivirus genome from an evolutionary point of view: The Flavivirus genome is very patchy structure from an evolutionary point of view. Proteins C, M, NS1, NS2A, NS2B, NS4A, and NS4B (in red) are Flavivirus ORFans. NS3 (in green) is the only Flavivirus protein that evolved linearly across the whole *Flaviviridae* family. Protein E and the methyltransferase domain of NS5 (NS5Met, in yellow) do not have homologues in other genera of *Flaviviridae* family, but they have close cellular homologues. The polymerase domain of Flavivirus NS5 (NS5Pol) is more closely related to the polymerases of several distant viruses than to the polymerase domain (NS5B) of other *Flaviviridae*. Size of individual proteins is not in scale.

SUPPLEMENTARY DATA:

Table S1 –Proteins used in phylogenetic analyses

Flavivirus protein (protein domain)	Flavivirus proteins	Homologous *Flaviviridae* proteins	Other homologues Flavivirus protein
E	TBEV (1SVB_A), MEAV (ABB90668.1), YFV (NP_041726.1), DENV1 (NP_056776.2), WNV (2HG0_A)	---	SFV E1 (1RER_A), CHIKV (3N43_F), *C.e.* EFF1 (4OJD_H), *B. e.* BRAFL (XP_002607817.1)
NS3Pro	WNV (2FOM_B), DENV (3U1I_B), YFV (NP_041726.1), MEAV (ABB90668.1), TBEV (NP_043135.1)	HCV (1A1R_A), PEGVA (NP_045010.1), BVDV1 (NP_040937.1)	SARS CoV (1UJ1_A), PolV (1L1N_A), CHIBAV (1WQS_A), HAV (2W5E_A), TEV (1LVM_A), PPV (P13529.2), SinV (2SNV_A), *B.t.* chymotrypsin (3T62_A), *R. r.* trypsin (1BRA_A), *S.g.* trypsin (1SGT_A), *H.s.* thrombin (1ABJ_L), *E. c.* AHP (1WXR_A), *H. e.* IgA1SP (3H09_A), *P.g.* DP7 (WP_005874121.1), *B. s.* SpoIVB (WP_015251736.1), *S. c.* Ssy5p (NP_012379.2), *E.c.* DegS (2QF3_A), *H.s.* HtrA2 (1LCY_A), *T.m.* HtrA (1L1J_A), *A.t.* Deg5 (4IC5_A), *S.a.* SplA (4MVN_A)
NS3Hel	DENV4 (2JLZ_A), KUNV (2QEQ_A), MEAV (ABB90668.1), TBEV (NP_043135.1), YFV (1YKS_A)0000	HCV (4B76_A), PEGVA (NP_045010.1), BVDV1 (NP_040937.1)	VV NPHII (YP_232959.1), TMV (NP_734217.1), *H.s.* BRR2 (4F92_B), *H.s.* RIGI (3TMI_A), *H.s.* BML (4CGZ_A), *H.s.* Ddx3x (2I4I_A), *H.s.* eIF4AIII (2XB2_X), *H.s.* Ddx10 (2PL3_A), *D.m.* VasA (2DB3_A), E.c. CsdA (1HV8_A), *S.c.* UPF1 (2XZ1_A), HHV1 UL5 (YP_009137079.1), SINV (NP_740671.1), *E.c.* RepA (1G8Y_A), T4 GP17 (3CPE_A), *E.c.* RecQ (1OYY_A), *K.p.* PriA (4NL4_A), TYMV (NP_733818.1), HEV (AAA03187.1), PVY (NP_734246.1), CHIKV (ADZ47896.1), *T.t.* HerA (4KBG_A), M.j. DEADBOX (1HV8_A), *S.t.* Hel (2Z0M_A), *H.s.* DDX6 (4CT5_A), A. m. PTHR18934 (XP_007259202.1), *Ch. a.* Icl (XP_006832768.1), *M. h.* HrpA (WP_032846000.1), *B. d.* HrpB (AP014685.1), *S. k.* Icl2 (XP_002734191.1), *S.c.* Bdp5 (3PEW_A), *H.s.* Ddx19B (3FHT_A), *B.m.* VasA (4D26_A), *S.c.* Mss116p (3I5Y_A), *A.p.* RigI (4A36_A)
NS5Met	DENV (2XBM_B), JEV (4K6M _A), YFV (3EVF_A), MEAV (ABB90668.1), TBEV (NP_043135.1)	---	MumpsV (AAT76834.1), NegevV (AFI24672.1), SARS CoV (2XYV_A), Reovirus (1EJ6_A), ASFV (P0C967.1), BaculoV (NP_054099.1), Mimivirus (YP_003987023.1), *H.s.* Ftsj (2NYU_A), E.c. Ftsj (1EJ0_A), VV CapE (4CKB_D), *H.s.* 2OmCap (XP_006715091.1), H.s. 2OtRNA (NP_036412.1), *S.c.* Spb1p (NP_009877.1), *S. c.* Trm7p (NP_009617.3), *S.c.* MRM2 (NP_011379.1), *A.t.* FtsJ (NP_196887.1), *S.te.* Hemolysin (3HP7_A), ChikV (3TRK_A), T.t. TrmN (3TMA_A), *P.h.* Met (2AS0_A), VEEV (2HWK_A), *P.fa.* Trm14 (3TM5_A), *E.c.* Fmu (1SQG_A), *M. a.* 23SrRNAm (WP_027329972.1), *T. v.* Gss1 (3DOU_A), *P. f.* rRNA Met (2PLW_A), *V.o.* 23rRNAm (GAJ70980.1) *H. s.* 2OmCAP (XP_005249012.1), P. p. CISIN (XP_001773849.1), *P. c.* 23SrRNAm (XP_742172.1), SARS CoV (AAS48581.1), *L. l.* Hemolysin (3OPN_A)
NS5Pol	DENV (4C11_A), WNV (2HCN_A), YFV (NP_041726.1), MEAV (ABB90668.1), TBEV (NP_043135.1)	HCV (1NB6_A), BVDV (1S49_A), PEGVA (NP_045010.1)	PolV (3OLB_A), NorV (3BSO_A), Qbeta (3AVX_A), IBDV (2PUS_A), Phi6 (1HI0_P), MOR3 (1N35_A), HIV (3V81_A), TYMV (), SARS CoV (ADC35510.1), HEV (AAA96139.1), AstroV (YP_003090286.1), PVY (NP_056759.1), ChikV (ADG95922.1)

Table S2 – MrBayes program parameters

parameter / protein (protein domain)	E	NS3Pro	NS3Hel	NS5Met	NS5Pol
amino acid residue substitution model	WAG+G	WAG+I+G	LG+I+G	Blossum62+G	ReRev+I+G+F
number of runs (cold/hot chains)	2 (1/3)	2 (1/3)	2 (1/3)	2 (1/3)	2 (1/3)
number of generations	20 000 000	13 000 000	20 000 000	13 000 000	13 000 000
burn in period	25%	25%	25%	25%	25%
sample frequency	500	500	500	500	500

File S1 – Alignments

A) Trimmed alignment of protein E homologues

```
                    10        20        30        40        50        60
                ....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          ----RSFPLEEKFDGLF-PPHCSKTVRAQTQNASIAGMQMQFSIGLHTAVCFRLY--ASQ
BfBRAFLDRAFT    MVVLRAVLLWASISGIHG--RCTKTVRYREAQALV-ETFIQSSIKPGETLCFVTNDDASS
TBEV            SRCTH-LENRDFVTGTQG------------------TTRVTLVLELGGCVTITAEGKPSM
DENV1           -RCVG-IGNRDFVEGLSG------------------ATWVDVVLEHGSCVTTMAKDKPTL
WNV             FNCLG--SNRDFLEGVSG------------------ATWVDLVLEGDSCVTIMSKDKPTI
YFV             AHCIG-ITDRDFIEGVHG------------------GTWVSATLEQDKCVTVMAPDKPSL
MeaV            SRCVH-LENRDFVTGTTG------------------SSRVSVVLEKHACVTIVAEGKPSL
RubellaV        ------------------------------------E-AFTYLCTAPGCATQT--------
ChikV           E--------GGGGSGGG-GY----------------EHVTVIPNTVGVPYKTLVN-----
SFV             --------------------Y----------------EHSTVMPNVVGFPYKAHIE-----

                    70        80        90       100       110       120
                ....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          EINDDENAGNQTSLLHTIRLEKLEHHHPITQRYTFGIPEVHAS-CICECD---STCTA--
BfBRAFLDRAFT    DVSISMATNGSVPLLWQLTYEGIEMD-YGVKNR-YSFYRPKYESKCVCDCPQYGDYCN--
TBEV            D----------------VWLDAIYQE-NPAKTREYCLHAKLSDTKVAARCPTMGPATL--
DENV1           D----------------IELLKTEVT-NPAVLRKLCIEAKISNTTTDSRCPTQGEATL--
WNV             D----------------VKMMNMEAA-NLAEVRSYCYLATVSDLSTKAACPTMGEAHN--
YFV             D----------------ISL-ETAID-RPAEVRKVCYNAVLTHVKINDKCPSTGEAHL--
MeaV            D----------------VWLDSIFQE-SPAPTREYCLDMGIFDQKVEARCPTMGEAHL--
RubellaV        --------------PVPVRLAGVRFESKIVDGGCFAPWDLEAT-----ICEIPTDVSC--
ChikV           ---------GYSPMVLEMELLSVTLEPTLSLDYITCEYKTVIPSPYV-KCC-------KN
SFV             ---------GYSPLTLQMQVVETSLEPTLNLEYITCEYKTVVPSPYV-KCCGASECSTKE

                   130       140       150       160       170       180
                ....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          D---S--TSSCYRTFFPNQTPIGCSEDDPKLCCDVRFKPY------KNMTFLAVKLEQPT
BfBRAFLDRAFT    -SKTNRCLDFCYNTYIPDQTAHGCTVY-WNESEVCCALYVG-------VKLALKVYDSRE
TBEV            -AEEHQGGTVCKRDQSDRGWGNHCGLF-GKGSIVACVKAAC---EA-KKKATGHVYDANK
DENV1           -VEEQDTNFVCRRTFVDRGHGNGCGLF-GKGSLITCAKFKC------VTKLEGKIVQYEN
WNV             -DKRADPAFVCRQGVVDRGWGNGCGLF-GKGSIDTCAKFAC------STKAIGRTILKEN
YFV             -AEENEGDNACKRTYSDRGWGNGCGLF-GKGSIVACAKFTC------AKSMS--LFEVDQ
MeaV            -DEEHQTGHLCRRDYSDRGWGNHCGLF-GKGSIVGCVRVNC---TA-GKTLKGLEFDSTK
RubellaV        ARIWNGTQRACTFWAVSDA-CWGFPTDT-VMSVFALASYVQ---H-PHKTVRVKFHTETR
ChikV           --LPDYSCKVFTGVYPFMWGGAYCFCDA-ENTQLSEAHVEKSESC-KTEFASAYRAHTAS
SFV             --KPDYQCKVYTGVYPFMWGGAYCFCDS-ENTQLSEAYVDRSDVC-RHDHASAYKAHTAS

                   190       200       210       220       230       240
                ....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          TYATFVY-AAYD----Y---VEK--DKTKIRSQLDGGTQDRHLDQKRRISLAVTAGGRAS
BfBRAFLDRAFT    VVY-----------------THRAQ---------D-------------------------
TBEV            IVYTVK-VEPHTGDYVAANETHS--GRKTASFTISSEKTILTMGEYGDVSLLCRVASGVD
DENV1           LKYSVI-VTVH-----------TTEHGTIATITPQAPTSEIQLTDYGALTLDCSPRTGLD
WNV             IKYEVA-IFVHG----STQVGAT--QAGRFSITPAAPSYTLKLGEYGEVTVDCEPRSGID
YFV             TKIQYV-IRAQ---------------G-IKTLKFDAGSQEVEFIGYGKATLECQVQTAVD
MeaV            ITYAVH-LEAH---------------RKTALVTVASEKHVSTIAGFGSVTIECRVSSGVD
RubellaV        TVWQLSV-A----------------GVSCNVTTEH--PFCNTP---HGQLEVQVPP---
ChikV           ASAKLRV-LYQ-----------G--NNITVTAYANG--HAVTVK---DAKFIVGPMSSA-
```

```
SFV             LKAKVRV-MYG-----------N--VNQTVDVYVNGD-HAVTI----GTQFIFGPLSSAW

                        250       260       270       280       290       300
                ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          H-QTGMYFSRT---ELRMQPLN-EITDNNFDRLGWYRMDDS-------------------
BfBRAFLDRAFT    -------------------------DY-----SETWQ--------P----------FT-
TBEV            L-AQTVILELD-KT-AWQVHRDW-FNDL-----ALPWKHEG-A-QN-WNNAERLV-EFGA
DENV1           F-NEMVLLTME-KK-SWLVHKQW-FLDL-----PLPWTSGASQ-ET-WNRQDLLV-TFKT
WNV             T-NAYYVMTVG-TK-TFLVHREW-FMDL-----NLPWSSA-GS--T-VWRNRETL-EFEE
YFV             F-GNSYIAEME-TE-SWIVDRQW-AQDL-----TLPWQSG-SG-GV-WREMHHLV-EFEP
MeaV            L-AKTMLIEMN-DN-VWSVHRDW-FEDL-----PYPWRH-G----NPWRDAGRLV-GFEP
RubellaV        P-GDLVEYIMNQ--QSRWGLGSPCHGPDWASPVCQRHSPDC-SR-L-VGATPER-PRLRL
ChikV           -PFDNKIVVYKG--DVYNMDYP-PFGAGRPGQFGDIQSRTP-ES-K-DVYANTQ-LVLQR
SFV             TPFDNKIVVYKD--EVFNQDFP-PYGSGQPGRFGDIQSRTV-ES-N-DLYANTA-LKLAR

                        310       320       330       340       350       360
                ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          -----QTYKSILSANHYMPGHFNLTRPLEVI-------------KPWIQ-----S-----
BfBRAFLDRAFT    ---HS--------------------------NRFANITVGSLVGSITPY----A------
TBEV            --PHA--------------------------VKMDVYNLGDQTGVLLKALAG-VP-----
DENV1           --AHA--------------------------KKQEVVVLGSQEGAMHTALTG-AT-----
WNV             --PHA--------------------------TKQSVIALGSQEGALHQALAG-AI-----
YFV             --PHA--------------------------ATIRVLALGNQEGSLKTALTG-AM-----
MeaV            --PHA--------------------------VKMVAYTLGDQTGTVLKILGD-AT-----
RubellaV        VDADDPLLRTAPGPGEVWVTPVIGSQARKCGLHIRAGPYGHATVEMPEWIHA-TT-----
ChikV           PAAGT--------------------------VHVPYSQAPSGFKYWLKERGAS-LQHTAP
SFV             PSPGM--------------------------VHVPYTQTPSGFKYWLKEKGTA-LNTKAP

                        370       380       390       400       410       420
                ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          ------S-RQAV--VTHAEGTNLQISIHLESQNLVFFHNASR-IRDFSG----IIVDSKS
BfBRAFLDRAFT    ---------------------------------NGT-SRMCL-S---------------
TBEV            ----VAHIEGTK--YHL-KSGHVTCEVGLEKLKMKGLTYTMCD-KTF-------------
DENV1           ----EIQTSGTT--TIF-A-GHLKCRLKMDKLTLKGMSYVMCT-GSFKLE----------
WNV             ----PVEFSSNT--VKL-TSGHLKCRVKMEKLQLKGTTYGVCS-KAFKFL----------
YFV             ----RV-TKDTN--NLY-G-GHVSCRVKLSALTLKGTSYKICT-DKMFFV----------
MeaV            ----KGRKTGNK--YEL-SGGHVSCSVGLEKLKLRGLTYGMCA-VGFSWK----------
RubellaV        ---------------PWHPPGPLGLKFKTVRPVALP--R------------TGCYQCGT
ChikV           FGCQIATN-PVR--AVNCAVGNMPISIDIPEAAFTRVVDAPS-LTDMSCEVPACTHSSDF
SFV             FGCQIKTN-PVR--AMNCAVGNIPVSMNLPDSAFTRIVEAPT-IIDLTCTVATCTHSSDF

                        430       440       450       460       470       480
                ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          NRLFNLTVYESGKIDGSVKMSTGFGSD--T--SDLHASNRSMIIPLPVGQGARA----AD
BfBRAFLDRAFT    ------------------------------------------------------------
TBEV            ------------------------------------------------------------
DENV1           ------------------------------------------------------------
WNV             ------------------------------------------------------------
YFV             ------------------------------------------------------------
MeaV            ------------------------------------------------------------
RubellaV        PALVEGLAPGNCHLT----------------EDVGAFPPGKFVTAALLNTPPPYQV-CGG
ChikV           GGVAIIKYAA--SKKGKCAVHSMTNAVTIREAEIEVEGNSQLQISFSTALASAEFRVQVC
SFV             GGVLTLTYKT--NKNGDCSVHSHSNVATLQEATAKVKTAGKVTLHFSTASASPSFVVSLC

                        490       500       510       520       530       540
                ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          SMADIDKICHVIEYFESPLEIDLVEGKWH---------INFNGMMKFLNPAHWIKGISS-
BfBRAFLDRAFT    ---------------QEALQDFFVPPTWR-NPEDAGPGFNFNWLFGFLNPAEWFDGIQG-
TBEV            ----------------------------RAPTDSGHD---------------------T
DENV1           ----------------------------KEVAETQHG---------------------T
WNV             ----------------------------GTPADTGHG---------------------T
YFV             ----------------------------KNPTDTGHG---------------------T
MeaV            ----------------------------RVPTDSQHD---------------------T
RubellaV        ESDRASARVIDPAAQS-TG-----------------------------------------
ChikV           STQVHCAAECHPPKDHIVN-----------------------------------------
SFV             SARATCSASCEPPK----------------------------------------------

                        550       560       570       580       590       600
                ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1          ------------------------------------------PF--------------
BfBRAFLDRAFT    ----------T-SPSRVT----V--HSR-CLSPLSRPDMQICKKVPFVRSNNGTILSFQE
```

```
TBEV          VVMEVTFS-G-TKPCRIPVRAV--GSPDV----------------------NVAMLITP-N
DENV1         VLVQVKY-EGTDAPCKIPFSSQDEK-VTQ----------------------NGRLITANP
WNV           VVLELQY-TGTDGPCKVPISSVASLNDLT----------------------PVGRLVTV-N
YFV           VVMQVKV-KG--APCRIPVIVADD--LTA--------------------AIN-GILVTVNP
MeaV          VVMEVTYT-G-SSPCRIPVRAY-HGTPE----------------------DVASVITA-N
RubellaV      ------------------------------------------------------------
ChikV         -------------------------H---GVQDISATAMSWVQKGPFE------------
SFV           ------------------------------------------------------------

                       610       620       630       640       650       660
              ....|....|....|....|....|....|....|....|....|....|....|....|
CeEFF1        ---------------------------------------------WS--HPQ------
BfBRAFLDRAFT  YLESAAIA-DVFLILLIAGFVASFSVC--CGGHLR-D---EPTEVGGTI--HSVKKEEEQ
TBEV          PTIEN---NGGGFIEMQLPPGDNIIYVGE--LSHQWFQ----------------------
DENV1         IVTDK---EKPVNIEAEPPFGESYIVVGAGEKALKLSWFKKGSSIGKMF------------
WNV           PFVSVATANAKVLIELEPPFGDSYIVV--G---------K--------------------
YFV           IASTN---DDEVLIEVNPPFGDSYIIVGRGDSRLTYQWHKEGSSIGKLFGIHTVFGSAFQ
MeaV          PVVES---THVKFIEMQLPPGDNVIAVGS--LRYQWFQK-S--TIG---AVH-ILGG--G
RubellaV      ----------------------------------------------A-GWS--HPQFEKGGG
ChikV         ----------------------------------------------A-GWS--HPQFEKGGG
SFV           ------------------------------------------------------------

                       670       680       690       700
              ....|....|....|....|....|....|....|....|.
CeEFF1        ------------------------GGS-----FE-------K
BfBRAFLDRAFT  A---KNGTVVTFTFVVVLYPLGG-------CFELGVKTTY-
TBEV          ----------------------------------------
DENV1         -------------------------------------R---
WNV           ----------------------------------------
YFV           GLFGGLNWITKVIMGAVLIWVGINGVIMM-FLSLGVGA---
MeaV          A-FGGLGSARNFTLSISLIAIGG---ILC-SLTLGVGADY-
RubellaV      S-GG-----------------SGGGSWSHPQFE-------K
ChikV         S-GG-----------------SGGGSWSHPQFE-------K
SFV           ----------------------------------------
```

B) Trimmed alignment of NS3Pro homologues

```
                        10        20        30        40        50        60
                    ....|....|....|....|....|....|....|....|....|....|....|
SARS_C30            --------------------TTTLNGLWLD--DTVYCPRHVICT--------SFLVQ-A-
PolV_C3             DYAV-AMAKRNIVTATT----GEFTMLGVHD-NVAILPTHASPG-----E-S-IVI----
ChibaV_C37          -----PTLWSRVVRFG------SGWGFWVSP-TVFITTTHVIPT-----G----------
DenV_S7             -TQK-AELEEGVYRIKQFGKTQVGVGVQKE--GVFHTMWHVTR-----GAV-L-------
WNV_S7              -PVG-AELEDGAYRIKQLGYSQIGAGVYKE--GTFHTMWHVTR-----GAV-L-------
YFV_S7              -------LEDGIYGIFQLGASQRGVGVAQG--GVFHTMWHVTR-----GAF-L-------
MeaV_S7             -------VKNGVYRIYEFGRRQIGVGYGNG--GVLHTMWHVTR-----GAA-I-------
TBEV_S7             -------VKDGVYRIFSFGQNQVGVGYGSK--GVLHTMWHVTR-----GAA-L-------
HCV_S29             -RDK-NQVEGEVQIVST---TQTFLATCIN--GVCWTVYHGAGT-----R------A-S-
PegiVA_S29          ---------GNVVVLGT---TTRSMGTCVN--GVMYATYHGTNG-----R------A-G-
BVDV1_MEROPS        ----------------L-RRGLETGWAYTHQGGISSVDHVTAG-----K----------
HAV_S1              HHHH-HIKPGALCVIDT----GKGTGFFSG--NDIVTAAHVVGN-----N-TFVNV-CYE
TEV_C4              -RDY-NPISSTICHLTN-GHTTSLYGIGFG--PFIITNKHLFRR-----N--TLLVQ-SL
PPV_S30             --------------------------------NRQVSNVHLL------------------
SinV_C3             ----------RLFDVKN--GDVIGHALAME--GKVMKPLHVKG-------T-I-------
BtChymotrypsin      EEAV-PGSWPWQVSLQD--GFHFCGGSLINE-NWVVTAAHCGV----T-T-S-DVV-VAG
RrTrypsin_S1        YTCQ-ENSVPYQVSLNS---YHFCGGSLIND-QWVVSAAHCYK----S-----IQV-RLG
SgTrypsin_S1        TRAA-QGEFPFMVRLS-----MGCGGALYAQ-DIVLTAAHCVS-SGNN-T-S-ITA-TGG
HsThrombin_S1       SDAE-IGMSPWQVMLFR--QELLCGASLISD-RWVLTAAHCLL-----EN-D-LLV-RIG
EcAHP_S6            -KAA-MPDFSAVDS-------EIGVATLINP-QYIASVKHNGG------TN-V---S-FG
NgIgAlSP_S6         --VP-MIDFSVADV-------NKRIATVVDP-QYAVSVKHAK--------------H-YG
PgDP7               -------IANAVVIF----GG--CTGITVSDQGLIFTNHHCGY-----GAI-Q-------
BsSpoIVB            --------------------DSAAGIG----P-----ALGHVIS-----------------
ScSsy5              ------------------------------------ITCAHVVL-----------------
EcDegs_S1           -LAV-RRAAPAVVNVYN-EIRTLGSGVIMDQRGYIITNKHVIN-----DAD-QIIVA-LQ
HsHtra2_S1          -DVV-EKTAPAVVYIEIEVPISNGSGFVVAADGLIVTNAHVVA-----DRR-RVRVR-LL
TmHtra_S1           -NVV-EACAPAVVKIDVRQVASLGSGFIFDPEGYILTNYHVVG-----GA-DNITVT-ML
AtDeg5_S1           -NLF-QKTSPSVVYIEA----GTGSGFVWDKLGHIVTNYHVIA-----KL--RCKVS-LV
SaSpla_S1           -DAT-KEPYNSVVAFV------GGTGVVVGK-NTIVTNKHIAK-----SND-KNRVS-AH
```

```
                          70        80        90       100       110       120
                    ....|....|....|....|....|....|....|....|....|....|....|....|
SARS_C30            -----------------GN-V-LR-VI-GHS-M------------QNCLLRLKVDTSPK
PolV_C3             -----------------GKE-VEIL-DA-KALE------------TNLEITIITLKRNKF
ChibaV_C37          -------------VREFFG-EPIE-SI-AIH-R------------AGEFTQFRFSRKRP
DenV_S7             ------------T---HN-GKRLE-PN-WAS-V------------KKDLISYG------
WNV_S7              ------------M---HK-GKRIE-PS-WAD-V------------KKDLISYG------
YFV_S7              ------------V---RN-GKKLI-PS-WAS-V------------KEDLVAYG------
MeaV_S7             ------------S---ID-GG-VQ-PS-WAD-V------------QKDLVAYG------
TBEV_S7             ------------S---ID-DAVAG-PY-WAD-V------------REDVVCYG------
HCV_S29             -----------------PK-GPVI-QM-YTN-V------------DQDLVGWPAPQG--
PegiVA_S29          -----------------PM-GPVN-AR-WWS-T------------SDDVCVYPLPMG--
BVDV1_MEROPS        -------------DLL-----RVV-CQ-SNN----------------DETEY-------
HAV_S1              ----------------GLM-YEAK-VR-YMPE--------------KDIAFITCPGDHP
TEV_C4              -------------VFKVKN-TTTL-QQ-HLI-D------------GRDMIIIRMPKDPP
PPV_S30             ----------E-------------------------------QFVN------------A---
SinV_C3             ----------------HP-VLSKL-KF-TKS-S------------AYDMEFAQLPV--R
BtChymotrypsin      --EFDQGSSS-EK---IQK-LKIA-KV-FKNS------------INNDITLLKLSTASF
RrTrypsin_S1        --EHNINVLE-GN---EQF-VNAA-KI-IKHP------------LNNDIMLIKLSSPKL
SgTrypsin_S1        --VVDLQ--S-GA---AVK-VRST-KV-LQAP------------TGKDWALIKLAQPN-
HsThrombin_S1       --KHSRTRYERNI---EKI-SMLE-KI-YIHP-YNWRE-------LDRDIALMKLKKPAF
EcAHP_S6            D-----------G---EN-RYNIV-DR-NNA-P------------SLDFHAPRLDKLTE
NgIgA1SP_S6         QD------VA-DK---EN-EYRVV-EQ-NNY-P------GA---GRLEDYNMARFNKFTE
PgDP7               ----------------QS--------PF-YSN----------------GDFSVFRVY----
BsSpoIVB            ------------------------------------------------KLARF-------
ScSsy5              ----------G-----------------------W-KKGQVVW-RLSDFAIIKVNSSKC
EcDegs_S1           ------------D---GR-VFEAL-LV-GSD-S------------LTDLAVLKINA--G
HsHtra2_S1          ------------S---GD-TYEAV-VT-AVD-P------------VADIATLRIQT-EP
TmHtra_S1           ------------D---GS-KYDAE-YI-GGD-E------------ELDIAVIKIKA-KK
AtDeg5_S1           -D------AK-GT---RF-SKEGK-IV-GLD-P------------DNDLAVLKIET-RE
SaSpla_S1           HS------KG-KG---GG-NYDVK-DI-VEYPG------------KEDLAIVHVHE-KN

                          130       140       150       160       170       180
                    ....|....|....|....|....|....|....|....|....|....|....|....|
SARS_C30            ---T-PKY-----K----------------------FVRIQPGQTFSVLACY---N---G
PolV_C3             ----R-DIRP-I------------------------TQITETNDGVLIVNTS------K
ChibaV_C37          -----TGM----VL----------------------EEGCPEGVVCSILIKR---D---S
DenV_S7             -------GGW--RL----------------------SAQWQKGEEVQVIAVE---P---G
WNV_S7              -------GGW--KL----------------------EGEWKEGEEVQVLALE---P---G
YFV_S7              -------GSW--KL----------------------EGRWDGEEEVQLIAAV---P---G
MeaV_S7             -------GDW--KL----------------------DKKW--GSDVQVHAFP---P---G
TBEV_S7             -------GAW--SL----------------------EEKW-KGETVQVHAFP---P---G
HCV_S29             ---S-RSL---------------------------------CGSSDLYLVTR---------
PegiVA_S29          ---A-TCL---------------------------------CSPQGVWVV-----------
BVDV1_MEROPS        ------------------------------------------------YVL-----------
HAV_S1              ---TA-RL--KLS-------------------------KNPDYSCVTVMAYVN------E
TEV_C4              ---FP-Q------K----------------------FREPQREERICLVTTN---F-TKS
PPV_S30             ---V-------------------------------------KANGQKVEIIGRKR-------
SinV_C3             ---S-EAF---------------------------TSEHPEG-FYNW------------
BtChymotrypsin      SQTVS-AV--CLPS----------------------SDDFAAGTTCVTTGWGLTR--NTP
RrTrypsin_S1        NARVA-TV--ALPS----------------------SCAPAGTQCLISGWGNTL-SNEP
SgTrypsin_S1        ---QP-TL--KIAT----------------------TTAYNQGTFTVAGWGANR-GSQQ
HsThrombin_S1       SDYIH-PV--CLPD----------------------ETALQAGYKGRVTGWGNLK-T-QP
EcAHP_S6            ---V-APTAV--TAVAG-------------------YLDKERYPVFYRLGSGTQY----Y
NgIgA1SP_S6         ---V-API-A--PTDAG-------------------YKDKNRFSSFVRIGAGRQLA---Y
PgDP7               -----------------QFANALAAHAGILKSKYKD-----------------------
BsSpoIVB            -----------------SER-TIGSPFG---TPIQN--EVKKGFDIEI-------------
ScSsy5              QNT-------------------------------------KPGMKVFKIGAST-------
EcDegs_S1           ---L-PTIPI--NA----------------------RRVPHIGDVVLAIGNP---Y---N
HsHtra2_S1          ---L-PTLPL--GR----------------------SADVRQGEFVVAMGSP---F---A
TmHtra_S1           ---F-PYLEF--GD----------------------SDKVKIGEWAIAIGNP---L---G
AtDeg5_S1           ---L-NPVVL--GT----------------------SNDLRVGQSCFAIGNP---Y---G
SaSpla_S1           ---V-SYTKF--AD----------------------GAKVKDRISVIGYP---K-QTK

                          190       200       210       220       230       240
                    ....|....|....|....|....|....|....|....|....|....|....|....|
SARS_C30            SP-SGVYQCAMRP-N---------------------------H-------T-IKGSFLN
PolV_C3             YP-NMYVPVGAV-TEQ--GYLNLGG--------RQT-A----RTLMYN--------FPTRA
ChibaV_C37          GE-LLPLAVRMGA-IA--SMKIQGR----------------HGQSGMLLT-G-MDLGTLP
DenV_S7             KN-PKNFQTMPGT-F-QT-T----------------TG---E-IGAI--------ALDFKP
WNV_S7              KN-PRAVQTKPGL-F-KT-N----------------TG---T-IGAV--------SLDFSP
```

```
YFV_S7          KN-VVNVQTKPSL-F-KV------------------G---E-IGAV--------ALDYPS
MeaV_S7         ---PHSVQTSPGV-L-RL-S---------------SG---E-KGAI--------HIDL-R
TBEV_S7         RA-HEVHQCQPGE-L-IL-D---------------TG--RK-LGAI--------PIDLVK
HCV_S29         ---ADVIPVRRRG-D-----------------------------SRGSL--SP-RPISYLK
PegiVA_S29      ---DGALC-----------------------------------L-------P-AELCDFR
BVDV1_MEROPS    ------------------------------------------------F-FDLKNLK
HAV_S1          -D-LVVSTAAAM-V----------------------G----NTLSYA--------VRTQD
TEV_C4          MS--MVSTSCTFP-S---------------------G----IFWKH--------WIQTKD
PPV_S30         ------------------------------------------------------GEVTP
SinV_C3         --------------------------HG-------G---R-FTIP--------RGVGGR
BtChymotrypsin  DR-LQQASLPLL-SNT--NCKKYWGT------KIK-D----AMICAG----A-SGVSSCM
RrTrypsin_S1    DL-LQCLDAPLL-PQA--DCEASYPG------KIT-D----NMVCVG----L-EGGGSCQ
SgTrypsin_S1    RY-LLKANVPFV-SDA--ACRSAYGN------ELV-N----EEICAG----YPGGVDTCQ
HsThrombin_S1   SV-LQVVNLPIV-ERP--VCKDSTRI------RIT-D----NMFCAG----KG-RGDACE
EcAHP_S6        SW-LTGGTVGSLS-S-----------------Y----GEM----ISTSSFDGA-MPIYGEA
NgIgAlSP_S6     RY-AIAGTPYKIN-I-----------------D----NGL----IGFG----A-LTNYGVL
PgDP7           --------------------------KGVLE-------Q---FLS--------NNDITG
BsSpoIVB        ----------------------------MVL----DP-----LLKE--------TGGIVQ
ScSsy5          -------------------------------------S----EF--------P-TPLFASA
EcDegs_S1       LG-QTITQGIISA-T-GR-I-----------------Q---N-FLQT--------DASINH
HsHtra2_S1      LQ-NTITSGIVSS-A-QR-P----------------NV---E-YIQT--------DAAIDF
TmHtra_S1       FQ-HTVTVGVVSA-T-NR-RIPKPDGS-----GY-YV---G-LIQT--------DAAINP
AtDeg5_S1       YE-NTLTIGVVSG-L-GR-EIPSPNGK------S-IS---E-AIQT--------DADINS
SaSpla_S1       YK-MFESTGTINH-I-S--------------------G---T-FMEF--------DAYAQP

                        250       260       270       280       290
                ....|....|....|....|....|....|....|....|....|....|....
SARS_C30        GSCGSVGFNIDY--DCVSFCYMHHMELP---------GVHAGTDLEGKFYGPFV
PolV_C3         GQCGGVITCT-----G-KVIGMHVGGNG---------SHGFAAALK--RSYFT-
ChibaV_C37      GDCGAPYVYKRN--NDWVVCGVHAAATKS-------GNTVVCAVQA--------
DenV_S7         GTSGSPIINR-----EGKVVGLYGNGVVT---K-NGGYVSGIAQTNAE------
WNV_S7          GTSGSPIVDK-----KGKVVGLYGNGVVT---R-SGAYVSAIANTE--------
YFV_S7          GTSGSPIVNR-----NGEVIGLYGNGIL--------SFVSAISQ----------
MeaV_S7         GTSGSPILDE-----NGNVVGLYGNGLR------YGNYVSCIAQG---------
TBEV_S7         GTSGSPILNA-----QGVVVGLYGNGLKT-----NETYVSSIAQG---------
HCV_S29         GSSGGPLLCP-----TGHAVGLFRAAVCT---R-GVAKAVDFIPVENLETTMR-
PegiVA_S29      GSSGSPILCD-----EGHAVGML-VSVLH---R-GVT-GIRY--TK-WETLPR-
BVDV1_MEROPS    GWSGLPIFEAS----SGRVVGRVKVGKNE---E---SKPTIMSGIQTVSK----
HAV_S1          GMSGAPVCDK-----YCRVLAVHQTNTG---------YTGGAVIID--PTDFHP
TEV_C4          GQCGSPLVSTR----DGFIVGIHSASNFT-------NTNNYFTSVPKNFMELLT
PPV_S30         GMSGFV-----------------------------------------------
SinV_C3         GDSGRPIMDN-----SGRVVAIVLGGADE----GTR-TALSVVTWNSKGKTIKT
BtChymotrypsin  GDSGGPLVCKKN--GAWTLVGIVSWGSST---C-STSTPGVYARVTALVNWVQQ
RrTrypsin_S1    GDSGGPVVCN------GELQGIVSWGYGC---A--PDNPDVYTKVCNYVDWIQD
SgTrypsin_S1    GDSGGPMFRKDNA-DEWIQVGIVSWGYGC---A--RPYPGVYTEVSTFASAIAS
HsThrombin_S1   GDSGGPFVMKSPF-NRWYQMGIVSWGEGC---D--DGKYGFYTHVFRLKKWIQK
EcAHP_S6        GDSGSPLFAFDTVQNKWVLVGVLTAGNGA---G--G-RGNNWAVI--PLDFIGQ
NgIgAlSP_S6     GDSGSPLFAFDK----WVFLGTYDYWAGY---G--KKSWQEWNIY--KKEFA--
PgDP7           GNSGSPVFDK-----NGRLIGLAFDGNWE----AMSGDIEFE---------VLF
BsSpoIVB        GMSGSPIIQ------NGKVIGAVTHVFVN----DPTSGYGVHIEWML-------
ScSsy5          GDSGAWILTK-----LEDRLGLGLVGMLH---S--QRQFGLFTPIGDILERLHD
EcDegs_S1       GNSGGALVNS-----LGELMGINTLSFD-SND--TPEGIGFAIPFQLATKIMDK
HsHtra2_S1      GNAGGPLVNL-----DGEVIGVNTMKVT--------AGISFAIPSDRLREFLHR
TmHtra_S1       GNSGGPLLNI-----HGEVIGINTAIVNP---Q-EAVNLGFAIPINTVKKFLDT
AtDeg5_S1       GNAGGPLLDS-----YGHTIGVNTATF--KGS------VNFAIPIDTVVRTVPY
SaSpla_S1       GNSGSPVLNS-----KHELIGILYAGSGK---D--ESEKNFGVYFTQLKEFIQN
```

C) Trimmed alignment of NS3Hel homologues

```
                    10        20        30        40        50        60
               ....|....|....|....|....|....|....|....|....|....|....|
VVNPHII        IN-SFD-----EYIL--------------RGL-LEIPL----AS-TPKAQR-EIFS-AWI
TMV            ---------------QLSRGFTIPHYR--TEG-KFMTF----TR-ATATEV-AGKI-AHE
HsBRR2         -L-PVEKLP----KY--------------AQ-AGFEGF----KT-LNRIQ-SKLYR-AAL
HsRIG1         ----------------------------------------K-PRNYQ-LELAL-PAM
HsBML          ----LSFPHTKEMMK--------------IF-HKKFG-L---HNF-RTNQL-EAIN-AAL
HsDdx3x        IES-FSDVEMGEIIM--------------GN-IELTR-Y---TRP-TPVQK-HAIP-IIK
HseIF4AIII     TPT-FDTMGLREDLL--------------RG-IYAYG-F---EKP-SAIQQ-RAIK-QII
HsDdx10        ITR-FSDFPLSKKTL--------------KG-LQEAQ-Y---RLV-TEIQK-QTIG-LAL
DmVasA         IQH-FTSADLRDIII--------------DN-VNKSG-Y---KIP-TPIQK-CSIP-VIS
```

```
EcCsdA      ---TFADLGLKAPIL--------------EA-LNDLG-Y---EKP-SPIQA-ECIP-HLL
ScUPF1      GH-QVVDISFDVPLP--------------KEF-SIPNF----AQ-LNSSQS-NAVS-HVL
HHV1UL19    ------------------------------------------------------------
HHV1UL5     ------------------------------------------------------------
SidV        ------------------------------------------------------------
EcRepA      -------------HK--------------PI-NILEAF----AA-APPPL-DYVLP-NMV
T4GP17      DD-IVYFAETYCAIT--------------HID-YGVIK----VQL-RDYQ-RDMLK-IMS
EcRecQ      ----EVLNLESGAKQ--------------VL-QETFG-Y---QQF-RPGQE-EIID-TVL
KpPriA      PI-GDVLFHALPVML--------------RQ-GKPASA-RSALR-LNTEQ-ATAVG-AIH
TYMV        ------------------------------------------------------------
DENV4       ---------------------------------------S----AMG-EPDY-EVDED-IFR
TBEV        ------------------------------------------------------------
KUNV        ------------------------------------------------------------
YFV         ------------------------------------------------------SH-MLK
MeaV        ----------------------------------------------------------RA-WMS
PegiVA      ------------------------------------------------------------
BVDV1       ------------------------------------------------------------
HCV         CT-RGVAKAVDFVPV--------------ESM-ETTMR----SPV-FTDN-SSPPA-VPQ
HEV         ------------------------------------------------------------
PVY         IK-NFDEFELSED--QIQMGHTLPHYR--TEG-HFMEF----TR-ATAVQV-ANDI-AHS
ChikV       ------------------------------------------------------------
DmDEAD      NP-SFEDLGLSPELL--------------KA-LKKLG-F---EKP-TPIQA-QAIP-LIL
TtHerA      -M-EFKDFPLKPEIL--------------EA-LHGRG-L---TTP-TPIQA-AALP-LAL
MjDDEADBOX  YX-NFNELNLSDNIL--------------NA-IRNKG-F---EKP-TDIQX-KVIP-LFL
StHel       ---------MNEKIE--------------QA-IREMG-F---KNF-TEVQS-KTIP-LML
HsDDX6      GNE-FEDYCLKRELL--------------MG-IFEMG-W---EKP-SPIQE-ESIP-IAL
PtHR18934   -------------Y--------------RDI-LKLKRR----L-V-HRQR-DEFLK--YQ
lcl         ----------------------------SI-QEQRES----LP-IYKYR-DQ---YAVE
HrpA        -------------N--------------VPDI-LEYRSG----LP-VTAVR-DEIL-KAIE
HrpB        -------------S-------------I--V-FMLRRA-MPALP-IEAVL-PD-L-RLAA
lcl2        ----------------------------SI-EEVRKS----LP-VYPYK-DELL-KAVK
ScBDP5      AKS-FDELGLAPELL--------------KG-IYAMK-F---QKP-SKIQE-RALP-LLL
HsDDX19B    VKS-FEELRLKPQLL--------------QG-VYAMG-F---NRP-SKIQE-NALP-LML
BmVasA      IES-FETANLRKYVL--------------DN-VLKAG-Y---RKP-TPIQK-NAIP-IIM
ScMss116p   EVT-LDSLVLDKEIH--------------KA-ITRME-F---PGL-TPVQQ-KTIK-PIL
ApRigI      ---------------------------------------T---KK-ARSYQ-IELAQ-PAI

                      70        80        90       100       110       120
            ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII     --S----H-RPVVLTGGTGVGKTSQ-VPKLLLWF---------------HE---RPVILS
TMV         --S----D-KDILLMGAVGSGKSTG-LPYHLSR---------------KG---NVLLLE
HsBRR2      E-T----D-ENLLLCAPTGAGKTNV-ALMCMLREIGKH-------INVDDF---KIIYIA
HsRIG1      --K----G-KNTIICAPTGCGKTFV-SLLICEHHLKKF-----------KG---KVVFFA
HsBML       --L----G-EDCFILMPTGGGKSLC-YQLPACVS---------------PG---VTVVIS
HsDdx3x     --E----K-RDLMACAQTGSGKTAA-FLLPILSQIYSDG-G-RYGRRKQYP---ISLVLA
HseIF4AIII  K------G-RDVIAQSQSGTGKTAT-FSISVLQCLD---------IQVRET---QALILA
HsDdx10     --Q----G-KDVLGAAKTGSGKTLA-FLVPVLEALYRLQW-----TSTDGL---GVLIIS
DmVasA      --S----G-RDLMACAQTGSGKTAA-FLLPILSKLLEDPHEL----ELGRP---QVVIVS
EcCsdA      --N----G-RDVLGMAQTGSGKTAA-FSLPLLQNL-----------DPEL-APQILVLA
ScUPF1      --Q----R-PLSLIQGPPGTGKTVT-SATIVYHLSKI----------HKD---RILVCA
HHV1UL19    -------R-CVTVVRAPMGSGKTTA-LIRWLREAIH----------SPDT---SVLVVS
HHV1UL5     -------F-AVYLITGNAGSGKSTC-VQTIN----------------------DCVVTG
SidV        ---------------GTPGSGKSAI-IKSTVTAR---------------------LVTS
EcRepA      --A----G-TVGALVSPGGAGKSML-ALQLAAQIAGG---------ELPTG---PVIYLP
T4GP17      --S----K-RMTVCNLSRQLGKTT-VVAIFLAHFVCFN----------DK---AVGILA
EcRecQ      --S----G-RDCLVVXPTGGGKSLC-YQIPALLL--------------NG---LTVVVS
KpPriA      --SAADRF-SAWLLAGITGSGKTEV-YLSVLENVLAQ-----------GR---QALVMV
TYMV        ----------VVHFAGFAGCGKTYP-IQQLLKTKLF---------------DFRVSC
DENV4       --K----K-RLTIMDLHPGAGKTKRILPSIVREALKR-----------RL---RTLILA
TBEV        -------KGQITVLDMHPGSGKTHRVLPELIRQCIDR-----------RL---RTLVLA
KUNV        -------KKQITVLDLHPGAGKTRRILPQIIKEAINR-----------RL---RTAVLA
YFV         --K----G-MTTVLDFHPGAGKTRRFLPQILAECARR-----------RL---RTLVLA
MeaV        --K----G-SITVVDMHPGSGKTHTVLPELVRRCIIE-----------RK---RTLVLA
PegiVA      --G----Y-REAPLYLPTGSGKSTR-IPAEY---AK-----------AGH---KVLVLN
BVDV1       ----------------ATGAGKTTE-LPKAVIEEIGR-----------HK---RVLVLI
HCV         --S----F-QVAHLHAPTGSGKSTK-VPAAY---AA-----------QGY---KVLVLN
HEV         --------------AGVPGSGKSRS-ITQAD----------------------VDVVVV
PVY         --E----H-LDFLVRGAVGSGKSTG-LPVHLSV----------------AG---SVLLIE
ChikV       -------------VFGVPGSGKSAI-IKNLVTRQ---------------------LVTS
DmDEAD      --E----G-RDVLAQAQTGSGKTLA-FLLPILQRL-----------LRQPNGPQALVLA
TtHerA      --E----G-KDLIGQARTGTGKTLA-FALPIAERLAPS--------RGRKP---RALVLT
```

```
MjDDEADBOX  --N---DE-YNIVAQARTGSGKTAS-FAIPLIELVN----------ENNGI---EAIILT
StHel       --Q----G-KNVVVRAKTGSGKTAA-YAIPILELG---------------M---KSLVVT
HsDDX6      --S----G-RDILARAKNGTGKSGA-YLIPLLERLD---------LKKDNI---QAMVIV
PtHR18934   --SAIREN-QVVVIVGETGSGKTTQ-IPQYLL--EA---------GYTKGG--GKIGCTQ
lcl         --D----N-QVLIVIGETGSGKSTQ-IPQYLA--EA---------GFASSG---KIACTQ
HrpA        --Q----N-QVVIIVGETGSGKTTQ-LPQFLL-EEG---------LGIAG----KIGCTQ
HrpB        --H----P-QV-VLEAPPGAGKTTA-VPLALL-DA----------PHADNAAGKKIIMLE
lcl2        --E----H-QVLIIVGETGSGKTTQ-IPQYLY--EA---------GYTKGK---KIGCTQ
ScBDP5      HNP----P-RNMIAQSQSGTGKTAA-FSLTMLTRVN---------PEDASP---QAICLA
HsDDX19B    AEP----P-QNLIAQSQSGTGKTAA-FVLAMLSQVE---------PANKYP---QCLCLS
BmVasA      --S----G-RDLMGCAQTGSGKTAA-FLVPIINMLLQDPKD-ISENGCAQP---QVIIVS
ScMss116p   SSE----D-HDVIARAKTGTGKTFA-FLIPIFQHLINTKF-----DSQYMV---KAVIVA
ApRigI      --N----G-KNALICAPTGSGKTFV-SILICEHHFQNM--------AGRKA---KVVFLA

                    130       140       150       160       170       180
            ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII     PRIALVRLHSNT---ILKL--FK---SPIS--RYG-------------------------
TMV         PTRPLAENVHKQ---LSQAPF-HQ---NTTL--MRG--------LTA-----FGSAPISV
HsBRR2      PMRSLVQEMVGS---FGKRLAT-YG-ITVA-ELTGDHQL----CKE----EI-SATQIIV
HsRIG1      NQIPVYEQQKSV---FSKYFER-HG-YRVT-GISGN------PVEQ----IV-ENNDIII
HsBML       PLRSLIVDQVQK---LTSL-----D-IPAT-YLTGDKTDSE-ATNIYLSKKD-PIIKLLY
HsDdx3x     PTRELAVQIYEE---ARKFSY-RSR-VRPC-VVYGGADIGQ-QIRD----LE-RGCHLLV
HseIF4AIII  PTRELAVQIQKG---LLALGDY-MN-VQCH-ACIGGTNVGE-DIRK----LD-YGQHVVA
HsDdx10     PTRELAYQTFEV---LRKVGK-NHD-FSAG-LIIGG--LKH-EA-E----RI-NNINILV
DmVasA      PTRELAIQIFNE---ARKFAF-ESY-LKIG-IVYGGTSFRH-QNEC----IT-RGCHVVI
EcCsdA      PTRELAVQVAEA---MTDFKH-MRG-VNVV-ALYGGQR-YD-QLRA-----L-RQGQIVV
ScUPF1      PSNVAVDHLAAK---LRDLG------LKVVR-LTAKSRERK-TEAE----IL-NKADVVC
HHV1UL19    CRRSFTQTLATR---FAESGL--------T-------------------------------
HHV1UL5     ATRIAAQNMYAK---LSG--------------------ELRN-EFR-LA----------LAP
SidV        GKKENCREIEADVLR-------------Q---------------------------------
EcRepA      AEDP-----------------------------------------EER----QA-VADGLLI
T4GP17      HKGSMSAEVLDR---TKQAIE-LLP-DF--------------GSIE----LD-NGSSIGA
EcRecQ      PLISLXKDQVDQ---LQAN-----G-VAAA-CLNSTQTREQ-QLEVXTGCRT-GQIRLLY
KpPriA      PEIGLTPQTIAR---FRQRFN-----APVE-VLHSGLNDSE-RLSA----WN-GEAAIVI
TYMV        PTTELRTEWKTA---MELH-------SQ--------------------------------
DENV4       PTRVVAAEMEEA---LRGL---------PIR-YQTP--------KSD----HT-GREIVDL
TBEV        PTRVVLKEMERA---LNGK---------RVR-FQQ-----------------A-GGAIVDV
KUNV        PTRVVAAEMAEA---LRGL---------PIR-YQTS------------------GNEIVDV
YFV         PTRVVLSEMKEA---FHGL---------DVK-FHTQ-----A-FSAH----GS-GREVIDA
MeaV        PTRVVLREMERA---LRGR---------NVR-FHSD-----S-VNVK----GE-GA-IVDV
PegiVA      PSIATVRAMGPY---MEKLTG-----QHPSV-YCGHD--------T----TT-TQSNLTY
BVDV1       PLRAAAESVYQY---MRLKHPS----ISFN-LRIGDM--------------D-MATGITY
HCV         PSVAATLGFGAY---MSKAHG-----IDPNI-RTGVR--------T----IT-TGAPVTY
HEV         PTRELRNAWRR-----------------------------------------------
PVY         PTRPLAENVFKQ---LSSEPF-FK---KPTL--MRG--------NSI-----FGSSPISV
ChikV       GKKENCQEITT-VMR-----------LEI-----------------------------
DmDEAD      PTRELAQQIYKV---LKKLGK-YLG-LRVA-LLIGGTSLKE-QIRR----LK-KGPDIVV
TtHerA      PTRELALQVASE---LTAVAP---H-LKVV-AVYGGTGYGK-QKEA----LL-RGADAVV
MjDDEADBOX  PTRELAIQVADE---IESLKG-NKN-LKIA-KIYGGKAIYP-QIKA-----L-KNANIVV
StHel       PTRELTRQVASH---IRDIGR-YMD-TKVA-EVYGGMPYKA-QINR-----V-RNADIVV
HsDDX6      PTRELALQVSQI---CIQVSK-HMGGAKVM-ATTGGTNLRD-DIMR----LD-DTVHVVI
PtHR18934   PRRVAAISVAER---VAEEMGEELG-EEVG-YQIRFE-------DC----TS-EKTRIKY
lcl         PRRVAAVSLAKR---VAEEMGCQLG-EEVG-YTIRFE-------DS----TS-KDTRIKY
HrpA        PRRLAARSVAER---VAEELGEKLG-ETVG-YSIRFE-------SK----VS-PRTRIKV
HrpB        PRRLAARAAARR---LAELLGERVG-ETVG-YRVRFE-------SK----VS-AKTRIEV
lcl2        PRRVAAMSVAAR---VAEEMGVKLG-HEVG-YSIRFE-------DC----TS-EKTVLKY
ScBDP5      PSRELARQTLEV---VQEMGKF-TK-ITSQ-LIVPDSF-----EKN----KQ-INAQVIV
HsDDX19B    PTYELALQTGKV---IEQMGKF-YPELKLA-YAVRGNK----LERG----QK-ISEQIVI
BmVasA      PTRELTLQIFNE---ARKFSY-GSV-LKVA-VAYGGTAVRH-QGDN----IA-RGCHILV
ScMss116p   PTRDLALQIEAE---VKKIHDGLKK-YACV-SLVGGTDFRA-AMNM----NK-LRPNIVI
ApRigI      TKVPVYEQQKNV---FKHHFER-QG-YSVQ-GISGENFSNV-SVEK----VI-EDSDIIV

                    190       200       210       220       230       240
            ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII     KKYGIVFSTHKL-SL--TKLFSYGTLIIDEVHEHD----------Q--IGDI-IIAVARK
TMV         MTSGFALNYFAN-NR--MRIEEFDFVIFDECHVHDANAM---------AMRC--------
HsBRR2      CTPEKWDIITRK-GGERTYTQLVRLIILDEIHLLHD-DR-----GP--VLEA-LVARAIR
HsRIG1      LTPQILVNNLKK-GT-IPSLSIFTLMIFDECHNTS-----------QHPYNM-IMFNYLD
HsBML       VTPEKICASNRL-ISTLYERKLLARFVIDEAHCVSQWG-DFRQDYK--RMNM--RQK---
HsDdx3x     ATPGRLVDMMER-GK--IGLDFCKYLVLDEADRMLDMGF-----EP--QIRRI-VEQ---
HseIF4AIII  GTPGRVFDMIRR-RS--LRTRAIKMLVLDEADEMLNKGF-----KE--QIYDV-YRYL--
```

```
HsDdx10      CTPGRLLQHMDETVS--FHATDLQMLVLDEADRILDMGF-----AD--TMNAV-IENL--
DmVasA       ATPGRLLDFVDR-TF--ITFEDTRFVVLDEADRMLDMGF-----SE--DMRRI-MTH---
EcCsdA       GTPGRLLDHLKR-GT--LDLSKLSGLVLDEADEMLRMGF-----IE--DVETI-MAQI--
ScUPF1       CTCVGAGDK---------RLDTKFRTVLIDESTQAS---------E--P----I------
HHV1UL19     -----SLHR--V-GP--NLLNNYDVLVLDEVMSTLGQLY-----SP--TMQQ---G--D-
HHV1UL5      ATHGA-----------LPAFTRSNVIVIDEAGLLGR-----HLL----------------
SidV         -------------------HKAVEVLYVDEAFACHA-----GALL---------------
EcRepA       QPLI-APEWFDG-LK--RAAEGRRLMVLDTLRRFHIEEE----------A--V----EA
T4GP17       YASSP---------A--VRGNSFAMIYIEDCAFIP---------N----HD-SWLAIQP
EcRecQ       IAPERLXLDNFL-E--HLAHWNPVLLAVDEAHCISQWG-DFRPEYA--ALGQ--RQR---
KpPriA       GTRSSL----------FTPFKDLGVIVIDEEHDSS----------YKQ--YH-ARDLAVW
TYMV         --------------S--SILKSSRILVIDEIYKMPRG-Y-----L---------------
DENV4        MCHATFTTRLLS-ST--R-VPNYNLIVMDEAHFTD----------P--CSVA-ARGYIST
TBEV         MCHATYVNRRLL-PQ--R--QNWEVAIMDEAHWTD----------P--HSIA-ARGHLYT
KUNV         MCHATLTHRLMS-PH--R-VPNYNLFVMDEAHFTD----------P--ASIA-ARGYIST
YFV          MCHATLTYRMLE-PT--R-VVNWEVIIMDEAHFLD----------P--ASIA-ARGWAAH
MeaV         MCHATYTHRRLL-PV--T-QVNYEVAIMDEGHWTD----------P--CSIA-ARG----
PegiVA       CTYGRFMANPR-------Y--GHDVVICDECHSTDG----------VSVL-GMGRLL-
BVDV1        ASYGYFCQMPQP-KL-RAAMVEYSYIFLDEYHCAT----------PEQ--LA-IIGKIH-
HCV          STYGKFLADGG------CSGGAYDIIICDECHSTDS----------TTIL-GIGTVLD
HEV          -----------------ARVTQGRRVVIDEAPSLP----------P--HL----------
PVY          MTSGFALHYFAN-NR--SQLAQFNFVIFDECHVLDPSAM---------AFRS--------
ChikV        --------------------RPVDVLYVDEAFACHS-----GTL----------------
DmDEAD       ATPGRLLDLLEN-GK--LNLKNLKYLVLDEADRMLDMGF-----EE--QIRKI-LRQL--
TtHerA       ATPGRALDYLRQ-GV--LDLSRVEVAVLDEADEMLSMGF-----EE--EVEAL-LSAT--
MjDDEADBOX   GTPGRILDHINR-GT--LNLKNVKYFILDEADEXLNXGF-----IK--DVEKI-LNAC--
StHel        ATPGRLLDLWSK-GV--IDLSSFEIVIIDEADLMFEMGF-----ID--DIKII-LAQT--
HsDDX6       ATPGRILDLIKK-GV--AKVDHVQMIVLDEADKLLSQDF-----VQ--IMEDI-ILTL--
PtHR18934    MTDGMLLRELLS----DPLLSKYSVIILDEAHERT----------L--NTDF-LLGLLKD
lcl          MTDGMLLREILK----DPLLSKYSVIILDEAHERS----------L--HTDI-LLGLLKK
HrpA         MTDGILLREIQN----DPLLSGYSVVIIDEAHERS----------L--NTDI-LLGLLKD
HrpB         VTEGVLTRMILD----DPELSGVGAVIFDEFHERS----------L--DADL---GLLAL
lcl2         MTDGMLLREFLS----EPDLASYSVIIVDEAHERT----------L--HTDI-LFGLVKD
ScBDP5       GTPGTVLDLMRR-KL--MQLQKIKIFVLDEADNMLDQQL-----GD--QCIRV-KRFL--
HsDDX19B     GTPGTVLDWCSKLKF--IDPKKIKVFVLDEADVMIATQH-----QD--QSIRI-QRML--
BmVasA       ATPGRLHDFVER-NR--VSFGSVRFVVLDQADCMLDMGF-----MP--SIEKM-MLH---
ScMss116p    ATPGRLIDVLEKY-SN-KFFRFVDYKVLDEADRLLEIGF-----RD--DLETI-SGILNE
ApRigI       VTPQILVNSFED-GT-LTSLSIFTLMIFDECHNTT----------NHPYNV-LMTRYLE

                      250       260       270       280       290       300
             ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII      -HH--------TKIDSMFLMTATLEDD----------R-------ERLKVFLP-NPAFIH
TMV          LLH-E---C--DYSGKIIKVSATPPGR----------EV------EFST-----------
HsBRR2       -NI-E---M--QEDVRLIGLSATLPNY----------E------D--VATFLDPAKGLFY
HsRIG1       -QK-LGG-S--GPLPQVIGLTASVGVGDAKNTDEALDY------ICKLCASLD-ASVIAT
HsBML        -----------FPSVPVMALTATANPR----------V------QKDILTKIL-RPQVFS
HsDdx3x      -D-M-----PPKGVRHTMMFSATFPKE----------I------QMLARDFLD-EYIFLA
HseIF4AIII   -----------PPATQVVLISATLPHE----------I------LEMTNKFMT-DPIRIL
HsDdx10      -----------PKKRQTLLFSATQTKS----------V------KDLARLSLK-NPEYVW
DmVasA       ---V-----TMRPEHQTLMFSATFPEE----------I------QRMAGEFLK-NYVFVA
EcCsdA       -----------PEGHQTALFSATMPEA----------I------RRITRRFMK-EPQEVR
ScUPF1       ------------GAKQVILVGDHQQLG----------PVILERKSLFERLISL-GHVPIR
HHV1UL19     --------L--RICPRIIAMDATANAQ----------LV------DF-------------
HHV1UL5      ------------------------------------------------------------
SidV         -----------RPRKKVVLC-----------------------------------------
EcRepA       -IA--------DTGCSIVFLHHAVLVD----------N------I----------RWQSY
T4GP17       -VI-S--S---GRRSKIIITTTPNGLN----------------WTAAVEGKS-GFEPYT
EcRecQ       -----------FPTLPFXALTATADDT----------T------RQDIVRGLN-DPLIQI
KpPriA       -RA-H------SEQIPIILGSATPALE----------T------LHNVRQ--G-KYRQLT
TYMV         ------------------------------------------------------------
DENV4        -RV----E---MGEAAAIFMTATPPGS----------ID--------------------P
TBEV         -LA----K---ENKCALVLMTATPPGK----------SE--------------------P
KUNV         -RV----E---LGEAAAIFMTATPPGT----------SD--------------------P
YFV          -RA----R---ANESATILMTATPPGT----------SD----------------------
MeaV         -LA----S---ANENAFVLMTATPPGT----------SD----------------------
PegiVA       --A-K---E--CRVRLLLFATATPPGA----------------------------------
BVDV1        -----------SESIRVVAMTATPAGS----------V------TTTGQ-----------
HCV          -QA-E---T--AGARLVVLATATPPGS----------V----------------------
HEV          ----------------------------------------DFEHA---------------
PVY          L---V---Y--HQACKVLKVSATPVGR----------EV------EFT-----------
ChikV        ------------------------------------------------------------
```

```
DmDEAD      -----------PPDRQTLLFSATLPKE----------V------EKLARKFLR-DPVRID
TtHerA      -----------PPSRQTLLFSATLPSW----------A------KRLAERYMK-NPVLIN
MjDDEADBOX  -----------NKDKRILLFSATXPRE----------I------LNLAKKYXG-DYSFIK
StHel       -----------SNRKITGLFSATIPEE----------I------RKVVKDFIT-NYEEIE
HsDDX6      -----------PKNRQILLYSATFPLS----------V------QKFMNSHLQ-KPYEIN
PtHR18934   -LL-R---K--RPDLKLILMSATLDAE----------K---------FSDYFG-NAPVIE
lcl         -IL-K---K--RPDLKLIIMSATLDAE----------K---------FSEYFN-NAPILT
HrpA        -LL-R---R--RDDLKLIIMSATLDAE----------R---------FSAYFG-NAPVIE
HrpB        -DV-Q---SALRDDLRLLVMSATLDGE----------R---------LASLLG-EAPVLE
lcl2        -IA-R---F--RPDLKLLISSATMDAE----------K---------FSAFFD-DAPIFR
ScBDP5      -----------PKDTQLVLFSATFADA----------V------RQYAKKIVP-NANTLE
HsDDX19B    -----------PRNCQMLLFSATFEDS----------V------WKFAQKVVP-DPNVIK
BmVasA      ---M-----VETTKRQTLMFSATFPED----------I------QHLAGRFLN-NYLFVA
ScMss116p   -KNS-----KSADNIKTLLFSATLDDK----------V------QKLANNIMNKECLFLD
ApRigI      -QK-FNSA---SQLPQILGLTASVGVGNAKNIEETIEH------ICSLCSYLD-IQAIST

                    310       320       330       340       350       360
            ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII     I-PGDT-------------FKISEVFI-----------------GSSGIVF---------
TMV         -------------------YPVS-----ISTEDTL---------GDNILVY---------
HsBRR2      F-DNSF------P------VPLEQTYV--------GIT--K-AIKNQVLVF---------
HsRIG1      V-KHNLEELE-------------------------------RISD-----LF---------
HsBML       M-SFN-------R------HNLKYYVL--------PKK-P----YDSGIIY---------
HsDdx3x     V-G--------TS------ENITQKVV--------WVE-E----DSLTLVF---------
HseIF4AIII  V-KRDEL----TL------EGIKQFFV--------AVE-R----ITQAVIF---------
HsDdx10     ------------------------------------------------------------
DmVasA      I-GIVGG----AC------SDVKQTIY--------EVN-K----ADGTIVF---------
EcCsdA      --SVTT--------------DISQSYW--------TV-RK----DFAAIIF---------
ScUPF1      L-EVQYR-----I-------RGIPMMFWA--------N-------PEQIGVI---------
HHV1UL19    ----------------------------------------------GDNICIF---------
HHV1UL5     ---------------------------------------------G-FVVF---------
SidV        ------------------------------------------------------TFYKYISRR
EcRepA      L-S-----------------------------------------------------------
T4GP17      A--WN-------S------V---LY-------------D----------------------
EcRecQ      S-SFD-------R------PNIRYXLX--------EKF------GKSGIIY---------
KpPriA      L-S-----------------PAQQHVL--------DLK------DNQVILF---------
TYMV        ------------------------------------------------------------
DENV4       FP-QS-------N------SPIEDIER--------EI-PE----QGKTVWF---------
TBEV        FP-ES-------N------GAITSEER--------QI-------EGRTAWF---------
KUNV        FP-ES-------N------APISDLQT--------EI-------IGKTVWF---------
YFV         ----------------------------------------------F-------------
MeaV        ------------------------------------------------------------
PegiVA      ---AA--------------DNITEEPL--------DT-EG----TGRHLLF---------
BVDV1       --------------------PIEEFIA--------P--------KGNMLVF---------
HCV         ---VP--------------PNIEEVAL--------SN-TG----GGRHLIF---------
HEV         ---------------------------------------------VGQKLVF---------
PVY         --------------------PVK-----LIVEDTL---------GSNVLVY---------
ChikV       ----------------------------------------------------VYHKSISRR
DmDEAD      V-GREEL----TP------EGLKQYYV--------VVE-E----IGKVIIF---------
TtHerA      V-IKDEP------------VTYEEEAV--------PA--V----PDRAMVF---------
MjDDEADBOX  A-KINA-------------NIEQSYV--------EV-NE----EFYGLVF---------
StHel       A-CIGL-------------ANVEHKFV--------H--------DKGVIVF---------
HsDDX6      L-MEELT------------KGVTQYYA--------YVT-E----INQSIIF---------
PtHR18934   I-PGR-------T------FPVEIFYL--------PE-------PGDILVF---------
lcl         I-PGR-------T------FPVEILYL--------KEP------PGDILVF---------
HrpA        I-EGR-------T------YPVEIRYLP--------EA------SGSILVF---------
HrpB        S-EGR-------S------FPVEIRYL--------PRT-------GSVLVF---------
lcl2        I-PGR-------R------YPVDIFYT--------PE-------LGDILVF---------
ScBDP5      L-Q----------TNEVNVDAIKQLYM--------DCK------IGSSIIF---------
HsDDX19B    L-K----------GAITIA-----------------------------MIF---------
BmVasA      V-GIVGG----AS------TDVEQIFI--------EVT-K----GKRILVF---------
ScMss116p   T-VDKN----EAH------ERIDQSVV--------ISE---F--NYKAIIF---------
ApRigI      V-REN---LQRFM------NKPEIDVR--------LVK-RRIHNQTRTLLF---------

                    370       380       390       400       410       420
            ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII     -VASVAQCHEY---K----SYLEK-R-L-PYDMYIIHG-KVLDIDE-IL-EKVYSS-PNV
TMV         -VASYNEVDAL---S----KLLI-ERD---FKVTKVDG-RTMKVGNIEI-TTSGTP-SKK
HsBRR2      -VHSRKETGKT---A----RAIR-DM-P-Y-GFAIHHA-GMTRVDRTLV-EDLFAD-KHI
HsRIG1      -VKTR----AL--------N-----------------T-----------------DH
HsBML       -CLSRRECDTM---A----DTLQ-RDG--L-AALAYHA-GLSDSARDEV-QQKWIN-DGC
```

```
HsDdx3x     -VETKKGADSL---E----DFLY-HE-G-Y-ACTSIHG-DRSQRDREEA-LHQFRS-GKS
HseIF4AIII  -CNTKRKVDWL---T----EKMR-EAN--F-TVSSMHG-DMPQKERESI-MKEFRS-GAS
HsDdx10     ------------------------------------------------------------
DmVasA      -VETKRGADFL---A----SFLS-EK-E-F-PTTSIHG-DRLQSQREQA-LRDFKN-GSM
EcCsdA      -VRTKNATLEV---A----EALE-N--G-Y-NSAALNG-DMNQALREQT-L---KD-GRL
ScUPF1      -TPYEGQRAYI---L----QYMQM-N-------------------------SLDKD-LYI
HHV1UL19    -SSTVSFAEIV---A----RFCR-QFT---DRVLLLHS-LT-------L-GDVTTW-GQY
HHV1UL5     ------------------------------------------------------HS-KQ
SidV        C-----------------------------------------------T----------
EcRepA      ------------------------------------------------------------
T4GP17      ------------------------------------------------------------
EcRecQ      -CNSRAKVEDT---A----ARLQ-SKG--I-SAAAYHA-GLENNVRADV-QEKFQRD-DL
KpPriA      -LNRRGGTEQL---E----QALA-PL-F-P--------------------LAAVHR-GGA
TYMV        ------------------------------------------------------------
DENV4       -VPSIKAGNDI---A----NCLRK-SG---KRVIQLSR-KTFDT----E-YPKTKL-TDW
TBEV        -VPSIAKGGAI---A----RTLRQ-KG---KSVICLNS-KTFE-----------RD-EKP
KUNV        -VPSVKMGNEI---A----LCLQR-AG---KKVIQLN----------------------DW
YFV         -LPSIRAANVM---A----ASLR---G--------------------------Q-KKP
MeaV        ------------------------------------------------------------
PegiVA      -CHSKVECERT---C----AALS-ALG--V-SAVTYYR-GRE--------TE-IP--AGD
BVDV1       -VPTRNMAVEV---A----KKLK-AK-G-Y-NSGYYYS--EDP-------LRVVTS-QSP
HCV         -CHSKKKCDEL---A----AKLS-GLG--I-NAVAYYR-GLD--------VSVIPT-IGD
HEV         -TQAAKAANP----S----V----------------------------------------
PVY         -VSSYNEVDTL---A----KLLT-DKN---MMVTKVDG-RTMKHGCLEI-VTKGTS-ARP
ChikV       C-----------------------------------------------T----------
DmDEAD      -VNTKKRADRL---A----ELLR-EL-G-F-PVLSLHG-DMSQEEREKI-LEEFRS-GKS
TtHerA      -TRTKAETEEI---A----QGLL-RL-G-H-PAQALHG-DLSQGERERV-LGAFRQ-GEV
MjDDEADBOX  -CKTKRDTKEL---A----SXLR-DI-G-F-KAGAIHG-DLSQSQREKV-IRLFKQ-KKI
StHel       -VRTRNRVAKL---V----RLFD---------AIELRG-DLPQSVRNRN-IDAFRE-GEY
HsDDX6      -CNSSQRVELL---A----KKIS-QL-G-Y-SCFYIHA-KMRQEHRNRV-FHDFRN-GLC
PtHR18934   -LTGQEEIETLCEL-QERARFLGD-V-PRL-LVLPLYS-SLPSEEQAKV-FEPPPP-GVR
lcl         -LTGQEEIEAACELLRERA-K-SL-E-P-E-LILPLYG-ALPSEEQSRV-FDPAPP-GKR
HrpA        -LPGQREIER-AEWL----EKAEL-D--DL-EILPLYG-ALSAEEQVRV-FEPAPG-GKR
HrpB        -LPGVAEIRRV---Q----ERLA--E-RGV-EVLPLYG-ELSPAEQDRA-IKPAPK-GRR
lcl2        -LTGQEEIETVKEN-KERCRRLGI-R-E-L-IVLPIYA-NLPSELQAKI-FEPTPP-GAR
ScBDP5      -VATKKTANVL---Y----GKLK-SEG--H-EVSILHG-DLQTQERDRL-IDDFRE-GRS
HsDDX19B    -CHTRKTASWL---A----AELS-KEG--H-QVALLSG-EMMVE-------ERFRE-GKE
BmVasA      -VETKRNADFI---A----AMLS-EQ-Q-L-LTSSIHG-DRMQREREEA-LQNFKS-GKH
ScMss116p   -APTVKFTSFL---C----SILK-NE-D-L-PILEFHG-KITQNKRTSL-VKRFKK-DES
ApRigI      -AKTRALVSAL---K----KCME-ENPY-I-KPGVLMGTGMTLPSQKGV-LDAFKS-KDN

                      430       440       450       460       470       480
            ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII     SIIISTPYLESSVTIRNVTHIYDMGKVFVPA----------------------PF-----
TMV         HFIVATNIIENGVTLD-IDVVADFGTKVLPY----------------------LDTD--S
HsBRR2      QVLVSTATLAWGVNLP-AHTVIIKGTQVYSP-----------------------------
HsRIG1      NILIATSVADEGIDIAQ----------------------------------------------
HsBML       QVICATIAFGMGIDKPDVRFVIHASLP-----------------------------------
HsDdx3x     PILVATAVAARGLDISNVKHVINFDLP-----------------------------------
HseIF4AIII  RVLISTDVWARGLDVPQVSLIINYDLPN----------------------------------
HsDdx10     ------------------------------------------------------------
DmVasA      KVLIATSVASRGLDIKNIKHVINYDMP-----------------------------------
EcCsdA      DILIATDVAARGLDVERISLVVNYDIP-----------------------------------
ScUPF1      KVEVASVDAFQGREK---DYIILSCV------------------------------------
HHV1UL19    RVVIYTTVVTVGLSFDPLHFDGMFAY-----------------------------------
HHV1UL5     QLVVARN----------VTYVLNSQI-----------------IFSGLISFY--------
SidV        ------------------------------ATKPKPGDIILTCFRGWVKQLQ-------
EcRepA      ------------------------------------------------------------
T4GP17      ------------------------------------------------------------
EcRecQ      QIVVATVAFGXGINKPNVRFVVHFDIP-----------------------------------
KpPriA      RILIGTQMLAKGHHFPDVTLVSLLDVDGAL------------------------A-------
TYMV        ----CTISSSQGLTFCDPAIIV--N-------------------------------------
DENV4       DFVVTTDISEMGANF-RAGRVIDPRRCLKPV----------------------ILTDGPE
TBEV        DFVVTTDISEMGANL-DVSRVIDGRTNIKPE----------------------EV-D--G
KUNV        DFVVTTDISEMGANF-KASRVIDSRKSVKPT----------------------IITEGEG
YFV         DFILATDIAEMGANL-CVERVLDCRTAFKPV----------------------LVDE--G
MeaV        ------------------------------------------------------------
PegiVA      VCVCATDALSTG-YSGNFDSVTDCGLMVEEV----------------------VE--L--
BVDV1       YVIVATNAIESGVTLPDLDTVIDTGLKCEKR----------------------VRVSS--
HCV         VVVVATDALMTGYTGD-FDSVIDCNTCVTQT----------------------VDFSLDT
HEV         ------------L-------------------------------------------------
```

```
PVY          HFVVATNIIENGVTLD-IDVVVDFGLKVSPF----------------------LDID--N
ChikV        PIVVDT-------------------------STKPDPGDLVLTCFRGWVKQLQ-------
DmDEAD       RVLVATDVAARGLDIPNVDLVINYDLP---------------------------------
TtHerA       RVLVATDVAARGLDIPQVDLVVHYRLP---------------------------------
MjDDEADBOX   RILIATDVXSRGIDVNDLNCVINYHLP---------------------------------
StHel        DMLITTDVASRGLDIPLVEKVINFDAP---------------------------------
HsDDX6       RNLVCTDLFTRGIDIQAVNVVINFDFP---------------------------------
PtHR18934    KVVLATNIAETSITIDGIVYVIDSGFVKEKV----------------------YNPR--T
lcl          KVILSTNIAETSLTIDGIRYVVDSGFVKQKK----------------------YNPR--T
HrpA         KVVLATNIAETSLTIPGIRYVIDSGLAKEKR----------------------YDPR--T
HrpB         KVVLATNIAETSLTIEGVRVVVDSGLARVPR----------------------FDPAS-T
lcl2         KVVLATNIAETSLTIDGIKYVIDPGFVKQNS----------------------YNPR--T
ScBDP5       KVLITTNVLARGIDIPTVSMVVNYDLPTLA-----------------------------
HsDDX19B     KVLVTTNVCARGIDVEQVSVVINFDLPVDKD----------------------------
BmVasA       CILVATAVAARGLDIKNVDIVVNYDLP---------------------------------
ScMss116p    GILVCTDVGARGMDFPNVHEVLQIGVP---------------------------------
ApRigI       RLLIATSVADEGIDIVQCNLVVLYEYS---------------------------------

                      490       500       510       520       530       540
             ....|....|....|....|....|....|....|....|....|....|....|....|
VVNPHII      ---G-GSQEFISKSMRDQRKGRVGRV---N---PGTYVYFYDLS-YMKSIQ----------
TMV          RMLS-TTKTSINYGERIQRLGRVGRH---K---PGHAL----------------------
HsBRR2       --EK-GRWTELGALDILQMLGRAGRPQYDT---KGEGILITSHG-ELQYYLS---------
HsRIG1       ---------------------------------GSKCFLLTSN------------------
HsBML        ----------KSVEGYYQESGRAGRDG--E---ISHCLLFYTYH-DVTRLK----------
HsDdx3x      ----------SDIEEYVHRIGRTGRVG--N---LGLATSFFNER-NINITKD---------
HseIF4AIII   -----------NRELYIHRIGRSGRYG--R---KGVAINFVKND-DIRILR----------
HsDdx10      ------------------------------------------------------------
DmVasA       ----------SKIDDYVHRIGRTGRVG--N---NGRATSFFDPE-KDRAIAA---------
EcCsdA       ----------MDSESYVHRIGRTGRAG--R---AGRALLFVENR-ERRLLRN---------
ScUPF1       ---Q-AIGFLRDPRRLNVGLTRAK----------YGLVILGNPR-SLARNTLW--NHL--
HHV1UL19     ----------PDMVSVYQSLGRVRTL---R---KGELLIYMDGS-G---------------
HHV1UL5      --------------------------G---PKS--------------------------
SidV         -----------------QGLTRKG-----------------------------------
EcRepA       ------------------------------------------------------------
T4GP17       ------------------------------------------------------------
EcRecQ       ----------RNIESYYQETGRAGRDG--L---PAEAXLFYDPA-DXAWLR----------
KpPriA       ---------ERFAQLYTQVSGRAGRAG--K---QGEVILQTHH----------------
TYMV         ------------------------------------------------------------
DENV4        RVIL-AGPIPVTPASAAQRRGRIGRNPAQE---DDQYVFSGDP-LKNDEDHAH--WTEAK
TBEV         KVELTGT-RRVTTASAAQRRGRVGRQDGR-----DEYIYSGQC-DDDDSG-VQ--WKEAQ
KUNV         RVIL-GEPSAVTAASAAQRRGRTGRNPSQ-----DEYCYGGHT-NEDDSNCAH--WTEAR
YFV          RKVAI-GPLR----SAAQRRGRIGRNPNR------------------------------
MeaV         ------------------------------------------------------------
PegiVA       ------------------------------------------------------------
BVDV1        KITG-LKRMAVTVGEQAQRRGRVGRV---K---PGRYYRSQE------------------
HCV          FTIE-TTTVPQDAVSRSQRRGRTGRG---R---RGIYRFVTPGE-RPSGMFD--SSVLCE
HEV          ------------------------------------------------------------
PVY          RSIA-YNKVSVSYGERIQRLGRVGRF---K---KGVALRIGHTE-KG-------SMVATE
ChikV        ---------------------------------QKV------------------------
DmDEAD       ----------RDIEDYIHRIGRTGRAG--R---KGLAITFVTPE-DRRLLRD--IEKLYE
TtHerA       ----------DRAEAYQHRSGRTGRAG--R---GGRVVLLYGPR-ERRDVEAL--ERAVG
MjDDEADBOX   ----------QNPESYXHRIGRTGRAG--K---KGKAISIINRR-EYKKLR----------
StHel        ----------QDLRTYIHRIGRTGRMG--R---KGEAITFILNE-YWLE-----------
HsDDX6       ----------KLAETYLHRIGRSGRFG--H---LGLAINLITYD-DRFNLKSI--EEQLG
PtHR18934    GMES-LVVTPISKASAEQRAGRAGRT---G---PGKCYRLYTEE-AFEN--P--EYTVPE
lcl          GLDS-LIVVPISKASANQRAGRAGRT---G---PGKCYRLYTES-AYDKMPL---QTVPE
HrpA         GLTR-LETEPISKASADQRAGRAGRT---G---PGICYRLYSEE-DF---LA-PEFTLPE
HrpB         RLET-VR---VSQASADQRAGRAGRT---E---PGVCIRLWSEE------RL-P--AFPE
lcl2         GMES-LLVTPISKASANQRAGRAGRT---G---PGKCFRLYTAW-AYEHELE--EMTVPE
ScBDP5       --------GQADPATYIHRIGRTGRFG--R---KGVAISFVHDK-NSFNILSA--IQKYF
HsDDX19B     ----------------YLHRIGRTGRFG--K---RGLAVN---------------------
BmVasA       ----------KSIDEYVHRIGRTGRVG--N---RGKAVSFYDSD-QDLALVAD--LSKIL
ScMss116p    ----------SELANYIHRIGRTARSG--K---EGSSVLFICKD-ELPFVRE---EDAKN
ApRigI       ----------GNVTKMIQVRGRGRA----A---GSKCILVTSKT-EVVENEK--CNRYKE


             ....|.
VVNPHII      ------
TMV          ------
HsBRR2       ------
```

```
HsRIG1      ------
HsBML       ------
HsDdx3x     ------
HseIF4AIII  ------
HsDdx10     ------
DmVasA      ------
EcCsdA      ------
ScUPF1      ------
HHV1UL19    ------
HHV1UL5     ------
SidV        ------
EcRepA      ------
T4GP17      ------
EcRecQ      ------
KpPriA      ------
TYMV        ------
DENV4       M-----
TBEV        I-----
KUNV        I-----
YFV         ------
MeaV        ------
PegiVA      ------
BVDV1       ------
HCV         CYDAGC
HEV         ------
PVY         AALA--
ChikV       ------
DmDEAD      EKLEEL
TtHerA      RRFKR-
MjDDEADBOX  ------
StHel       ------
HsDDX6      TEIKPI
PtHR18934   ILRTNL
lcl         IQRVNL
HrpA        ILRTDL
HrpB        ILQADL
lcl2        IQRTNL
ScBDP5      GDIEMT
HsDDX19B    ------
BmVasA      RQADQS
ScMss116p   IVIAKQ
ApRigI      EMMNKA
```

D) Trimmed alignment of NS5Met homologues

```
                    10        20        30        40        50        60
             ....|....|....|....|....|....|....|....|....|....|....|....|
MumpsV       ------------------------------------------------------------
NegevV       --------CRSGMKTAEMFVRY-FS-------TR-EYE-SAVSIG-G----PG-GEVQYL
SARS         ---------------------------------VP-YNM-RVIHFGAGSDKGV-APGTAVL
Reovirus     SLAR-KIGDRSLVKDTAVLKHA-YY-------LR-SRQ-SVAYFGAS----A-P---E--
DENV         ETTH-HAVSRGSAKLQWFVERN-MV-------IP-E-G-RVIDLGCG----RG-GWSYYC
JEV          IVGG-HPVSRGSAKLRWLVEKG-FV-------SP-I-G-KVIDLGCG----RG-GWSYYA
YFV          VDTG-VAVSRGTAKLRWFHERG-YV-------KL-E-G-RVIDLGCG----RG-GWCYYA
TBEV         TNVG-LAVSRGTAKLAWLEERG-YA-------TL-K-G-EVVDLGCG----RG-GWSYYA
MeaV         GKTG-LSVSRGTAKLAWMEERG-YV-------EL-T-G-RVVDLGCG----RG-GWSYYA
ASFV         ------IVTNAWLKMYELLNTM-NF-------NN-TSQ-A-FCNCEL----PG-GFISAI
Baculovirus  --RP-TRRPRCWRKLSEIDKKFH-V-------CR-HVD-TFLDLCGG----PG-EFANYT
Mimivirus    -----EMITTAWIKLYEILNEF-PD-------II-PSV-KSFHLCEA----PG-AFVSAT
HsFtsj       -----SYRSRSAFKLLEVNERHQ-I-------LR-PGL-RVLDCGAA----PG-AWSQVA
EcFtsj       -----GLRSRAWFKLDEIQQSDK-L-------FK-PGM-TVVDLGAA----PG-GWSQYV
VVCapE       ---G-PLGILSNYVKTLLISMYCLD-------DS-NKR-KVLAIDFG----NG-ADLEKY
Hs2OmCap     ------------------------KPLVKDR---ELL-YFADVCAG----PG-GFSEYV
Hs2OtRNA     -KEN-GWRARSAFKLLQLDKEFQ-L-------FQ--GVTRAVDLCAA----PG-SWSQ-V
Hs2OrRNA     --EK-GYRARSSFKIIQINEKYG-F-------LE-KSK-VVIDLCAA----PG-SWCQVA
Hs2OtRNAm2   -KEQ-GYRARSAFKLLQLNDQFH-F-------LDDPNLKRVVDLCAA----PG-SWSQVL
MRM2         --VQ-NLRSRAAFKLMQIDDKYR-L-------FSKTDQ-RILDLGYA----PG-AWSQVA
AtFtsj       --R--GYVARSAFKLLQIQKQY-KL-------IK-PGS-SVLDLGCA----PG-AWLQVA
StHemolysin  GEKL-RYVSRGGLKLEKALAVFN-L-------SV-EDX-ITIDIGAS----TG-GFTDVX
ChikVPro     ----E-HRPVKGERXEWLVNKI--------------GH-HVLLVSGY----N--------
```

```
TtTrmN        PKAA-LRGSLTPVLAQALLRLA--D-------AR-PGM-RVLDPFTG----SG-TIALEA
PhMet         -----QKTGFFLDQRENRLALE-KW-------VQ-PGD-RVLDVFTY----TG-GFAIHA
VEEV          LPH-A-LVLHHNEHPQSDFSSF-VS-------KL-KGR-TVLVVGEK----L--------
PfTrm14       RVYD-HPAHLKASIANAMIEL---A-------EL-DGG-SVLDPMCG----SG-TILIEL
EcFmu         GFED-GWVTVQDASAQGCMT-W-LA-------PQ-NGE-HILDLCAA----PG-GKTTHI
23SrRNAm      --EG-GYRSRAAYKLLEINEKFK-L-------FK-PGM-RVLDLGAA----PG-SWSQ-V
TvGss1        -----QLRSRAAFKLEFLLDRY-RV-------VR-KGD-AVIEIGSS----PG-GWTQVL
Pf            -----NYRSRAAYKLIELDNKYL-F-------LK-KNK-IILDIGCY----PG-SWCQVI
Hs2OCapm      IRGV-FFLNRAAMKMANMDFVFD-RKPLVKDREA-ELL-YFADVCAG----PG-GFSEYV
PpCISIN       -----GWRARSAFKLLQIDEEFQ-I-------FE--GVKRVVDLCAA----PG-SWSQVL
Mh23SrRNAm    ------YRARSAFKLIQIDEKFN-L-------LK-PGE-IVVDLCAA----PG-GWSQVA
CoroVNSP13    -------------------------------VP-HNM-RVLHLGAGSDKGV-APGSAVL
LlHemolysin   --KL-RYVSRGGLKLEKALKEFH-L-------EI-NGK-TCLDIGSS----TG-GFTDVX
Vo23SrRNAm    KFPA-DAPSRSTLKLEEAFHTFI-R-------LA-PGM-RAVDLGAC----PG-GWTYQL

                       70        80        90       100       110       120
              ....|....|....|....|....|....|....|....|....|....|....|....|
MumpsV        ------------------------------------------------------------
NegevV        TNK-G-------------ITKTDLI-----------------------------DTYD---
SARS          RQW-L--P--TGTL----LVDSDLNDFVS---------------------D-ADSTLI---
Reovirus      IQ-----I---PSS----VRQFGYDVA-------------------------GAI---
DENV          AGL-K-----KVTE----VRGYTKGGPGHEEPVP------M--STYG-WN-IVKLMS---
JEV           ATL-K-----KVQE----VRGYTKGGAGHEEPML------M--QSYG-WN-LVSLKS---
YFV           AAQ-K-----EVSG----VKGFTLGRDGHEKPMN------V--QSLG-WN-IITFKD---
TBEV          ASR-P-----AVMS----VRAYTIG--GHEAPKM------V--TSLG-WN-LIKFRS---
MeaV          ASR-P-----HVMD----VRAYTLGVGGHEVPRI------T--ESYG-WN-IVKFKS---
ASFV          NH---------------------------------TDALED----------K-MDYNN---
Baculovirus   MSL-NP-----LCK----AYGVTLTN---C---P--------------RK-NFTTIT---
Mimivirus     HHY-M--Y---E-D----WYAQTLN---------NKALDD----------G---NNT---
HsFtsj        VQK-VN--A-PVG----FVLGVDLLHIFP--------------------LE-GATFLC---
EcFtsj        VTQ-IG--GK--G----RIIACDLLPMDP--------------------IV-GVDFLQ---
VVCapE        FYG-E--I----AL----LVATDPDADAIARGNE------RYN----KYY-KFDYIQ---
Hs2OmCap      LWR-KK--W--HAK----GFGMTLK-------------------------------GEG
Hs2OtRNA      LSQKIG--GQ--G--SGHVVAVDLQAMAP--------------------LP-GVVQIQ---
Hs2OrRNA      SKL-CP--VN--S----LIIGVDIVPMKP--------------------MP-NVITFQ---
Hs2OtRNAm2    SRK-L-D-PS--SDEDRKIVSVDLQPMSP--------------------IP-HVTTLQ---
MRM2          RQR-SS----PNS----MILGVDILPCEP--------------------PH-GVNSI----
AtFtsj        CQS-L--GP-SGGI----VVGMDIKKVKV--------------------DS-RVQTIA---
StHemolysin   LQN-G------AKL----VYAVDVGTNQL---VW----------LRQ-DD-RVRSXE---
ChikVPro      -------L--PTKR----VTWVAPLGVRG-------------------------DYT---
TtTrmN        AST-L--G--PTSP----VYAGDLDEKRLGLARE------AA-LASG-LS-WIRFLR---
PhMet         AIA-G------ADE----VIGIDKSPRAIETAKE------NAKLNGV-ED-RXKFIV---
VEEV          -------V--PGKM----VDWLSDRP---------------------------ATFR---
PfTrm14       ALR----R--YSGE----IIGIEKYRKHLIGAEM------NALAAGV-LD-KIKFIQ---
EcFmu         LEV-A--P---EAQ----VVAVDIDEQRLSRVYD------NLK----LGM-KATVKQ---
23SrRNAm      ASQKVG--AK--G----RVIAVDLQPID-----P--------------IE-GVTFIQ---
TvGss1        NSL-A-------RK----IISIDLQEXEE--------------------IA-GVRFIR---
Pf            LER-TKN-YK--N----KIIGIDKKIMDP--------------------IP-NVYFIQ---
Hs2OCapm      LWR-KK--W--HAK----GFGMTLKGPNDFKLED------F--YSAS-SE-LFEPYYGEG
PpCISIN       SRK-LYK-PLSSGERDKKIVAVDLQPMAP--------------------IE-GVIQLQ---
Mh23SrRNAm    AKL-LG--KK--G----KVIAVDLQPIRP--------------------DE-GVKTLK---
CoroVNSP13    RQW-L--P--KGTL----LVDNDLVDYVS--------------------D-ADASVL---
LlHemolysin   LQN-G------AKL----VYALDVGTNQL---AW----------IRS-DE-RVVVXE---
Vo23SrRNAm    VRR-G-------MF----VTAVDNGPMA-----E-----------MD-TG-QVEHLR---

                      130       140       150       160       170       180
              ....|....|....|....|....|....|....|....|....|....|....|....|
MumpsV        ------------T-LSTKTSV-IHK------GA--DTCALVHVDL-----EG-V---MNS
NegevV        -----GDI----TKECNLVAFR-DS----I-RN--C----FGGDVATS----SDSNHNFP
SARS          -----GDC----AT----------V----H-TA--NKWDLIISDMYDP---------END
Reovirus      -----VDL----AR----------------PS--GDYQFVYSDVDQVV-DG---HDDLS
DENV          -----GKDV---FY---------------L-PP--EKCDTLLCDIGE-SS-P-SPTVEES
JEV           -----GVDV---FY---------------K-PS--EPSDTLFCDIGE-SS-P-SPEVEEQ
YFV           -----KTDI---HR---------------L-EP--VKCDTLLCDIGE-SS-S-SSVTEGE
TBEV          -----GMDV---FS---------------M-P---HRADTVMCDIGE-SS-P-DAAVEGE
MeaV          -----RVDI---HT---------------L-PV--ERTDVIMCDVGE-SS-P-KWSVESE
ASFV          -----GDV----TIASNVKNLA-AT------RL--TPIHLYTADGGINVG-H-DYNKQEE
Baculovirus   GPDKSGDV----FDKNVVFEIS--------KCG--NACDLVLADGSVDV-NG-RENEQER
Mimivirus     -----GDI----TSSEIIKSYA-SN----K-QL--SNIDFMTGDAGIYCR-PNCLNEQET
HsFtsj        ----PADVTDPR----TSQRIL-EV-----LPG--RRADVILSDMAPNA-TG-FRDLDHD
EcFtsj        -----GDFRDEL----VMKALL-ER-----VGD--SKVQVVMSDMAPNM-SG-TPAVDIP
```

```
VVCapE          -----ETI----RSDTFVSSVR-EV----F-YF--GKFNIIDWQFAIHY-SF-HPRHYAT
Hs2OmCap        GIDGDGDI----TRPENISAFR-NF-VLD-TDR--KGVHFLMADGGFSV-EG-QENLQEI
Hs2OtRNA        -----GDI----TQLSTAKEII-QH-----F--KGCPADLVVCDGAPDV-TG-LHDVDEY
Hs2OrRNA        -----SDI----TTEDCRSKLR-GY-----M--KTWKADTVLHDGAPNV-GL-GWVQDAF
Hs2OtRNAm2      -----ADI----THPKTLARIL-KL-----F--GNEKADFVCSDGAPDV-TG-LHDLDEY
MRM2            ----QANILAKR----THDLIR-LF------EK--FPVDVIISDMYENT-SG-VSIRDHY
AtFtsj          -----ADV----LNF-PRQKIRL------P-QQ--LGFSVILSDMCHSV-SG-ITTRDAA
StHemolysin     -----QYNF---RY-----AEP-VD----F-TE--GLPSFASIDVSF-------------
ChikVPro        -----YNL------------LG-LP----A-TL--GRYDLVVINIHTPF-RI-HHYQQCV
TtTrmN          -----ADA----RH-----------P-----FF--PEVDRILANPPHG----------KE
PhMet           -----GSA----FE-----------E---K-KG--EKFDIVVLDPPAFV-QH--------
VEEV            -----ARL----------DLG-IP----G-DV--PKYDIIFVNVRTPY-KY-HHYQQCE
PfTrm14         -----GDA----TQ-----------------YV--DSVDFAISNLPY--------IGKKS
EcFmu           -----GDG----RY---------QW----C-GE--QQFDRILLDAPCSA-TG-VIRRHPD
23SrRNAm        -----GDVKDEL-EK-TLNKIL-EE-----LGKPLKKVDVVISDMAPNT-TG-NRAIDHA
TvGss1          -----CDI----FKETIFDDID-RA----LEGI--EKVDDVVSDAXAKV-SG-IPSRDHA
Pf              -----GEI----GKDSVDYKLK-EI-----L--QDKKIDIILSDAAVPC-IG-NKIDDHL
Hs2OCapm        GIDGDGDI----TRPENISAFR-NF-VLD-TDR--KGVHFLMADGGFSV-EG-QENLQEI
PpCISIN         -----GDI----TSASTAEAII-EH-----F--GGEKADLVVCDGAPDV-TG-LHDLDEY
Mh23SrRNAm      -----GDI----TSPETLDALK-EL-----L--KDEKVDVVLSDGAPNV-SG-TRSKDSF
CoroVNSP13      -----GDC----AT----------L----Y-TE--DKFDLIISDMYDGR--------ENN
LlHemolysin     -----QFNF---RN-----AVL-AD----F-EQ--GRPSFTSIDVSF-------------
Vo23SrRNAm      --------G---FK---------------P-PR--KNVDWLVCDMVE-------------

                        190       200       210       220       230       240
                ....|....|....|....|....|....|....|....|....|....|....|....|
MumpsV          MLERAQVHALLITVTVLKPGGLLILKASW-PFN--------------------RFSFLLT
NegevV          ALAKLVAWEIVLCTTVLKNGGDAYFKVLD-LLSD-------------------QMPYNIE
SARS            SKEGFFTYLCGFIKQKLALGGSIAVKITE-H----------------------SWNADLYK
Reovirus        ISSGLVESLLSSCMHATAPGGSFVVKINF-P----------------------TRPVWHYI
DENV            RTIRVLKMVEPWL-----KNNQFCIKVLN-PYM--------------------PTVIEHLE
JEV             RTLRVLEMTSDWLHR---GPREFCIKVLC-PYM--------------------PKVIEKME
YFV             RTVRVLDTVEKWLAC---GVDNFCVKVLA-PYM--------------------PDVLEKLE
TBEV            RTRKVILLMEQWKNR--NPTAACVFKVLA-PYR--------------------PEVIEALH
MeaV            RTIKILELLEKWKV---NPSADFVVKVLC-PYS--------------------VEVMERLS
ASFV            LNLKLHFGQALTGLLSLSKGGNMILKHYT-LNHA--------------------FTLSLIC
Baculovirus     LNFDLIMCETQLILICLRPGGNCVLKVFD-AFEH--------------------ETIQMLN
Mimivirus       VMAKINMGQIVCILACLSKGRSAVFKTFL-PLTE--------------------PLNISLLN
HsFtsj          RLISLCLTLLSVTPDILQPGGTFLCKTWA-GS----------------------QSRRLQR
EcFtsj          RAMYLVELALEMCRDVLAPGGSFVVKVFQ-GE----------------------GFDEYLR
VVCapE          VMNNLSELTASGGKVLITTMDGD---KLT-------------------------KTFIIHK
Hs2OmCap        LSKQLLLCQFLMALSIVRTGGHFICKTFD-LFTP--------------------FSVGLVY
Hs2OtRNA        MQAQLLLAALNIATHVLKPGGCFVAKIFR-GR----------------------DVTLLYS
Hs2OrRNA        TQSQLTLQALKLAVENLVVNGTFVTKIFR-SK----------------------DYNKLIW
Hs2OtRNAm2      VQQQLIMSALQLTACILKKGGTFVAKIFR-GR----------------------DIDMLYS
MRM2            QSIDLCDAALVTAIDLLRPLGSFVCKLYT-GE----------------------EENLFKK
AtFtsj          LSAELGMRALDLAVGVLRHGGHLVIKLLE-SED---------------------AQDFAR
StHemolysin     ---ISLNLILPALAKILVDGGQVVALVKP-Q---GREQIGKNGIVRESSIHEKVLETVTA
ChikVPro        DHAXKLQXLGGDSLRLLKPGGSLLIRAYG-YAD--------------------RTSERVIC
TtTrmN          GLFHLYWDFLRGALALLPPGGRVALLTLR--------------------------------
PhMet           ----AYFNVNFAGLNLVKDGGILVTCSCS-QHV--------------------DLQXFKDX
VEEV            DHAIKLSMLTKKACLHLNPGGTCVSIGYG-YAD--------------------RASESIIG
PfTrm14         MIPDLYMKFFNELAKVL---KRGVFITTE---A--------------------IEE-----
EcFmu           IKWLR-D-DI-AIWPHLKTGGTLVYATCS-VLP--------------------ENSLQIK
23SrRNAm        RSIELCESALKIAKEVLKKGGSFVVKIFQ-GE----------------------EFDELLK
TvGss1          VSYQIGQRVXEIAVRYLRNGGNVLLKQFQ-G----------------------XTNDFIA
Pf              NSCELTLSITHFMEQYINIGGTYIVKMYL-GS----------------------QTNNLKT
Hs2OCapm        LSKQLLLCQFLMALSIVRTGGHFICKTFD-LFTP--------------------FSVGLVY
PpCISIN         VQAQLLLAALNIATCVLKPGGSFVAKIFR-GR----------------------DTSLLYS
Mh23SrRNAm      NQLELVLAALKLALKLLKPGGRFVTKTFR-SE----------------------EEESLIW
CoroVNSP13      SKEGFFTYICGFIREKLALGGSIAIKITE-F----------------------SWNADLYE
LlHemolysin     ---ISLDLILPPLYEILEKNGEVAALIKP-QFE-GREQVGKNGIIRDPKVHQ-TIEKVLK
Vo23SrRNAm      -------------------CREAIFNLKL-P-----------------------------

                        250       260       270
                ....|....|....|....|....|....|.
MumpsV          ILWQF-FSTIR-ILRS-SY--SDPNNHEVYIIA---
NegevV          FLNNS-FDSVEI-VKL-E-T-SRAASTELHLICRGF
SARS            LM-GH-FSWWT-AFVT-NV---NASSSEAFLIGANY
Reovirus        EQKIL-PNITS-Y----------------------
DENV            RLQRK-HGG-ML-VRN-P-L-SRNSTHEMYWISNGT
```

```
JEV           VLQRR-FGG-GL-VRL-P-L-SRNSNHEMYWVSGAA
YFV           LLQRR-FGG-TV-IRN-P-L-SRNSTHEMYYVSGAR
TBEV          RFQLQ-WGG-GL-VRT-P-F-SRNSTHEMYYSTAVT
MeaV          VMQRK-WGG-GL-VRN-P-Y-SRNSTHEMYFTSRAG
ASFV          VFSHF-FEELYI-TKP-T-S-SRPTNSETYIVGKNR
Baculovirus   KFVNH-FEKWVL-YKP---PSSRPANSERYLICFNK
Mimivirus     LLSSI-FEELIF-YKP-G-A-SNGSNSEIYIVLKSY
HsFtsj        RLTEE-FQNVRI-IKP---EASRKESSEVYFLATQY
EcFtsj        EIRSL-FTKVKV-RKP---DSSRARSREVYIVATGR
VVCapE        NLPSS----MS-----P-------------------
Hs2OmCap      LLYCC-FERVC-------------------------
Hs2OtRNA      QLQVF-FSSVLC-AKP---RSSRNSSIEAFAVCQGY
Hs2OrRNA      VFQQL-FEKVEA-TKP---PASRNVSAEIFVVCKGF
Hs2OtRNAm2    QLGYL-FDKIVC-AKP---RSSRGTSLEAFIVCLGY
MRM2          RMQAV-FTNVHK-FKP---DASRDESKETYYIGLKK
AtFtsj        ICKPI-FNKAS-WLRP-KA--TRPSSREIYLICQGF
StHemolysin   FAV--G----G-LDFS-P-I--GHGNIEFLAHLEK-
ChikVPro      VLGRK-FRSSR-ALKP-PC--VTS-NTEXFFLFSNF
TtTrmN        ------------A-----------GVYPRVFVLE--
PhMet         IIAAG-A-----Y-----------TEYLKCLFLY--
VEEV          AIARQ-FKFSR-VCKP-KS--SLE-ETEVLFVFIGY
PfTrm14       ------------H------------LMVHLYVVKL-
EcFmu         AFLQR-----------P--------DGFFYAKLI--
23SrRNAm      ELRKH-FSKVKI-FKP---KASRKESAEVYIVALGF
TvGss1        IWRKN-FSSYKI-SKP-P--ASRGSSSEIYIXFFGF
Pf            YLKGM-FQLVHT-TKP---KASRNESREIYLVCKNF
Hs2OCapm      LLYCC-FERVCL-FKP---ITSRPANSERYVVCKGL
PpCISIN       QLRKF-FKKVTC-AKP---RSSRNSSIEAFVVCLGY
Mh23SrRNAm    VLKKL-FGKVKV-LKP---KASRKKSSEGFIVCLGK
CoroVNSP13    LM-QY-FSFWT-MFCT-NV---NTSSSEAFLIGINY
LlHemolysin   TATQ-G----G-LTFS-P-I-GGAGNVEFLVHLLKD
Vo23SrRNAm    ------------AKQL-Y--------EEITVHIRRK
```

E) Trimmed alignment of NS5Pol homologues

```
                10        20        30        40        50        60
          ....|....|....|....|....|....|....|....|....|....|....|....|
DENV      -EKKLGEFGKAK-SRAIWYMWLGVRYLEFEALGFL-NEDHWFSRENSYSGVEGEGLH-KL
TBEV      -EKKLGEFGV--GSRAIWYMWLGSRFLEFEALGFL-NEDHWASRESSGAGVEGISLN-YL
YFV       -EKKLSEFGKAKGSRAIWYMWLGARYLEFEALGFL-NEDHWASRENSGGGVEGIGLQ-YL
WNV       ---------KAK-SRAIWFMWLGARFLEFEALGFL-NEDHWLGRKNSGGGVEGLGLQ-KL
HCV       P------------ARLIVFPDLGVRVCEKMALYD-VVST-LPQAVMGSSYGFQYSPK-QR
MeaV      ------------GSRAIWYMWLGSRFLEFEALGFL-NEDHWASREKSGGGVEGMGLH-YL
PegiVA    P------------PRFIVFPPLDFRIAEKMILGD-IVAK----AVLGSAYLFQYTPN-QR
BVDV      R------------PRVIQYPEAKTRLAITKVXYNW-VKQQPVV---IPGYEGKTPLF-NI
PolV      K------------SRLIEASSLNDSVAMRMAFGNL-YAAFHKNPGVITGSAVGCDPD-LF
NorV      K------------KRLLWGSDLATMIRCARAFGGL-MDELKTHCVT-LPIRVGMNMN-ED
Qbeta     T------------DRCIAIEPGWNMFFQLGIGGIL-RDR--------WGIDLND-----T
IBDV      K------------TRNIWSAPSPTHLMISMITWPV-MSNSPNNVLNIEPSLYKFNPFRGG
Phi6      R------------RRTAMGGPFALNAPIMAVAQP-VRNK---------------TRL-NK
MORV3     R------------PRSIMPLNVPQQQVSA-PHTLT-ADY--INYHM---SPTSSAVI-EK
HIV1      K------------WRKLVDF-RELNKRT--------------------------------
TYMV      --------------------W---------VLGPVDNADRPPN-----------TPN-QL
SARS      ------------------------------------------------------------
HEV       -----------------FGPW--FRAIEKAILALL-PQ----------------------
AstroV    --------------------PIFSR------------------------QCGWSPFM--
PVY       K------------TRTFTAAPLDTLLGGKV-VDD-FNNQ--YSKNIECCWTVGMTKF-YG
ChikV     --------------QVIQ-------------------------------------------

                70        80        90       100       110       120
          ....|....|....|....|....|....|....|....|....|....|....|....|
DENV      GYILRDIS-K-IPGGAMYADDTAGWDTRITE-DDLHNEEKIIQQMDP--EHRQLANAIFK
TBEV      GWHLKKLS-T-LNGGLFYADDTAGWDTKVTN-ADLEDEEQILRYMEG--EHKQLATTIMQ
YFV       GYVIRDLA-A-MDGGGFYADDTAGWDTRITE-ADLDDEQEILNYMSP--HHKKLAQAVME
WNV       GYILREVG-T-RPGGRIYADDTAGWDTRITR-ADLENEAKVLELLDG--EHRRLARAIIE
HCV       VEFLVNTWKS-KKCPMGFSYDTRCFDSTVTES-DIRVEESIYQCCDLAPEARQAIRSLTE
MeaV      GWLVKDLA-E-LEGGKLYADDTAGWDTRVTN-SDLEDEEEILNHLEG--EHKKLAEAIMK
PegiVA    VKALVAAWEG-KKHPAAITVDATCFDSSIDEH-DMQVEAAIFAAAS-DD-VR--VHALC-
BVDV      FDKVRKEW-DSFNEPVAVSFDTKAWDTQVTS-KDLQLIGEIQKYYYK-KEWHKFIDTITD
PolV      WSKIPVLM-----EEKLFAFDYTGYDASLSP-AWFEALKMVLEKIGF-----GDRVDYID
```

```
NorV    GPIIF------SRYRYHYDADYSRWDSTQQR-AVLAAALEIMVKF-----EPHLAQVVAE
Qbeta   INQRRAHE-G-SVTNNLATVDLSAASDSISL-ALCELLLP---------------FEVLM
IBDV    ---IVEWI--A-EEPTWYSIDLEKGEANCTR-QHMQAAMYYILTRGW-------------
Phi6    EE------K--KEWSLCVATDVSDHDTFWPGW-LRDLICDELLNMGYAP------VKLFE
MORV3   VIPLGVYA-S-SPPNQSINIDISACDASITWDFFLSVIMAAIHE-----GMQNMIQHLSK
HIV1    --------GL-KKKKSVTVLDVGDAYFSVPLD----------------------------
TYMV    R---Q-------HSTPKIANDYTAFDQSQHGE-SVVLEALKMKRL--------IPSHLIQ
SARS    ------------------------------------------------------------
HEV     ------VA-A-ARASMVFENDFSEFDSTQNN-FSLGLECAIMVEC----G-PQWLIRLYH
AstroV  ------------GNDYFIEFDWTRYDGTIPN-EVFKAIKDFRFSCL----NRDVYNWYCE
PVY     -----LLR-R-LPENVYCDADGSQFDSSLTPY-LINAVLTIRSTYM-E---WDVGLQMLR
ChikV   -------H-F-KPGDTVLETDIASFDKSQDD-SLALTALMLLEDLGV-----------I-

                130       140       150       160       170       180
        ....|....|....|....|....|....|....|....|....|....|....|....|
DENV    LTY-QN-KVVKV-QRPTPTG-TVMDIISRKDQRGSGQVGTYGLNTFTNMEAQLVRQMEGE
TBEV    KAY--HAKVVKV-ARPSRDGGCIMDVITRRDQRGSGQVVTYALNTLTNIKVQLIRMMEGE
YFV     MTY-KN-KVVKV-LRPAPGGKAYMDVISRRDQRGSGQVVTYALNTITNLKVQLIRMAEAE
WNV     LT-----------------G------ISREDQRGSGQVVTYALNTFTNLAVQLVRMMEGE
HCV     RLY-IG-GPLTNS---------KGQNCGYRRCRASGVLTTSCGNTLTCYLKATAACRAAK
MeaV    LAY--HAKVVKV-ARPASDGGTVMDIISRRDQRGSGQVVTYALNTITNIKVQLIRMMEGE
PegiVA  RYY-VE-GPMVSP---------DGVMLGHRACRSSGVLTTSSANSXTCYIKXXAAXXRAG
BVDV    H-X-TE-VPVIT-----AD----GEVYIRNGQRGSGQPDTSAGNSXLNVLTXXYAFCEST
PolV    Y-L-NH-SHHLY------K----NKTYCVKGGMPSGCSGTSIFNSMINNLIIRTLLLKTY
NorV    DLL-SP-SVVDV-G---------DFTISINEGLPSGVPCTSQWNSIAHWLLTLCALSEVT
Qbeta   D-L-RS-PKGRL-----PD----GSVVTYEKISSMGNGYTFELESLIFASLARSVCEILD
IBDV    --V-VD-SSCLI-----------NL-QIKTYGQGSGNAATFINNHLLSTLVLDQWNLMRQ
Phi6    ----LK-LPVYVG-EQG---LGDPSNPDLEVGLSSGQGATDLMGTLLMSITYLVMQLDHT
MORV3   LYK-RG-FSYRV-NDSF-----GNDFTHMTTTFPSGSTATSTEHTANNSTMMETFLTVWG
HIV1    ------------------NNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFKKQNP
TYMV    LHVH------------------------PLTCMRLTGEPGTYDDNTDYNLAVIYSQYD---
SARS    ----------------------GSLYVKPGGTSSGDATTAYANSVFNICQAVTANVNAL
HEV     LIR----WILQA-P---------KESLRGFWKKHSGEPGTLLWNTVWNMAVITHCY----
AstroV  N---FR-RYVML-----PS----GEVTIQDRGNPSGQISTTMDNNICNVFFQAFEFAYLN
PVY     NLYTVY-TPISTP---------DGTIVKKFRGNNSGQPSTVVDNSLMVVLAMHYALIKEC
ChikV   --------SCHL-----PT----GTRFKFGAMMKSGMFLTLFVNTLLNITIASRVLED--

                190       200       210       220       230       240
        ....|....|....|....|....|....|....|....|....|....|....|....|
DENV    GVLTKADLENP-HLLEKKITQWLETKGVERLKRMAISGDDCVVKPI--D-------DRFA
TBEV    GVIEAADAHR-LR-----VERWLKEHGEERLGRMLVSGDDCVVRPL--D-------DRFG
YFV     MVIHHQHVQDCDESVLTRLEAWLTEHGCDRLKRMAVSGDDCVVRPI--D-------DRFG
WNV     GVIGPDDVEKLTKGKGPKVRTWLSENGEERLSRMAVSGDDCVVKPL--D-------DRFA
HCV     L---------------------------QDCTMLVNGDDLVVICE--SAGTQEDAAALR
MeaV    GVIGPADMTE-PRI--IRVERWLERHGEERLGRLLVSGDDCVVKPI--D-------DRFA
PegiVA  V-----------------------------KEPTFXIAGDDCLIIYE--NDGTDPCAR-LK
BVDV    GV--------------------------NRVARIHVCGDDGFLITEK---------KFAN
PolV    KG-------------------------LDHLKMIAYGDDVIASY--PHE------VDAS
NorV    NL---------------------SPDIIQANSLFSFYGDDEIVSTDI-K-------LDPE
Qbeta   ---------------------------LDSSEVTVYGDDIILPS-----------CAVP
IBDV    -------------------------------KALVYADNIYIVH---R--------DSE
Phi6    APHLNSR----------KDMPSACRFLDHEEIRQISKSDDAMLGWT--KGR-------AL
MORV3   PEH--------------DPDVLRLMKSLTIQRNYVCQGDDGLMIIDG-Q-------NDLE
HIV1    DI-------------------------------VIYQYMDDLYVGSD--------------
TYMV    ----------------------------VGSCPIMVSGDDSLIDHPL-P-------TRHD
SARS    ------------------------------HFSMMILSDDAVVCYN--------------
HEV     ---------------------------DLQVAAFKGDDSIVLCSE--------------
AstroV  TELDSDEL-E----------N--------DKYDSLIYGDDRLTTT--------------
PVY     ----------------------------STCVFFVNGDDLLIAVN--PEKE--------
ChikV   ----------------------------TKSACAAFIGDDNIIHGVV-S-------DELM

                250       260       270       280       290       300
        ....|....|....|....|....|....|....|....|....|....|....|....|
DENV    NA-LLALNDMGKVRKDIPQWQPSKGWHDWQQVPFCSHHFHELIMKDGRKLVVPC-RPQDE
TBEV    KA-LYFLNDMAKTRKDIGEWEHSAGFSSWEEVPFCSHHFHELVMKDGRTLVVPC-RDQDE
YFV     LA-LH-LNAMSKVRKDISEWQPSKGWNDWENVPFCSHHFHELQLKDGRRIVVPC-REQDE
WNV     TS-LHFLNAMSKVRKDIQEWKPSTGWYDWQQVPFCSNHFTELIMKDGRTLVTPC-RGQDE
HCV     A-FTEAMTRYSAPPG------PPQPEYDLELITSCSSNVSVAHDASGKRVYYLT-RDPTT
MeaV    EA-VHFLNDMSKTRKDIGEWSPSVGYTNWEEVPFCSHHFHRLVMKDGRELIVPC-RDQDE
PegiVA  A----ALADY-----D-----PVK-HASLDTAECCSAYLAVA----GKKRWWLS-TDMRK
BVDV    KG-XQILHEAGKPQKITEG-EKXKVAYRFEDIEFCSHTPVPVRWSDNTSSHXAGR-DTAV
```

```
PolV    LL-AQSGKDYGLTMTPADKSATFETV-TWENVTFLKRFFRADE--KYPFLIHPV-MPMKE
NorV    KL-TAKLKEYGLKPTRP----PLVISEDLNGLTFLRRTVTRDP-----AGWFGKL-EQSS
Qbeta   AL-REVFKYVGFTTNTK-------TFSEGPFRESCGKHYYSGVDVTPFYIRHRIV-SPAD
IBDV    EF-KSIEDKLGINFKYLSGGVEPEQSSPTVELDLLGWSATYSK-----GIYVPV-LDKER
Phi6    V---------YMKISY-------------HGGAFLGDILLYDSRREPG-SAIFV-GNINS
MORV3   LI-SKYGEEFGWKYDIA----------YDGTAEYLKLYFIFG----------RIP-NLSR
HIV1    ------------------------------------------------------------
TYMV    -----VLK---------------------------------L------------------
SARS    ------------------------------------------------------------
HEV     ------------------------------------------------------------
AstroV  --------------------V--KVSNEINGLTFCGFT-----------LFVP-------
PVY     ---SQHFSDLGLNYD-------SS----KEELWFMSH------------MYVPK-LEEER
ChikV   AA----MNMEVKIIDAV------------QKAYFCGGFIHDIVTGLGKPLA--GD-EQDE

              310
        ....|....|....
DENV    LI----GRARISQG
TBEV    LV----GRARISPG
YFV     LI----GRGRVSPG
WNV     LV----GRARISP-
HCV     PL----ARAAWETA
MeaV    LI----GRARVSPG
PegiVA  PL----ARASSE--
BVDV    IL----SKXATRLD
PolV    IH----ESIRWTKD
NorV    IL--D---------
Qbeta   LI----LVLNNLYR
IBDV    LF----CSAAYPKG
Phi6    M-----NNQF----
MORV3   HPRAN-SAEE----
HIV1    --------------
TYMV    --PLF---------
SARS    --------------
HEV     ----G---------
AstroV  --------------
PVY     I-------------
ChikV   DR----RRALADE-
```

File S2 – Phylogenetic trees of Flavivirus proteins homologues in Newick format

A) Protein E homologues – Cell fusion proteins class II

(CeEFF1:1.460115,(RubellaV:1.580933,(ChikV:0.1775973,SFV:0.2791707):1.573846):0.8505151,(BfBRAFLDRAFT:0.4766951,((TBEV:0.2305624,MeaV:0.2923539):0.5212072,((DENV1:0.3143197,WNV:0.5056887):0.2164969,YFV:0.4264118):0.09781294):1.17196):0.5521507);

B) NS3Pro homologues – PA proteases

(SARS:1.575817,PolV:1.087531,((((ChibaV:1.249973,(((BtChymotrypsin:0.3427255,(RrTrypsin:0.3716837,HsThrombin:0.7423175):0.1213948):0.1675317,SgTrypsin:0.7215007):0.6889323,(EcAHP:0.258818,NgIgA1SP:0.8355711):0.9301418):0.2544907):0.4097438,((EcDegs:0.4402425,((HsHtra2:0.5011075,AtDeg5:0.5058617):0.1529979,TmHtra:0.3276787):0.2165272):0.6384487,SaSpla:0.6628864):0.3568464):0.3188891,((((DENV:0.1331282,WNV:0.1746989,YFV:0.2928628):0.1955023,(MeaV:0.1482132,TBEV:0.2742729):0.2593673):0.832589,(HCV:0.4807529,PegiVA:0.4992751):0.8231502):0.5787507,SinV:1.30398):0.3778801,HAV:0.9482968):0.3808645,TEV:0.9723378):0.2265748);

C) NS3Hel homologues – SF1/SF2 helicases

(VVNPHII:2.21305,(((PtHR18934:0.1168464,(lcl:0.2023179,lcl2:0.2616023):0.05190452):0.3052608,HrpA:0.04740705):0.309296,HrpB:0.4322393):1.106223,(((TMV:0.2539305,PVY:0.2807512):1.392059,(((((HsBRR2:1.288411,((HsRIG1:0.2880521,ApRigI:0.1922275):1.272879,EcRepA:4.734329):1.070742):0.613683,(HsBML:0.7225154,EcRecQ:0.628395):1.970617):0.3289334,((((((HsDdx3x:0.4183594,(DmVasA:0.2651189,BmVasA:0.3313262):0.3615112):0.4965892,((EcCsdA:0.58857254,(MjDDEADBOX:0.5660469,StHel:0.9415699):0.1366844):0.1158066,TtHerA:0.732202):0.192605,ScMss116p:1.3904):0.1213705,((HseIF4AIII:0.4952791,(ScBDP5:0.3576733,HsDDX19B:0.5112569):0.6441769):0.09747424,HsDDX6:0.9505962):0.2464013):0.1043519,DmDEAD:0.0474572):0.2484937,HsDdx10:0.9907007):0.5319457):0.371298,KpPriA:2.305292):0.404488,ScUPF1:2.733074,HHV1UL19:2.163068,HHV1UL5:1.809946,(SidV:0.2403329,ChikV:0.1752967):1.688481,T4GP17:4.077006,(TYMV:1.647808,HEV:1.252906):1.065153):0.249964,(((DENV4:0.2819152,KUNV:0.1245947):0.2341034,(TBEV:0.1748888,MeaV:0.2922182):0.2982291,YFV:0.4253623):1.96314,((PegiVA:0.7636222,HCV:0.2610942):1.635553,BVDV1:0.9847205):0.5556438):0.2411252):0.48226);

D) NS5Met homologues – Ftsj-like methyltransferases

(MumpsV:0.6996368,((((DENV:0.1682111,JEV:0.1619898):0.111694,YFV:0.2347749):0.1365287,(TBEV:0.2437508,MeaV:0.1595617):0.09871555):0.8234756,Vo23SrRNAm:0.7820906):0.3607747,(NegevV:0.7744162,(((ASFV:0.4526622,Mimivirus:0.4581867):0.4846836,Baculovirus:0.5647551):0.1549755,(Hs2OmCap:0.003791735,Hs2OCapm:0.003950054):0.6046178):0.3477147):0.2019818,(((SARS:0.1237352,CoroVNSP13:0.0801189):0.8343802,Reovirus:1.072845):0.4349217,(HsFtsj:0.3998248,MRM2:0.5196736):0.1496911,(EcFtsj:0.3839636,23SrRNAm:

0.03818875):0.09840878,((((Hs2OtRNA:0.2107124,PpCISIN:0.02461776):0.1127614,Hs2OtRNAm2:0.2098716):0.2673496,Hs2OrRNA:0.4593706):0.09184749,Mh23SrRNAm:0.146827):0.1157771,AtFtsj:0.5641959,(TvGss1:0.5569959,Pf:0.596015):0.1441094):0.1586821,((VVCapE:1.422596,(((TtTrmN:0.4781312,PfTrm14:0.8036866):0.3590191,PhMet:0.7641935):0.4031617,EcFmu:0.8940841):0.3073212):0.351583,(StHemolysin:0.2356279,LlHemolysin:0.1006174):0.9932651,(ChikVPro:0.4195942,VEEV:0.5076508):0.7923391):0.2334819);

E) NS5Pol homologues – viral RNA-dapandent RNA polymerases

((NorV:0.9850886999999999,Qbeta:1.4913910000000001,IBDV:1.001757,(Phi6:1.060286,HIV1:1.246267):0.5325600000000001,MORV3:1.142357,SARS:0.5570598000000002,AstroV:0.7700803,PVY:0.7186563000000001,(BVDV:0.8134816000000002,((HCV:0.42063419999999985,PegiVA:0.3113287):0.5929465999999999,((TYMV:0.6714199000000001,(HEV:0.7012287000000001,ChikV:0.9577031000000003):0.2960191999999999):0.34889749999999964,(DENV:0.15700609999999982,(((TBEV:0.13745619999999992,MeaV:0.0845408299999999):0.09729925000000028,YFV:0.17346850000000025):0.07793996999999964,WNV:0.1600777):0.06088041000000022):0.6060611000000002):0.4710502999999999):0.17446929999999972):0.37028060000000007):0.38436065,PolV:0.7687213000000002);

7. CURRICULUM VITAE

Personal data:

Name:	**Jiří Černý**
Address:	Kout na Šumavě 70, 345 02, Czech Republic
Citizenship:	Czech Republic
Date and place of birth:	15. 9. 1984, Klatovy
Contact phone number:	+420 603 978 071
E-mail:	cerny@paru.cas.cz (academic), JiriCernyPost@seznam.cz (personal)

Education:

- 2009 - now: Ph. D. student at the University of South Bohemia, field of study: Cellular and molecular biology and genetics, Ph. D. thesis topic: Structural evolution of flaviviral genes, supervisors: Prof. RNDr. Libor Grubhoffer, CSc., Doc. RNDr. Daniel Růžek, Ph.D.)
- 2009: Master degree at Charles University in Prague, supervisor: RNDr. Martin Pospíšek, Ph.D.)
- 2007: Bachelor degree at Charles University in Prague, supervisor: RNDr. Martin Pospíšek, Ph.D.)

Scientific experiences:

- 2014 – now: member of the Laboratory of Arbovirology (PI: Daniel Růžek), joint research unit of Biology Centre, Academy of Sciences of the Czech Republic and Veterinary Research Institute
- 2009 - now: member of the Laboratory of Molecular Ecology of Vectors and Pathogens (PI: Libor Grubhoffer), joint research unit of Biology Centre, Academy of Sciences of the Czech Republic and Faculty of Science, University of South Bohemia
- 2005 - 2009: member of the Laboratory of RNA Biochemistry (PI: Martin Pospíšek), Faculty of Science, Charles University

Main research interests:

- Structural evolution of viral proteins
- Structure-function relationship of viral polymerases
- Arbovirus ecology and epidemiology in polar areas

Publications:

Published:

- Černý J.,Černá Bolfíková B., Vadés J. V., Grubhoffer L., Růžek D. (2014): Evolution of Tertiary Structure of Viral RNA Dependent Polymerases, PLoS One, 9(5), doi: 10.1371/journal.pone.0096070
- Petra Formanová; Jiří Černý; Barbora Černá Bolfíková; James J. Valdés; Irina Kozlova; Yuri Dzhioev; Daniel Růžek: Full genome sequences and molecular characterization of tick-borne encephalitis virus strains isolated from human patients. Ticks and Tick-Borne Diseases, Ticks and Tick Borne Diseases, 2015, 6(1):38-46. doi:10.1016/j.ttbdis.2014.09.002
- Luděk Eyer, James J. Valdés, Victor A. Gil, Radim Nencka, H.Hřebabecký, Michal Šála, Jiří Salát, Jiří Černý, Martin Palus, Erik De Clercq, and Daniel Růžek. 2015. Nucleoside inhibitors of tick-borne encephalitis virus. Antimicrobial Agents and Chemotherapy 06/2015; DOI:10.1128/AAC.00807-15
- Jiří Černý, Barbora Černá Bolfíková, Paolo M. de A. Zanotto, Libor Grubhoffer, Daniel Růžek: A deep phylogeny of viral and cellular right-hand polymerases. Infection, Genetics and Evolution, 36:275-286. doi: 10.1016/j.meegid.2015.09.026.
- Jana Elsterová; Jiří Černý; Radek Šíma; Jana Müllerová; Steven Coulsen; Erlend Lorentzen; Libor Grubhoffer: Tick-borne pathogens on Jan Mayen and Svalbard. Polar Research, 34, 27466, http://dx.doi.org/10.3402/polar.v34.27466

Submitted:

- Jiří Černý, Martin Selinger, Martin Palus, Zuzana Vavrušková, Hana Tykalová, Lesley Bell-Sakyi, Libor Grubhoffer, Daniel Růžek: Expression of a second open reading frame present in the genome of tick-borne encephalitis virus strain Neudoerfl is not detectable in infected cells. under revision process in Virus Genes

- Jiří Černý, Barbora Černá Bolfíková, Libor Grubhoffer, Daniel Růžek: Genomes of viruses classified in genus Flavivirus (family *Flaviviridae*) evolved via multiple recombination events. under revision process in BMC Evolutionary Biology

Grants and awards:

- Award for excellent master thesis at the Department of Genetics and Microbiology, Faculty of Science, Charles University in 2009
- Co-investigator of the grant of the Grant Agency of the Charles University
- Principal investigator of the grant of the Grant Agency of the University of South Bohemia

Research internships:

- 2008 - 2009: RNA Research Group (PI: Henrik Nielsen), Department of Cellular and Molecular Medicine, Faculty of Health Sciences, University in Copenhagen (4 months)
- 2013: Institute of Virology (PI: Manfred Weidmann), University of Medicine in Gottingen (1 month)
- 2013: Institute of Biochemistry (PI: Rolf Hilgenfeld), University of Lubeck (3 months)

Conferences:

Jiří Černý attended numerous Czech and international conferences where he presented results of his research. Here are mentioned only a few of the most important conferences: Hot Topics in Microbiology, 2015, Štrbské pleso, Slovensko (oral presentation); VIII International Conference on Ticks and Tick-borne Pathogens, 2014, Cape Town, South Africa (oral presentation); 5th European Congress of Virology, 2013, Lyon, France (poster presentation); VII International Conference on Ticks and Tick-borne Pathogens, 2011, Zaragoza, Spain (poster presentation); Arbo-zoonet annual meeting 2011, Rabat, Morocco (poster presentation); 34th FEBS Congress, 2009, Prague, CZ (poster presentation).

Skills:

- classical methods of gene engineering (DNA, recombinant DNA techniques, sequencing, biophysical and biochemical analysis of DNA molecules)
- classical microbiological methods (sterile work, cell culture cultivation, cell culture analysis)
- work with RNA(RNA isolation, reverse transcription, *in vitro* transcription, real time PCR, catalytic RNA experiments, radioactive and "cold" labeling of RNA molecules)
- protein handling (protein isolation, enzymatic activity measurements, western blotting, protein crystallization)
- basic immunological methods (ELISA, polyclonal antibodies preparation)
- next generation (454, Roche) sequencing
- work with radioisotopes
- work with infectious material (BSL2, BSL 3/4 course is planned for November 2014)
- user skills in work with bioinformatics databases
- computational processing of biological sequences
- basic skills in work with phylogenetic programs
- protein structure modelling and model evaluation

Teaching experiences:

- 2009 - 2011: participation on biochemistry laboratory course at the Faculty of Science, University of South Bohemia
- 2014-now: participation on virology lectures at the Faculty of Science, University of South Bohemia

cerny@paru.cas.cz

Molecular Evolution of Flaviviral Genes

Ph.D. Thesis Series, 2015, No. 9

Printed in the Czech Republic by Typodesign
Edition of 20 copies

University of South Bohemia in České Budějovice
Faculty of Science
Branišovská 1760
CZ-37005 České Budějovice, Czech Republic

Phone: +420 387776 201
www.prf.jcu.cz, e-mail: sekret-fpr@prf.jcu.cz