DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
359

MIKK PUUSTUSMAA

On the origin of papillomavirus proteins

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
359

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
359

MIKK PUUSTUSMAA

On the origin of papillomavirus proteins

UNIVERSITY OF TARTU
Press

Institute of Molecular and Cell Biology, University of Tartu, Estonia

This dissertation is accepted for the commencement of the degree of Doctor of Philosophy in Gene technology on June 28, 2019 by the Council of the Institute of Molecular Cell Biology, University of Tartu.

Supervisor: Aare Abroi, PhD
Institute of Technology, University of Tartu, Tartu, Estonia

Prof. Maido Remm, PhD
Chair of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

Reviewer: Prof. Juhan Sedman, PhD
Chair of General and Microbial Biochemistry, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

Opponent: Dr Andrew E. Firth, PhD
Department of Pathology, Division of Virology, University of Cambridge, Cambridge, United Kingdom

Commencement: Room No. 105, 23B Riia St., Tartu, on August 28, 2019, at 14:15 pm.

The publication of this dissertation is granted by the Institute of Molecular and Cell Biology at the University of Tartu.

This research was funded by ETF8812 during the years 2012–2014. The development of the cRegions webpage was supported by the European Regional Development Fund through the Research Internationalization Programme (ELIXIR).

ISSN 1024-6479
ISBN 978-9949-03-136-8 (print)
ISBN 978-9949-03-137-5 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The current thesis is based on the following original publications, referred to in the text by Roman numerals (Ref. I to Ref. III):

I **Puustusmaa M.***, Kirsip H.*, Gaston K., Abroi A. 2017. The enigmatic origin of papillomavirus protein domains. Viruses 9. DOI: 10.3390/v9090240.

II **Puustusmaa M.**, Abroi A. 2016. Conservation of the E8 CDS of the E8^E2 protein among mammalian papillomaviruses. J. Gen. Virol 97:2333–2345. DOI: 10.1099/jgv.0.000526.

III **Puustusmaa M.,** Abroi A. 2019. cRegions – a tool for detecting conserved cis-elements in multiple sequence alignment of diverged coding sequences. PeerJ. 2019 Jan 10;6:e6176. doi: 10.7717/peerj.6176.

My contributions to the listed publications were as follows:

Ref. I Performed the HMM search on bacterial data and participated in the writing of the manuscript.

Ref. II Performed the analysis, wrote the manuscript and designed the algorithm used in the publication.

Ref. III Developed the cRegions software including the web application, performed the analysis and wrote the manuscript.

LIST OF ABBREVIATIONS

BLAST	The basic local alignment search tool
CDS	Protein-coding sequence
H2V	Host to virus gene transfer
HGT	Horizontal gene transfer
HMM	Hidden Markov model
ICTV	The International Committee on Taxonomy of Viruses
MYA	Million years ago
NCBI	The National Center for Biotechnology Information
ORF	Open reading frame
PaVE	The Papillomavirus Episteme (PaVE) is a resource for papillomaviruses' sequences, annotations, and analysis.
PDB	Protein Data Bank
Pfam	Pfam resource is a collection of protein domain families, each represented by multiple sequence alignments and hidden Markov models.
Profile-HMM	A variant of hidden Markov model used for representing a profile of a multiple sequence alignment.
PVs	Papillomaviruses
SCOP	Structural Classification of Proteins
SF	Superfamily (SCOP hierarchical level)
SUPERFAMILY	SUPERFAMILY is a database of structural and functional annotation for proteins based on a collection of hidden Markov models.
TOL	Tree of life
UniProt	The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data.
UniProtKB	The UniProt Knowledgebase is the central database of core information and annotations on proteins.
V2H	Virus to host gene transfer

INTRODUCTION

Viruses are obligatory intracellular parasites harbouring enormous genetic and biological diversity. Viruses are the most abundant biological entities on Earth. Viruses have captured our interest due to their association with many diseases and their importance in our environment and to our economy. However, despite decades of research, the exact origin of viruses is still a mystery.

Currently, three main scenarios exist how viruses might have emerged: the virus-first hypothesis, the reduction hypothesis, and the escape hypothesis. The last two scenarios have one important implication – most of the genes found in viruses should have their distant homologs in cellular genomes. However, the similarity between homologous sequences may have decreased to the point where the homology is not detectable with pairwise sequence comparison methods like BLAST, especially in the case of viruses due to their high mutation rate. Fortunately, profile hidden Markov models (profile-HMMs) combined with structural information of proteins may allow us to overcome the limitations of pairwise sequence comparison methods in distant homology detection.

Still, not all genes found in viruses have homologs in cellular organisms. Some of the protein-coding sequences originate *de novo*, i.e., the genesis of these sequences take place in viruses. One of the mechanisms how *de novo* genes can emerge is overprinting – mutations lead to a new protein-coding gene overlapping the ancestral gene. Overlapping genes have been described in many viruses. In addition, protein-coding genes of viruses often contain various non-coding embedded elements including internal promoters, viral packaging signals, subgenomic promoters, and splice sites. In order to fully understand the molecular biology and functioning of a virus, we need to be able to identify these embedded elements.

In the current thesis, papillomaviruses (PVs) are used as an example to study the potential origin of a viral family. PVs infect many mammalian species, but also birds, turtles, snakes, and fish. PVs have been of interest due to their association with various cancers. Oncogenic human papillomaviruses (HPVs) are responsible for almost all cases of cervical and anal cancers. A typical PV genome encodes eight proteins on average. It has been estimated that PV protein-coding genes evolve 5–10 times faster compared to their mammalian host nuclear protein-coding sequences, confirming the need to use more sensitive approaches to detect distant homologs in other organisms. In this thesis, profile-HMMs from Pfam and SUPERFAMILY resources were used to detect distant homologs to PV protein domains in cellular organisms and other viruses.

In addition, the existence of dual-coding regions and other embedded elements in papillomaviruses were studied. In this thesis, over 300 PV genomes were analysed *in silico* to detect an embedded E8 CDS inside the E1 protein-coding gene. Also, a web tool called cRegions was developed to detect dual-coding regions and other embedded elements in protein-coding genes of viruses.

1. REVIEW OF THE LITERATURE

1.1. Virosphere

Viruses are the most abundant entities on Earth. It is estimated that the total number of viral particles is about 10^{31} (Cobián Güemes et al., 2016) which is an order of magnitude higher than prokaryotic cells (Whitman, Coleman, & Wiebe, 1998). Viruses are obliged to invade hosts and parasitize their subcellular machinery. Viruses are often referred to as pseudo-living entities that are borderline between inanimate and living matter. Nevertheless, they play a major role in the marine and terrestrial ecosystems. For instance, oceanic viruses are the major pathogens of planktonic organisms (a crucial source of food to many large aquatic organisms) and thus, a fundamental factor in nutrient and energy cycle (Suttle, 2005).

For decades, scientists have been limited to studying viruses which are easy to work with (e.g. M13, T7, ΦX174 bacteriophage), have a major impact on human health (e.g. HPV, HIV) or cause diseases in animals or plants of economic value (e.g. TMV). Fortunately, metagenomic studies have revealed us the stunning world of diverse viral genes and genomes (Breitbart et al., 2004; Chen, Suttle, & Short, 1996; Cobián Güemes et al., 2016; A. I. Culley, Lang, & Suttle, 2003; Jameson, Mann, Joint, Sambles, & Mühling, 2011; Labonté & Suttle, 2013; Li et al., 2015; Rohwer, 2003; S. M. Short & Suttle, 2002). Culture-independent techniques like shotgun sequencing of marine and terrestrial environments have shown that we are just scraping the surface of viral life (Suttle, 2005). A study focused on the analysis of marine sediment demonstrated that three-quarters of the resulting sequences were not related to anything previously reported (Breitbart et al., 2004). It should be noted that marine sediments are one of the largest biotopes in the world and 97% of viruses live in soil and sediments (Cobián Güemes et al., 2016; Whitman et al., 1998). Even today, the majority of sequences acquired in metagenomic studies of viruses do not have homologs in databases (Gregory et al., 2019).

Bacteriophages have been extensively studied for decades and bacteriophages with DNA genomes are thought to represent the majority of marine viruses (Steward et al., 2013). However, this claim is rivaled by some studies, showing that the abundance of RNA viruses equals or even exceeds that of DNA viruses in samples of coastal seawater (Steward et al., 2013). RNA viruses in the marine environment are mainly composed of positive-sense single-stranded RNA ((+)ssRNA) and double-stranded RNA (dsRNA) viruses with an apparent predominance of viruses that infect eukaryotes (A. Culley, 2018; Gregory et al., 2019). Still, the diversity and abundance of RNA viruses remain largely unknown (Gregory et al., 2019). Even simple flaws in commonly used methods affect our assessment of viral diversity by excluding some of the viral subgroups, like the case with non-tailed double-stranded DNA (dsDNA) viruses (Kauffman et al., 2018). Thus, we have little knowledge about viral diversity in

different environments and there is immense information still to be discovered about viruses.

Nevertheless, even the little we know, the diversity of viruses is staggering compared to cellular organisms. Viruses use different replication strategies and their genomes could be either DNA or RNA, single-stranded or double-stranded, linear or circular. Also, their genome size varies tremendously, from a tiny 1759 nucleotide genome of Porcine circovirus (excluding viroids and satellites) (Meehan, Creelan, McNulty, & Todd, 1997; Tischer, Gelderblom, Vettermann, & Koch, 1982) to 2.47 Mb genome of *Pandoravirus salinus* (Philippe et al., 2013). Also, virion size differs hugely between viruses. A virion of a Porcine circovirus is about 17 nm in diameter (Tischer et al., 1982), an order of magnitude smaller than *Pithovirus sibericum*, which is approximately 1.5 μm in length and 0.5 μm in diameter (Legendre et al., 2014). The virion of *Pithovirus sibericum* is bigger than the smallest free-living eukaryote *Ostreococcus tauri* (Courties et al., 1994) and almost as large as a typical prokaryotic cell, reducing the gap in size between viruses and cellular organisms. In conclusion, virosphere is a complex and diverse world. This makes the taxonomy of viruses a crucial part of the discipline of virology, helping us to make the world of viruses comprehensible.

1.1.1. Taxonomy of viruses

Nature is a continuum in which adjacent elements are similar, but the extremes are quite distinct. The purpose of taxonomy is to draw boundaries within this continuum – an artificial task, but necessary nevertheless. Viruses are physical entities, whereas taxa are abstract concepts that facilitate communication among virologists and between other stakeholders (investors, government regulators, and farmers).

Viruses were historically characterised by their ability to pass through filters that retained most of the bacteria. Dimitri Ivanofsky (1864–1920), commissioned by the Russian Department of Agriculture to investigate the cause of a tobacco disease on plantations in Ukraine, reported to the Academy of Sciences on February 12, 1892; “The sap of leaves infected with tobacco mosaic disease retains its infectious properties even after filtration through Chamberland filter candles” (Knipe, 2013). However, Martinus Willem Beijerinck (1851–1931) was the first to call these incitants of tobacco a “virus” in 1898 (Knipe, 2013). Since then, the number of different viruses has grown tremendously and there have been many efforts to create a unified taxonomy of viruses. One of the first was the Baltimore classification (Baltimore, 1971), which still co-exists with the International Committee on Taxonomy of Viruses (ICTV).

1.1.1.1. Baltimore classification

David Baltimore developed a virus classification scheme in the early 1970s, which grouped viruses into classes, depending on the nature of nucleic acid packaged in virions (Baltimore, 1971). The initial publication defined six different classes. Later, the classification has been extended by adding a seventh class. Baltimore classification contains the following classes:

- Class I: Double-stranded DNA (**dsDNA**) viruses (e.g., *Pandoravirus salinus*, *Pithovirus sibericum*, *Papillomaviridae* family, *Polyomaviridae* family, *Herpesviridae* family)
- Class II: Single-stranded DNA (**ssDNA**) viruses DNA (e.g., Porcine circovirus, *Parvoviridae* family, *Geminiviridae* family)
- Class III: Double-stranded RNA (**dsRNA**) viruses (e.g., *Reoviridae* family)
- Class IV: Positive-sense single-stranded RNA [**(+)ssRNA**] viruses (e.g., *Alphavirus* genus)
- Class V: Negative-sense single-stranded RNA [**(–)ssRNA**] viruses (e.g., Influenza Virus)
- Class VI: Positive-sense single-stranded RNA reverse transcribing (**ssRNA-RT**) viruses with DNA intermediate in life-cycle (e.g., *Retroviridae* family)
- Class VII: Double-stranded DNA reverse transcribing (**dsDNA-RT**) viruses with RNA intermediate in life-cycle (e.g., Hepatitis B virus)

1.1.1.2. ICTV taxonomy

Nowadays, the classification of viruses is handled by the ICTV. It is solely responsible for naming viruses and classifying them into a taxon. The lowest taxonomic rank is species, defined as "a monophyletic group of viruses whose properties can be distinguished from those of other species by multiple criteria – virion morphology, replication strategy, genome type, host range, pathogenicity and epidemiology (Peter Simmonds et al., 2017). The majority of viral species are assigned to a genus and genera in turn into a family. Relatively few families are assigned to an order (Peter Simmonds et al., 2017). Current ICTV release (2018b) includes 1 realm, 14 orders, 150 families, 1019 genera and 5560 species [https://talk.ictvonline.org/taxonomy/p/taxonomy_releases, 12.04.2019]. A realm is the highest taxonomic rank established by the ICTV. To date, only *Riboviria* is described at this rank
[https://talk.ictvonline.org/ictv/proposals/2017.006G.A.v3.Riboviria.zip, 12.04.19].

In recent years, metagenomic data has changed our view on virus diversity and the way we classify viruses (Peter Simmonds et al., 2017). Many metagenomic studies have exposed the "missing" diversity of viruses and even increased the number of viral genes many times over (Brum et al., 2015; Paez-Espino et al., 2016; Roossinck, 2012; Steward et al., 2013). For example, a study that assessed viral community patterns from 43 Tara Oceans expedition samples (collected from different seas and oceans around the world) showed

that only a tiny fraction, 39 out of 5476 distinct dsDNA virus clusters, corresponded to cultured viruses in databases (Brum et al., 2015). This result shows the dearth of reference genomes in databases. However, the solution is not as easy as just including all metagenomic findings into the ICTV taxonomy. There are many challenges. First, most of the viruses found in metagenomic studies lack biological properties (e.g., virion morphology and host). Second, the risk of incorporating incomplete or chimeric genomes into taxonomy increases. Third, assembling a segmented or multipartite (segments are in different capsids that are independently transmitted) viral genome from short sequence reads is difficult. (Peter Simmonds et al., 2017)

Biological properties of viruses are largely encoded in their genomes, except for some examples of viral epigenetics (Milavetz & Balakrishnan, 2015). Therefore, the classification based on sequence information alone is not limited by the absence of biological attributes, but by our inability to infer virion structure or other phenotypic attributes from its genome (Peter Simmonds et al., 2017). Bioinformatics' tools and machine learning methods can help us solve this problem. For instance, the work done in Google DeepMind (AlphaFold, https://deepmind.com/blog/alphafold/, 12.04.2019) has shown unprecedented progress in the ability to predict protein structure using artificial neural networks (Hou, Wu, Cao, & Cheng, 2019). In the future, machine learning methods could hold the key to determining structures for the vast number of different viral proteins.

1.1.2. The origin of viruses

Neither the Baltimore classification nor the ICTV taxonomy at higher ranks (orders, realm) claims a common origin of viruses in these taxa. A common origin can only be assumed with confidence at species and genus level, likely at the family level as well, with some exceptions. For instance, *Myoviridae*, *Podoviridae* and *Siphoviridae* families from order *Caudovirales* (the tailed bacteriophages) each contain multiple highly divergent lineages (Aiewsakun, Adriaenssens, Lavigne, Kropinski, & Simmonds, 2018). Only 22 currently assigned subfamilies in order *Caudovirales* are clearly monophyletic (Aiewsakun et al., 2018). In higher ranks, the relationship between viral families is vague at best. Still, that does not mean a common origin can be ruled out (Low, Džunková, Chaumeil, Parks, & Hugenholtz, 2019).

Unfortunately, unlike cellular organisms, viruses leave no fossil records. Their evolutionary origin and relationships with other organisms must be deduced from "surviving" viral features (Nasir, Kim, & Caetano-Anollés, 2012). However, it is suggested that RNA-dependent RNA polymerases (RdRp) and reverse transcriptases in viruses are the relics of the primordial world (Krupovic, Dolja, & Koonin, 2019). For instance, the analysis of 4617 RNA virus RdRp sequences showed that (−)ssRNA viruses probably evolved from dsRNA viruses and dsRNA viruses in turn evolved from (+)ssRNA viruses (Wolf et al., 2018).

Reconstruction of RNA virus evolution suggested that the last common ancestors of (+)ssRNA viruses encoded only the RdRp and a single jelly-roll capsid protein (Wolf et al., 2018). However, the exact origin of RNA and DNA viruses is still unknown. At the present time, we are left with three main scenarios: the virus-first hypothesis, the reduction hypothesis and the escape hypothesis (Forterre, 2006a).

1.1.2.1. The virus-first hypothesis

The virus-first hypothesis states that viruses predated modern cells and coexisted with ancestral cells (predated LUCA) or were even direct descendants of the first replicons and existed during the precellular stage of life (Bamford, 2003; Holmes, 2011; Eugene V Koonin, Senkevich, & Dolja, 2006, 2009; Krupovic et al., 2019). This suggests that viruses are billions of years old and may have even contributed some of the fundamental architectures to cellular life, including DNA itself (Forterre, 2006b; Eugene V Koonin et al., 2006, 2009). Multiple findings support the virus-first hypothesis:

- The emergence of selfish replicating elements, in a system, having a resource that can be potentially exploited, is almost inevitable (Bansho, Furubayashi, Ichihashi, & Yomo, 2016; Ichihashi, 2019; Iranzo, Puigbò, Lobkovsky, Wolf, & Koonin, 2016; Eugene V Koonin, Wolf, & Katsnelson, 2017). A long-term *in vitro* replication experiment has provided experimental evidence that replicating systems can be viable even in the presence of parasitic replicators (Ichihashi et al., 2013). However, the presence of cell-like compartments seems to be an important factor for continuous host-parasite co-replication as the parasitic RNAs that spontaneously appear in the artificial replication systems collapse host's RNA replication under bulk condition (Bansho et al., 2016).
- Another convincing evidence for primordial origin is the fact that viruses use many genome types (ssDNA, dsDNA, (−)ssRNA, (+)ssRNA and dsRNA) compared to cellular organisms, which only use one – dsDNA. In addition, viruses benefit from different replication strategies, for instance, rolling circle replication (e.g. geminiviruses (Rizvi, Choudhury, & Tuteja, 2015)), protein-primed replication (e.g. bacteriophage Φ29 (Mendez, Blanco, & Salas, 1997; Salas & de Vega, 2016)) and the classic bidirectional theta replication (HPV16 (Flores & Lambert, 1997)) in dsDNA viruses, not to mention strategies in RNA viruses. In some viruses (HPV16, bacteriophage lambda, Epstein Barr virus) there is even a switch from one replication to another (Flores & Lambert, 1997; Hammerschmidt & Sugden, 1988; Narajczyk, Barańska, Wegrzyn, & Wegrzyn, 2007).
- The existence of several genes central to virus replication and structure in virus genomes with different replication strategies, such as large DNA viruses and positive-strand RNA viruses (Eugene V Koonin et al., 2006), without any indication of horizontal gene transfer (HGT) between these

viruses suggests the model of an ancient virus world (Eugene V Koonin et al., 2009). These genes are called viral hallmark genes (VHGs). The phrase "viral hallmark genes" was coined by Koonin et. al indicating genes shared by many diverse groups of viruses, with only distant or no homologs in cellular organisms (Eugene V Koonin et al., 2006). Also, Abroi and Gough have shown that the existence of virosphere-specific protein domains is not an artefact of missing data and it will not be overturned in the future by the increasing number of sequenced genomes and knowledge of protein structures (Abroi & Gough, 2011). It can be reasoned that the existence of VHGs in an enormous range of viruses is a relic of precellular evolution.

- Structural analyses of virion architecture and capsid protein topology of icosahedral viruses have revealed evidence of putative ancient viral lineages that co-evolved with ancestral cells (Bamford, Grimes, & Stuart, 2005). The fact that the convergence is not a viable option for the evolution of the capsid protein of icosahedral viruses only strengthens the claim (Krupovic & Bamford, 2008). Convergence is also a debatable issue for other homologous VHGs as they often have a high sequence similarity (Eugene V Koonin et al., 2006, 2009).
- Some capsid proteins from viruses infecting phylogenetically distant hosts have shown to have a common ancestry. For instance, PRD1 protein from adenoviruses (eukaryotic virus), STIV from archaea viruses, PRD1 from bacteriophages, and PBCV from an algae virus (Fu & Johnson, 2012). Their abundance in different types of viruses with respect to the range of their hosts indicates ancestral origin (Abroi & Gough, 2011; Bamford, 2003; Fu & Johnson, 2012).

The virus-first hypothesis has been challenged mainly by reasoning that all of the present-day viruses need a cellular host to replicate, therefore, requiring the existence of cells before viruses (Forterre, 2006a). In the absence of cells, virus particles are nothing but inanimate complex organic matter as virus particles are "not living, but lived entities" – viruses are produced and evolved by the cells, viruses do not self-reproduce or evolve by themselves (Guerrero, Piqueras, & Berlanga, 2002; Moreira & López-García, 2009). Also, HGT seems to be rampant in viruses (Eugene V Koonin & Dolja, 2006; E V Koonin, Makarova, & Aravind, 2001; Moreira & Brochier-Armanet, 2008), therefore the claim about the existence of ancient viral lineages, just because different viruses encode one or a few common genes, might be misguided (Moreira & López-García, 2009).

1.1.2.2. The reduction hypothesis

The reduction hypothesis ("regressive" hypothesis) postulates that viruses are regressed copies of parasitic cellular species that have lost the majority of their genes that are provided by the host (Krupovic et al., 2019; Nasir & Caetano-Anollés, 2015). The reductive evolution works as follows: initially, two free-

living organisms developed a symbiotic relationship. Over time, one of the organisms became more dependent on the other and the relationship turned to parasitic. Eventually, the previously free-living organism was unable to replicate independently anymore and it became an obligate intracellular parasite. There are many examples of reductive genomic evolution in nature, for instance, mitochondria in eukaryotic cells and several bacteria species (e.g Rickettsia) that are obligate intracellular parasites, evolved from free-living ancestors (Sagan, 1967; Weinert, Werren, Aebi, Stone, & Jiggins, 2009; Williams, Sobral, & Dickerman, 2007).

However, in viruses, the hypothesis is mainly considered in case of giant protist-infecting dsDNA viruses (Nasir, Kim, & Caetano-Anollés, 2012), but can be also considered for several bacterial viruses which encode ribosomal proteins (Krupovic et al., 2019; Mizuno et al., 2019). Some studies even suggest that giant dsDNA viruses should form the fourth domain of life next to Bacteria, Archaea, and Eukarya as the genomes of large dsDNA viruses contain many genes present in cells including elements from translation system (Desnues, Boyer, & Raoult, 2012; Legendre, Arslan, Abergel, & Claverie, 2012; Nasir, Kim, & Caetano-Anolles, 2012; Raoult et al., 2004). Still, host to virus (H2V) gene transfer combined with accelerated evolution of viral genes is probably a more likely explanation than large dsDNA viruses being the fourth domain of life (Yutin, Wolf, & Koonin, 2014). Also, among many proteins shared with cellular organisms (aaRS, RNAP II, translation factors like ElF1) only IleRS showed some support for fourth domain theory (Yutin et al., 2014). In addition, Gao et al. found that giant viruses have the largest number of duplicated genes indicating that giant viruses might evolve by complexification from smaller viruses not by reduction (Gao, Zhao, Jin, Xu, & Han, 2017). However, previous points do not render reductive evolution invalid as a process of how viruses can evolve. For instance, the loss of the core genes of a putative ancestral virus of orthopoxviruses played a critical role in speciation (Hendrickson, Wang, Hatcher, & Lefkowitz, 2010).

1.1.2.3. The escape hypothesis

The parasitic nature of viruses implies that cells predated viruses and viruses could have emerged from these cells as “escaped genes” that acquired the ability to replicate and later evolved via HGT (Forterre, 2006a; Nasir, Kim, & Caetano-Anollés, 2012). The escape hypothesis (escaped host’s gene hypothesis or progressive hypothesis) implies that these “escaped genes” might have been pieces of genetic material capable of moving within a genome (e.g. retrotrans-posons) that acquired the ability to exit the cells. The escape event may have happened from modern cells (e.g., hepatitis delta virus (Radjef et al., 2004; J. M. Taylor, 2014; J. Taylor & Pelchat, 2010)) but is possible also from primordial cells (Krupovic et al., 2019).

The most interesting implication of this hypothesis is that the majority of genes in viruses should have homologs in cellular organisms. However, the

presence of structures that are unique to viruses has put a challenge to this hypothesis (Abroi & Gough, 2011; Forterre, 2006a; Eugene V Koonin et al., 2009). For instance, RdRp, reverse transcriptase and protein-primed DNA polymerase in viruses do not have cellular homologs other than horizontally acquired counterparts (Krupovic et al., 2019). It should be noted that cellular RdRp (involved in the formation of telomeres and small RNAs) are homologous to DNA-dependent RNA polymerases involved in transcription, not to the viral RdRp (Iyer, Koonin, & Aravind, 2003; Krupovic et al., 2019).

1.1.2.4. Implications of the origin of viruses hypotheses

Viruses have different replication strategies, gene content, capsid architecture, and genome types which suggest various evolutionary origins – viruses are polyphyletic (Bamford, 2003; Eugene V Koonin et al., 2006; Moreira & López-García, 2009). Thus, we do not have to pick one single hypothesis and discard others, as all of them might be correct at the same time (but for different viruses). In addition, there is no reason that any of these events (e.g. gene escape) only happened once. Also, a chimeric scenario has been proposed in which the virus replication machinery originates from the primordial pool of genetic elements, but the capsid proteins were acquired from the ancestors of modern cells at different stages of evolution (Krupovic et al., 2019). In conclusion, it can be reasoned that different viral families could have emerged through different paths.

The escape hypothesis, the reduction hypothesis and partly also the chimeric hypothesis create one important prediction – many genes found in viruses should have their ancestries (homologs) in cellular genomes. Investigating the provenance of viral genes may give us insights into the matter of viral evolution and origin.

1.1.3. Papillomaviruses

In the current thesis, the origin of papillomaviruses was studied. Papillomaviruses (PVs) infect many mammalian species (including marine mammals), birds, turtles, snakes and fish (Van Doorslaer, Li, et al., 2017). PVs have been of interest due to their association with cancers. Oncogenic human papillomaviruses (HPVs) are responsible for almost all cases of cervical (99%) and anal (88%) cancers, as well as about 70% vagina, 50% penile, 13–56% oropharynx (depending on the geographical location) and 43% of vulvar cancers (De Vuyst, Clifford, Nascimento, Madeleine, & Franceschi, 2009; Forman et al., 2012).

PVs have a circular double-stranded DNA genome between 5748–8809 bp (pave.niaid.nih.gov, 23.06.2019) which is packed in a non-enveloped icosahedral capsid (Van Doorslaer et al., 2013). The PV genome organization is highly conserved (Van Doorslaer & McBride, 2016). A typical mammalian PV genome encodes at least 8 proteins (E1, E2, L1, L2, E6, E7, E8^E2, E1^E4).

The “E” stands for early and the “L” stands for late – proteins that are expressed in the early or late phase of viral infection (Van Doorslaer, 2013). At the present time, a total of 405 PV reference genomes are available in The Papillomavirus Episteme (PaVE) database (pave.niaid.nih.gov, 23.06.2019), including 198 HPVs (Van Doorslaer, Li, et al., 2017). The PaVE database (pave.niaid.nih.gov) is reliable and widely used resource by PV researchers. It contains highly organised and curated papillomavirus genomics information including many tools for the scientific community (Van Doorslaer, Li, et al., 2017).

1.1.3.1. The origin of papillomaviruses

In 1933 Shope et al. published work on infectious papillomatosis of wild cottontail rabbits found in northwestern Iowa (Shope & Hurst, 1933). Now, almost 90 years later after decades of research, scientists have acquired a wealth of information about the molecular biology of papillomaviruses and viruses in general. However, the evolutionary origin of papillomaviruses is still enigmatic.

PVs have been isolated from various mammalian species and sauropsids, but also from four different bony fish: gilthead seabream, rainbow trout, red snapper and haddock (López-Bueno et al., 2016; Willemsen & Bravo, 2019). These PVs exhibit a unique genome organization, encoding only the minimal PV backbone (E1, E2, L1 and L2) while lacking any of oncogenes (E5, E6, and E7) (López-Bueno et al., 2016; Willemsen & Bravo, 2019). Also, these PVs form a mono-phyletic clade in the E1-E2-L2-L1 concatenated tree at the nucleotide level and are suggested as a new root to the phylogenetic tree of papillomaviruses (Willemsen & Bravo, 2019). The analysis of the phylogenetic tree of papilloma-viruses has dated the root around 481 MYA (656–326 MYA) in one study (Van Doorslaer et al. 2017) and 424 MYA (446–402 MYA) in another (Willemsen & Bravo, 2019). The gain of ancestral E6 and E7 gene has been dated much later about 184 MYA (Willemsen & Bravo, 2019).

The occurrence of PVs in fish gives an indication that PVs were already infecting the earliest Euteleostomi (Van Doorslaer, Ruoppolo, et al., 2017). The Euteleostomi clade includes more than 90 percent of the living vertebrate species (Van Doorslaer, Ruoppolo, et al., 2017). Also, the lack of E5, E6 and E7 genes from genomes of fish PVs, reinforce the proposed evolutionary scenario that ancestral PV genome contained only four core genes (E1, E2, L1 and L2) and did not contain any of the oncogenes (García-Vallvé, Alonso, & Bravo, 2005; Willemsen & Bravo, 2019). Investigating the occurrence of PV gene homologs, especially the core gene homologs, in cellular organisms may give us clues for PV origin. This was the task in Ref. I of current thesis.

Still, not all genes (protein folds) in viruses can be traced back to cellular organisms (Abroi & Gough, 2011). There are multiple potential scenarios that explain the missing homologs: ancestral viral origin (virus-first hypothesis); cellular origin but later lost by cells; cellular origin but the respective taxon has become extinct; or the genes could have been evolved *de novo* in viruses (Abroi & Gough, 2011; Sabath, Wagner, & Karlin, 2012). Some of these *de novo*

evolved proteins, like tombusvirus' p19, have been structurally and functionally characterised, showing the previously unknown structure and an unknown mechanism of action (Pavesi, Magiorkinis, & Karlin, 2013; Vargason, Szittya, Burgyán, & Hall, 2003). However, almost all proteins identified as evolved *de novo* in viruses have a "secondary" function, e.g. related to pathogenicity not to replication or structure (Pavesi et al., 2013). At the same time, the inability to find cellular homologs to a viral protein does not prove that it has originated *de no* or has an ancestral origin. The evolution rate in viruses is much higher, sometimes up to five orders of magnitude higher compared to cellular organisms. Thus, the sequence similarity can be so low that the homology is not confidently detectable by pairwise sequence analysis (Aiewsakun & Katzourakis, 2016; Duffy, Shackelton, & Holmes, 2008; Sanjuán, Nebot, Chirico, Mansky, & Belshaw, 2010). The high mutation rate in viruses is not the only difficulty that scientists face in the field of the deep phylogenetic studies of viruses. Rooting phylogenetic trees, distant homology detection, HGT are just a few of these difficulties.

1.1.4. The phylogenetic studies of viruses

Traditionally, species phylogenies are inferred from a single gene tree or from a concatenated nucleotide sequence tree (Gadagkar, Rosenberg, & Kumar, 2005). In order to infer deeper relationships, protein multiple sequence alignments are used. For instance, from a set of core genes (e.g. genes involved in the protein synthesis), which are nearly universal protein-coding genes in cellular organisms, a universal tree of life (TOL) can be constructed (O'Malley & Koonin, 2011). However, viruses have always been left out from the TOL (Claverie & Ogata, 2009; Hegde, Maddur, Kaveri, & Bayry, 2009; Ludmir & Enquist, 2009; Moreira & López-García, 2009) and therefore there is no viral equivalent to the cellular tree of life. In fact, it is not even reasonable to construct one single tree of viruses as viruses are thought to be polyphyletic and no single gene has been identified that is shared by all viruses. Therefore, constructing a unified "gene tree" of all viruses is impossible (Holmes, 2011; Eugene V Koonin & Dolja, 2013; Eugene V Koonin et al., 2006). Furthermore, even between different Baltimore classes, very few genes are shared (Nasir and Caetano-Anollés, 2015).

However, is it possible to give a rough estimate to the number of different monophyletic groups in viruses (viral origins)? The number of viral origins should not exceed the number of genera in virus taxonomy. Thus, based on ICTV release 2018b, there should be less than 1019 monophyletic groups. However, the number is probably closer to the number of viral families as the majority of them are monophyletic except families from *Caudovirales*. In ICTV taxonomy (release 2018b) there are 150 viral families and 12 genera which are not signed into a family. The five families in *Caudovirales* order are divided into 26 subfamilies and 271 genera are not assigned into a subfamily. However,

a bipartite network of viral genera shows less than 20 unconnected clusters (Fig. 1). Bipartite networks have been successfully used for researching virosphere in several studies (Iranzo et al., 2016c, 2016b). It consists of two types of nodes: a virus genome or higher taxon (in our example genera) and genes or protein domains (in our example assigned Pfam protein domain families).

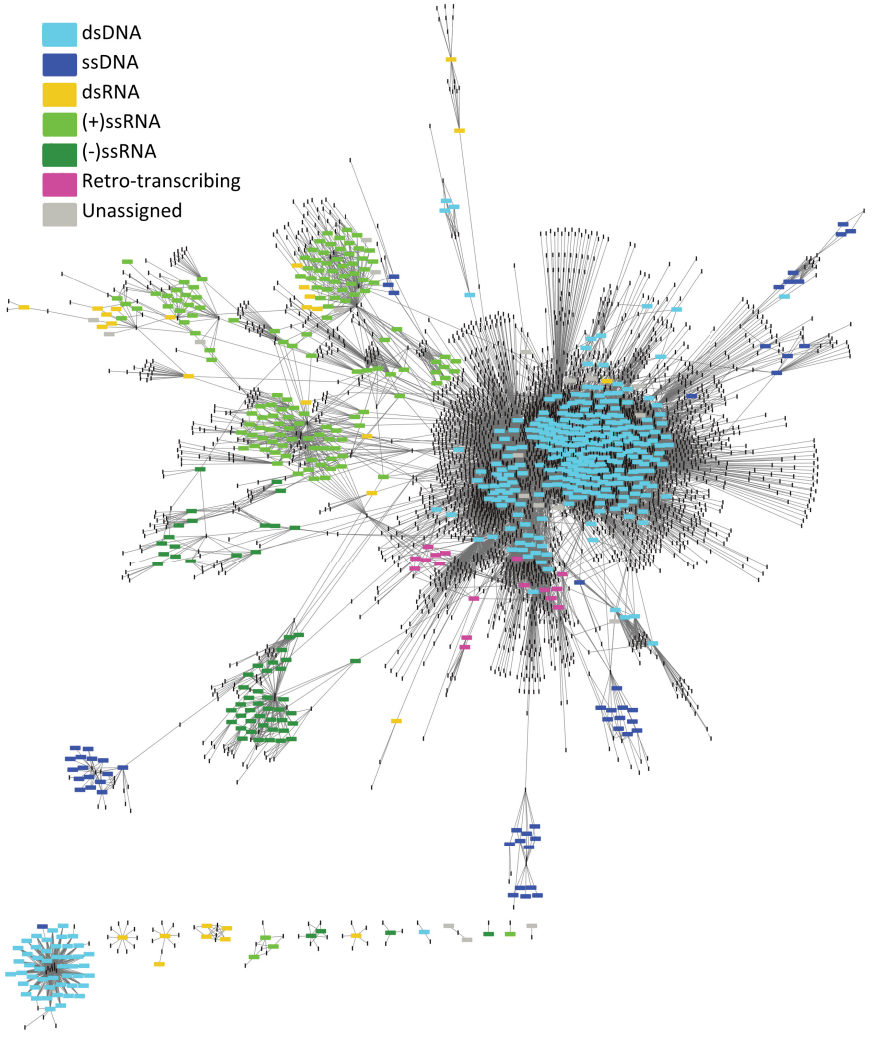

Figure 1. **A bipartite network of viral genera.** The network is based on Pfam 32 assignments in UniProtKB "Reference proteomes". Black dots represent Pfam 32.0 protein domain families. Virus genera are coloured based on the genome type. Protein domains are connected to a viral genus if in at least one proteome a protein domain is assigned (a genus may contain multiple reference proteomes). Viruses are positioned on the graph using the Fruchterman-Reingold force-directed layout algorithm. Visualization is done in Cytoscape 3.7.1 (Shannon et al., 2003) and layout is calculated with the AllegroLayout plugin. Maximum iterations were 2000 with the option "no overlap iterations" enabled. Edges were weighted using the proportion of genomes in a genus having a domain assigned. (Data not published)

Of course, virus genera in the major cluster in figure 1 do not share one single gene and separate subclusters can be observed. In addition, virus to virus horizontal gene transfer through a common host and gene loss should be taken into account before making any conclusions. It has been suggested that gene loss plays an important role in speciation and evolution in some viruses (Hendrickson et al., 2010; Van Doorslaer & McBride, 2016). There are also other problems which virologists face in the field of deep phylogenetic studies of viruses discussed in the following chapter.

1.1.4.1. Peculiarities in deep evolutionary studies of viruses

One of the most troublesome features that affect the deep phylogenetic studies of viruses is the viruses' ability to pickpocket genes from their hosts. Without considering horizontal gene transfer (HGT), drawing conclusions can be erroneous (Yutin et al., 2014). Some studies even suggest that viruses could have been an engine for the genesis of protein structures in cellular organisms through host-to-virus (H2V) and virus-to-host (V2H) HGT (Abroi & Gough, 2011). In addition, it has been shown that the mechanisms applied for creating new genes in viruses and in cellular organisms differ. Emergence of new protein-coding genes in cellular organisms is mainly attributed to gene duplication, which is a major mechanism of evolutionary change in bacteria and eukaryotes (Conant & Wolfe, 2008; Gao et al., 2017; He & Zhang, 2005; Magadum, Banerjee, Murugan, Gangapur, & Ravikesavan, 2013; Panchy, Lehti-Shiu, & Shiu, 2016; Simon-Loriere & Holmes, 2013). Gene duplication also plays a role in the evolution of some dsDNA viruses (Gao et al., 2017). However, in RNA viruses, ssDNA viruses and in dsDNA-RT viruses gene duplication is rare (Gao et al., 2017; Simon-Loriere & Holmes, 2013).

Another feature, which is characteristic of viruses, is the high mutation rate. It is typically much higher than in bacteria, archaea or in eukaryotes. What makes the situation even more complex is that the mutation rate differs between viruses with different genomes, especially if we compare ssRNA and dsDNA viruses (Aiewsakun & Katzourakis, 2016; Duffy et al., 2008; Sanjuán et al., 2010). The mutation rate in dsDNA viruses is about 10^{-7}–10^{-8} mutations per replication and in ssRNA viruses, it is about 10^{-3}–10^{-5} mutations per replication (Duffy et al., 2008). This corresponds to the fidelity of the polymerases – RNA-dependent RNA polymerase (RdRp) is more error-prone than DNA polymerase (Gout, Thomas, Smith, Okamoto, & Lynch, 2013; Lynch, 2010). Nonetheless, it is remarkable that nearly identical sequences at the nucleotide level occur in such far-reaching environments as the Southern Ocean, the Gulf of Mexico, an Arctic freshwater cyanobacterial mat and Lake Constance, Germany (C. M. Short & Suttle, 2005; Suttle, 2005). However, not all generated mutations will be fixed in a population.

The overall nucleotide substitution rate (fixed mutations in a population) varies also between viruses. For instance, it falls in the range between 10^{-2} to

10^{-5} nucleotide substitutions per site per year in nearly all RNA viruses (Duffy et al., 2008; Hanada, Suzuki, & Gojobori, 2004; Jenkins, Rambaut, Pybus, & Holmes, 2002). In papillomaviruses (dsDNA viruses), it has been estimated that the viral genes evolve about 5–10 times faster compared to their mammalian host nuclear protein-coding sequences which are thought to acquire about 2×10^{-9} substitutions per site per year (Kumar & Subramanian, 2002; Rector et al., 2007; Shah, Doorbar, & Goldstein, 2010; Van Doorslaer, 2013). In addition, it has been shown that the short-term substitution rate of viruses is much higher than the long-term substitution rate (Aiewsakun & Katzourakis, 2016). Hence, the sequence space sampled by viruses is even larger than that expected from long-term substitution rates.

The mutation saturation may destroy phylogenetic signals in viral sequences affecting the validity of deep phylogenetic inference (G. Caetano-Anollés & Nasir, 2012; Sober & Steel, 2002). Therefore, multiple sequence alignment of viral genes may not be sufficiently robust to draw conclusions about the early moments of viral evolutions and we should always interpret the results with extreme caution (Holmes & Duchêne, 2019; Wolf et al., 2018, 2019). Also, due to the high substitution rate in viral genomes, the similarity between homologous sequences in viruses and cellular organisms may be too low to detect homology. Fortunately, profile hidden Markov models (profile-HMMs) may allow us to detect these distant homologous sequences, which may be problematic with traditional pairwise sequence comparison methods.

1.2. Methods for homology detection

1.2.1. Pairwise sequence comparison methods

Homology is the existence of shared ancestry between two sequences. Pairwise sequence comparison methods have been the traditional approach to find best-matching alignments between the two sequences from which homology can be inferred. The alignment between the two sequences can be global or local. A global alignment is achieved by aligning two sequences end-to-end which may include large stretches of low similarity regions. In the case of local alignments, only regions with high similarity are aligned. Often, local alignments are preferred as proteins are built of distinct regions called domains.

One of the most used methods for producing pairwise local alignment is the word method. The word method identifies all possible non-overlapping words (subsequences) in the query sequence that are then matched to a sequence in a database. These words must have an identical match or have a similarity score of at least some threshold *T*. Word method is a heuristic method that does not guarantee an optimal solution (alignment) but is more efficient than dynamic programming (e.g., Smith-Waterman algorithm) which guarantees to find an optimal solution. BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) and FASTA (W R Pearson & Lipman, 1988) are two well-known pairwise sequence

comparison algorithms that identify the similarity between protein or nucleotide sequences using the word method. These algorithms can be used to infer functional and evolutionary relationships between sequences as well as help identify members of protein families. (Altschul et al., 1997; William R Pearson, 2014, 2016)

1.2.1.1. FASTA

One of the first protein sequence alignment programs was FASTP developed by David J. Lipman and William R. Pearson in 1985 (Lipman & Pearson, 1985). Later, FASTP evolved into a FASTA package (W R Pearson & Lipman, 1988). The name FASTA stands for "FAST-All" as it works with protein and nucleotide sequences. FASTA algorithm searches for word-to-word matches (aligned identical amino acids) of a given length *k*, before performing a more time-consuming search with a local alignment algorithm. Focusing only on small identical regions between two sequences requires fewer comparisons resulting in a faster algorithm. The word size *k* controls the sensitivity and the speed of the algorithm. The method is faster but less sensitive at higher values of *k* (*ktup* parameter). By default, k=2 in the case of protein sequences, for nucleotide sequences, the *k* is higher (k=4 or k=6). Only a small set of highest scoring local regions, which exceed a given threshold, are selected to the alignment step. The scoring is based on PAM (initially PAM250) or BLOSUM (BLOSUM50 in the latest versions) substitution matrix. The BLOSUM (BLOcks SUbstitution Matrix) substitution matrix is derived from about 2000 blocks of aligned sequence segments, however, the PAM (point accepted mutation) matrices are based on evolutionary rates. In general, a substitution matrix describes the rate at which one amino acid is replaced with another. Amino acids with similar properties (e.g., charge or polarity) are replaced more easily. The number after the BLOSUM matrix shows the maximum pairwise identity of blocks from which the matrix is built. The number behind the PAM matrix shows the number of mutations per 100 amino acids. (Henikoff & Henikoff, 1992). Wilbur and Lipman algorithm (Wilbur & Lipman, 1983) computes the final similarity score allowing insertions and deletions. FASTA also provides tools for evaluating the statistical significance of an alignment. (Lipman & Pearson, 1985; William R Pearson, 2016; W R Pearson & Lipman, 1988)

1.2.1.2. BLAST

The Basic Local Alignment Search Tool (BLAST) was developed in the 90s and was an order of magnitude faster than FASTP (Altschul et al., 1990). Similar to FASTP it uses the word method to find initial similar local regions. However, instead of finding identical matches, a similarity score is used to select the best matching words. Each of these matches must have a similarity score of at least some threshold *T.* A higher value of *T* yields greater speed, but weak similarities between sequences may be missed. BLOSUM62 substitution matrix is used by default (in the initial implementation PAM120 substitution matrix was used). In the next step, dynamic programming is used to extend the best matching words in both directions and allow gaps in the resulting alignments. In addition, BLAST calculates the statistical E-value of matches that can be used to filter significant hits. The E-value shows the number of hits that could be expected by chance when searching a database of a particular size. (Altschul et al., 1997, 1990)

Both BLAST and FASTA provide a variety of similarity measurements (bit score, E-value, percent identity, and percent similarity) from which one can infer homology or distinguish biologically significant results from randomly occurring high scoring alignments. The difference is in the procedure of finding matching words (identical matching words in FASTA vs substitution matrix based scoring in BLAST). Also, the default word size is larger in BLAST (6 vs 2). The default parameters in FASTA allow higher sensitivity for very distantly related sequences but require longer alignments. However, the BLAST algorithm is faster than the FASTA algorithm. (William R Pearson, 2014)

1.2.2. Hidden Markov models

Pairwise sequence comparison methods for homology searches like BLAST or FASTA work well only with protein sequences whose identities are larger than 30%, but fail to find more distantly related proteins at lower identity (Brenner, Chothia, & Hubbard, 1998). Thus, detection of distant homologs is problematic with pairwise sequence comparison methods, especially in deep viral phylogenies. A more sensitive approach is to use hidden Markov models (HMMs) to detect remote homologs (Kirsip & Abroi, 2019; Kuchibhatla et al., 2014; Park et al., 1998).

Markov models are statistical models that are well-known for their performance in modeling the correlations between adjacent symbols on time series or on a linear sequence (Eddy, 1998, 2004; Yoon, 2009). A hidden Markov model is used to describe observable symbols (e.g., amino acids) that depend on hidden states. In other words, an HMM consists of two stochastic processes – an invisible process of hidden states and a visible process of observable symbols. The hidden states form a Markov chain. A Markov chain is a stochastic model that experiences transitions from one state to another according to certain

probabilities. However, no matter how a present state is achieved, all possible future states are fixed. I.e., the probability of transitioning to any next state is dependent only on the state attained in the previous event (Sean R Eddy, 2004; S R Eddy, 1998; Yoon, 2009). In biology, HMMs have been used in gene prediction (Munch and Krogh 2006), modeling DNA sequencing errors (Lottaz et al. 2003), protein secondary structure prediction (Won et al. 2007) and modeling protein domains (Gough et al., 2001; Lewis et al., 2018; Sonnhammer et al., 1997).

There exist a large number of HMM variants that modify and extend the basic model and one of these variants is profile-HMMs which is used to model a multiple sequence alignment (Sean R Eddy, 2004; S R Eddy, 1998; Yoon, 2009). A profile-HMM uses three types of hidden states: match states (M_n), insert states (I_n), and delete states (D_n). As a simple example, let's consider an HMM that models a small alignment of amino acid sequences (Fig. 2). The sequence alignment contains different observed symbols (amino acids) at each position. The amino acid frequencies at the n-th position are emission probabilities for the n-th match state. The transition probabilities (match state to match state, match state to insert state, etc.) are calculated from the alignment. Now, given a new amino acid sequence, we can compute the most likely hidden state sequence (alignment) based on observed amino acids. For that, we could construct all possible alignments and calculate probabilities for each hidden state sequence. However, this is computationally very expensive, therefore, more efficient algorithms are used, for instance, the *Viterbi* (Forney, 1973) algorithm. (Sean R Eddy, 2004; S R Eddy, 1998; Yoon, 2009).

Compared to pairwise sequence comparison methods, a profile-HMM can include information from many sequences into one model, which allows it to be more sensitive and find more distant homologs (Kuchibhatla et al., 2014; Park et al., 1998). Also, profile-HMM are able to model gaps using insertion and deletion states whereas pairwise sequence comparison methods use some fixed function to penalize for opening and extending gaps without distinguishing between them. Another very important aspect, why HMMs are popular in biology, is the availability of tools like HMMER (Sean R Eddy, 2009; Mistry, Finn, Eddy, Bateman, & Punta, 2013) and the existence of high-quality models in different resources like Pfam and SUPERFAMILY (J Gough, Karplus, Hughey, & Chothia, 2001; Sonnhammer, Eddy, & Durbin, 1997).

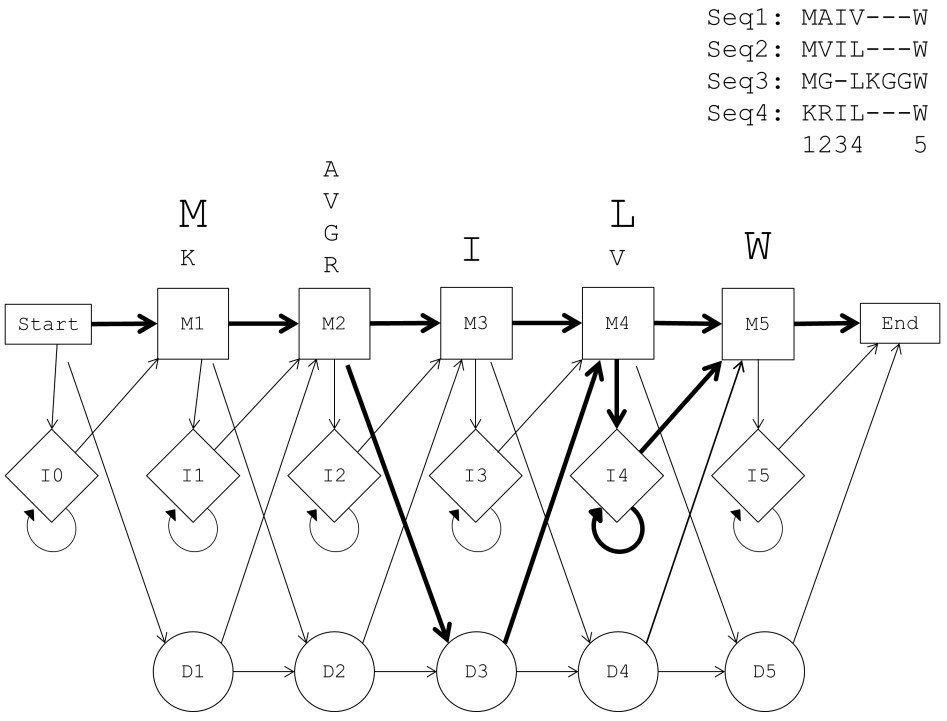

Figure 2. The architecture of a profile-HMM. Building a profile-HMM starts with a multiple sequence alignment (right corner). Profile-HMMs of biological sequence analysis have three hidden states – match state (M), insert state (I) and delete state (D). The emission probabilities in the match states are based on amino acid frequencies in the alignment. Transition probabilities are marked with arrows. All possible paths in the current example have been highlighted.

1.2.2.1. HMMER

Currently, one of the most popular software suites for protein sequence analysis, which implements profile-HMMs, is HMMER. It is designed to detect remote homologs as sensitively as possible using profile-HMMs. In addition to homology searches, HMMER can be used to make sequence alignments, build profile-HMM models and work with single query sequences like BLAST. HMMER can handle both protein and nucleotide sequences. The latest version of HMMER (HMMER3) is essentially as fast as BLAST. (Sean R Eddy, 2009; Mistry et al., 2013)

HMMER consists of many individual programs: *alimask, hmmalign, hmmbuild, hmmconvert, hmmemit, hmmfetch, hmmlogo, hmmpgmd, hmmpress, hmmscan, hmmsearch, hmmsim, hmmstat, jackhmmer, makehmmerdb, nhmmer, nhmmscan* and *phmmer* [http://eddylab.org/software/hmmer/Userguide.pdf, 20.06.2019]. Each of these has a specific task. In Ref. I, *hmmscan* and *hmmsearch* were used via https://www.ebi.ac.uk/Tools/hmmer/ webpage (Finn,

Clements, & Eddy, 2011). HMMER *hmmscan* program allows you to scan a sequence against a profile database (e.g., Pfam and SUPERFAMILY) to divide the sequence into its components (domains). HMMER *hmmsearch* searches profile-HMM against a sequence database looking for homologs to the model.

1.3. Resources of protein domains families

Protein space we know today is the result of billions of years of continuous evolution. Proteins are composed of one or more regions, known as domains. However, a polypeptide chain of a protein can be divided into domains on multiple criteria, therefore domain borders (length) may differ comparing various resources like Pfam, SCOP, CATH or ECOD (will be discussed in the following chapter) (Cheng et al., 2014; Dawson et al., 2017; Finn et al., 2006; Murzin et al., 1995). Usually, domains are defined based on reuse (Narunsky et al., 2019). However, there is no consensus on what is the exact definition of a domain (Day, Beck, Armen, & Daggett, 2003; Hadley & Jones, 1999; Holland, Veretnik, Shindyalov, & Bourne, 2006). For example, it has been estimated that only 60% of CATH domains have a similar SCOP counterpart (Kelley & Sternberg, 2015).

The number of proteins in nature is much higher than the number of domain families. A large number of proteins is achieved by different combinations of domains i.e. architectures of protein domains (Green et al., 1993; Murzin, Brenner, Hubbard, & Chothia, 1995; Sonnhammer et al., 1997). Combinations could occur between domains with a different phylogenetic origin. Therefore, protein domains are more monophyletic than whole proteins. For instance, papillomavirus E1 protein consists of an E1 DNA binding domain (DBD) and a P-loop containing nucleoside triphosphate hydrolase domain. The latter is found in all cellular organisms, but the former exists only in few, implying different origin. Hence, protein domains are one of the fundamental units of evolution and can be used to trace the evolutionary history of proteins. Currently, Pfam (Sonnhammer et al., 1997) is among the most popular protein annotation tools.

1.3.1. Pfam

The Pfam database is a collection of protein domain families. A protein domain family is a group of evolutionarily-related protein domains. Thus, protein domains in a family descended from a common ancestor. The primary use of Pfam is to identify and classify domains in protein sequences. In Pfam, each domain family is represented by a curated multiple sequence alignment from which a profile-HMM is built. (El-Gebali et al., 2019; Sonnhammer et al., 1997)

Originally Pfam consisted of two parts A and B. Pfam-A was a set of manually curated protein domain families with high-quality alignments, whereas Pfam-B contained automatically generated families. As

of version 28.0 (released in 2015), Pfam-B is discontinued [ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam28.0/relnotes.txt, 18.04.2019]. The novelty of Pfam's approach was that it used two alignments: a high-quality seed alignment (non-redundant dataset) and a full alignment. The latter is built automatically by aligning all members to a profile-HMM which is built from the seed alignment (Sonnhammer et al., 1997). In addition, Pfam contains assignments to protein sequences available in the UniProtKB. The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt Knowledgebase (UniProtKB) is the central collection of information on proteins (from amino acid sequence and taxonomic data to biological ontologies). The UniProtKB consists of two sections: manually-annotated records (UniProtKB/Swiss-Prot) and computationally analysed (unreviewed) records (UniProtKB/TrEMBL). Proteomes in the UniProtKB can include protein sequences from both UniProtKB/Swiss-Prot and UniProtKB/TrEMBL sections of the UniProtKB. A proteome in this context is a set of protein sequences that can be acquired by translating all protein-coding genes of a completely sequenced genome. UniProt includes two subsets of UniProtKB called "Complete proteomes" and "Reference proteomes". The first contains a full set of protein sequences from completely sequenced and annotated genomes [https://www.uniprot.org/keywords/KW-0181, 13.05.2019]. The "Reference proteomes" subset is, in turn, a subset of the "Complete proteomes" subset, providing a non-redundant selection of species representing a broad coverage of the tree of life [https://www.uniprot.org/help/reference_proteome, 13.05.2019]. (The UniProt Consortium, 2017)

Pfam 28.0, which was used in Ref. I, contains a total of 16230 protein domain families. About 81% of all proteins in UniProtKB (version 2014_07) contain a match to at least one model (sequence coverage) [ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam28.0/relnotes.txt, 18.04.2019]. In the most recent version of Pfam 32.0, there are 17929 protein domain families and about 77% of protein sequences in UniProtKB "Reference proteomes" (version 2018_04) have at least one match to a Pfam model (El-Gebali et al., 2019).

Pfam database also includes the hierarchical classification of protein families into clans. A clan contains more than one Pfam families that are assumed to be evolutionarily related. Classification of Pfam families into the same clan is ensured by different data: related structure, related function, significant matching of the same sequence to HMMs from different families and profile-profile comparisons. (Finn et al., 2006). In past years, the scientists behind Pfam are trying to ensure that Pfam entries and clan relationships are consistent with structural classifications like CATH (Dawson et al., 2017), SCOP (Murzin et al., 1995) and ECOD (Cheng et al., 2014). The most common tags for HMM models are *family* (62.3%) and *domain* (34.9%), comprising over 97.2% of all entries in Pfam 32.0 (others are a *motif*, *repeat*, *coiled-coil* or *disordered*). The type *domain* is usually distinguished from type *family* by a known structure that indi-

cates that the entry represents a single globular domain. So, there is experimental evidence only for 1/3 of protein domain families in Pfam release 32.0 that they exist as structural globular entities. (El-Gebali et al., 2019)

1.3.2. Classification of protein domains based on the structure

The sequence similarity between distant homologs can be so low that the homology is not detectable by pairwise sequence similarity analysis. For instance, sequence similarity among viral capsid proteins may be very low even at short evolutionary distances (Abrescia, Bamford, Grimes, & Stuart, 2012; Krupovic et al., 2019). Fortunately, the structure of a protein is more conserved than the polypeptide sequence (Abroi & Gough, 2011; Balaji & Srinivasan, 2001; Chothia & Lesk, 1986; Holm & Sander, 1996; Hubbard & Blundell, 1987; Murzin et al., 1995; Todd, Orengo, & Thornton, 1999). It has been shown that structural cores of protein domains evolve much slower than sequences (Illergård, Ardell, & Elofsson, 2009) and active sites of distantly related proteins can have very similar geometrics (Chothia & Lesk, 1986). In addition, Challis and Schmidler have demonstrated that the inclusion of structural information enables us to study deeper phylogenetic relationships that are not attainable with sequence evolution models (Challis & Schmidler, 2012). Also, some studies have shown that structure-based methods compute more reliable alignments (Carpentier & Chomilier, 2019; Rozewicki, Li, Amada, Standley, & Katoh, 2019). Thus, protein structure allows us to see even further back in time compared to analysing sequence similarity alone (Holm & Sander, 1996).

Currently, three leading hierarchical classifications of protein domains based on the structure are CATH (Class, Architecture, Topology, Homology), ECOD (Evolutionary Classification of protein Domains), and SCOP (Structural Classification of Proteins). These resources provide functional inference for homologous structures and differentiate between homologs and analogs (Cheng et al., 2014; Dawson et al., 2017; Murzin et al., 1995). All three are widely used in analysing protein sequence, structure, function, and evolution.

In the CATH database, protein domains are hierarchically classified into four groups: C, A, T, and H (Dawson et al., 2017). Protein domains are grouped together into a single homologous superfamily “H” if there is sufficient evidence that they share a clear common ancestor (Ian Sillitoe 2015). However, CATH is largely automatic with added manual curation and emphasises more on geometry, while SCOP (Murzin et al., 1995) is mainly manual and focuses on the function and evolution (Nasir & Caetano-Anollés, 2015). In the SCOP hierarchical classification, related protein domains are grouped into Families. The Family level is defined as a cluster of proteins having residue identities of 30% and greater or whose functions and structures are very similar. Families are grouped into Superfamilies (SFs) and SFs into Folds. Finally, Folds with similar secondary structure compositions are classified into Classes. However, the highest level indicating confident common ancestry is Superfamily level. (Murzin et al., 1995).

ECOD (Cheng et al., 2014) is distinct from CATH and SCOP as it groups domains primarily by evolutionary relationships (homology), rather than polypeptide chain topology. ECOD tries to extend distant evolutionary relationships beyond the SCOP SF level using different state of the art homology-inference algorithms (Cheng et al., 2014). For example, Pfam used ECOD database in their pipeline which led to the creation of 825 new families in their latest release (El-Gebali et al., 2019).

Still, SCOP is considered the "gold standard" in the classification of protein domains with known structure and provides useful evolutionary information (Nasir & Caetano-Anollés, 2015). Since the last version of SCOP (1.75 from 2009), it has diverged into two variants: SCOP2 (Andreeva, Howorth, Chothia, Kulesha, & Murzin, 2014) and SCOPe (Fox, Brenner, & Chandonia, 2014). One of the resources that use SCOP classification to build protein domain models is the SUPERFAMILY resource (J Gough et al., 2001). It should be noted that the name of the resource (SUPERFAMILY) is written with all capital letters, but the SCOP hierarchical level (Superfamily) with only the first letter capitalised.

1.3.3. SUPERFAMILY

The SUPERFAMILY resource is a collection of profile-HMMs representing SCOP protein domains (J Gough et al., 2001). Protein domain families in SUPERFAMILY are classified based on the SCOP hierarchical classification (Murzin et al., 1995). The SUPERFAMILY database focuses on the Superfamily level (a group of families with common ancestry), but also provides Family level annotations (Oates et al., 2015). In the SUPERFAMILY HMM library each SCOP SF is represented by one or more profile-HMMs, depending on how many sequences are available with less than 95% identity with known structure (J Gough et al., 2001). SFs are suitable for deep evolutionary studies (Abroi & Gough, 2011; D. Caetano-Anollés, Kim, Mittenthal, & Caetano-Anollés, 2011; G. Caetano-Anollés & Nasir, 2012; Nasir & Caetano-Anollés, 2015). Also, the structural methodology is robust against many artefacts that may occur in sequence-based phylogenetic studies (G. Caetano-Anollés & Nasir, 2012).

The procedure of creating a profile-HMM in SUPERFAMILY starts with a single sequence seed with a known structure followed by a BLAST search with strict criteria. This approach solves the practical problem of accurately aligning distantly related sequences for the purpose of generating good HMMs. In SUPERFAMILY, the model library is also curated – models that consistently give a significant score to sequences that are not homologs (model-building errors) were re-run with more restrictive parameters and re-checked until they were behaving properly. (J Gough et al., 2001)

The SUPERFAMILY version 1.75, which was used in Ref. I, is based on SCOP 1.75 containing 15 438 families and about 2000 distinct protein domain SFs (Oates et al., 2015). About 64% of all proteins in

UniProtKB (version 2018_03) contain a match to at least one model [http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi?genome=up;listtype=sf , 18.04.2019]. The latest version of SUPERFAMILY 2.0 contains 27 623 HMMs and is based on SCOPe and SCOP2 (Pandurangan, Stahlhacke, Oates, Smithers, & Gough, 2019).

Working with various resources like UniProtKB, SUPERFAMILY or Pfam and drawing seemingly genuine conclusions may be erroneous if we do not consider different biases and possible annotation errors. For instance, all of the previously mentioned resources are affected by the bias in the sequenced genomes. Not all taxa and environments (e.g. terrestrial vs marine) are covered equally. E.g., viral genomes have been subject to selection bias to medically and economically important viruses. Also, often sequence collections are redundant – containing multiple copies of one species (isolates). Fortunately, some collections like UniProt "Reference proteomes" try to solve the problem by providing a representative cross-section of the taxonomic diversity. In addition, SUPERFAMILY through SCOP and other similar resources are also biased towards available structures in Protein Data Bank (Berman et al., 2000). Protein Data Bank (PDB) is an archive of structural data of biological macromolecules (Berman et al., 2000). For instance, SCOP 1.75 is based on PDB from February 2009 (http://scop.mrc-lmb.cam.ac.uk/scop/ 20.12.17). Fortunately, protein domain structures of papillomaviruses are quite well covered even in the older version of PDB used in SCOP 1.75.

1.4. Embedded elements in protein-coding sequences of viruses

In order to keep the genome size small, genomes of viruses have a high gene density and non-coding regions are usually very small. Therefore, functional cis-elements are often embedded in protein-coding genes of viruses. Many different non-coding embedded elements have been found in protein-coding genes of viruses, like internal promoters, viral packaging signals, transcription factor binding sites, microRNAs, splice sites, frameshifting signals, etc. (Firth, 2014; Grundhoff & Sullivan, 2011; Kim, Firth, Atasheva, Frolova, & Frolov, 2011). In addition to non-coding overlapping elements, dual-coding regions are also common in viruses (Belshaw, Pybus, & Rambaut, 2007; Chirico, Vianelli, & Belshaw, 2010; Rancurel, Khosravi, Dunker, Romero, & Karlin, 2009; Veeramachaneni, Makałowski, Galdzicki, Sood, & Makałowska, 2004). A dual-coding region of a protein-coding gene is an area which partially overlaps with another protein-coding gene or which fully embeds another gene. For instance, in many papillomaviruses, E1^E4 and E8^E2 mRNA are generated via splicing by using dual-coding regions. The E4 ORF of the E1^E4 protein is embedded inside the E2 gene and the E8 ORF of the E8^E2 is embedded inside the E1 gene (Van Doorslaer et al., 2013). As the existence of E8 was studied in Ref. II

of this thesis, a small overview is given of the E8^E2 protein in the following chapter.

There are several valuable methods available for detecting dual-coding regions and other embedded functional elements in protein-coding genes of viruses (Firth, 2014; Gog et al., 2007; Mayrose et al., 2013; Sealfon et al., 2015; P Simmonds & Smith, 1999). One approach is to measure codon variability. For instance, Gog et al. developed a method that calculates mean pairwise distance (MDP) of codons in a multiple sequence alignment and uses normalised MDP as a proxy for variation. The low normalised MDP score indicated less variability than expected. However, the method developed by Gog et. al does not measure the statistical significance and does not have an available implementation. The second approach is to use synonymous substitutions rates as a proxy for variation. The idea is that in a genetic region encoding an overlapping functional element, synonymous substitutions are selectively disfavoured, as these are likely to disrupt the embedded element. For instance, A. E. Firth developed a tool called SynPlot2 which identifies regions in a protein-coding gene where there is a statistically significant reduction in the degree of variability of synonymous sites (Firth, 2014). SynPlot2 uses statistical tests and is shown to work well with RNA viruses, but is applicable to nearly any coding-sequence alignment, including DNA viruses, bacteria or eukaryotic protein-coding sequences (Firth, 2014).

General methods (e.g. SynPlot2) are ideal for detecting previously unknown embedded elements. Often, however, we need to apply very specific criteria to pinpoint the location of an embedded element. Therefore, developing programs for a single purpose is a necessity in some cases (Ref. II).

1.4.1. The E8^E2 protein

The E2 protein is a major regulator of PV gene transcription and replication (Alison A McBride, 2013). The E2 protein can act as a repressor or an activator of transcription, depending on the location of the E2 binding sites within the enhancer/promoter region (Alison A McBride, 2013). In addition to the full-length E2, several PVs express an alternatively spliced protein known as E8^E2, which is generated by fusing the E8 exon to the splice-acceptor site (3' ss) located at the beginning of the hinge region of the E2 gene (Fig. 3). The E8 coding sequence (CDS) overlaps with the E1 gene. As a result, the E8^E2 protein consists of an E8 peptide, the E2 hinge region and the E2 DNA binding domain (DBD). Thus, E8^E2 is a DNA-binding protein that is able to compete with full-length E2 for binding to E2-binding sites and to form heterodimers with full-length E2. (Kurg et al., 2009; Kurg, Tekkel, Abroi, & Ustav, 2006; A A McBride, Byrne, & Howley, 1989). E8^E2 plays a role in viral gene expression and controls the viral genome copy number (Isok-Paas, Männik, Ustav, & Ustav, 2015; Kurg, Uusen, Võsa, & Ustav, 2010; Stubenrauch, Hummel, Iftner, & Laimins, 2000). Also, the E8 peptide, at least in BPV1, is the

shortest known nuclear matrix targeting signal (Sankovski et al., 2015). To date, transcripts corresponding to the E8^E2 protein have been described experimentally for 12 PV types – BPV1 (Choe, Vaillancourt, Stenlund, & Botchan, 1989), HPV11 (Rotenberg, Chow, & Broker, 1989), HPV1 (Palermo-Dilts, Broker, & Chow, 1990), HPV16 (Doorbar et al., 1990), HPV33 (Snijders et al., 1992), HPV31 (Stubenrauch et al., 2000), SfPV (Jeckel, Loetzsch, Huber, Stubenrauch, & Iftner, 2003), HPV18 (Kurg et al., 2010), HPV5 (Sankovski, Männik, Geimanen, Ustav, & Ustav, 2014), EcPV2 (Ramsauer, 2015), MmuPV1 (Xue et al., 2017) and MfPV1 (Tombak et al., 2019). As there are very few experimentally confirmed E8^E2 transcripts known, it was decided to analyse all available E1 protein-coding genes of papillomaviruses to search the existence of E8 in these genes (Ref. II).

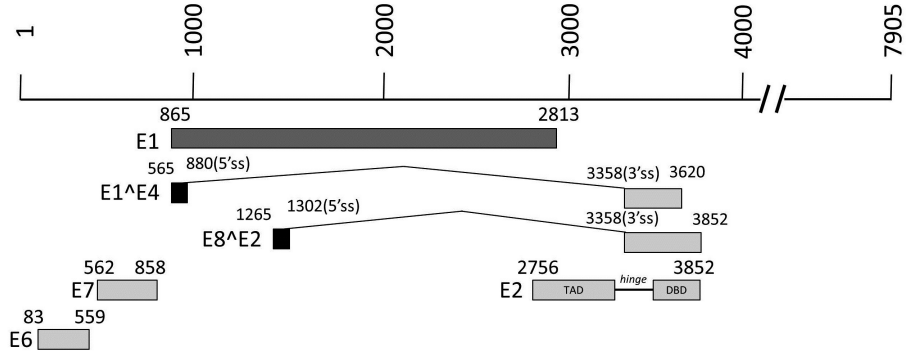

Figure 3. Early stage protein-coding genes of the human papillomavirus 16 (HPV16). Numbers indicate a nucleotide position in the HPV16 genome (PaVE HPV16REF) (Ref. II).

2. AIMS OF THE STUDY

The aim of this thesis was:

- To analyse the occurrence of papillomavirus protein domains in other organisms.
- To detect and analyse E8 dual-coding region in papillomaviruses.
- To develop a tool for detecting embedded functional elements in protein-coding sequences of viruses.

3. RESULTS AND DISCUSSION

3.1. Protein domain families found in papillomaviruses (Ref. I)

The main goal of this research (Ref. I) was to study the origin of papillomaviruses. It is not an easy task to find evidence for any of the viral origin scenarios; however, clues for viral origin most likely emerge from homologous relationships or absence of it. Thus, distant homologs were searched for protein domain families found in papillomaviruses from cellular organisms and also from genomes of other viruses. In the study, two resources were used: Pfam 28.0 and SUPERFAMILY 1.75. Both of these resources contain thousands of profile-HMMs for different protein domain families. However, not all protein-coding genes have assignments to protein domain families. Fortunately, protein-coding genes of PVs are almost fully covered with protein domain families (Pfam family and Superfamily) found in both resources (Fig. 4).

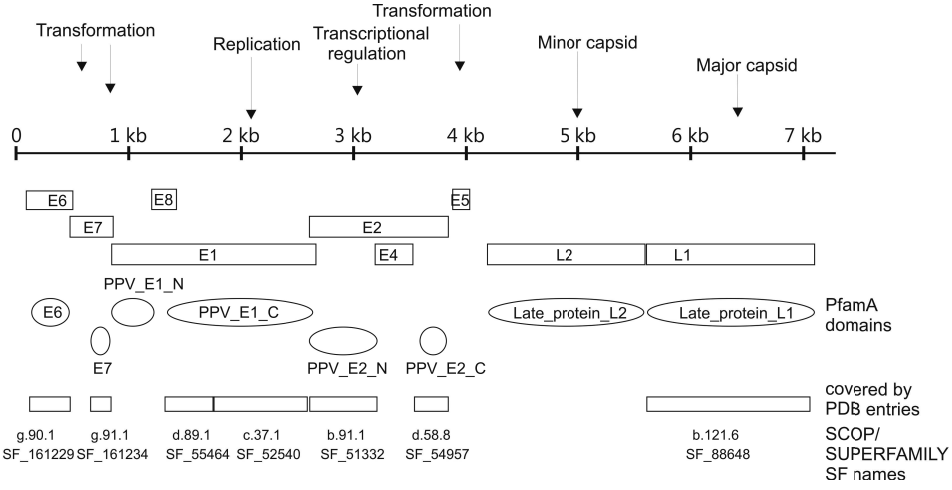

Figure 4. Protein-coding genes of BPV1 and respective Pfam and SUPERFAMILY protein domain families/Superfamilies. Bovine PV type 1 encodes 9 proteins: oncoproteins E6, E7 and E5, the viral helicase E1, the helicase-loading factor and transcription factor E2, the coat proteins L1 and L2 and the E8^E2 and E1^E4 proteins. (Ref. I)

In Pfam 28.0, there is a total of 12 different protein domain families found in protein-coding genes of papillomaviruses (Table 1, Table 2 in Ref. I), collectively named PV_PfamA in Ref. I. Pfam also classifies protein domain families into clans. However, the classification level clan was not used, as only the E1 helicase C-terminal domain (PF00519) is assigned into a clan (CL0023) out of all 12 families in Pfam 28.0 (Table 1). About 84% of amino acids of PV proteins are covered (residue coverage) with Pfam HMMs, which is a relatively

high number compared to viruses in general, for which the same number is ~66% (Table 1 in Ref. I). Only two regions are not assigned to Pfam protein domain families – the E2 hinge region which also partly contains the E4 and the short region between PPV_E1_N and PPV_E1_C (Fig. 4). The hinge region is not conserved between different PV genera (Alison A McBride, 2013).

Table 1. Pfam 28.0 protein domain families found in papillomaviruses.

Accession	**Identifier**	**Clan**
PF00500	Late_protein_L1	–
PF00508	PPV_E2_N	–
PF00511	PPV_E2_C	–
PF00513	Late_protein_L2	–
PF00518	E6	–
PF00519	PPV_E1_C	CL0023
PF00524	PPV_E1_N	–
PF00527	E7	–
PF02711	Pap_E4	–
PF03025	Papilloma_E5	–
PF05776	Papilloma_E5A	–
PF08135	EPV_E5	–

In SUPERFAMILY 1.75, seven Superfamilies(SFs) are assigned to PV protein-coding sequences, collectively called PV_SFs (Table 2, Table 3. Ref. I). Unlike Pfam protein domain families, a Superfamily may consist of multiple protein domain families (Table 2). Furthermore, each protein domain family may contain multiple profile-HMMs depending on the number of available structures and sequence similarity. Nevertheless, the SF hits are collected as a union combining all of the HMM results under a Superfamily (J Gough et al., 2001). If we compare available protein domain families/Superfamilies in Pfam 28.0 and SUPERFAMILY (SCOP) 1.75, one important difference is that the C-terminal domain of the E1 protein in SCOP is divided into two Superfamilies (55464 and 52540) instead of one family in Pfam (PF00519). In addition, there is no Superfamily for the L2 protein and for the N-terminal region of the E1 protein (Fig. 4). Both are also missing from the SUPERFAMILY 2.0 (Pandurangan et al., 2019). The reason is very simple – in both cases, there is no protein structure available in the PDB database.

In the SUPERFAMILY 1.75, the average residue coverage of proteins found in viruses is about 28%, it is much lower compared to Pfam for which the same

number was about 66% (Table 1 in Ref. I). The reason for the low residue coverage comes partially from the fact that the existing assignments were based on the NCBI viral sequence collection (2014-08-20) which is a non-redundant dataset. The residue coverage of proteins found in papillomaviruses is much higher, about 58%, but still not comparable with Pfam (84%) (Table 1 in Ref. I). Again, the reason is the missing structures for the L2 and N-terminal region of the E1 protein.

In the UniProt viral sequence collection, the average residue and sequence coverage of viral proteins is more similar to Pfam – 61% of proteins have at least one protein domain assigned to it and about 57% of amino acids are covered by protein domain in viruses [http://supfam.org/SUPERFAMILY/cgi-bin/gen_list.cgi].

Table 2. SUPERFAMILY 1.75 protein domain Superfamilies found in papillomaviruses.

Superfamily accession	Description	Number of families
55464	Origin of replication-binding domain, RBD-like (E1 DBD)	5
52540	P-loop containing nucleoside triphosphate hydrolases (E1 helicase)	24
51332	E2 regulatory, transactivation domain (E2 TAD)	1
54957	Viral DNA-binding domain (E2 DBD)	1
88648	Group I dsDNA viruses (L1)	1
161229	E6 C-terminal domain-like	1
161234	E7 C-terminal domain-like	1

3.1.1. The occurrence of papillomavirus protein domains in the biosphere

Pfam 28.0 and SUPERFAMILY 1.75 resources contain existing assignments to many sequences from various sequence collections (the “Complete proteomes” subset of UniProtKB, full UniProtKB, NCBI viral genomes, and Ensembl genomes). In Ref. I, the first step was to analyse the existing assignments in both resources. The second step was to scan up to date sequence collections using the HMMER toolkit (Sean R Eddy, 2004; S R Eddy, 1998; Yoon, 2009) to detect remote homologs for PV protein domain families. Several criteria were used to improve the quality of the results and to reduce the number of false positives. First, the evidence for viral contamination was checked by analysing the annotation and the size of the cellular contig/scaffold. I.e., if the contig contained only genes from a virus (sometimes a whole viral genome) and did

not have any adjacent cellular genes, it was considered as viral contamination. Second, a reciprocal search was performed by taking the protein-coding sequence which got a positive hit to a PV protein domain family (e.g., using *hmmsearch*) and scanning it against all HMM models with *hmmscan*. Reciprocal search should give the best hit to the same exact protein domain family or a Superfamily. Third, LOMETS (Wu and Zhang, 2007) 3D structure prediction meta-server was used to validate our findings. The potential hit should give the best modeling templates from PV structures with at least one non-HMM algorithm used in LOMETS (Material and methods in Ref. I). LOMETS generates protein structure predictions by ranking and selecting models from multiple state-of-the-art threading programs. These programs identify structural templates from the PDB library. The top templates are ranked and selected (Wu & Zhang, 2007). LOMETS combined with I-TASSER ('Zhang-server') has been one of the best structure prediction servers for several years in the CASP challenges (Cozzetto et al., 2009; Kryshtafovych et al., 2018).

3.1.1.1. Papillomavirus protein domain homologs according to Pfam

First, we applied our search criteria to existing HMM assignments of the non-redundant subset of UniProtKB (UniProtKB "Complete proteomes"). "Complete proteomes" subset of UniProtKB was preferred because of the quality of the data and interpretability. The analysis showed that only three potential homologs passed our quality checks (Table 2 in Ref. I). PPV_E1_C family HMM, which is a helicase (incl. DBD domain), gave one hit in a *Dickeya dadantii* (Bacteria) protein E0SH87_DICD3. PPV_E1_N protein domain family was found in two fungi proteins (C4V8V5_NOSCE and R0MJR2_NOSB1) from *Nosema* species. *Nosema* is a genus of microsporidian parasites. However, as there is no structure available for PPV_E1_N, homology could not be confirmed with LOMETS. Also, in the newer version of Pfam, release 32.0, the *Dickeya dadantii* protein sequence E0SH87_DICD3 gives a better hit to the PF03288 model (Pox_D5, https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan on 13.03.2019) Therefore, in all three cases, the homology is questionable. In addition, distant homologs from viruses were not detected from the "Complete proteomes" subset (Table 2 in Ref. I). Thus, at least from UniProtKB "Complete proteomes", used in Pfam 28.0, we were unable to detect any distant homologs to PV protein domains in other organisms (Table 2 in Ref. I).

Next, existing assignments in the full UniProt were analysed. The primary results must be interpreted with caution as the full UniProt may contain more misannotations and partial sequences than the "Complete proteomes" subset. In general, the results were similar to the "Complete proteomes" subset – only the PPV_E1_C and the PPV_E1_N family gave significant hits, which passed our criteria (Table 2, Ref. I). PPV_E1_C gave hits to sequences from 20 different Bacteria species, mostly from *Enterobacteriaceae* family and PPV_E1_N gave hits from four eukaryotic sequences (three fungi and one *Viridiplantae*). Other PV_PfamA models did not give any significant hits to cellular sequences which

passed our quality checks. In viruses, PPV_E1_C gave highly significant hits to *Polyomaviridae* Large-T and *Parvoviridae* NS1 proteins. This similarity has been observed previously, mostly based on shared common helicase motifs (Astell, Mol, & Anderson, 1987). Another sequence that also passed our criteria was Q91S73_9VIRU, which belongs to the small segment of Planaria asexual strain-specific virus-like element type 1. Planarian is a free-living flatworm from which extrachromosomal DNA-containing virus-like elements have been discovered (Rebrikov, Bulina, Bogdanova, Vagner, & Lukyanov, 2002). The similarity to the papillomavirus E1 helicase domain was also reported by the authors who discovered the element (Rebrikov et al., 2002).

Last, as the UniProtKB version in Pfam 28.0 was not up to date, a *hmmsearch* with HMMER toolkit was performed on a newer version of UniProt (2017_03). Again, in viruses PPV_E1_C gave highly significant hits to *Polyomaviridae* Large-T and *Parvoviridae* NS1 proteins and also to the previously mentioned Planaria asexual strain-specific virus-like element type 1. However, unlike previous results, *hmmsearch* did not return any positive hits among *Enterobacteriaceae*, only one protein sequence (A0A177Q2P3_9PLAN) from bacteria (*Planctomycetaceae bacterium*) gave true positive hits with PPV_E1_C which passed our criteria.

3.1.1.2. Papillomavirus protein domain homologs according to SUPERFAMILY

Results with Pfam models gave us very few and weak connections with cellular organisms. In order to find deeper evolutionary connections that are "lost" in sequence similarity but are still present in the protein structure, we decided to use SUPERFAMILY resource. In our research, we used the assignments at Superfamily level, which is the highest level with confident homologous relationships.

Our analysis of existing assignments showed that out of seven Superfamilies (Table 2) only the domains from the E1 protein (E1 DBD and E1 helicase) are found in cellular organisms similar to the results obtained with Pfam. The E1 helicase belongs to the "Extended AAA-ATPase domain" family which in turn belongs to the "P-loop containing nucleoside triphosphate hydrolases" Superfamily (SF_52540). As expected, the SF_52540 (helicase domain with P-loop NTPase) is present in all cellular organisms as the P-loop NTPase is a very widespread domain. The E1 DBD domain belongs to the "Replication initiation protein E1" family, which in turn belongs to the "Origin of replication-binding domain" Superfamily (SF_55464). The SF_55464 was found in 8 eukaryotes (5 fungi, 1 *Alveolata*, 1 *Amoebozoa*, and 1 *Viridiplantae*) and in 134 bacterial genomes (Table 3 in Ref. I). Most likely, all these 8 occurrences in eukaryotic genomes are relatives to geminiviral Rep protein (Family 82728) not to E1 DBD because all assignments are based on HMM 0040363, not on HMMs 0037306 or 0043184 which are models for replication initiation protein E1 family (Supplementary Materials in Ref. I). It can be reasoned that the geminiviral *Rep*

gene ended in their host's genomes through V2H gene transfer. It has been shown that sequences related to the *Rep* gene of geminiviruses, nanoviruses, and circoviruses have been frequently transferred to a broad range of eukaryotic species, including plants, fungi, animals, and protists (Liu et al., 2011). All of the hits in bacterial genomes are mostly the relaxase domain family, which also belongs to SF_55464 Superfamily. The relaxase domain is responsible for site-specific and strand-specific nicks in double-stranded DNA and plays an essential role in the initiation and termination of conjugative DNA transfer (Byrd & Matson, 1997). SF_55464 was also found in more than 400 bacterial plasmids (again, mostly the model of relaxase domain) including one eukaryotic plasmid pPT4-NU with red algal host *Pyropia tenera* and notably, only in a single bacterial virus. Thus, the E1 DBD connects PVs confidently only with bacteria and bacterial plasmids.

Among viruses, SF_52540 and SF_55464 are present in all members of *Polyomaviridae*. In addition, the SF_55464 was found in several viral sequences from *Parvoviridae*, *Geminiviridae*, in two viruses from *Betaherpesvirinae*, in one member of *Circoviridae* and *Siphoviridae* (relaxase domain), and in *Genomoviridae*.

The extended analysis of the full UniProt sequences (existing assignments and *hmmsearch* against a newer version of the UniProt database) increased the number of positive hits of SF_55464 within the bacterial and eukaryotic sequences. In addition, potential homologs to E6 (SF_161229), L1 protein domain (SF_88648), and E2 DBD (SF_54957) were found in cellular organisms. However, according to LOMETS, the sequence containing E6 homolog fits equally well into ferredoxin structures making the result questionable. Distant homologs of the L1 protein were found in *Polyomaviridae* and distant homologs of the E2 DBD were found in a subset of gammaherpesviruses.

Evidence has been presented that at least the protein architectures (protein domain combinations) rarely evolve by convergent evolution (Julian Gough, 2005). The E1 protein contains SF_55464:SF_52540 domain pair, from which SF_52540 is abundant in nature and the SF_55464 was present in several genomes. Therefore, it was decided to search for the presence of the domain pair from the rest of the biosphere. The architecture was detected in all polyomaviruses (except one incomplete genome), 40% parvoviruses, 4% geminiviruses, three members of *Genomoviridae*, one member of *Circoviridae,* one member of *Siphoviridae* (Table 4 in Ref. I). The domain pair was found in some eukaryotes, bacterial plasmids and in more than 100 bacterial species. It should be noted that often bacterial chromosome and plasmid are not discriminated in the databases. Nevertheless, most of the plasmids containing the SF_55464 also had SF_52540 assigned. Thus, at least according to SUPERFAMILY, the PV replication protein E1 is confidently evolutionarily connected with *Polyomaviridae*, *Parvoviridae*, conjugative plasmids, and probably to bacteria.

In conclusion, domains from the E1 protein, the major capsid protein L1 and the E2 DBD show confident deeper evolutionary connections to other viruses.

However, in cellular organisms and bacterial plasmids, only homologs of the E1 protein were found.

3.1.2. The origin of papillomaviruses

The major capsid protein L1, E2 DBD and both domains from E1 had distant homologs in the rest of the biosphere. Out of these, only domains from the E1 replication protein had homologs in cellular organisms (Fig. 4 in Ref. I). However, the presence of SF_52540 (P-loop NTPase) in cellular proteins is non-informative as it is a very widespread domain. Thus, the informative connections to eukaryotic proteins are almost non-existent.

PVs are clearly related to *Polyomaviridae*, sharing structural homologs of capsid protein L1 and two domains of replication protein E1 at SCOP Superfamily level (Fig. 5 in Ref. I). Both viral families have dsDNA viral genomes packed into nucleosomes inside the viral particle. In addition, members of *Parvoviridae* (ssDNA viruses) share two replication related domains and, including extended structural similarity, also the capsid protein with PVs and with *Polyomaviridae*. The extended structural similarity comes from the fact that the PV L1, the *Polyomaviridae* VP1, *Parvoviridae* VP2 belong into the same Fold level (single jelly-roll) in SCOP. The Fold level in SCOP joins Superfamilies into a common fold if their proteins have the same major secondary structures (α-helices and β-strands) in the same arrangement with the same topological connections (Murzin et al., 1995). The Fold level does not guarantee a common ancestor; however, it does not rule it out either. Most likely, the last common ancestor of *Papillomaviridae*, *Polyomaviridae*, and *Parvoviridae* inhabited a marine environment. However, only very few marine eukaryotic organisms outside fungi and vertebrates are sequenced. Thus, most likely, we have an unexplored sequence and structure space in both cellular and viral taxa, as well as in other types of mobile elements in marine environments, which could reveal more information about the origin of PVs.

The E2 DBD domain connects PVs to members of genus *Lymphocryptovirus*, which belongs to the *Gammaherpesvirinae* subfamily. E2 DBD functional and structural homologs are shown to be present also in *Rhadinovirus* genus, which also belongs to the *Gammaherpesvirinae* subfamily (Correia et al., 2013; Domsic, Chen, Lu, Marmorstein, & Lieberman, 2013; Hellert et al., 2013). Therefore, it can be reasoned that the ancestor of the *Gammaherpesvirinae* subfamily may have had the E2 DBD "relative" in its genome. So, is the origin of the E2 DBD in ancestors of gammaherpesviruses?

Herpesviruses (HVs) are a group of DNA viruses which have been extensively studied. The order *Herpesvirales* includes three families: *Malacoherpesviridae* (viruses of Molluscs), *Alloherpesviridae* (viruses of amphibians and fish), and *Herpesviridae* (viruses of reptiles, birds, and mammals) (Grose, 2012). *Herpesviridae* consists of three subfamilies (the *Alpha*-, *Beta*-, and *Gammaherpesvirinae*), which are all related and have a common ancestor. It has been

estimated that the common ancestor of *Herpesviridae* family existed at least 400 million years ago (McGeoch & Gatherer, 2005). *Betaherpesvirinae* and *Gammaherpesvirinae* diverged about 350 MYA (McGeoch & Gatherer, 2005). However, the ancestor of PVs have been estimated to exist at least 400 MYA containing at least four core genes (E1-E2-L1-L2) (Van Doorslaer, Ruoppolo, et al., 2017; Willemsen & Bravo, 2019). Therefore, it can be reasoned that PV E2 DBD does not originate from gammaherpesviruses.

Our research in Ref. I showed that the majority of protein domains in PVs did not have homologs in cellular genomes. In general, there are three explanations for missing homologs in genomes of cellular organisms in addition to virus-first hypothesis:

- The absence of genomes from databases containing the homologs.
- The gene has been lost from all current cellular species.
- Primordial cellular lineages that contained the homologs are now extinct.

In addition, *de novo* gene generation in viruses can also be one explanation for missing homologs in cellular organisms. One of the mechanisms how *de novo* genes can emerge in viruses is called overprinting – mutations lead to a new protein-coding gene by overlapping an ancestral gene (Rancurel et al., 2009; Sabath et al., 2012). Detecting these overlapping genes and other functional embedded elements calls for a specialised method. In the ref. II, we studied the presence of E8 inside the E1 gene of papillomaviruses and developed a method to detect overlapping genes and other embedded elements in viruses.

3.2. The conservation of the E8 CDS in the E1 gene of papillomaviruses (Ref. II)

At the time of writing of Ref. II, the E8^E2 was annotated and mRNA experimentally confirmed only in nine PVs and our goal was to examine the prevalence of E8 in other PVs. Fortunately, the distribution of these nine PVs was phylogenetically sparse, which gave us a good base for building an algorithm. We used multiple parameters, inferred from existing data, to detect E8 in PV genomes (Methods in Ref. II). These parameters included restriction to E8 length, E8 location inside the E1 gene and a consensus sequence for splicing donor site (5' ss). However, the restrictions were not very strict. The only strict restriction used in the model was that the length of the E8 CDS divided by three must produce a residual of two. The restriction was needed to keep the E2 reading frame as the E1^E4 and E8^E2 use the same splicing acceptor site inside the E2 hinge region.

We predicted putative E8 in 308 papillomavirus genomes out of 318 analysed PV genomes (Fig. 5, Fig. 2 in Ref. II and Table 2 in Ref. II). The average length of predicted putative E8 sequences was 34 bp and the average distance from the E1 initiation codon was 376 bp, which matches initial data well (Supplementary Table S1 in Ref. II).

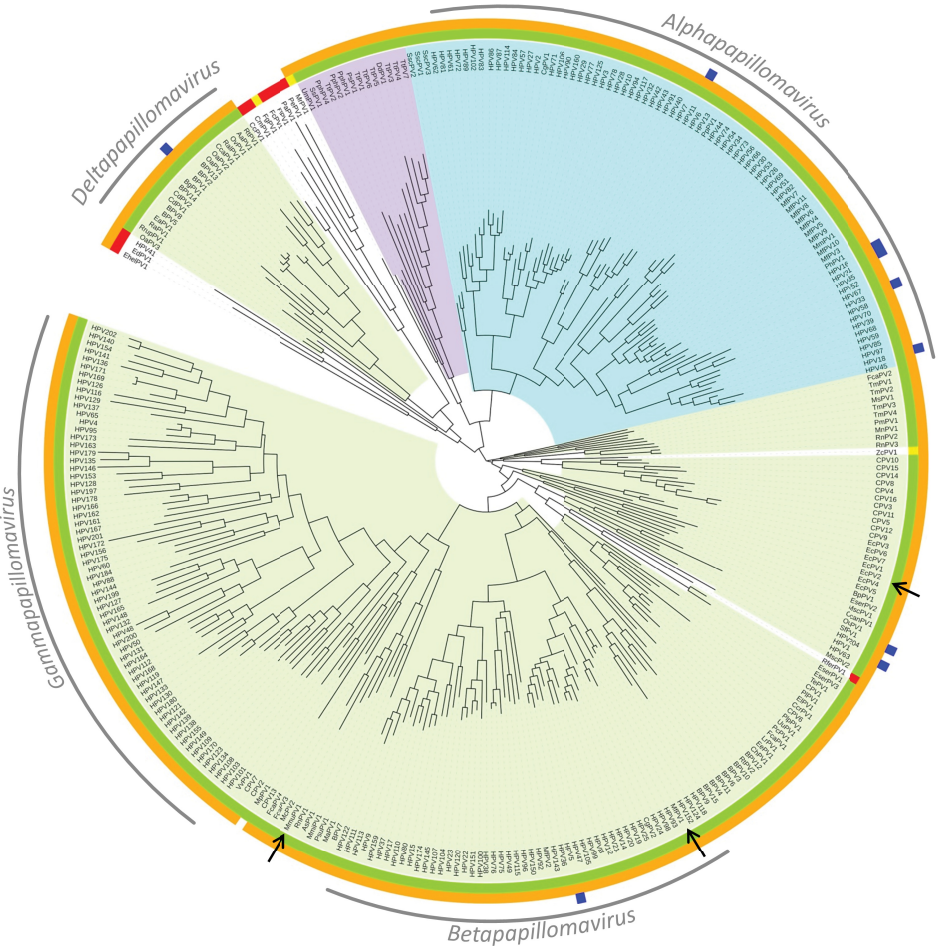

Figure 5. Predicted putative E8 in an E1-based phylogenetic tree. The blue squares represent PVs in which the E8^E2 mRNA sequence has been experimentally confirmed. Arrows are pointing to PVs (MmuPV1, MfPV1, EcPV2) in which E8 was experimentally confirmed after the Ref. II was published. Green arcs of the circle represent PVs in which a potential E8 CDS was detected. Red sections represent E8-deficient PV types, and yellow squares represent PVs in which a predicted E8 CDS was questionable. Orange arcs show PVs that encode E1^E4 (based on the PaVE database). Each coloured full clade (sector) represents a group of PVs with similar putative E8 peptide sequences. The E1 protein alignment was performed with muscle v3.8.31 (default settings, -refine option). The analysis involved 318 E1 amino acid sequences. Evolutionary analyses were conducted in MEGA6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013) (maximum likelihood method, bootstrap 100). The tree was visualised in iTOL (Letunic & Bork, 2007) and is available at
http://itol.embl.de/tree/1934012106546145215056 80 (Ref. II)

Another property observed among predicted E8 sequences was that in only very few cases the splicing donor site diverged from AG|GTA (Table 2 in Ref. II). In addition, all predicted E8 sequences were in the +1 reading frame with respect to E1 (E1 is considered 0 frame). In the initial dataset of 9 PVs, the E8 CDSs were also in the +1 reading frame. Also, the E1 alignment showed that all predicted E8 sequences were coded in the same region (Fig. S10 in Supplementary material in Ref. II). The observation that all experimentally confirmed and our predicted putative E8 sequences located in the +1 reading frame with respect to E1 and all of these were coded in the same region indicate homologous nature of E8.

The E8 was not detected in a total of 10 PV genomes – the greater horseshoe bat PV 1 (*Rhinolophus ferrumequinum* PV 1, RfPV1), the straw-coloured fruit bat PV 1(*Eidolon helvum* PV 1, EhPV1 aka EhelPV1), North American porcupine PV 1 (*Erethizon dorsatum* PV 1, EdPV1), human papillomavirus type 41 (HPV41), and in six sauropsids' PVs. These sauropsids' PVs form a monophyletic clade in a phylogenetic tree of the E1 protein sequence (Fig. 5, Fig. 2 in Ref. II). A putative E8 CDS was detected in the Northern fulmar papillomavirus 1 (*Fulmarus glacialis* PV 1, FgPV1) which also belongs to Sauropsida taxonomic group. However, it is probably a false positive as FgPV1 does not have an annotated E1^E4 splice site (similarly to other sauropsids' PVs in that clade) and predicted E8 peptide is different from others. The fact that these E8-deficient PVs infect Sauropsida, which is a much older taxonomic clade than Mammalia, suggest that E8^E2 emerged later in papillomavirus evolution. This is also confirmed by the fact that PVs recovered from the fish (*Sparus aurata* PV 1 (SaPV1) and GenBank accessions: MH510267, MH616908, MH617143, MH617579) do not have an annotated E4 ORF (López-Bueno et al., 2016; Willemsen & Bravo, 2019) and our algorithm did not detect E8 CDS in the E1 protein-coding genes of these PVs (data not published).

In several cases, our predictions have been confirmed experimentally by other scientists giving credibility to our predictions. For instance, the existence of an mRNA capable of producing E8^E2 has been experimentally confirmed in a mouse PV (*Mus musculus* PV type 1, MmuPV1) from *Pipapillomavirus* genus (Xue et al., 2017), in a macaque PV (*Macaca fascicularis* PV type 1, MfPV1) from *Betapapillomavirus* genus (Tombak et al., 2019) and in a horse PV (*Equus caballus* PV type 2, EcPV2) from *Dyoiotapapillomavirus* genus (Ramsauer, 2015). In all cases, the location of the E8 was the same as we predicted (Supplementary Table S1 in Ref. II).

3.2.1. Distinct E8 groups

The analysis of E8 peptide sequences allowed us to divide them into three distinct groups (Fig. 5 in Ref. II). The smallest group contained only PV types isolated from species of the infraorder *Cetacea*. A slightly bigger group contained PVs mainly from *Alphapapillomavirus* genus. The third, largest group (light

green sector of Fig. 5 and Fig. 2 in Ref. II) is likely the most ancestral and is about 150 millions of years old according to the divergence of PV genera within the group (Shah, Doorbar, & Goldstein, 2010).

Analysis of the embedded functional elements like the previously mentioned E8 can give us more information about the evolutionary relationship between species inside a family. Therefore, correct annotation and detection of these "hidden" elements is a crucial task to fully understand the evolution of viral species. Also, it is important that all of the findings from a research end up in databases used by other scientists. Our results in Ref. II contributed to the E8 annotations update in the papillomavirus episteme (PaVE) database (Van Doorslaer, Li, et al., 2017).

3.3. Identifying embedded elements in protein-coding sequences of viruses (Ref. III)

A protein-coding sequence of a virus may contain various functional embedded elements that play an important role in gene expression and/or in replication. Therefore, it is crucial to detect these elements in viruses to fully understand the molecular biology of an organism. However, these elements can go unnoticed and may be missing from genome annotations like the case with the E8 (Van Doorslaer, Li, et al., 2017). Fortunately, the "comparative analysis" of sequences, produced by the widespread use of second-generation sequencing, allow us to detect these elusive cis-elements. In Ref. III of this thesis, we set out to develop a web tool called cRegions, which is capable of detecting embedded elements at single nucleotide resolution in protein-coding sequences of DNA and RNA viruses.

3.3.1. Developing cRegions

The idea behind cRegions is to compare expected nucleotide proportions to observed nucleotide frequencies at each position in a codon alignment. cRegions uses PAL2NAL (Suyama, Torrents, & Bork, 2006) to convert a protein MSA into a codon alignment, allowing researchers to use all available protein alignment tools which do not support generating codon alignments directly. The result of this setup is that cRegions requires two inputs: an MSA of protein sequences and their respective coding sequences.

The first step is to calculate the expected nucleotide proportions for each position in the codon alignment. The expected values are based on observed amino acid frequencies and preferred codon usage (Fig. 6, Supplementary Materials Fig. S1 in Ref. III and Supplementary Materials Fig. S2 in Ref. III). Incorporating codon usage bias of the same set of protein-coding sequences into a model is reasonable. For instance, it has been shown that viral proteins originated *de novo* by overprinting can be identified by codon usage (Pavesi et al.,

2013). There are also other aspects which are discussed in Ref. III. The only drawback is that the codon usage estimation may be affected by the presence of long dual-coding areas. Also, it is impossible to assess conservation at the nucleic acid level if an amino acid is encoded by a single codon (e.g. methionine and tryptophan). Acquired codon usage bias is adjusted using the Henikoff position-based sequence weights (Supplementary Materials Fig. S1 and S2 in Ref. III). cRegions web tool also provides results with uniform codon usage (all codons of an amino acid have equal expected proportions).

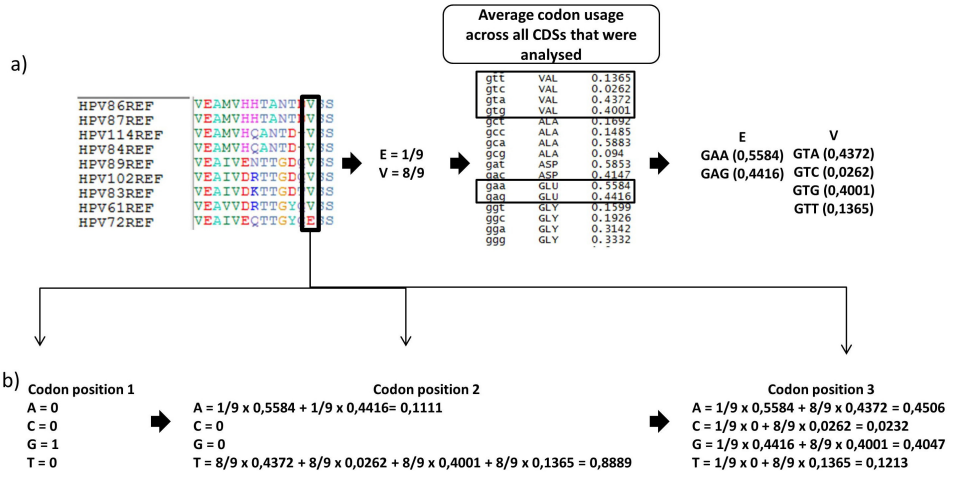

Figure 6. Calculating expected values for single codon in an E1 MSA with cRegions. (a) First, an MSA is generated from the E1 protein sequences. Two different amino acids (valine V and glutamic acid E) are observed in a position highlighted with a black box. Valine is encoded by four codons, and glutamic acid is encoded by two codons (standard codon table). The table in the centre shows the preferred codon usage in E1 genes. Codon usage is calculated from the same set of sequences analysed with cRegion. (b) Second, the expected nucleotide proportions are calculated for each codon position in the MSA based on observed amino acid frequencies and preferred codon usage. (Ref. II)

Next, observed nucleotide frequencies are compared with the expected values. cRegions calculates three different metrics: chi-square goodness of fit test (1), root-mean-square deviation (RMSD) (2) and maximum difference (MAXDIF) (3). The chi-square goodness of fit test evaluates whether the observed distribution of nucleotides is significantly different from the expected distribution. RMSD measures the root mean difference of expected and observed nucleotide proportion. It is frequently used to measure differences between predicted and observed values. MAXDIF selects only one nucleotide from each position that has the highest absolute difference between predicted and observed values. Out of all these, the chi-square goodness of fit test allows us to assess the significance. The method *chisq.test* (1) from R was used to acquire p-values for each position in the codon alignment. The metric displayed on cRegions graphs is the

negative logarithm of the p-value. As the chi-square goodness of fit test is applied on all positions in the codon alignment (or on one-third in the sliding window mode), multiple testing correction is needed. cRegions uses Bonferroni correction (marked with a red horizontal line in Fig. 7).

$$chisq.test(c(\ A_{obs}, C_{obs}, G_{obs}, T_{obs}), p = c(A_{exp}, C_{exp}, G_{exp}, T_{exp}\ )) \quad (1)$$

* The subscript "obs" indicates observed frequencies; the subscript "exp" indicates expected proportions.

$$RMSD = \sqrt{\frac{1}{4}[(A_{obs} - A_{exp})^2 + (C_{obs} - C_{exp})^2 + (G_{obs} - G_{exp})^2 + (T_{obs} - T_{exp})^2]} \quad (2)$$

$$MAXDIF\ = max(|A_{obs} - A_{exp}|, |C_{obs} - C_{exp}|, (G_{obs} - G_{exp}|, |T_{obs} - T_{exp}|) \quad (3)$$

* The subscript "obs" and "exp" indicates observed and expected nucleotide proportions respectively.

Multiple sequence alignments often contain gaps due to deletions or insertions in sequences. By default, cRegions calculates the metric values only for positions in the codon alignment that do not contain more than 20% of gaps. The relatively high threshold is applied in order to guarantee better estimations for expected values. However, the *Allowed Gaps* parameter can be changed by the user. In the sliding window mode, multiple consecutive positions are combined to produce one metric value. In the case of RMSD and MAXDIF, arithmetic mean is calculated. The p-value in the chi-square goodness of fit test is acquired by joining observed and expected values from consecutive positions. By default, if a column in the codon alignment has over 90% of gaps (i.e., there is an insertion in very few sequences) then this position is skipped and the next is included to the current window. Skipping can happen several times in a row. The threshold can be changed through the *Skip Gaps* parameter.

3.3.2. Performance of cRegions

The first version of cRegions (without Henikoff position-based sequence weights) was applied to the E1 gene of papillomaviruses (Fig. 4 in Ref. I and Supplementary Fig. S4–S8 in Ref. I). As we used a non-redundant set of sequences, the effect of weighting would have been negligible. In all PV genera, our method was able to detect E1^E4 splicing donor site, a conserved region first described by Campione-Piccardo (Campione-Piccardo, Montpetit, Grégoire, & Arella, 1991), the E8 CDS, and an E1–E2 overlap. The conserved region described by Campione-Piccardo turned out to be an E8 promoter in HPV16 and 18 (Straub, Fertey, Dreer, Iftner, & Stubenrauch, 2015).

In addition to these common signals, we were able to detect a *Deltapapillomaviruses*-specific signal before the E1–E2 overlap (Fig. 7, Supplementary Fig. S5

in Ref. III). To our best knowledge, it corresponds to the P_{2443} promoter in BPV1 (Hermonat, Spalholz, & Howley, 1988) and a fifth unknown signal in *Gammapapillomaviruses* after the E1 splicing site (Supplementary Fig. S6 in Ref. III). All the analyses were compared to the SynPlot2 (Firth, 2014) which gave identical results (Fig. 4 in Ref. II and Supplementary Fig. S4–S8 in Ref. II).

cRegions was also applied to two different protein-coding genes of Alphaviruses. Alphaviruses are positive-sense single-stranded RNA viruses. We analysed the non-structural and structural polyproteins (Fig. 1 in Ref. III and Fig. 2 in Ref. III). In the non-structural polyprotein, we successfully detected a wide variety of functional elements known in Alphaviruses, including packaging signals in both subgroups and the subgenomic promoter. In the structural polyprotein, our method detected a known frameshift signal (Fig. 2 in Ref. III). Again, all the results were compared and confirmed with SynPlot2 which gave identical results (Supplementary Fig. S4–S7 in Ref. III).

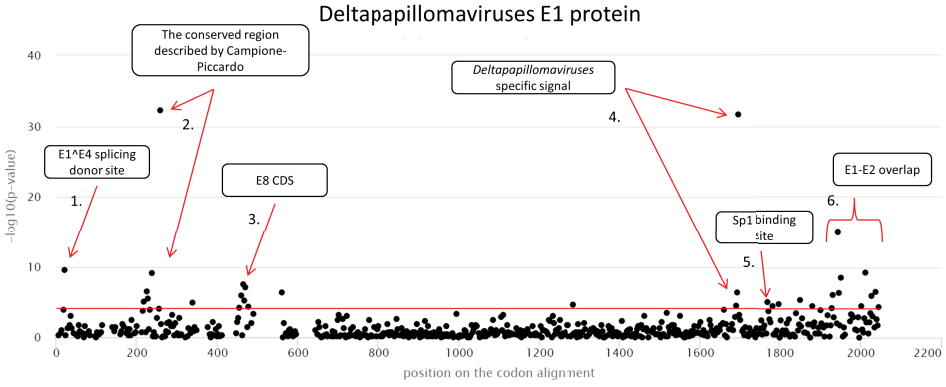

Figure 7. Embedded functional elements in the E1 gene of *Deltapapillomaviruses*. 17 E1 protein sequences were downloaded from PaVE database. Protein sequences were aligned with MAFFT v7.397 using default settings. The results from the chi-square goodness of fit test are displayed using window size 1.

3.3.3. Prerequisites of cRegions

The most important requirement of cRegions is that the codon alignment must be based on a non-redundant set of diverged sequences. It is crucial because non-diverged sequences do not contain enough information and may produce false positive signals. Henikoff position-based sequence weighting reduces potential false positives only if small subset of sequences in a codon alignment is similar (Supplementary Fig. S1and S2 in Ref. III). Further, the quality of the codon alignment is also important. Incorrect alignment of an embedded element may make it undetectable. Another critical requirement for detection of an

embedded element is that it must have been under selection. Furthermore, this element has to be conserved with respect to amino acid sequences.

Still, cRegions can be a valuable tool for a bioinformatician in the field of virology as it is capable of detecting many different non-coding and coding elements in protein-coding genes of viruses. We have shown that cRegions is able to detect: dual-coding regions, splicing sites, internal promoters, packaging signals, and frameshift signals.

CONCLUSION

Viruses harbour enormous genetic and biological diversity and are the most abundant biological entities on Earth. However, the exact origin of viruses is still unknown. Three classic scenarios exist: the virus-first hypothesis, the reduction hypothesis, and the escape hypothesis. The last two have one important implication – most of the genes found in viruses should have their distant homologs in cellular genomes. The occurrence of viral protein domains in cellular organisms may give us information about the origin of viruses. In the current thesis, papillomaviruses (PVs) were used as an example to study the potential origin of a viral family.

We found that PVs have very weak connections to cellular proteins, as only domains from the E1 replication protein had homologs in cellular organisms. However, our study showed that the PVs are clearly evolutionarily related to the *Polyomaviridae* and possibly to *Parvoviridae* family. Polyomaviruses shared structural homologs of capsid protein L1 and two domains of replication protein E1 at SCOP Superfamily level. Members of *Parvoviridae* (ssDNA viruses) shared two replication related domains and, including the extended structural similarity, also the capsid protein with PVs and with *Polyomaviridae*.

In addition, the presence of embedded E8 ORF inside E1 gene of papillomaviruses was studied. The E8 was detected in almost all PV E1 genes, except PVs infecting Sauropsida and fish. These hosts are evolutionarily older than mammalian species, confirming that the E8 emerged after the divergence of the mammalian ancestor.

In the current thesis, a web tool called cRegions was developed for detecting functional embedded elements in protein-coding genes of viruses. We have shown that cRegions is capable of detecting dual-coding regions like E8 and other elements: splicing sites, internal promoters, packaging signals and frameshift signals in protein-coding sequences of DNA and RNA viruses.

SUMMARY IN ESTONIAN

„Papilloomiviirustes esinevate valkude päritolu"

Viirused on parasiitse eluviisiga bioloogilised objektid, mis kasutavad peremeesraku ressursse endi paljundamiseks. Viiruseid võib leida kõikidest biotoopidest ning nende arvukus on enamikes biotoopides suurusjärgu võrra suurem kui prokarüootsetel rakkudel. Viirused on võimelised nakatama organisme kõigist kolmest elusloodu domeenist: arhedest, bakteritest ja eukarüootidest. Lisaks on viirused ka geneetiliselt ja bioloogiliselt väga mitmekesised – nende genoom võib olla RNA või DNA, ühe- või kaheahelaline, lineaarne või tsirkulaarne ja nad kasutavad väga palju erinevaid strateegiaid endi paljundamiseks. Samas on mitmed metagenoomide analüüsid näidanud, et väga suur osa viiruste mitmekesisusest on siiski veel teadmata ja uurimata.

Erinevalt rakulistest organismidest puuduvad usaldusväärsed tõendid viiruste monofüleetilisuse kohta ja nende täpne päritolu on tänini ebaselge. Eksisteerib kolm klassikalist versiooni, kuidas viiruseid võisid tekkida: „viirused esmalt" hüpotees, „reduktsiooni" hüpotees ja „põgenemise" hüpotees. „Viirused esmalt" hüpotees väidab, et viirused tekkisid enne, kui ilmusid esimesed rakud. Antud hüpoteesi toetab mitmekesiste replikatsioonimehhanismide olemasolu viirustes. Vastuväiteks on argument, et kuna kõik tänapäevased viirused vajavad paljunemiseks peremeesrakku, siis viiruste eksisteerimine enne rakke näib ebatõenäoline. „Reduktsiooni" hüpoteesi järgi on viirused kunagi elanud parasiitsete rakuliste organismide järeltulijad. Seda hüpoteesi pakutakse tihti suurte kaheahelalise DNA genoomiga viiruste tekkemehhanismiks. „Põgenemise" hüpotees väidab, et viirused tekkisid DNA või RNA järjestustest, mis saavutasid osaliselt autonoomse paljunemise ja omandasid võime rakust väljuda ning teise rakku siseneda. Viimased kaks hüpoteesi loovad eelduse, et paljud tänapäeva viirustes esinevad geenid võivad omada ühist päritolu mõnede rakulistes organismides leiduvate geenidega.

Antud doktoritöös keskenduti papilloomiviiruse (PV) sugukonna päritolu uuringutele. PV-d on võimelised nakatama mitmesuguseid imetajaid, sealhulgas ka mereimetajaid, linde, roomajaid ja ka kalu. Kõrge riskiga inimese PV-d on vastutavad peaaegu kõigi emakakaelavähi juhtude eest ning on ka paljude teiste kasvajate tekitajad. Tänapäeval on teada üle 400 erineva PV tüübi, sealhulgas ~200 inimese PV-st. Tüüpiline PV genoom kodeerib kaheksat valku (E1, E2, L1, L2, E6, E7, E8^E2, E1^E4). Eelpool nimetatud valkude homoloogide tuvastamine rakulistes organismides võib anda meile informatsiooni PV päritolu kohta.

Järjestuste paariviisiline võrdlemine (nagu BLAST) on olnud klassikaline meetod homoloogide tuvastamiseks, kuid see toimib hästi ainult valgu järjestustega, millede identsus on suurem kui 30%. Suure mutatsioonikiiruse tõttu viirustes võib homoloogsete järjestuste tuvastamine olla keeruline. Varjatud Markovi mudelid (HMM) kombineerituna struktuurse infoga annavad meile siiski võimaluse tuvastada ka kaugemaid homolooge. Struktuurse info

kasutamine on väga oluline, sest valgu struktuur on ajas püsivam kui valgu aminohappeline järjestus.

Töö käigus analüüsiti mitmeid erinevad järjestuste andmebaase tuvastamaks papilloomiviirustes leiduvate valgudomeenide homolooge teistest organismides. Analüüsi käigus leiti rakulistest organismidest ja plasmiididest homolooge vaid papilloomiviiruse replikatsioonivalgule E1, jättes papilloomiviiruste päritolu siiski veel ebaselgeks. Samas näitasid meie tulemused, et papilloomiviirused on evolutsiooniliselt suguluses polüoomiviiruste, aga ka parvoviiruste sugukonnaga. Seosele viitasid nii L1 kapsiidivalk kui ka mõlemad domeenid E1 valgust.

Enamikule PV valkudele ei suudetud homolooge tuvastada. Nende puudumisel võib olla mitmeid põhjuseid lisaks „viirused esmalt" hüpoteesile. Esiteks, andmebaasid sisaldavad ainult sekveneeritud genoome ning PV geenide homolooge omavaid organisme pole veel sekveneeritud. Teiseks põhjuseks võib olla geenide kadumine (gene loss) ehk antud geenid on kõigist tänapäeval eksisteerivatest organismidest kadunud. Kolmandaks, antud liigid, kust viirused pärinesid, on välja surnud. Homoloogide puudumise organismidest võib põhjustada ka *de novo* geenide tekkimine viirustes. Üheks mehhanismiks, kuidas *de novo* geenid tekivad, on topeltkodeerimine (overprinting). Selle protsessi käigus tekib mutatsioonide tõttu uus lugemisraam teise, eelnevalt eksisteerinud geeni, sisse. Mitmed tööd on eksperimentaalselt näidanud, et ka osade papilloomiviiruste E1 geen sisaldab ülekattuvat ehk topeltkodeerivat lugemisraami nimega E8. Töös analüüsiti üle 300 PV genoomi eesmärgiga tuvastada E8 lugemisraam nendes genoomides. E8 tuvastati peaaegu kõigis PV genoomides, välja arvatud PV-des, mis nakatavad roomajaid, linde ja kalu. Antud peremeesorganismid on evolutsiooniliselt vanemad kui imetajad, seega tekkis E8 imetajate PV-des, pärast imetajate lahknemist teistest selgroogsetest.

Eelpool nimetatud topeltkodeeriva lugemisraami, aga ka paljude teiste geenisiseste funktsionaalsete elementide tuvastamine nõuab spetsiifilisi lahendusi. Antud doktoritöö käigus loodi veebitööriist nimega cRegions [http://bioinfo.ut.ee/cRegions/], mis on võimeline tuvastama topeltkodeerivaid lugemisraame viiruslikest geenidest. Lisaks ülekattuvatele lugemisraamidele suudab cRegions tuvastada ka teisi elemente viiruste genoomides, näiteks splaiss-saidid, kapsiidi pakkimise signaalid, subgenoomsed promootorid ja raaminihke signaalid. cRegions, aga ka teised sarnased tööriistad on olulised viiruslike järjestuste uurimisel *in silico*, mille tulemusi saab rakendada hilisemates eksperimentaalsetes analüüsides.

REFERENCES

Abrescia, N. G. A., Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2012). Structure unifies the viral universe. *Annual Review of Biochemistry*, *81*, 795–822.

Abroi, A., & Gough, J. (2011). Are viruses a source of new protein folds for organisms? – Virosphere structure space and evolution. *Bioessays: News and Reviews in Molecular, Cellular and Developmental Biology*, *33*, 626–635.

Aiewsakun, P., Adriaenssens, E. M., Lavigne, R., Kropinski, A. M., & Simmonds, P. (2018). Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *The Journal of General Virology*, *99*, 1331–1343.

Aiewsakun, P., & Katzourakis, A. (2016). Time-Dependent Rate Phenomenon in Viruses. *Journal of Virology*, *90*, 7184–7195.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.

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 Research*, *25*, 3389–3402.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, *42*, D310-4.

Astell, C. R., Mol, C. D., & Anderson, W. F. (1987). Structural and functional homology of parvovirus and papovavirus polypeptides. *The Journal of General Virology*, *68 (Pt 3)*, 885–893.

Balaji, S., & Srinivasan, N. (2001). Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Engineering*, *14*, 219–226.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, *35*, 235–241.

Bamford, D. H. (2003). Do viruses form lineages across different domains of life? *Research in Microbiology*, *154*, 231–236.

Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2005). What does structure tell us about virus evolution? *Current Opinion in Structural Biology*, *15*, 655–663.

Bansho, Y., Furubayashi, T., Ichihashi, N., & Yomo, T. (2016). Host-parasite oscillation dynamics and evolution in a compartmentalized RNA replication system. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 4045–4050.

Belshaw, R., Pybus, O. G., & Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research*, *17*, 1496–1504.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*, 235–242.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P., & Rohwer, F. (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings. Biological Sciences / the Royal Society*, *271*, 565–574.

Brenner, S. E., Chothia, C., & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 6073–6078.

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., … Sullivan, M. B. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science*, *348*, 1261498.

Byrd, D. R., & Matson, S. W. (1997). Nicking by transesterification: the reaction catalysed by a relaxase. *Molecular Microbiology*, *25*, 1011–1022.

Caetano-Anollés, D., Kim, K. M., Mittenthal, J. E., & Caetano-Anollés, G. (2011). Proteome evolution and the metabolic origins of translation and cellular life. *Journal of Molecular Evolution*, *72*, 14–33.

Caetano-Anollés, G., & Nasir, A. (2012). Benefits of using molecular structure and abundance in phylogenomic analysis. *Frontiers in Genetics*, *3*, 172.

Campione-Piccardo, J., Montpetit, M. L., Grégoire, L., & Arella, M. (1991). A highly conserved nucleotide string shared by all genomes of human papillomaviruses. *Virus Genes*, *5*, 349–357.

Carpentier, M., & Chomilier, J. (2019). Protein Multiple Alignments: Sequence-based vs Structure-based Programs. *Bioinformatics*. doi:10.1093/bioinformatics/btz236

Challis, C. J., & Schmidler, S. C. (2012). A stochastic evolutionary model for protein structure alignment and phylogeny. *Molecular Biology and Evolution*, *29*, 3575–3587.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., … Grishin, N. V. (2014). ECOD: an evolutionary classification of protein domains. *PLoS Computational Biology*, *10*, e1003926.

Chen, F., Suttle, C. A., & Short, S. M. (1996). Genetic diversity in marine algal virus communities as revealed by sequence analysis of DNA polymerase genes. *Applied and Environmental Microbiology*, *62*, 2869–2874.

Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings. Biological Sciences / the Royal Society*, *277*, 3809–3817.

Choe, J., Vaillancourt, P., Stenlund, A., & Botchan, M. (1989). Bovine papillomavirus type 1 encodes two forms of a transcriptional repressor: structural and functional analysis of new viral cDNAs. *Journal of Virology*, *63*, 1743–1755.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*, 823–826.

Claverie, J.-M., & Ogata, H. (2009). Ten good reasons not to exclude giruses from the evolutionary picture. *Nature Reviews. Microbiology*, *7*, 615; author reply 615.

Cobián Güemes, A. G., Youle, M., Cantú, V. A., Felts, B., Nulton, J., & Rohwer, F. (2016). Viruses as winners in the game of life. *Annual Review of Virology*, *3*, 197–214.

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews. Genetics*, *9*, 938–950.

Correia, B., Cerqueira, S. A., Beauchemin, C., Pires de Miranda, M., Li, S., Ponnusamy, R., … McVey, C. E. (2013). Crystal structure of the gamma-2 herpesvirus LANA DNA binding domain identifies charged surface residues which impact viral latency. *PLoS Pathogens*, *9*, e1003673.

Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M. J., Neveux, J., … Claustre, H. (1994). Smallest eukaryotic organism. *Nature*, *370*, 255–255.

Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., & Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. *Proteins*, *77 Suppl 9*, 18–28.

Culley, A. (2018). New insight into the RNA aquatic virosphere via viromics. *Virus Research*, *244*, 84–89.

Culley, A. I., Lang, A. S., & Suttle, C. A. (2003). High diversity of unknown picorna-like viruses in the sea. *Nature*, *424*, 1054–1057.

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., … Sillitoe, I. (2017). CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, *45*, D289–D295.

Day, R., Beck, D. A. C., Armen, R. S., & Daggett, V. (2003). A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Science*, *12*, 2150–2160.

Desnues, C., Boyer, M., & Raoult, D. (2012). Sputnik, a virophage infecting the viral domain of life. *Advances in Virus Research*, *82*, 63–89.

De Vuyst, H., Clifford, G. M., Nascimento, M. C., Madeleine, M. M., & Franceschi, S. (2009). Prevalence and type distribution of human papillomavirus in carcinoma and intraepithelial neoplasia of the vulva, vagina and anus: a meta-analysis. *International Journal of Cancer*, *124*, 1626–1636.

Domsic, J. F., Chen, H.-S., Lu, F., Marmorstein, R., & Lieberman, P. M. (2013). Molecular basis for oligomeric-DNA binding and episome maintenance by KSHV LANA. *PLoS Pathogens*, *9*, e1003672.

Doorbar, J., Parton, A., Hartley, K., Banks, L., Crook, T., Stanley, M., & Crawford, L. (1990). Detection of novel splicing patterns in a HPV16-containing keratinocyte cell line. *Virology*, *178*, 254–262.

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews. Genetics*, *9*, 267–276.

Eddy, Sean R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*, 1315–1316.

Eddy, Sean R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, *23*, 205–211.

Eddy, S R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*, 755–763.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., … Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*, D427–D432.

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*, W29–37.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., … Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Research*, *34*, D247–51.

Firth, A. E. (2014). Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Research*, *42*, 12425–12439.

Flores, E. R., & Lambert, P. F. (1997). Evidence for a switch in the mode of human papillomavirus type 16 DNA replication during the viral life cycle. *Journal of Virology*, *71*, 7167–7179.

Forman, D., de Martel, C., Lacey, C. J., Soerjomataram, I., Lortet-Tieulent, J., Bruni, L., … Franceschi, S. (2012). Global burden of human papillomavirus and related diseases. *Vaccine*, *30 Suppl 5*, F12–23.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*, 268–278.

Forterre, P. (2006a). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Research*, *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. *Proceedings*

of the National Academy of Sciences of the United States of America, *103*, 3669–3674.

Fox, N. K., Brenner, S. E., & Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins – extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, *42*, D304–9.

Fu, C., & Johnson, J. E. (2012). Structure and cell biology of archaeal virus STIV. *Current Opinion in Virology*, *2*, 122–127.

Gadagkar, S. R., Rosenberg, M. S., & Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, *304*, 64–74.

Gao, Y., Zhao, H., Jin, Y., Xu, X., & Han, G.-Z. (2017). Extent and evolution of gene duplication in DNA viruses. *Virus Research*, *240*, 161–165.

García-Vallvé, S., Alonso, A., & Bravo, I. G. (2005). Papillomaviruses: different genes have different histories. *Trends in Microbiology*, *13*, 514–521.

Gog, J. R., Afonso, E. D. S., Dalton, R. M., Leclercq, I., Tiley, L., Elton, D., … Digard, P. (2007). Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Research*, *35*, 1897–1907.

Gough, Julian. (2005). Convergent evolution of domain architectures (is rare). *Bioinformatics*, *21*, 1464–1471.

Gough, J, Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, *313*, 903–919.

Gout, J.-F., Thomas, W. K., Smith, Z., Okamoto, K., & Lynch, M. (2013). Large-scale detection of in vivo transcription errors. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 18584–18589.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., & Claverie, J. M. (1993). Ancient conserved regions in new gene sequences and the protein databases. *Science*, *259*, 1711–1716.

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., … Sullivan, M. B. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*. doi:10.1016/j.cell.2019.03.040

Grose, C. (2012). Pangaea and the Out-of-Africa Model of Varicella-Zoster Virus Evolution and Phylogeography. *Journal of Virology*, *86*, 9558–9565.

Grundhoff, A., & Sullivan, C. S. (2011). Virus-encoded microRNAs. *Virology*, *411*, 325–343.

Guerrero, R., Piqueras, M., & Berlanga, M. (2002). Microbial mats and the search for minimal ecosystems. *International Microbiology*, *5*, 177–188.

Hadley, C., & Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, *7*, 1099–1112.

Hammerschmidt, W., & Sugden, B. (1988). Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. *Cell*, *55*, 427–433.

Hanada, K., Suzuki, Y., & Gojobori, T. (2004). A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Molecular Biology and Evolution*, *21*, 1074–1080.

Hegde, N. R., Maddur, M. S., Kaveri, S. V., & Bayry, J. (2009). Reasons to include viruses in the tree of life. *Nature Reviews. Microbiology*, *7*, 615; author reply 615.

Hellert, J., Weidner-Glunde, M., Krausze, J., Richter, U., Adler, H., Fedorov, R., … Lührs, T. (2013). A structural basis for BRD2/4-mediated host chromatin interaction

and oligomer assembly of Kaposi sarcoma-associated herpesvirus and murine gammaherpesvirus LANA proteins. *PLoS Pathogens*, *9*, e1003640.
Hendrickson, R. C., Wang, C., Hatcher, E. L., & Lefkowitz, E. J. (2010). Orthopoxvirus genome evolution: the role of gene loss. *Viruses*, *2*, 1933–1967.
Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*, 10915–10919.
Hermonat, P. L., Spalholz, B. A., & Howley, P. M. (1988). The bovine papillomavirus P2443 promoter is E2 trans-responsive: evidence for E2 autoregulation. *The EMBO Journal*, *7*, 2815–2822.
He, X., & Zhang, J. (2005). Gene complexity and gene duplicability. *Current Biology*, *15*, 1016–1021.
Holland, T. A., Veretnik, S., Shindyalov, I. N., & Bourne, P. E. (2006). Partitioning protein structures into domains: why is it so difficult? *Journal of Molecular Biology*, *361*, 562–590.
Holmes, E. C. (2011). What does virus evolution tell us about virus origins? *Journal of Virology*, *85*, 5247–5251.
Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science*, *273*, 595–603.
Holmes, E. C., & Duchêne, S. (2019). Can sequence phylogenies safely infer the origin of the global virome? *MBio*, *10*. doi:10.1128/mBio.00289-19
Hou, J., Wu, T., Cao, R., & Cheng, J. (2019). Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins*. doi:10.1002/prot.25697
Hubbard, T. J., & Blundell, T. L. (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Engineering*, *1*, 159–171.
Ichihashi, N. (2019). What can we learn from the construction of in vitro replication systems? *Annals of the New York Academy of Sciences*. doi:10.1111/nyas.14042
Ichihashi, N., Usui, K., Kazuta, Y., Sunami, T., Matsuura, T., & Yomo, T. (2013). Darwinian evolution in a translation-coupled RNA replication system within a cell-like compartment. *Nature Communications*, *4*, 2494.
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.
Iranzo, J., Puigbò, P., Lobkovsky, A. E., Wolf, Y. I., & Koonin, E. V. (2016). Inevitability of genetic parasites. *Genome Biology and Evolution*, *8*, 2856–2869.
Isok-Paas, H., Männik, A., Ustav, E., & Ustav, M. (2015). The transcription map of HPV11 in U2OS cells adequately reflects the initial and stable replication phases of the viral genome. *Virology Journal*, *12*, 59.
Iyer, L. M., Koonin, E. V., & Aravind, L. (2003). Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Structural Biology*, *3*, 1.
Jameson, E., Mann, N. H., Joint, I., Sambles, C., & Mühling, M. (2011). The diversity of cyanomyovirus populations along a North-South Atlantic Ocean transect. *The ISME Journal*, *5*, 1713–1721.
Jeckel, S., Loetzsch, E., Huber, E., Stubenrauch, F., & Iftner, T. (2003). Identification of the E9/E2C cDNA and functional characterization of the gene product reveal a new

repressor of transcription and replication in cottontail rabbit papillomavirus. *Journal of Virology*, *77*, 8736–8744.
Jenkins, G. M., Rambaut, A., Pybus, O. G., & Holmes, E. C. (2002). Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of Molecular Evolution*, *54*, 156–165.
Kauffman, K. M., Hussain, F. A., Yang, J., Arevalo, P., Brown, J. M., Chang, W. K., … Polz, M. F. (2018). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*, *554*, 118–122.
Kelley, L. A., & Sternberg, M. J. E. (2015). Partial protein domains: evolutionary insights and bioinformatics challenges. *Genome Biology*, *16*, 100.
Kim, D. Y., Firth, A. E., Atasheva, S., Frolova, E. I., & Frolov, I. (2011). Conservation of a packaging signal and the viral genome RNA packaging mechanism in alphavirus evolution. *Journal of Virology*, *85*, 8022–8036.
Kirsip, H., & Abroi, A. (2019). Protein Structure-Guided Hidden Markov Models (HMMs) as A Powerful Method in the Detection of Ancestral Endogenous Viral Elements. *Viruses*, *11*. doi:10.3390/v11040320
Knipe, D. M. (2013). Chapter 1 | Virology: From Contagium Fluidum to Virome. In P. M. Howley (Trans.), *Fields Virology Volume I* (6th ed., pp. 1–20).
Koonin, Eugene V, & Dolja, V. V. (2006). Evolution of complexity in the viral world: the dawn of a new vision. *Virus Research*, *117*, 1–4.
Koonin, Eugene V, & Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Current Opinion in Virology*, *3*, 546–557.
Koonin, Eugene V, Senkevich, T. G., & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biology Direct*, *1*, 29.
Koonin, Eugene V, Senkevich, T. G., & Dolja, V. V. (2009). Compelling reasons why viruses are relevant for the origin of cells. *Nature Reviews. Microbiology*, *7*, 615; author reply 615.
Koonin, Eugene V, Wolf, Y. I., & Katsnelson, M. I. (2017). Inevitability of the emergence and persistence of genetic parasites caused by evolutionary instability of parasite-free states. *Biology Direct*, *12*, 31.
Koonin, E V, Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology*, *55*, 709–742.
Krupovic, M., & Bamford, D. H. (2008). Virus evolution: how far does the double beta-barrel viral lineage extend? *Nature Reviews. Microbiology*, *6*, 941–948.
Krupovic, M., Dolja, V. V., & Koonin, E. V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews. Microbiology*. doi:10.1038/s41579-019-0205-6
Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Moult, J., Schwede, T., & Tramontano, A. (2018). Evaluation of the template-based modeling in CASP12. *Proteins*, *86 Suppl 1*, 321–334.
Kuchibhatla, D. B., Sherman, W. A., Chung, B. Y. W., Cook, S., Schneider, G., Eisenhaber, B., & Karlin, D. G. (2014). Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently "orphan" viral proteins. *Journal of Virology*, *88*, 10–20.
Kumar, S., & Subramanian, S. (2002). Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 803–808.

Kurg, R., Tekkel, H., Abroi, A., & Ustav, M. (2006). Characterization of the functional activities of the bovine papillomavirus type 1 E2 protein single-chain heterodimers. *Journal of Virology*, *80*, 11218–11225.

Kurg, R., Uusen, P., Sepp, T., Sepp, M., Abroi, A., & Ustav, M. (2009). Bovine papillomavirus type 1 E2 protein heterodimer is functional in papillomavirus DNA replication in vivo. *Virology*, *386*, 353–359.

Kurg, R., Uusen, P., Võsa, L., & Ustav, M. (2010). Human papillomavirus E2 protein with single activation domain initiates HPV18 genome replication, but is not sufficient for long-term maintenance of virus genome. *Virology*, *408*, 159–166.

Labonté, J. M., & Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *The ISME Journal*, *7*, 2169–2177.

Legendre, M., Arslan, D., Abergel, C., & Claverie, J.-M. (2012). Genomics of Megavirus and the elusive fourth domain of Life. *Communicative & Integrative Biology*, *5*, 102–106.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., … Claverie, J.-M. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 4274–4279.

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*, 127–128.

Lipman, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, *227*, 1435–1441.

Liu, H., Fu, Y., Li, B., Yu, X., Xie, J., Cheng, J., … Jiang, D. (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology*, *11*, 276.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., … Zhang, Y.-Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *ELife*, *4*. doi:10.7554/eLife.05378

López-Bueno, A., Mavian, C., Labella, A. M., Castro, D., Borrego, J. J., Alcami, A., & Alejo, A. (2016). Concurrence of Iridovirus, Polyomavirus, and a Unique Member of a New Group of Fish Papillomaviruses in Lymphocystis Disease-Affected Gilthead Sea Bream. *Journal of Virology*, *90*, 8768–8779.

Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H., & Hugenholtz, P. (2019). Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nature Microbiology*. doi:10.1038/s41564-019-0448-z

Ludmir, E. B., & Enquist, L. W. (2009). Viral genomes are part of the phylogenetic tree of life. *Nature Reviews. Microbiology*, *7*, 615; author reply 615.

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*, 345–352.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*, 155–161.

Mayrose, I., Stern, A., Burdelova, E. O., Sabo, Y., Laham-Karam, N., Zamostiano, R., … Pupko, T. (2013). Synonymous site conservation in the HIV-1 genome. *BMC Evolutionary Biology*, *13*, 164.

McBride, Alison A. (2013). The papillomavirus E2 proteins. *Virology*, *445*, 57–79.

McBride, A A, Byrne, J. C., & Howley, P. M. (1989). E2 polypeptides encoded by bovine papillomavirus type 1 form dimers through the common carboxyl-terminal domain: transactivation is mediated by the conserved amino-terminal domain.

Proceedings of the National Academy of Sciences of the United States of America, *86*, 510–514.
McGeoch, D. J., & Gatherer, D. (2005). Integrating reptilian herpesviruses into the family herpesviridae. *Journal of Virology*, *79*, 725–731.
Meehan, B. M., Creelan, J. L., McNulty, M. S., & Todd, D. (1997). Sequence of porcine circovirus DNA: affinities with plant circoviruses. *The Journal of General Virology*, *78 (Pt 1)*, 221–227.
Mendez, J., Blanco, L., & Salas, M. (1997). Protein-primed DNA replication: a transition between two modes of priming by a unique DNA polymerase. *The EMBO Journal*, *16*, 2519–2527.
Milavetz, B. I., & Balakrishnan, L. (2015). Viral epigenetics. *Methods in Molecular Biology*, *1238*, 569–596.
Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*, e121.
Mizuno, C. M., Guyomar, C., Roux, S., Lavigne, R., Rodriguez-Valera, F., Sullivan, M. B., … Krupovic, M. (2019). Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nature Communications*, *10*, 752.
Moreira, D., & Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evolutionary Biology*, *8*, 12.
Moreira, D., & López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews. Microbiology*, *7*, 306–311.
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. *Journal of Molecular Biology*, *247*, 536–540.
Narajczyk, M., Barańska, S., Wegrzyn, A., & Wegrzyn, G. (2007). Switch from theta to sigma replication of bacteriophage lambda DNA: factors involved in the process and a model for its regulation. *Molecular Genetics and Genomics*, *278*, 65–74.
Nasir, A., & Caetano-Anollés, G. (2015). A phylogenomic data-driven exploration of viral origins and evolution. *Science Advances*, *1*, e1500527.
Nasir, A., Kim, K. M., & Caetano-Anolles, G. (2012). Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evolutionary Biology*, *12*, 156.
Nasir, A., Kim, K. M., & Caetano-Anollés, G. (2012). Viral evolution: Primordial cellular origins and late adaptation to parasitism. *Mobile Genetic Elements*, *2*, 247–252.
Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J. L., Sardar, A. J., … Gough, J. (2015). The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic Acids Research*, *43*, D227–33.
O'Malley, M. A., & Koonin, E. V. (2011). How stands the Tree of Life a century and a half after The Origin? *Biology Direct*, *6*, 32.
Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., … Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*, 425–430.
Palermo-Dilts, D. A., Broker, T. R., & Chow, L. T. (1990). Human papillomavirus type 1 produces redundant as well as polycistronic mRNAs in plantar warts. *Journal of Virology*, *64*, 3144–3149.
Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, *171*, 2294–2316.

Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B., & Gough, J. (2019). The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Research*, *47*, D490–D494.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, *284*, 1201–1210.

Pavesi, A., Magiorkinis, G., & Karlin, D. G. (2013). Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. *PLoS Computational Biology*, *9*, e1003162.

Pearson, William R. (2014). BLAST and FASTA similarity searching for multiple sequence alignment. *Methods in Molecular Biology*, *1079*, 75–101.

Pearson, William R. (2016). Finding Protein and Nucleotide Similarities with FASTA. *Current Protocols in Bioinformatics*, *53*, 3.9.1–25.

Pearson, W R, & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *85*, 2444–2448.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., … Abergel, C. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, *341*, 281–286.

Radjef, N., Gordien, E., Ivaniushina, V., Gault, E., Anaïs, P., Drugan, T., … Dény, P. (2004). Molecular phylogenetic analyses indicate a wide and ancient radiation of African hepatitis delta virus, suggesting a deltavirus genus of at least seven major clades. *Journal of Virology*, *78*, 2537–2544.

Ramsauer, A. S. (2015). *Viral and cellular transcription profiles in Equine PapillomavirusType 2 positive squamous cell carcinomas* (Doctoral dissertation). University of Zurich, Institute of Virology. Retrieved from https://doi.org/10.5167/uzh-129304

Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R., & Karlin, D. (2009). Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *Journal of Virology*, *83*, 10719–10736.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., … Claverie, J.-M. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science*, *306*, 1344–1350.

Rebrikov, D. V., Bulina, M. E., Bogdanova, E. A., Vagner, L. L., & Lukyanov, S. A. (2002). Complete genome sequence of a novel extrachromosomal virus-like element identified in planarian Girardia tigrina. *BMC Genomics*, *3*, 15.

Rector, A., Lemey, P., Tachezy, R., Mostmans, S., Ghim, S.-J., Van Doorslaer, K., … Van Ranst, M. (2007). Ancient papillomavirus-host co-speciation in Felidae. *Genome Biology*, *8*, R57.

Rizvi, I., Choudhury, N. R., & Tuteja, N. (2015). Insights into the functional characteristics of geminivirus rolling-circle replication initiator protein and its interaction with host factors affecting viral DNA replication. *Archives of Virology*, *160*, 375–387.

Rohwer, F. (2003). Global phage diversity. *Cell*, *113*, 141.

Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annual Review of Genetics*, *46*, 359–369.

Rotenberg, M. O., Chow, L. T., & Broker, T. R. (1989). Characterization of rare human papillomavirus type 11 mRNAs coding for regulatory and structural proteins, using the polymerase chain reaction. *Virology*, *172*, 489–497.

Rozewicki, J., Li, S., Amada, K. M., Standley, D. M., & Katoh, K. (2019). MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Research*. doi:10.1093/nar/gkz342

Sabath, N., Wagner, A., & Karlin, D. (2012). Evolution of viral proteins originated de novo by overprinting. *Molecular Biology and Evolution*, *29*, 3767–3780.

Sagan, L. (1967). On the origin of mitosing cells. *Journal of Theoretical Biology*, *14*, 255–274.

Salas, M., & de Vega, M. (2016). Protein-Primed Replication of Bacteriophage Φ29 DNA. *The Enzymes*, *39*, 137–167.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral mutation rates. *Journal of Virology*, *84*, 9733–9748.

Sankovski, E., Karro, K., Sepp, M., Kurg, R., Ustav, M., & Abroi, A. (2015). Characterization of the nuclear matrix targeting sequence (NMTS) of the BPV1 E8/E2 protein – the shortest known NMTS. *Nucleus (Austin, Tex.)*, *6*, 289–300.

Sankovski, E., Männik, A., Geimanen, J., Ustav, E., & Ustav, M. (2014). Mapping of betapapillomavirus human papillomavirus 5 transcription and characterization of viral-genome replication function. *Journal of Virology*, *88*, 961–973.

Sealfon, R. S., Lin, M. F., Jungreis, I., Wolf, M. Y., Kellis, M., & Sabeti, P. C. (2015). FRESCo: finding regions of excess synonymous constraint in diverse viruses. *Genome Biology*, *16*, 38.

Shah, S. D., Doorbar, J., & Goldstein, R. A. (2010). Analysis of host-parasite incongruence in papillomavirus evolution using importance sampling. *Molecular Biology and Evolution*, *27*, 1301–1314.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*, 2498–2504.

Shope, R. E., & Hurst, E. W. (1933). Infectious papillomatosis of rabbits : with a note on the histopathology. *The Journal of Experimental Medicine*, *58*, 607–624.

Short, C. M., & Suttle, C. A. (2005). Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Applied and Environmental Microbiology*, *71*, 480–486.

Short, S. M., & Suttle, C. A. (2002). Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Applied and Environmental Microbiology*, *68*, 1290–1296.

Simmonds, P, & Smith, D. B. (1999). Structural constraints on RNA virus evolution. *Journal of Virology*, *73*, 5787–5794.

Simmonds, Peter, Adams, M. J., Benkő, M., Breitbart, M., Brister, J. R., Carstens, E. B., … Zerbini, F. M. (2017). Consensus statement: Virus taxonomy in the age of metagenomics. *Nature Reviews. Microbiology*, *15*, 161–168.

Simon-Loriere, E., & Holmes, E. C. (2013). Gene duplication is infrequent in the recent evolutionary history of RNA viruses. *Molecular Biology and Evolution*, *30*, 1263–1269.

Snijders, P. J., van den Brule, A. J., Schrijnemakers, H. F., Raaphorst, P. M., Meijer, C. J., & Walboomers, J. M. (1992). Human papillomavirus type 33 in a tonsillar carcinoma generates its putative E7 mRNA via two E6* transcript species which are terminated at different early region poly(A) sites. *Journal of Virology*, *66*, 3172–3178.

Sober, E., & Steel, M. (2002). Testing the hypothesis of common ancestry. *Journal of Theoretical Biology*, *218*, 395–408.

Sonnhammer, E. L., Eddy, S. R., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*, 405–420.
Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M., & Poisson, G. (2013). Are we missing half of the viruses in the ocean? *The ISME Journal*, *7*, 672–679.
Straub, E., Fertey, J., Dreer, M., Iftner, T., & Stubenrauch, F. (2015). Characterization of the human papillomavirus 16 E8 promoter. *Journal of Virology*, *89*, 7304–7313.
Stubenrauch, F., Hummel, M., Iftner, T., & Laimins, L. A. (2000). The E8E2C protein, a negative regulator of viral transcription and replication, is required for extra-chromosomal maintenance of human papillomavirus type 31 in keratinocytes. *Journal of Virology*, *74*, 1178–1186.
Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*, 356–361.
Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*, W609–12.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*, *30*, 2725–2729.
Taylor, J. M. (2014). Host RNA circles and the origin of hepatitis delta virus. *World Journal of Gastroenterology*, *20*, 2971–2978.
Taylor, J., & Pelchat, M. (2010). Origin of hepatitis delta virus. *Future Microbiology*, *5*, 393–402.
The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*, D158–D169.
Tischer, I., Gelderblom, H., Vettermann, W., & Koch, M. A. (1982). A very small porcine virus with circular single-stranded DNA. *Nature*, *295*, 64–66.
Todd, A. E., Orengo, C. A., & Thornton, J. M. (1999). Evolution of protein function, from a structural perspective. *Current Opinion in Chemical Biology*, *3*, 548–556.
Tombak, E.-M., Männik, A., Burk, R. D., Le Grand, R., Ustav, E., & Ustav, M. (2019). The molecular biology and HPV drug responsiveness of cynomolgus macaque papillomaviruses support their use in the development of a relevant in vivo model for antiviral drug testing. *Plos One*, *14*, e0211235.
Van Doorslaer, K. (2013). Evolution of the papillomaviridae. *Virology*, *445*, 11–20.
Van Doorslaer, K., Li, Z., Xirasagar, S., Maes, P., Kaminsky, D., Liou, D., … McBride, A. A. (2017). The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic Acids Research*, *45*, D499–D506.
Van Doorslaer, K., & McBride, A. A. (2016). Molecular archeological evidence in support of the repeated loss of a papillomavirus gene. *Scientific Reports*, *6*, 33028.
Van Doorslaer, K., Ruoppolo, V., Schmidt, A., Lescroël, A., Jongsomjit, D., Elrod, M., … Varsani, A. (2017). Unique genome organization of non-mammalian papillomaviruses provides insights into the evolution of viral early proteins. *Virus Evolution*, *3*, vex027.
Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S., Gopalan, V., Mohamoud, Y., … McBride, A. A. (2013). The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. *Nucleic Acids Research*, *41*, D571–8.
Vargason, J. M., Szittya, G., Burgyán, J., & Hall, T. M. T. (2003). Size selective recognition of siRNA by an RNA silencing suppressor. *Cell*, *115*, 799–811.

Veeramachaneni, V., Makałowski, W., Galdzicki, M., Sood, R., & Makałowska, I. (2004). Mammalian overlapping genes: the comparative perspective. *Genome Research*, *14*, 280–286.

Weinert, L. A., Werren, J. H., Aebi, A., Stone, G. N., & Jiggins, F. M. (2009). Evolution and diversity of Rickettsia bacteria. *BMC Biology*, *7*, 6.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 6578–6583.

Wilbur, W. J., & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences of the United States of America*, *80*, 726–730.

Willemsen, A., & Bravo, I. G. (2019). Origin and evolution of papillomavirus (onco)genes and genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*, 20180303.

Williams, K. P., Sobral, B. W., & Dickerman, A. W. (2007). A robust species tree for the alphaproteobacteria. *Journal of Bacteriology*, *189*, 4578–4586.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., … Koonin, E. V. (2018). Origins and evolution of the global RNA virome. *MBio*, *9*. doi:10.1128/mBio.02329-18

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., … Koonin, E. V. (2019). Reply to holmes and duchêne, "can sequence phylogenies safely infer the origin of the global virome?": deep phylogenetic analysis of RNA viruses is highly challenging but not meaningless. *MBio*, *10*. doi:10.1128/mBio.00542-19

Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, *35*, 3375–3382.

Xue, X.-Y., Majerciak, V., Uberoi, A., Kim, B.-H., Gotte, D., Chen, X., … Zheng, Z.-M. (2017). The full transcription map of mouse papillomavirus type 1 (MmuPV1) in mouse wart tissues. *PLoS Pathogens*, *13*, e1006715.

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, *10*, 402–415.

Yutin, N., Wolf, Y. I., & Koonin, E. V. (2014). Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology*, *466–467*, 38–52.

ACKNOWLEDGMENTS

First and foremost, I wish to thank my supervisor Aare Abroi, who showed me the world of viruses and my co-supervisor prof. Maido Remm, who welcomed me to the bioinformatics workgroup. Thank you both for your patience, helpful ideas, and guidance.

Science is not a field for individuals. There is always a team behind an idea. Therefore, I would like to thank Aare, Heleri, and Kevin for their contribution to our publications. Often, the team is much larger than the authors on a paper. Märt I have to thank you for being my roommate for all these years, sharing ideas, giving feedback, showing me the field of finance, taking me along to orienteering events, and of course our unforgettable journeys in the mountains. Similarly, I would like to thank Mihkel, for giving me ideas, inspiration, great feedback on my writings, and our memorable journeys in the field of martial arts. I would like to thank both of them in our joint venture with the MOOC "Geenid – müüdid ja tegelikkus" and foremost being my friends. Now, two M-s out of M3 have (hopefully) defended their thesis – you are next Mihkel.

All the other people in the bioinformatics workgroup are also amazing. I have to thank Lauris for his excellent talks from quantum physics to black holes during our lunch breaks. In addition, I thank Tarmo for the memorable stories about the times in the Soviet Union. I would like to thank (another) Märt for giving me guidance in the world of statistics. In addition, I would like to thank Meelis and Markus from the Institute of Computer Science for introducing me to the world of machine learning. And last but not least, I would like to thank all the other members (incl. previous members) in the bioinformatics workgroup: Reidar, Kairi, Age, Triinu, Erki, Fanny-Dhelia, Oliivika, Aneth, Maarja, Tõnu, Mikk – for the good advice and company. Special thanks go to the people in the Estonian Genome Center whom we shared the building, thank you for the company and memorable discussions.

And of course, I would like to thank my partner Keidy and my brother Siim who have given me support and guidance in my journey to PhD.

PUBLICATIONS

CURRICULUM VITAE

Name: Mikk Puustusmaa
Date of birth: 17.03.1989
E-mail: mikk.puustusmaa@gmail.com

Education:

2014–…	PhD, Gene technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia
2012–2014	Master's studies (*cum laude*), Gene technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia
2008–2011	Bachelor's studies, Gene technology, Institute of Molecular and Cell Biology, Faculty of Science and Technology, University of Tartu, Estonia

Professional employment:

08.2018–01.2019	University of Tartu, Faculty of Science and Technology, Institute of Computer Science, Assistant of Data Science (1,00)
09.2014–01.2015	University of Tartu, Faculty of Mathematics and Computer Science, Institute of Computer Science, Programmer (1,00)

Publications:

1. **Puustusmaa M**., Abroi A. (2019). cRegions – a tool for detecting conserved cis-elements in multiple sequence alignment of diverged coding sequences. PeerJ. 2019 Jan 10;6:e6176. doi: 10.7717/peerj.6176.
2. Roosaare, M., **Puustusmaa, M**., Möls, M., Vaher, M., and Remm, M. (2018). PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ.10.7717/peerj.4588.
3. **Puustusmaa M**.*, Kirsip H.*, Gaston K., Abroi A. (2017). The enigmatic origin of papillomavirus protein domains. Viruses 9. DOI: 10.3390/v9090240.
4. **Puustusmaa M**., Abroi A. (2016). Conservation of the E8 CDS of the E8^E2 protein among mammalian papillomaviruses. J. Gen. Virol 97:2333–2345. DOI: 10.1099/jgv.0.000526.

Supervised dissertations:

2018 Co-supervision of the bachelor's thesis of Brigitta-Robin Raudne "Various pipeline analysis for virus detection" (Gene Technology)

Additional information:
One of the main authors of the online course GEENID – MÜÜDID JA TEGELIKKUS [P2TP.TK.072] in collaboration with Märt Roosaare and Mihkel Vaher.

Awards and stipends:
2018 Lydia ja Felix Krabi Scholarship
2018 Online course GEENID – MÜÜDID JA TEGELIKKUS [P2TP.TK.072] got an award "E-kursuse kvaliteedimärk 2018"
2015 DoRa Programme Activity 6 Scholarship "Semester abroad for Doctoral students".
Collaboration with prof. Julian Gough in University of Bristol
2014 First place in the student project competition at the Institute of Computer Science

ELULOOKIRJELDUS

Nimi: Mikk Puustusmaa
Sünniaeg: 17.03.1989
E-post: mikk.puustusmaa@gmail.com

Haridus:

2014–…	PhD, Geenitehnoloogia eriala, Molekulaar- ja rakubioloogia instituut, Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti
2012–2014	Magistrikraad (*cum laude*), Geenitehnoloogia eriala, Molekulaar- ja rakubioloogia instituut, Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti
2008–2011	Bakalaureusekraad, Geenitehnoloogia eriala, Molekulaar- ja rakubioloogia instituut, Loodus- ja täppisteaduste valdkond, Tartu Ülikool, Eesti

Teenistuskäik:

08.2018–01.2019	Tartu Ülikool, Loodus- ja täppisteaduste valdkond, Arvutiteaduse instituut, andmeteaduse assistent (1,00)
09.2014–01.2015	Tartu Ülikool, Matemaatika-informaatikateaduskond, Arvutiteaduse instituut, programmeerija (1,00)

Teaduspublikatsioonid:

1. **Puustusmaa M**., Abroi A. (2019). cRegions – a tool for detecting conserved cis-elements in multiple sequence alignment of diverged coding sequences. PeerJ. 2019 Jan 10;6:e6176. doi: 10.7717/peerj.6176.
2. Roosaare, M., **Puustusmaa, M**., Möls, M., Vaher, M., and Remm, M. (2018). PlasmidSeeker: identification of known plasmids from bacterial whole genome sequencing reads. PeerJ.10.7717/peerj.4588.
3. **Puustusmaa M**.*, Kirsip H.*, Gaston K., Abroi A. (2017). The enigmatic origin of papillomavirus protein domains. Viruses 9. DOI: 10.3390/v9090240.
4. **Puustusmaa M**., Abroi A. (2016). Conservation of the E8 CDS of the E8^E2 protein among mammalian papillomaviruses. J. Gen. Virol 97:2333–2345. DOI: 10.1099/jgv.0.000526.

Juhendatud väitekirjad:

2018 Brigitta-Robin Raudne bakalaureusetöö „Erinevate töövoogude analüüs viiruste tuvastamisel“ kaasjuhendamine (Geenitehnoloogia)

Loometöö:

Veebikursuse GEENID – MÜÜDID JA TEGELIKKUS [P2TP.TK.072] üks loojatest ja läbiviijatest koostöös Märt Roosaare ja Mihkel Vaheriga.

Teaduspreemiad ja tunnustused:

2018 Lydia ja Felix Krabi stipendiumi

2018 Veebikursuse GEENID – MÜÜDID JA TEGELIKKUS [P2TP.TK.072] E-kursuse kvaliteedimärk 2018

2015 DoRa programmi tegevus 6 stipendium „Doktorantide semester välismaal“. Välislähetus Bristoli Ülikoolis, koostöö prof. Julian Gough'iga sealses bioinformaatika töögrupis.

2014 Tartu Ülikooli arvutiteaduse instituudi tudengiprojektide konkursi I koht.

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand**. The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro.* Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan**. Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.
41. **Sulev Ingerpuu**. Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.

42. **Veljo Kisand**. Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa**. Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa**. Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik**. Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo**. Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo**. Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots**. Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero**. Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees**. Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks**. Cholecystokinin (CCK) – induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild**. Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva**. Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna**. Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro**. Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane**. Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm**. Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg**. Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild**. The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu**. Studies of the TOL plasmid transcription factor XylS. Tartu, 2000, 88 p.
61. **Dina Lepik**. Modulation of viral DNA replication by tumor suppressor protein p53. Tartu, 2000, 106 p.

62. **Kai Vellak**. Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu, 2000, 122 p.
63. **Jonne Kotta**. Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu, 2000, 160 p.
64. **Georg Martin**. Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000, 139 p.
65. **Silvia Sepp**. Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira**. On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000, 96 p.
67. **Priit Zingel**. The role of planktonic ciliates in lake ecosystems. Tartu, 2001, 111 p.
68. **Tiit Teder**. Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu, 2001, 122 p.
69. **Hannes Kollist**. Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu, 2001, 80 p.
70. **Reet Marits**. Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu, 2001, 112 p.
71. **Vallo Tilgar**. Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002, 126 p.
72. **Rita Hõrak**. Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002, 108 p.
73. **Liina Eek-Piirsoo**. The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002, 74 p.
74. **Krõõt Aasamaa**. Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002, 110 p.
75. **Nele Ingerpuu**. Bryophyte diversity and vascular plants. Tartu, 2002, 112 p.
76. **Neeme Tõnisson**. Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002, 124 p.
77. **Margus Pensa**. Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003, 110 p.
78. **Asko Lõhmus**. Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003, 168 p.
79. **Viljar Jaks**. p53 – a switch in cellular circuit. Tartu, 2003, 160 p.
80. **Jaana Männik**. Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003, 140 p.
81. **Marek Sammul**. Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003, 159 p
82. **Ivar Ilves**. Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003, 89 p.

83. **Andres Männik**. Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003, 109 p.
84. **Ivika Ostonen**. Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003, 158 p.
85. **Gudrun Veldre**. Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003, 199 p.
86. **Ülo Väli**. The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004, 159 p.
87. **Aare Abroi**. The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004, 135 p.
88. **Tiina Kahre**. Cystic fibrosis in Estonia. Tartu, 2004, 116 p.
89. **Helen Orav-Kotta**. Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004, 117 p.
90. **Maarja Öpik**. Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004, 175 p.
91. **Kadri Tali**. Species structure of *Neotinea ustulata*. Tartu, 2004, 109 p.
92. **Kristiina Tambets**. Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004, 163 p.
93. **Arvi Jõers**. Regulation of p53-dependent transcription. Tartu, 2004, 103 p.
94. **Lilian Kadaja**. Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004, 103 p.
95. **Jaak Truu**. Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004, 128 p.
96. **Maire Peters**. Natural horizontal transfer of the *pheBA* operon. Tartu, 2004, 105 p.
97. **Ülo Maiväli**. Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004, 130 p.
98. **Merit Otsus**. Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004, 103 p.
99. **Mikk Heidemaa**. Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004, 167 p.
100. **Ilmar Tõnno**. The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004, 111 p.
101. **Lauri Saks**. Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004, 144 p.
102. **Siiri Rootsi**. Human Y-chromosomal variation in European populations. Tartu, 2004, 142 p.
103. **Eve Vedler**. Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.

104. **Andres Tover**. Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 126 p.
105. **Helen Udras**. Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005, 100 p.
106. **Ave Suija**. Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005, 162 p.
107. **Piret Lõhmus**. Forest lichens and their substrata in Estonia. Tartu, 2005, 162 p.
108. **Inga Lips**. Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005, 156 p.
109. **Kaasik, Krista**. Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005, 121 p.
110. **Juhan Javoiš**. The effects of experience on host acceptance in ovipositing moths. Tartu, 2005, 112 p.
111. **Tiina Sedman**. Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005, 103 p.
112. **Ruth Aguraiuja**. Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005, 112 p.
113. **Riho Teras**. Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005, 106 p.
114. **Mait Metspalu**. Through the course of prehistory in India: tracing the mtDNA trail. Tartu, 2005, 138 p.
115. **Elin Lõhmussaar**. The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006, 124 p.
116. **Priit Kupper**. Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006, 126 p.
117. **Heili Ilves**. Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006, 120 p.
118. **Silja Kuusk**. Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006, 126 p.
119. **Kersti Püssa**. Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006, 90 p.
120. **Lea Tummeleht**. Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006, 94 p.
121. **Toomas Esperk**. Larval instar as a key element of insect growth schedules. Tartu, 2006, 186 p.
122. **Harri Valdmann**. Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers**. Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli**. Gata3 and Gata2 in inner ear development. Tartu, 2007, 123 p.

125. **Kai Rünk**. Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007, 143 p.
126. **Aveliina Helm**. Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007, 89 p.
127. **Leho Tedersoo**. Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007, 233 p.
128. **Marko Mägi**. The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007, 135 p.
129. **Valeria Lulla**. Replication strategies and applications of Semliki Forest virus. Tartu, 2007, 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007, 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007, 171 p.
132. **Tatjana Krama**. Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007, 112 p.
133. **Signe Saumaa**. The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007, 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007, 96 p.
135. **Priit Kilgas**. Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007, 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007, 95 p.
137. **Kärt Padari**. Protein transduction mechanisms of transportans. Tartu, 2008, 128 p.
138. **Siiri-Lii Sandre**. Selective forces on larval colouration in a moth. Tartu, 2008, 125 p.
139. **Ülle Jõgar**. Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008, 99 p.
140. **Lauri Laanisto**. Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008, 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008, 105 p.
142. **Birgot Paavel**. Bio-optical properties of turbid lakes. Tartu, 2008, 175 p.
143. **Kaire Torn**. Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg**. Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd**. Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.

146. **Lauri Saag**. Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.
147. **Ulvi Karu**. Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm**. Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks**. Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu**. Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš**. Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko**. Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast**. Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats**. Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova**. The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik**. Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver**. Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja**. Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast**. Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak**. Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.
162. **Triinu Remmel**. Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe**. Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe**. Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.

165. **Liisa Metsamaa**. Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.
166. **Pille Säälik**. The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil**. Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik**. Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark**. Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap**. Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan**. Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe**. Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi**. Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson**. Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts**. Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis**. Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov**. Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster**. Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap**. Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar**. Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül**. Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.
182. **Arto Pulk**. Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü**. Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla**. Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.

185. **Gyaneshwer Chaubey**. The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.
186. **Katrin Kepp**. Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber**. The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro**. The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold**. Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert**. Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu**. Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik**. ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida*. Tartu, 2010, 118 p.
193. **Siim Sõber**. Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper**. Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak**. Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo**. Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel**. Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus**. Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius**. Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv**. Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk**. Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.
202. **Arno Põllumäe**. Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht**. Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.

205. **Teele Jairus**. Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.
206. **Kessy Abarenkov**. PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova**. Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar**. The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild**. Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag**. Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla**. Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve**. Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar**. Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp**. Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero**. Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram**. Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents**. Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.
221. **Katrin Männik**. Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous**. Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu**. Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.

224. **Nele Tamberg**. Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.
225. **Tõnu Esko**. Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula**. Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu**. Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem**. The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen**. Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv**. The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi**. Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais**. Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja**. Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis**. Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme**. Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla**. Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke**. Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan**. Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm**. Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.
240. **Liina Kangur**. High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik**. Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski**. The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja**. Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus**. Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.

245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.
246. **Mario Plaas**. Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov**. Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks**. Arbuscular mycorrhizal fungal diversity patterns in boreo-nemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina**. The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau**. Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg**. Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis**. Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja**. Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg**. Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon**. A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.
259. **Andrei Nikonov**. RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets**. Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and inter-annual patterns. Tartu, 2014, 115 p.

263. **Külli Lokko**. Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.
265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun**. Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick**. Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme**. The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm**. Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve**. Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov**. The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook**. Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi**. Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu**. Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson**. Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.
277. **Priit Adler**. Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas**. Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.
279. **Silja Laht**. Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices. Tartu, 2015, 100 p.
280. **Martin Kesler**. Biological characteristics and restoration of Atlantic salmon *Salmo salar* populations in the Rivers of Northern Estonia. Tartu, 2015, 97 p.
281. **Pratyush Kumar Das**. Biochemical perspective on alphaviral nonstructural protein 2: a tale from multiple domains to enzymatic profiling. Tartu, 2015, 205 p

282. **Priit Palta**. Computational methods for DNA copy number detection. Tartu, 2015, 130 p.
283. **Julia Sidorenko**. Combating DNA damage and maintenance of genome integrity in pseudomonads. Tartu, 2015, 174 p.
284. **Anastasiia Kovtun-Kante**. Charophytes of Estonian inland and coastal waters: distribution and environmental preferences. Tartu, 2015, 97 p.
285. **Ly Lindman**. The ecology of protected butterfly species in Estonia. Tartu, 2015, 171 p.
286. **Jaanis Lodjak**. Association of Insulin-like Growth Factor I and Corticosterone with Nestling Growth and Fledging Success in Wild Passerines. Tartu, 2016, 113 p.
287. **Ann Kraut**. Conservation of Wood-Inhabiting Biodiversity – Semi-Natural Forests as an Opportunity. Tartu, 2016, 141 p.
288. **Tiit Örd**. Functions and regulation of the mammalian pseudokinase TRIB3. Tartu, 2016, 182. p.
289. **Kairi Käiro**. Biological Quality According to Macroinvertebrates in Streams of Estonia (Baltic Ecoregion of Europe): Effects of Human-induced Hydromorphological Changes. Tartu, 2016, 126 p.
290. **Leidi Laurimaa**. *Echinococcus multilocularis* and other zoonotic parasites in Estonian canids. Tartu, 2016, 144 p.
291. **Helerin Margus**. Characterization of cell-penetrating peptide/nucleic acid nanocomplexes and their cell-entry mechanisms. Tartu, 2016, 173 p.
292. **Kadri Runnel**. Fungal targets and tools for forest conservation. Tartu, 2016, 157 p.
293. **Urmo Võsa**. MicroRNAs in disease and health: aberrant regulation in lung cancer and association with genomic variation. Tartu, 2016, 163 p.
294. **Kristina Mäemets-Allas**. Studies on cell growth promoting AKT signaling pathway – a promising anti-cancer drug target. Tartu, 2016, 146 p.
295. **Janeli Viil**. Studies on cellular and molecular mechanisms that drive normal and regenerative processes in the liver and pathological processes in Dupuytren's contracture. Tartu, 2016, 175 p.
296. **Ene Kook**. Genetic diversity and evolution of *Pulmonaria angustifolia* L. and *Myosotis laxa sensu lato* (Boraginaceae). Tartu, 2016, 106 p.
297. **Kadri Peil**. RNA polymerase II-dependent transcription elongation in *Saccharomyces cerevisiae*. Tartu, 2016, 113 p.
298. **Katrin Ruisu**. The role of RIC8A in mouse development and its function in cell-matrix adhesion and actin cytoskeletal organisation. Tartu, 2016, 129 p.
299. **Janely Pae**. Translocation of cell-penetrating peptides across biological membranes and interactions with plasma membrane constituents. Tartu, 2016, 126 p.
300. **Argo Ronk**. Plant diversity patterns across Europe: observed and dark diversity. Tartu, 2016, 153 p.

301. **Kristiina Mark**. Diversification and species delimitation of lichenized fungi in selected groups of the family Parmeliaceae (Ascomycota). Tartu, 2016, 181 p.
302. **Jaak-Albert Metsoja**. Vegetation dynamics in floodplain meadows: influence of mowing and sediment application. Tartu, 2016, 140 p.
303. **Hedvig Tamman**. The GraTA toxin-antitoxin system of *Pseudomonas putida*: regulation and role in stress tolerance. Tartu, 2016, 154 p.
304. **Kadri Pärtel**. Application of ultrastructural and molecular data in the taxonomy of helotialean fungi. Tartu, 2016, 183 p.
305. **Maris Hindrikson**. Grey wolf (*Canis lupus*) populations in Estonia and Europe: genetic diversity, population structure and -processes, and hybridization between wolves and dogs. Tartu, 2016, 121 p.
306. **Polina Degtjarenko**. Impacts of alkaline dust pollution on biodiversity of plants and lichens: from communities to genetic diversity. Tartu, 2016, 126 p.
307. **Liina Pajusalu**. The effect of CO_2 enrichment on net photosynthesis of macrophytes in a brackish water environment. Tartu, 2016, 126 p.
308. **Stoyan Tankov**. Random walks in the stringent response. Tartu, 2016, 94 p.
309. **Liis Leitsalu**. Communicating genomic research results to population-based biobank participants. Tartu, 2016, 158 p.
310. **Richard Meitern**. Redox physiology of wild birds: validation and application of techniques for detecting oxidative stress. Tartu, 2016, 134 p.
311. **Kaie Lokk**. Comparative genome-wide DNA methylation studies of healthy human tissues and non-small cell lung cancer tissue. Tartu, 2016, 127 p.
312. **Mihhail Kurašin**. Processivity of cellulases and chitinases. Tartu, 2017, 132 p.
313. **Carmen Tali**. Scavenger receptors as a target for nucleic acid delivery with peptide vectors. Tartu, 2017, 155 p.
314. **Katarina Oganjan**. Distribution, feeding and habitat of benthic suspension feeders in a shallow coastal sea. Tartu, 2017, 132 p.
315. **Taavi Paal**. Immigration limitation of forest plants into wooded landscape corridors. Tartu, 2017, 145 p.
316. **Kadri Õunap**. The Williams-Beuren syndrome chromosome region protein WBSCR22 is a ribosome biogenesis factor. Tartu, 2017, 135 p.
317. **Riin Tamm**. In-depth analysis of factors affecting variability in thiopurine methyltransferase activity. Tartu, 2017, 170 p.
318. **Keiu Kask**. The role of RIC8A in the development and regulation of mouse nervous system. Tartu, 2017, 184 p.
319. **Tiia Möller**. Mapping and modelling of the spatial distribution of benthic macrovegetation in the NE Baltic Sea with a special focus on the eelgrass *Zostera marina* Linnaeus, 1753. Tartu, 2017, 162 p.
320. **Silva Kasela**. Genetic regulation of gene expression: detection of tissue- and cell type-specific effects. Tartu, 2017, 150 p.

321. **Karmen Süld**. Food habits, parasites and space use of the raccoon dog *Nyctereutes procyonoides*: the role of an alien species as a predator and vector of zoonotic diseases in Estonia. Tartu, 2017, p.
322. **Ragne Oja**. Consequences of supplementary feeding of wild boar – concern for ground-nesting birds and endoparasite infection. Tartu, 2017, 141 p.
323. **Riin Kont**. The acquisition of cellulose chain by a processive cellobiohydrolase. Tartu, 2017, 117 p.
324. **Liis Kasari**. Plant diversity of semi-natural grasslands: drivers, current status and conservation challenges. Tartu, 2017, 141 p.
325. **Sirgi Saar**. Belowground interactions: the roles of plant genetic relatedness, root exudation and soil legacies. Tartu, 2017, 113 p.
326. **Sten Anslan**. Molecular identification of Collembola and their fungal associates. Tartu, 2017, 125 p.
327. **Imre Taal**. Causes of variation in littoral fish communities of the Eastern Baltic Sea: from community structure to individual life histories. Tartu, 2017, 118 p.
328. **Jürgen Jalak**. Dissecting the Mechanism of Enzymatic Degradation of Cellulose Using Low Molecular Weight Model Substrates. Tartu, 2017, 137 p.
329. **Kairi Kiik**. Reproduction and behaviour of the endangered European mink (*Mustela lutreola*) in captivity. Tartu, 2018, 112 p.
330. **Ivan Kuprijanov**. Habitat use and trophic interactions of native and invasive predatory macroinvertebrates in the northern Baltic Sea. Tartu, 2018, 117 p.
331. **Hendrik Meister**. Evolutionary ecology of insect growth: from geographic patterns to biochemical trade-offs. Tartu, 2018, 147 p.
332. **Ilja Gaidutšik.** Irc3 is a mitochondrial branch migration enzyme in *Saccharomyces cerevisiae.* Tartu, 2018, 161 p.
333. **Lena Neuenkamp**. The dynamics of plant and arbuscular mycorrhizal fungal communities in grasslands under changing land use. Tartu, 2018, 241 p.
334. **Laura Kasak.** Genome structural variation modulating the placenta and pregnancy maintenance. Tartu, 2018, 181 p.
335. **Kersti Riibak.** Importance of dispersal limitation in determining dark diversity of plants across spatial scales. Tartu, 2018, 133 p.
336. **Liina Saar.** Dynamics of grassland plant diversity in changing landscapes. Tartu, 2018, 206 p.
337. **Hanna Ainelo.** Fis regulates *Pseudomonas putida* biofilm formation by controlling the expression of *lapA*. Tartu, 2018, 143 p.
338. **Natalia Pervjakova.** Genomic imprinting in complex traits. Tartu, 2018, 176 p.
339. **Andrio Lahesaare.** The role of global regulator Fis in regulating the expression of *lapF* and the hydrophobicity of soil bacterium *Pseudomonas putida*. Tartu, 2018, 124 p.

340. **Märt Roosaare.** *K*-mer based methods for the identification of bacteria and plasmids. Tartu, 2018, 117 p.
341. **Maria Abakumova.** The relationship between competitive behaviour and the frequency and identity of neighbours in temperate grassland plants. Tartu, 2018, 104 p.
342. **Margus Vilbas.** Biotic interactions affecting habitat use of myrmecophilous butterflies in Northern Europe. Tartu, 2018, 142 p.
343. **Liina Kinkar**. Global patterns of genetic diversity and phylogeography of *Echinococcus granulosus* sensu stricto – a tapeworm species of significant public health concern. Tartu, 2018, 147 p.
344. **Teivi Laurimäe.** Taxonomy and genetic diversity of zoonotic tapeworms in the species complex of *Echinococcus granulosus* sensu lato. Tartu, 2018, 143 p.
345. **Tatjana Jatsenko.** Role of translesion DNA polymerases in mutagenesis and DNA damage tolerance in Pseudomonads. Tartu, 2018, 216 p.
346. **Katrin Viigand.** Utilization of α-glucosidic sugars by *Ogataea* (*Hansenula*) *polymorpha.* Tartu, 2018, 148 p.
347. **Andres Ainelo.** Physiological effects of the *Pseudomonas putida* toxin grat. Tartu, 2018, 146 p.
348. **Killu Timm**. Effects of two genes (DRD4 and SERT) on great tit (*Parus major*) behaviour and reproductive traits. Tartu, 2018, 117 p.
349. **Petr Kohout**. Ecology of ericoid mycorrhizal fungi. Tartu, 2018, 184 p.
350. **Gristin Rohula-Okunev**. Effects of endogenous and environmental factors on night-time water flux in deciduous woody tree species. Tartu, 2018, 184 p.
351. **Jane Oja.** Temporal and spatial patterns of orchid mycorrhizal fungi in forest and grassland ecosystems. Tartu, 2018, 102 p.
352. **Janek Urvik**. Multidimensionality of aging in a long-lived seabird. Tartu, 2018, 135 p.
353. **Lisanna Schmidt**. Phenotypic and genetic differentiation in the hybridizing species pair *Carex flava* and *C. viridula* in geographically different regions. Tartu, 2018, 133 p.
354. **Monika Karmin.** Perspectives from human Y chromosome – phylogeny, population dynamics and founder events. Tartu, 2018, 168 p.
355. **Maris Alver.** Value of genomics for atherosclerotic cardiovascular disease risk prediction. Tartu, 2019, 148 p.
356. **Lehti Saag**. The prehistory of Estonia from a genetic perspective: new insights from ancient DNA. Tartu, 2019, 171 p.
357. **Mari-Liis Viljur**. Local and landscape effects on butterfly assemblages in managed forests. Tartu, 2019, 115 p.
358. **Ivan Kisly.** The pleiotropic functions of ribosomal proteins eL19 and eL24 in the budding yeast ribosome. Tartu, 2019, 170 p.