universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

"Algorithmic Approaches on Structural RNA Evolution in Viruses"

verfasst von / submitted by

Roman Wilhelm Ochsenreiter, BSc MSc

angestrebter akademischer Grad /
in partial fulfillment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796 605 419
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Chemie
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

ABSTRACT

Viruses are a ubiquitous and evolutionary ancient component of earth's biosphere. Their enormous significance is illustrated by the fact that all cellular organisms, from unicellular organisms to the largest mammals, are infected by viruses, where for each cellular species there is at least one specific virus. In order to succeed, Over the course of evolution viruses have invented a multitude of strategies to adapt to this enormous number of hosts and environmental conditions. Of particular interest are RNA viruses, as they have developed unique ways to carry out otherwise classically protein-based functionalities directly through RNA, where functionality mediated by RNA molecules is often strongly determined by the structure of the RNA.

Despite the remarkable impact of functional RNA structure, only a comparably small number of viral RNA elements have been characterized in detail. A comparative analysis of RNA structure, within or between viral species, can be performed *in silico* and provide valuable insights into the characteristics of structural elements and their evolutionary dynamics. Crucial for a thorough understanding of the structural evolution of viral RNA is the choice of a suitable algorithmic approach, which is able to account for the extreme rate of evolution of RNA viruses.

In this thesis, detailed *in silico* studies on the relevance and influence of RNA structure in different viruses and virus families are presented. In the first paper, the Chikungunya virus, which is endemic to tropical areas, is used demonstrate the *de novo* discovery of structural elements in a virus. Particular focus is placed on differences between individual lineages of the Chikungunya virus and their implications for its epidemiological history and the emergence of today's dominant subtypes.

In a further work, the structural architecture of the untranslated regions of all flaviviruses is examined and deconstructed in detail. Existing models of the family of exoribonuclease-blocking RNAs, which are essential to flaviviruses, are improved or replaced with more accurate models. In addition, new structural elements are identified, and a large number of known or new elements are annotated in many previously poorly studied flaviviruses. In the process, the suitability of comparative methods for analysis of viral structural RNA is demonstrated and evidence for the relevance and validity of the *in silico* results is presented.

Finally, a systematic analysis of the frequency and impact of conserved RNA structure in all known viruses is presented. General evolutionary trends in the development of viral RNA structure are presented to a broad audience in the form of the RNASIV database for further analysis.

ZUSAMMENFASSUNG

Viren sind ein allgegenwärtiger und evolutionär sehr alter Bestandteil der Biosphäre der Erde. Ihre enorme Bedeutung wird dadurch verdeutlicht dass sämtliche zelluläre Organismen, von Einzellern bis zu den größten Säugetieren, von Viren identifiziert werden wobei für jede zelluläre Spezies mindestens ein spezifisches Virus existiert. Im Verlauf der Evolution haben Viren mannigfaltige Strategien zur Anpassung an diese enorme Zahl an Wirten und Umweltbedingungen entwickelt um sich effizient vermehren zu können. Von besonderem Interesse sind dabei RNA-Viren, da sie einzigartige Methoden entwickelt haben um sonst klassisch Protein-basierte Funktionalitäten direkt durch RNA auszuführen. Die Funktionalität von RNA wird dabei entscheidend durch die Struktur des Moleküls geprägt.

Trotz der bemerkenswerten Bedeutung von funktioneller RNA-Struktur, wurde bisher nur eine überschaubare Anzahl an viralen RNA-Elementen eingehend charakterisiert. Eine vergleichende Analyse von RNA-Struktur, innerhalb oder zwischen Virenspezies, kann dabei *in silico* durchgeführt werden und wertvolle Erkenntnisse über die Charakteristika von Strukturelementen und ihre evolutionäre Dynamik liefern. Entscheidend für ein eingehendes Verständnis der strukturellen Evolution von viraler RNA ist dabei die Wahl eines geeigneten algorithmischen Ansatzes, welcher die extreme Evolutionsrate von RNA-Viren berücksichtigen kann.

Es werden detaillierte *in silico* Studien zur Relevanz und dem Einfluss von RNA Struktur in verschiedenen Viren bzw. Virenfamilien präsentiert. In der ersten Arbeit wird das *Chikungunya Virus*, ein in tropischen Gebieten endemisches Virus, als Beispiel für eine Studie zur *de-novo* Entdeckung von Strukturelementen präsentiert. Besonderer Fokus wird darauf gelegt Unterschiede zwischen einzelnen Varianten des Chikungunya-Virus aufzuzeigen und daran seine epidemiologische Geschichte und die Entstehung der aktuell dominanten Subtypen zu rekonstruieren.

In einer weiteren Arbeit wird die strukturelle Architektur der Untranslatierten Regionen aller Flaviviren eingehend untersucht. Bestehende Modelle der in Flaviviren bedeutsamen Familie der Exoribonuklease-blockierenden RNAs werden stark verbessert oder ersetzt. Darüber hinaus werden neue Strukturelemente identifiziert, und eine Vielzahl von bestehenden und neuen Elementen in zahlreichen Vertretern bisher schwach untersuchter Flaviviren annotiert. Dabei wird die Eignung vergleichender Analysemethoden für derartige Probleme demonstriert und stichhaltige Indizien für die Relevanz der *in silico* Ergebnisse dargelegt.

Zuletzt wird ein eine systematische Analyse der Häufigkeit von konservierter RNA-Struktur in allen bekannten Viren präsentiert. Es werden generelle evolutionäre Trends in der Entwicklung viraler RNA Struktur aufgezeigt und einem breiten Publikum in Form der RNASIV-Datenbank für weitere Analyse zur Verfügung gestellt.

CONTENTS

1 PREFACE 1

2 BACKGROUND 2
2.1 Viruses . . . 2
2.2 Virus Structure . . . 5
2.3 Virus Systematics . . . 7
2.4 RNA and Viruses . . . 11
2.5 Flaviviruses . . . 13
2.6 Chikungunya Virus . . . 19

3 METHODS 21
3.1 RNA folding . . . 21
3.2 Homology Search . . . 24
3.3 Evaluating RNA structure . . . 28

4 GOALS OF THIS THESIS 31

5 RNA ARCHITECTURE OF CHIKUNGUNYA VIRUS LINEAGES 34
5.1 Motivation . . . 34
5.2 Added value . . . 34
5.3 Reprint of: Viruses 2019, 11, 798. . . . 35

6 FUNCTIONAL STRUCTURES IN THE UTRS OF FVS 52
6.1 Moviation . . . 52
6.2 Added value . . . 52
6.3 Reprint of: Viruses, 2019, 11(3), 298. . . . 53

7 CONSERVED SECONDARY STRUCTURES IN VIRAL RNAS 73
7.1 Motivation . . . 73
7.2 Added value . . . 73
7.3 Reprint of: Viruses 2019, 11, 401. . . . 74

8 SUMMARY AND CONCLUSION 97
8.1 Central Findings . . . 97
8.2 Outlook . . . 99

BIBLIOGRAPHY 101

LIST OF PUBLICATIONS 109

ACKNOWLEDGMENTS 111

ACRONYMS

AV	Alphavirus
BC	Baltimore Classification
CHIKV	Chikungunya Virus
CM	Covariance Model
DENV	Dengue Virus
FV	Flavivirus
HMM	Hidden Markov Model
ICTV	International Commitee on Taxonomy of Viruses
ISFV	Insect Specific Flavivirus
cISFV	*Classic* Insect Specific Flavivirus
dISFV	*Dual-Host affiliated* Insect Specific Flavivirus
JEV	Japanese Encephalitis Virus
MBFV	Mosquito Borne Flavivirus
MFE	Minimum Free Energy
MGE	Mobile Genetic Element
NKV	No Known Vector Flavivirus
ncRNA	Non-Coding RNA
sfRNA	subgenomic flaviviral RNA
xrRNA	Exoribonuclease resistant RNA element
TBEV	Tick-Borne Encephalitis Virus
TBFV	Tick Borne Flavivirus
UTR	Untranslated Region
VOG	Viral Orthologous Group
WNV	West Nile Virus
YFV	Yellow Fever Virus
ZIKV	Zika Virus

1

PREFACE

Already shortly after starting to work on viruses, I realized that virology and research related to the nature of viruses diverges from other life science disciplines. Many foundational principles of biology, which are normally beyond reproach, do not strictly hold true in the realm of viruses and must be examined and re-interpreted from novel angles. For example, several of the most fundamental characteristics of life, are not encountered in viruses: Viruses have no cells, Viruses do not synthesize proteins, Viruses even display no metabolic activity - just to name a few. Other aspects of virology seemed outright absurd, best illustrated by the large topic of morphology-based virus taxonomy - it struck me as profoundly odd that an entire kingdom of biology would be organized simply on the basis of visual shape, similar to how taxonomy had been approached in the earliest days of biology.

While categorizing life based on its shape seems archaic from the present perspective, the reality is that taxonomy in those early days of biology there were little other options other than morphology. Given the back then limited scientific knowledge on life in general, morphology seemed to be the most apt basis for a taxonomy. I quickly realized that today's virology faces a comparable dilemma: many parts of the existing knowledge on organismic biology cannot be transferred to viruses without modifications and yet our knowledge on viruses is still fragmentary. Beyond that, obtaining new knowledge on viruses has always been difficult - by their ultra-microscopic nature, viruses defy many experimental approaches that are otherwise well suited for investigation of microorganisms. On a historic notice, this elusive nature of viruses is nicely illustrated through their discovery process, where scientists first described viruses as a form of "poison" (Latin *virus*) before learning their true nature. Fortunately, the staggering success of sequencing-based methods has given us a new types of instruments for investigation and opens up a new era in virus research.

At the same time, viruses are environmentally omnipresent and infect every organism. Viruses are an integral and ubiquitous part of our biosphere and critically influence all other domains of life. While commonly discussed in the context of pathogens and diseases, viruses in fact assume a central role in the maintenance and equilibrium of many ecosystems. More importantly, over the course of billions of years, viruses have shaped the evolution of all other domains of life and their footprint can be found in the genome of every organism. In the context of this thesis, especially their ability to employ RNA structure for a broad array of biological functions is a fascinating evolutionary development that sets viruses apart from other domains of life. It is especially fascinating to see how the very nature of RNA and its structure, characterized by a high degree of evolutionary plasticity, beautifully integrates with the extreme rates of evolution typically found in viruses.

I was immediately captivated by the prospect of investigating viral evolution from the angle of RNA structure, not only because of the insight that could be gained on individual virus species, but because of the opportunity to better understand evolution itself. This thesis explores how evolution has shaped the RNA structure in viruses, both generally but also in specific virus species. It aims to better understand the role RNA structure can play in this exotic realm of viruses and how it emerged, seeking to to add a slice of knowledge to the vast field of virology.

2 BACKGROUND

The work presented in this thesis is tightly connected to the subject matters of RNA secondary structure prediction and viruses. This chapter provides an overview over both major background topics. The reader is familiarized with the fundamentals of virus biology, virus evolution and the significance of RNA in the viral life cycle, followed by a more detailed discussion of the *Flaviviridae* family and the *Chikungunya* virus.

2.1 VIRUSES

Viruses are microscopic acellular entities that proliferate as obligate intracellular parasites within other organisms. In contrast to any other organisms that use cells as their fundamental biological entity, a virus is considered to be a Mobile Genetic Element (MGE). Compared to cellular organisms, viruses are drastically reduced in complexity and lack even the most foundational biological capabilities such as independent self-replication, protein synthesis or machinery for catabolic and anabolic processes. For reproduction, viruses instead take control over a host's cellular machinery for replication, assembly and spread of new virus particles. Viruses infect all organisms, with each virus usually specifying on a narrow range of host species.

The unique nature of viruses hinders their straightforward definition and over the course of history the formal definition of viruses has undergone many drastically different iterations[1]. It is regularly subject to change according to the current state of knowledge on the viral world. In 2021, the International Commitee on Taxonomy of Viruses (ICTV) agreed on a revised definition which for the first time focuses on the role of viral phylogeny and evolutionary history:

> "Viruses *sensu stricto* are defined operationally by the ICTV as a type of MGE that encodes at least one protein that is a major component of the virion encasing the nucleic acid of the respective MGE and therefore the gene encoding the major virion protein itself or MGEs that are clearly demonstrable to be members of a line of evolutionary descent of such major virion protein-encoding entities. Any monophyletic group of MGEs that originates from a virion protein-encoding ancestor should be classified as a group of viruses." - Kuhn et al.[2]

Due to their acellular nature and absence of any metabolic activity, viruses are not considered to be alive in the same way as eukaryotes, prokaryotes or archea. Theories on how such a form of quasi-life might arise are plentiful but mostly of speculative nature (see section 2.1.3) and there is no straightforward consensus on how viruses are intertwined with other domains of life. Some consider viruses to form their own domain of life that has evolved in parallel with all other organisms and possibly pre-dates the origin of the first cell.

Despite their microscopic dimensions and radically different life style, viruses are among the most abundant biological entities on earth. They infect every organism, can be encountered in all ecosystems, and in marine ecosystems the number of virus particles is estimated to outnumber all other species combined by at least an order

of magnitude[3,4]. Furthermore, viruses serve crucial roles in keeping ecosystems in balance and marine environments rely on viruses to keep the cyanobacteria population in check to prevent overgrowth.

In addition to their role as parasites and pathogens, viruses have shaped the evolution of their hosts[5]. In bacteria, lateral gene transfer is frequently mediated by viruses[6]. In humans, the genome has been demonstrated to shelter numerous sequences of viral origin that have integrated into the human genome over the course of primate evolution and are now an integral part of the genome[7,8]. Many of those elements are now defunct, but some are still active as retrotransposable elements, which are considered a major force determining genome stability and evolutionary pace, and are regulated by a specific family of proteins in humans and other higher organisms[9].

2.1.1 *The viral life cycle*

Viruses have a life cycle that strongly differs from cellular organisms as a result of their acellular nature. Unlike cellular organisms, viruses cannot reproduce through cell division and require a host organism to provide them with energy, proteins, and other necessary building blocks of life. While the specifics of the life cycle can vary between species, it always involves four main stages: recognition of a host, entering the host, replication, and leaving the host. Prior to entering their host organisms, viruses are inert entities devoid of any characteristics of a living organism. Outside of a host cell, their time of survival is limited but heavily dependent on the species and environmental conditions such as temperature, surface, and humidity[10].

The life cycle of a virus can be divided into several main stages:

1. **Attachment**. The virus initiates contact with a potential host cell by binding surface proteins on the host cell with the help of receptor proteins present on the virus particle's surface. These viral receptors are shaped in a manner that specifically recognizes suitable host cells. The attachment stage is crucial because viruses typically target a specific type of cell, and binding to an incompatible cell will reduce their ability to replicate.

2. **Entry** - Once the virus has attached to a suitable host cell, it gains entry into the cell through endocytosis or membrane fusion. In the case of certain viruses like bacteriophages, they can directly inject their DNA into the host cell, while leaving the empty virus particle outside. After entering the host cell, the viral genome may be transported to various cellular compartments, such as the nucleus or endoplasmic reticulum, depending on the specific virus. Some viruses may integrate into the host genome before initiating the replication of their genome.

3. **Replication** - Inside the host cell, the virus takes control of the cell's machinery to replicate its DNA or RNA genome. DNA viruses further transcribe mRNA and utilize on the host's protein synthesis machinery to produce viral proteins. RNA viruses can immediately start synthesizing proteins using their own genome as mRNA. During replication, the virus often distinguishes between "early" and "late" genes. "Early" genes primarily produce proteins that assist in replication, while "late" genes primarily produce structural proteins that aid in assembling new virus particles.

4. **Assembly** - Once the viral components (genome, proteins) have been replicated, they assemble to form new viruses inside the host cell. This process of viral assembly is often self-guided and facilitated by the structural proteins of the virus.

5. **Exit** - After assembly, the newly formed viruses are released from the host cell into the environment. The specific mechanism of release varies depending on the virus species. Bacteriophages often destroy their host cells through lysis, whereas many viruses that infect eukaryotic cells leave through exocytosis or direct budding from the cell membrane. In some cases, viruses can induce apoptosis in their host cells, leading to the death of the cell and subsequent ingestion of the virus particles by other cells.

2.1.2 *Virus Transmission*

As a parasitic organism, reaching its host organisms to initiate a new replication cycle plays a crucial role in the viral life cycle. Viruses have developed numerous different strategies to achieve this goal.

Some viruses use the natural environment as a transmission medium to reach new hosts. For instance, measles or chickenpox viruses are transmitted in an **airborne** way directly via the air and are ingested by the host through their respiratory tracts. Similarly, viruses such as Rhinoviruses (common cold) are transmitted via air in small fluid droplets, which are dispersed as an aerosol. Some viruses have adapted to a **waterborne** transmission route and can survive in liquids for a prolonged period of time until they come in contact with a suitable host. Other viruses spread via indirect contact, where they are deposited on surfaces, where they eventually come into contact with new hosts that touch those surfaces, as seen in with the Hepatitis A virus.

Moreover, **direct contact** is also a common transmission mode, requiring a direct physical connection between an infected individual and a potential new host. For example, sexually transmitted viruses like HIV are transmitted through direct contact, and, more rarely, viruses can be directly transmitted *in utero* from mother to an unborn child.

A large group of viruses use other organisms or **vectors** as a means of transmission. Vector species are specialized hosts, which are little affected by an infection and allow the virus to persist until it is transmitted to another host. *Arboviruses* such as Flaviviruses or Alphaviruses use arthropod vectors such as Mosquitoes from the *Aedes* family or ticks from the *Culex* family, in order to spread to new hosts. The actual transmission then occurs via bites, where through the insects saliva a large quantity of viruses is injected into the host.

2.1.3 *Virus Origins*

There exists no universally accepted or experimentally verified theory on the origin of viruses. Their unique physiology, structure and genome has inspired three major theories on their origin;

1. **Reduction** or **Regression** Hypothesis: Viruses are the remnants of earlier prokaryotic or eukaryotic cells that have gradually lost all genes and functionalities that were not essential to their parasitism. This theory is inspired by

the observation that certain parasitic prokaryotes with a similar obligate intracellular life style such as *Carsonella* or *Rickettsia* have also strongly reduced genomes and transferred essential functionality to their hosts. None of them though has yet undergone gene loss as radical as viruses, arriving at a point where even the translational machinery is not present any more.

2. **Escape** Hypothesis: Viruses are independent parts of higher cells that have moved out of the confines of their original cells and developed novel ways of replication. In Bacteria, Plasmids are considered to be a likely candidate for such an imagined process, whereas in Eukaryotes, Transposable elements are considered to be a possible candidate for the support of the Escape Hypothesis.

3. **Primordial** Hypothesis: Viruses originate from a period that pre-dates even the most ancient cells and have co-evolved with life already from a pre-biotic age, possibly as part of a (hypothetical) primordial *RNA world*. This hypothesis is inspired by the existence of *Viroids*, RNA Viruses that consist only of an infectious (and highly structured) strand of RNA. Further support for the Primordial Hypothesis is drawn from the findings that there exist viral proteins that are not homolog to any other non-viral protein, suggesting an independent origin[11].

At the same time, it is likely that the origin of viruses might be due to a combination of all three above factors. Viruses are known to be a non-monophyletic group that has multiple origins[12], and also have sometimes acquired significant portions of their genes from their hosts. While it is possible to reconstruct the evolutionary history of some viruses, there exist no universally conserved genes in *all* viruses that could enable the construction of a common phylogeny, such as ribosomal RNAs. Furthermore, observed similarities of viruses today also could rather be a result of convergent evolution, than having their a common origin.

2.2 VIRUS STRUCTURE

The composition of viruses differs drastically from cellular organisms or the substructures found in cellular organisms. Moreover, the structure of a virus depends heavily on the stage of the life cycle. Outside of its host organism, almost all viruses exist in the form of biologically inactive but infectious particles, termed **Virions**. Inside their host cell, virions disintegrate into their individual components and begin proliferation. A notable exception are *viroids*, which consist of only of a strand of RNA, regardless whether inside or outside their host cell.

2.2.1 *Virion components*

In contrast to cells of higher organisms, virions are tightly packed structures that lack any cytoplasm, nor do viruses display any degree of compartmentalization or complex organelles such as a nucleus, endoplasmatic reticulum or mitochondria. Typical virions consist of a core of one or multiple strands of DNA or RNA which are enclosed by a shell of structural proteins, often in a highly symmetric shape. The entire protein shell is referred to as the viral **capsid**, while the parts of the protein shell that are in direct contact with the nucleic acids are specifically referred to as the **Nucleocapsid**. Multiple capsid proteins frequently self-assemble into **capsomeres**, which in return then constitute the building blocks that self-assemble into the capsid.

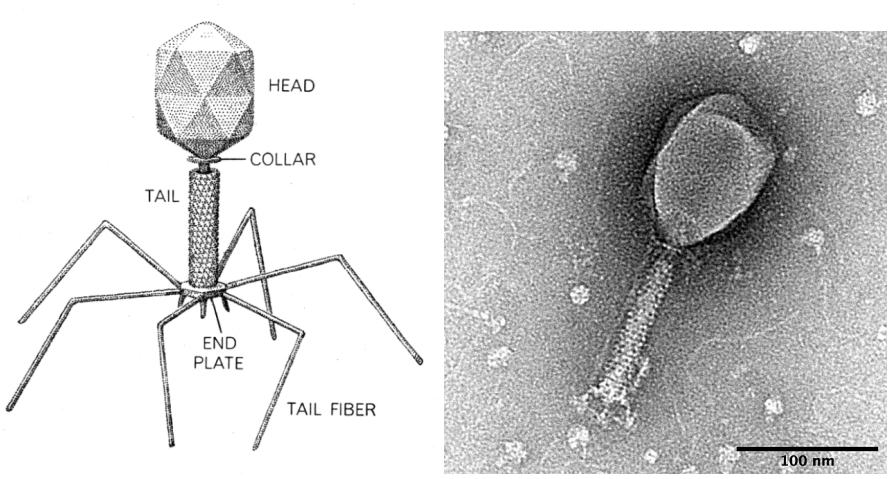


Figure 1: Schematic drawing[13] of Bacteriophage T4 and EM micrograph[14] of Phage T2. The head-tail architecture of the virion is characteristic for bacteriophages, but not representative for the majority of viruses.

Many virus species further engulf the capsid in an **Envelope**, a lipid mono- or bilayer derived from the host cell upon exit, which further protects the virion from environmental influences.

Virions of many virus species contain further components. These proteins are tailored to the needs of the individual species but commonly include proteins that mediate functionalities in the viral life cycle, such as polymerases for genome replication or specialized proteins that facilitate entry into a host cell.

Many virus species also provide surface-bound glycoproteins or **Receptors**, which facilitate the process of docking to and penetrating the surface of their host cells. Viral receptors can also be molecules that are normally found on cell surfaces of their hosts and have specialized physiological functions, which the viruses have evolved to use for their own purposes.

2.2.2 *Genome Structure*

Viruses have developed various strategies for genome structure and replication. Depending on the species, the genetic material is DNA or RNA, in either single-stranded or double-stranded configuration. Single-stranded RNA viruses are further categorized into positive-sense RNA viruses, which can immediately use the RNA genome as mRNA for translation, or negative-sense RNA viruses, which require the synthesis of a complementary RNA strand before translation.

A special case are Retroviruses, which carry single-stranded RNA in the virion, but inside the host cell, transcribe it into DNA with the help of a special enzyme called Reverse Transcriptase. The DNA copy then integrates into the host genome and subsequently produces more copies of the RNA genome through regular transcription. Some DNA viruses (*Caulimoviridae*) also transcribe their genome into an intermediate RNA stage, which serves as a template for replication of mRNAs, but also for the reverse transcription of new DNA genomes.

The shape and organization of viral genomes significantly vary across species. Common genome shapes are linear molecules, such as those found in *Flaviviridae*, or circularized molecules, such as those found in *Polyomaviridae*. In most species, the

genome is organized in one single continuous DNA/RNA molecule. However, some species distribute their genome over multiple chromosome-like segments, which are also packaged separately into the virion, such as in Influenza.

There is a wide range of variation between species regarding genome size, both in terms of gene number and the total count of bases. On one end of the spectrum, there are viroids (e.g., Potato Spindle tuber viroid[15]), which typically have genome sizes of no more than 400 nucleotides and no protein coding or otherwise detectable genes[16]. On the other end of the spectrum, there is a growing number of giant viruses that exceed even small prokaryotes in genome size such as *Mamavirus*[17] which has a genome of over 1.2 Mb and more than 1000 protein coding genes.

2.3 VIRUS SYSTEMATICS

A formal description of the phylogeny of the entire viral world is a challenging endeavour, compared to any other domain of life. Major difficulties arise through large quantities of lateral gene transfer, extreme rates of evolution and the polyphyletic origin of viruses. While all higher organisms can be traced back to the Last Common Universal Ancestor (LUCA) by the presence of an universally conserved core of genes almost exclusively coding for protein and RNA components of the translation system[18], none of these genes exist in viruses, nor do viruses share any comparable universal gene core. The absence of such a core further illustrates not only the difficulties of relating viruses to other domains of life but also the problems innate to establishing a comprehensive description of the viral world.

Still, several attempts have been undertaken at ordering the relationships between all species of the viral world. Each method utilized a different set of criteria as determinants, such as virion shape[19], genome type and distinct features of the virion[20], type of nucleic acid encapsidated into virions and virion morphology[21] or a combination thereof. These approaches often reflected the then current state of the art of methodologies that were available for investigation of viruses in a pre-genomic era.

2.3.1 *Morphological Classification*

Classification of viruses on the basis of their morphology comprised one of the earliest attempts at systematizing the viral world[19]. Motivated by the observation that many viruses from distinct shapes, viruses are commonly categorized into distinct morphological categories:

- **Icosahedral** Viruses form a capsid in the shape of a regular icosahedron. There is variation between species regarding the number of capsomeres that constitute the faces of the icosahedron, with 3 capsomere units being the most frequent case. Notable examples of icosahedral viruses are Rotaviruses or Flaviviruses.

- **Helical** Viruses have a long cylindrical capsid with a radial symmetry. The nucleic acid genome is coiled inside the capsid in a helical pattern, forming an overall rigid, rod-like structure. The length of a helical capsid is determined by the overall length of the virus genome. Examples of viruses with helical morphology include the tobacco mosaic virus and the influenza virus.

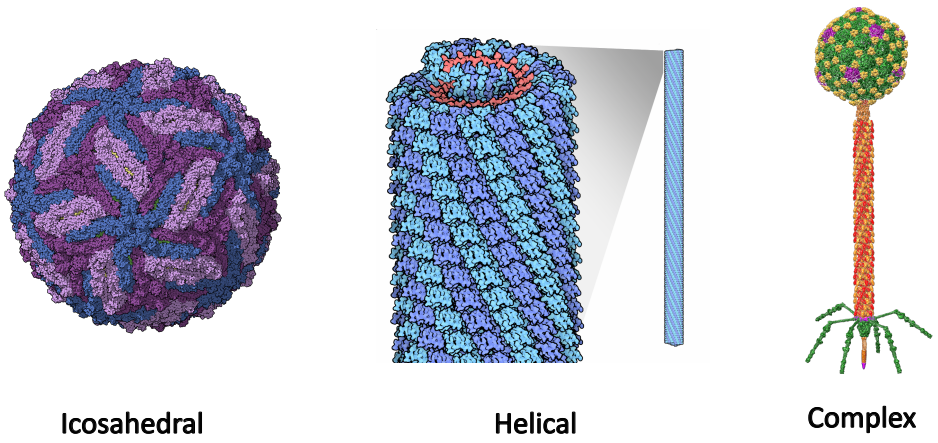


Figure 2: The three most common groups of Virus morphologies: Icosahedral (ZIKV cryo EM[22]), Helical (TMV[23]), Complex (PhageLambda[24])

- **Complex** Viruses either form capsids that do not belong to the helical or icosahedral groups. Those viruses may exhibit capsids made out of multiple parts (e.g. bacterophage T4), have highly structured capsids that do not resemble any classical symmetry (HIV-1), or have capsids with irregular shapes (Hepatitis-C).

In addition to the three major classes of virus morphology, some viruses have additional structures or features that are not strictly considered part of their morphology. A notable example, are many enveloped viruses that have additional proteins which were acquired from the host cell during the process of viral assembly and exit from the cell. Some viruses also have spikes or other surface proteins that help them to bind to and infect host cells.

It is important to note that the morphological classification of viruses does not take into account any genetic properties of the species, but assumes that the form of the virus will also determine its function. Despite not being necessarily true, virus morphology still provides a useful framework for description and understanding the structure and function of viruses in the absence of any genomic data.

2.3.2 *Baltimore Classification*

The Baltimore Classification (BC) scheme was proposed by David Baltimore in 1971 in order to categorize viruses on the basis of similar strategies of genome replication[25] and classified the all viruses in one of originally 6, now 7 distinct classes. Central rationale behind this approach to classification is the assumption that the only biological mechanism valid for all viruses is the central dogma of molecular biology, which dictates the information flow from DNA to Protein, with an RNA intermediate. Therefore every virus must implement a strategy of genome replication, which can be used as a uniting basis for taxonomy. The BC distinguishes between seven classes, based on configuration and replication of the viral genome:

I. Double stranded (ds) DNA viruses

II. Single stranded (ss) DNA viruses

III. dsRNA viruses

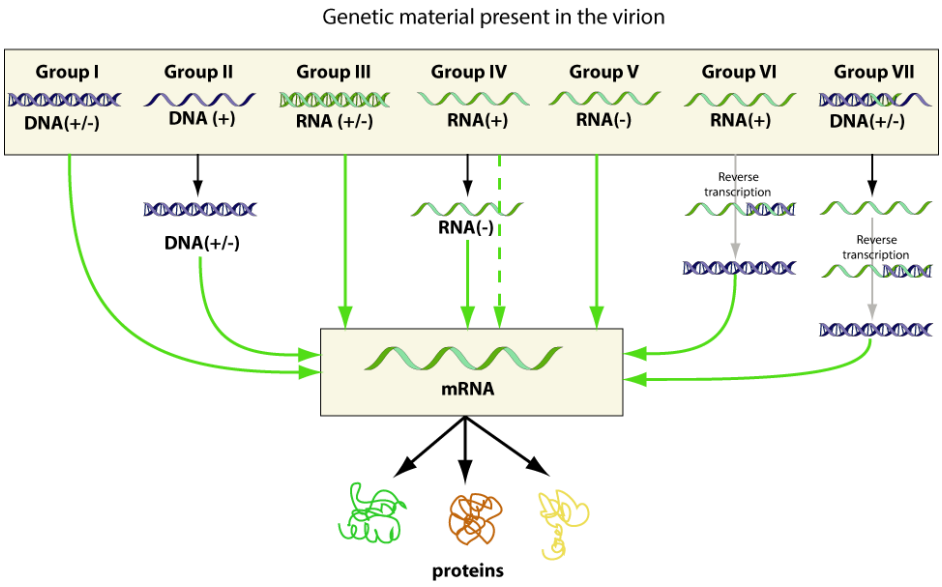


Figure 3: Baltimore Classification of Viruses. Each of the 7 classes is characterized by a unique strategy that ultimately leads to the replication of the viral genome, the transcription of viral mRNA and expression of proteins. Figure adapted from ViralZone.[26]

IV. ssRNA viruses, with sense (+) genome that also acts as mRNA

V. ssRNA viruses, with antisense (-) genome that acts as template for mRNA synthesis

VI. RNA viruses with a (+) genome which replicate by reverse-transcribing a DNA intermediate

VII. DNA viruses which replicate by reverse transcribing from an RNA intermediate

The Baltimore Classification acts as the *de facto* standard framework for the description of viruses. It is tempting to assume that Baltimore Classes resemble quasi-monophyletic groups, but closer examination of the various genomes within each BC shows that it does not accurately reflect the phylogenetic relationships between the viruses. While recent studies on virus metagenomes indeed could identify a set of viral hallmark genes (VHGs), which are characteristic for each class[12], it furthermore identified a set of 6 super-VHGs which are present in an enormously diverse set of viruses which spans over two or even three BCs. Super-VHGs include a set of 6 proteins such as Polymerases, Capsid proteins, Helicases and Rolling-circle replication initiation endonucleases.

The origins of super-VHGs are assumed to be widely different. In particular, replication-relevant enzymes most likely derive from a primordial, precellular pool of genes, further emphasized by the lack of any orthologs of these proteins in cellular life-forms. Other proteins were likely acquired at an early stage of virus evolution from cellular hosts or evolved from other proteins in an ancient viral common ancestor.

With the help of VHG phylogenies, efforts are being made to reconcile and integrate the BCs with the general virus taxonomy. Baltimore classes III, IV, and V (RNA viruses) have already been formally consolidated as a *taxon*[27,28], a rank equivalent to a *domain* as used for cellular organisms. Recent propositions of a viral megataxon-

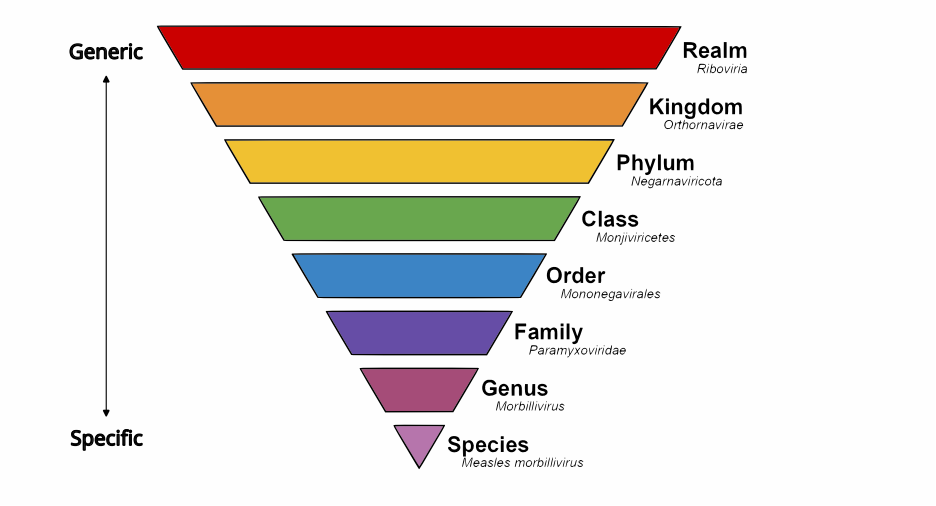


Figure 4: Current (as of 2023) ICTV taxonomy levels of viruses, with examples. Figure adapted from ICTV.[32]

omy[29] further replace the BCs from their current (informal) position as highest rank taxonomic units and introduce new taxa that unify all viruses under one framework.

Despite the new (i.e. genomic) direction of virus classification, the BCs remain a central framework for the study of virus biology, comparable to concepts like prokaryotes and eukaryotes, which also do not comprehensively reflect the phylogeny of cellular life-forms but still are widely used in comparative cell biology.

2.3.3 *Virus Taxonomy*

Virus taxonomy is the classification of viruses into a system of categories called taxa, and the development and implementation of standardized methods for nomenclature of taxa. Whereas virus taxonomy has a long standing history, it was not until recently that the field gained importance due to the rapidly increasing number of new viruses discovered by next-generation sequencing methods. Virus taxonomy traditionally had been a process that used a mixture of morphological, biophysical or other arbitrarily chosen characteristics for classification, frequently using different sets of characteristics for different groups of viruses in an inherently subjective way[30].

In recent years, the availability of large metagenomic datasets and the identification of Viral Hallmark Genes (VHGs) have brought a paradigm shift viral taxonomy, best illustrated by the joint efforts of the community to establish a framework[31] that recognizes novel viruses solely on the basis of full genome sequences and disregards any biological or biophysical characteristics. The genomic approach not only opens up the possibility to classify species on the basis of phylogeny, but more importantly eliminates the requirement for a virus to be cultured or otherwise characterized in the laboratory which previously rendered the majority of viruses by default unclassifiable.

From a practical perspective, the International Commitee on Taxonomy of Viruses (ICTV) acts as the international and authoritative body that establishes the virus taxonomy. The ICTV is a global organization that consists of various working groups which are responsible for the recognition of new viruses. The ICTV itself does not decide on any matters regarding creation or organization of taxa, but acts as

Rank	Number of Taxa
Realm	6
Kingdom	10
Phylum	17
Class	39
Order	65
Family	233
Genus	2606
Species	10434

Table 1: Major viral taxa currently (2023) recognized by the ICTV. Species that belong to sub-taxa are not shown.

a platform to the virology community for a structured discourse and consensus finding. It operates primarily through review and voting on proposals, which can be raised at any time by any member of the ICTV. The decisions of the ICTV in practice do not follow a universal strategy, while proposals from some working groups might suggest changes based on algorithmic arguments or formally defined procedures, others might proposed changes based on a different but equally valid rationale. Ultimately, the process of taxonomy through the ICTV is decided by a consensus vote.

The latest comprehensive Viral Taxonomy classifies all virus species under a system of 15 hierarchically organized taxa (Figure 4). The ICTV recognizes 10434 virus species organized under 8 major taxa (Table 1), where all taxa above the 'species' level are potentially allowed to have further sub-taxa, even though just a handful actually have been assigned. It is important to note that the hierarchical manner of the taxonomy tree not necessarily reflects the phylogenetic relations between individual sub-groups of a taxon.

2.4 RNA AND VIRUSES

The Baltimore Classification schema illustrates central role of RNA in the viral life cycle (Figure 3) and emphasizes that regardless the type of virus, RNA is an indispensable part for successful replication. More than other organisms, viruses have evolved to utilize RNA in numerous different ways - not only in the traditional fashion as mRNA, but also as information storage medium for the genome, a role that is classically associated with DNA.

Viruses furthermore utilize RNA's capability mediate funcionality via RNA structure in order to thrive within a host. Some viruses actively recruit ribosomes with the help of RNA structure elements despite the lack of traditional ribosome binding sites in the genome[33]. Other viruses have developed RNA-structure mediated defense mechanisms against degradation by nucleases[34].

The unique plasticity of RNA combined with the extreme evolution rates of RNA viruses, have made them also a target of intensive study by computational and theoretical biologists interested in evolutionary processes. RNA structure has been shown to impose constraints on viral evolution[35] and formal investigations have been conducted that show that there is an upper limit to the rate of evolution, above which RNA structure cannot be maintained any more[36]. RNA and viruses have also served as an important model for the study of evolutionary robustness[37].

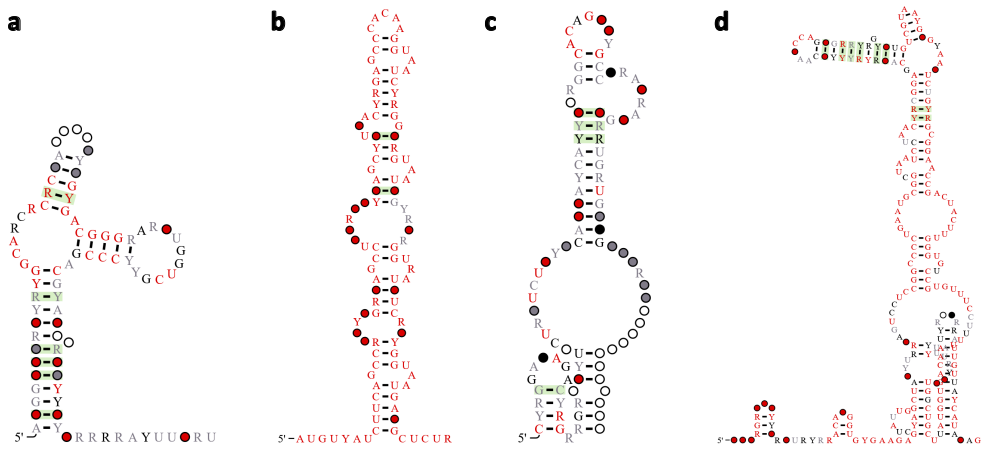


Figure 5: Examples of viral RNA families with functional RNA structures. **a** Tick Borne Flavivirus xrRNA (RF03536) **b** Coronavirus packaging signal (RF00182) **c** Flavivirus Cis-regulatory element (RF00185) **d** Picornavirus IRES (RF00229) . Colours indicate the conservation of the nucleotide sequence. Nucleotide colour indicates sequence conservation over the RNA family, red: ≥97%, black: ≥90%, grey: ≥75%. Images adapted from respective Rfam families, RF03536, RF00182, RF00229 and RF03536.

2.4.1 *RNA structure in viruses*

Viruses have developed numerous ways to utilize RNA structure for biological functionalities that are otherwise performed via proteins. A wide spread example are the Internal Ribosome Entry Sites (IRES) found in many RNA viruses (RF00229), which allow the recruitment of ribosomes to the viral mRNA. Though the secondary structure is critical for a functioning IRES, it also shows stretches of high sequence conservation.

Other elements, such as the Selenocystein Insertion Sequence (SECIS) (RF00031) show a high degree of structure conservation, despite diverging significantly in sequence. The SECIS element thereby acts as the trigger for the incorporation of the non-canonical amino acid selenocystein into the nascent protein chain.

Other structure elements allow viruses to ignore and read through stop codons during translation, enabling the virus to express additional isoforms of a protein.[38]

Viruses have furthermore adapted RNA structure to help with replication initiation of Flavivirus genomes (RF00185) or help in packaging of the virions in Coronaviruses (RF00182). An eminent example of functional RNA structures especially relevant for this thesis are exoribonuclease resistance mediating elements in Flaviviruses (RF03536) which are discussed in detail as part of chapter chapter 6.

2.4.2 *RNA groups*

RNA groups are collections of RNA molecules that share a set of characteristics. Depending on the type of shared characteristics of an RNA group it is referred to as either *RNA family*, *RNA clan* or *RNA class*. RNA groups can be regarded as a form of "taxonomy" for RNA molecules and facilitate the standardized description and identification of RNA molecules (Table 2). There exist 3 major quasi-hierarchic levels of RNA groups:

RNA families[39] are sets of RNAs which perform a specific biological function in different organisms. These RNAs are homologous and can be traced back to a common

Group properties	Family	Clan Type 1	Clan Type 2	Class
Shared function	+	+	-	+
Homologous	+	+	-	-
Diverse	-	+	-	+
Alignable	+	-	+	-

Table 2: Overview of RNA groups and the presence (-) or absence (-) of characteristic properties.

ancestor and mostly share a common secondary structure. A notable examples of an RNA family are tRNAs (Rfam ID: RF00005), which are ubiquitous in (non-viral) genomes but share a common functionality and structure. Sequence divergence in an RNA family can vary strongly, but must still allow for the construction of a reasonable multiple sequence alignment.

RNA clans[40] are groups of RNA families. There are 2 types of RNA clans: Families part of a Type 1 clan share the same functionality and ancestor, but are diverged to a point where no meaningful alignment can be established any more. Many varieties of tRNAs (e.g. mt-tRNAs) form a type 1 clan (Rfam ID: CL00001), which contains of 7 families and more than 1.4 million individual sequences. Type 2 clans on the contrary stem from a common ancestor but do not share common functionality but still can be aligned with each other.

RNA classes[41] are in a way complementary to type 2 RNA clans. While its member families mostly share a common biological functionality, they do not have a common ancestor or any alignability. A notable example of an RNA class are the various types of micro RNAs, which share a common mechanism of function, but are not always homologous.

It is important to note that RNA families are built from a pragmatic perspective. There are no quantitative rules that determine the assignment of individual RNAs to families, even though the formal definitions of all RNA groups may suggest otherwise. Frequently, the ability to build a functioning Covariance Model (CM) is used to determine the scope of an RNA family, but ultimately family curation is carried out and agreed on by the RNA community and may undergo frequent revisions. The Rfam (RNA families) database collects and curates families of Non-Coding RNA (ncRNA)s and acts as a central data hub. RNA families deposited in Rfam are provided with a CM (see chapter 3.2.3), a multiple sequence alignment, consensus secondary structure and documentation via a summary article in Wikipedia. Its latest major release (Rfam 14[42]) contained a total of 3444 families and included a special focus on viral RNA families, including several RNAs identified through the work in this thesis, see chapter 6.

2.5 FLAVIVIRUSES

Flaviviruses are RNA viruses organized under the genus *Flavivirus*. Together with the genera *Pestivirus*, *Pegivirus* and *Hepacivirus* they are part of the large *Flaviviridae* family. The *Flavivirus* genus is the largest in the family and includes more than 50 arthopod-borne viruses, including several prominent disease causing viruses such as Dengue Virus (DENV), Tick-Borne Encephalitis Virus (TBEV), West Nile Virus (WNV) and Japanese Encephalitis Virus (JEV).

Primary hosts of Flaviviruses are mammals and birds, while arthropods such as members of the *Aedes* or *Culex* family act as vectors. While a small group of recently discovered Flaviviruses exclusively replicates without the help of any arthropod vector[43] in insects (Sokuluk Virus, SOKV) or bats (Tamana Bat Virus, TABV), the majority of Flaviviruses is vector-dependent. Infections with Flaviviruses manifest depending on the virus on a broad range, from asymptomatic to severe or fatal hemorrhagic fever (DENV) or neurological disease (WNV).

Severe Flavivirus infections regularly affect a wide range of humans or animals. It is estimated that each year DENV alone accounts for 390 million infections, of which 96 million manifest in clinical symptoms[44]. In 2002, a large part of the austrian blackbird (*Turdus merula*) population was decimated by a newly adapted strain of the Usutu Virus[45], a flavivirus endemic to southern africa. Zika Virus, a flavivirus considered hitherto unremarkable, surprisingly caused in 2016 an outbreak of epidemic scale and led to numerous cases of microcephaly in unborn children[46], a novel symptom not associated with ZIKV before.

The combination of a global spread, widespread endemic distribution and rapid adaptability render Flaviviruses a significant threat to public health systems. Many flaviviruses are therefore considered ""emerging" or ""re-emerging" diseases and receive increased attention from a public health institutions.

2.5.1 *Flavivirus evolution and host-tropism*

As vector-dependent viruses, Flaviviruses constantly switch between reservoir (or *amplification*) hosts and vector hosts. Prominent members of the FV genus such as Zika, Dengue or West Nile Virus circulate between primates as reservoir hosts and arboreal mosquitoes as vectors that enable the virus to spread to new primate hosts[47]. Depending on the environment, transmission cycles are referred to as either *sylvatic* (primate-mosquito-primate) or *urban* (human-mosquito-human), with intersection between both cycles being rare but possible.

Continous cycling between hosts puts evolutionary pressure on the virus, as it needs to maintain its fitness in both host and vector species. Flavivirus evolution has been significantly shaped by host-tropism, leading to striking genomic differences between Flavivirus (FV)s depending on host specificity. The differences especially manifest in the structure and sequence of the 3'UTRs, where the biggest differences can be found between the groups (Section 2.5.3).

The major categories of FVs are:

- Mosquito Borne Flavivirus (MBFV) are transmitted by mosquitoes, most importantly by individuals of the *aedes* family. Especially relevant for human transmission in urban cycles are *aedes albopictus* and *aedes aegypti aegypti*.

- Tick Borne Flavivirus (TBFV) use ticks as vectors to cycle between animal hosts but also for infection of humans. Notable vectors are ticks from the *Ixodes* genus.

- Insect Specific Flavivirus (ISFV) form a phylogenetically distinct group of flaviviruses that have adapted to a life cycle without mammal hosts and persist only in insect hosts. The group is further subdivided into classic ISFVs and dual-host affiliated ISFVs, which could be shown to persist in multiple insect hosts, even though the mechanisms of transmission are not clear[48].

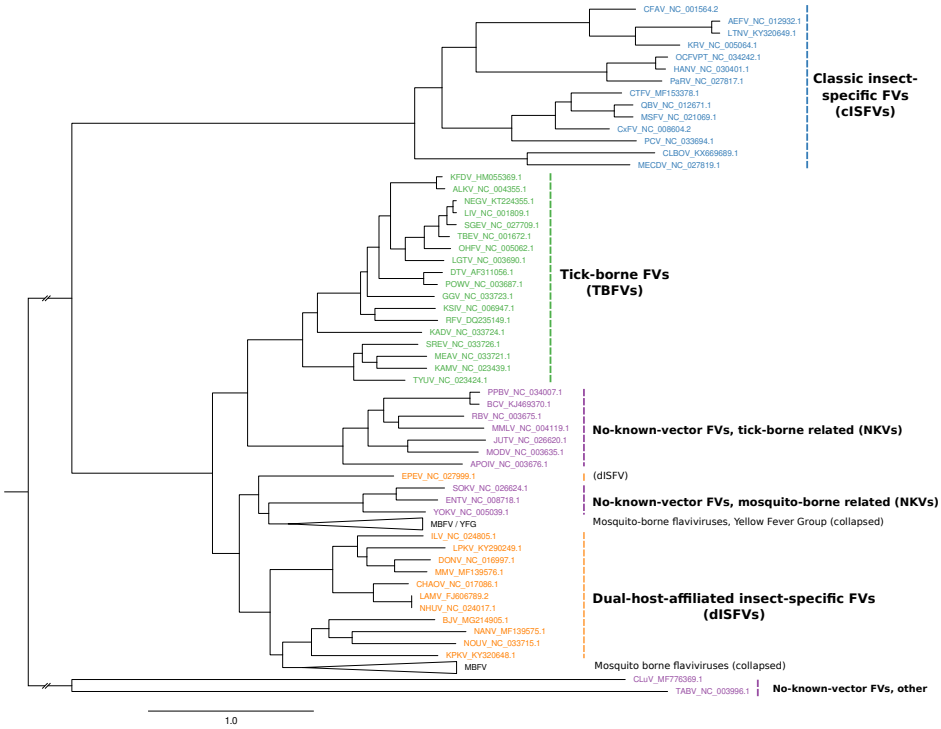


Figure 6: Maximum-likelihood phylogenetic tree of the genus *Flavivirus*, highlighting the Tick-Borne FVs, Insect-Specific FVs and No-known-vector FVs. The large group of Mosquito-Borne FVs is collapsed under the 'MBFV' clade for better visibility. Figure taken from Ochsenreiter et al.[49]

- No Known Vector Flavivirus (NKV) do not have a known vector species (yet). While for many NKVs the vector has not yet been discovered, some (TABV, CLuV) might not need an arthopod vector at all, but have adapted to purely in-vertebrate life cycles.

Interestingly, the UTR architectures of MBFVs, TBFVs and ISFVs are highly specific and can be clearly separated from each other via simple phylogenetic analysis (Figure 6. Topology of the tree further suggests that some FVs classified as No-Known vector viruses might actually be classified as Tick-borne or Mosquito-borne.

2.5.2 *Flavivirus genome*

The flavivirus genome (Figure 7) consists of a single, positive-sense strand of RNA of approximately 10-12 kB length. The genome consists of one large open reading frame which encodes a single polyprotein which is cleaved during and after translation into 10 mature proteins. The coding region is flanked by a 100 nt 5'UTR and a 400-700 nt 3'UTR at the end. Both UTRs are highly structured and shelter a set of structure elements that are necessary for replication of the viral genome and are discussed in more detail in section 2.5.3. In contrast to other genera of the *Flaviviridae* family, the genomic RNA of flaviviruses is capped at the 5' end and has no poly-A tail at the 3' end.

The flavivirus genome encodes 10 proteins, of which 3 are structural (C, prM, E) and 7 (NS1 NS2a/b, NS3, NS4a/b,NS5) serve non-structural purposes. Structural proteins include the main capsid protein C, the membrane protein prM, which is involved in the formation of the viral envelope derived from the host cell membrane

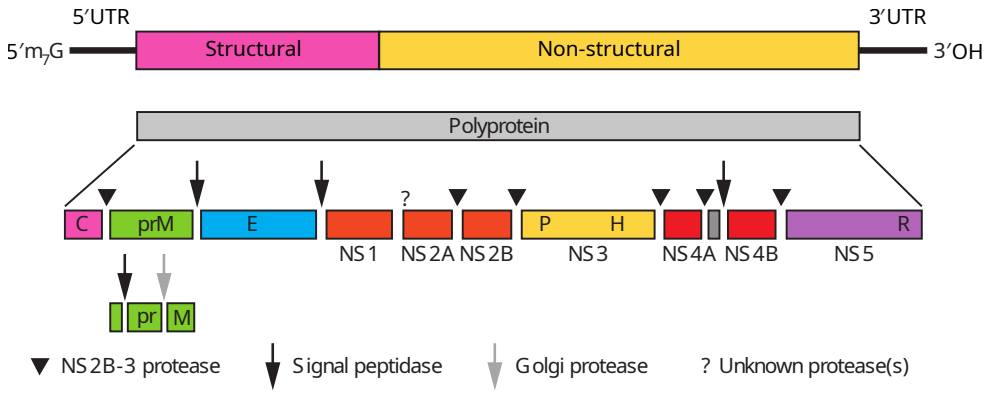


Figure 7: Schematic drawing of a general Flavivirus genome. Boxes indicate the coding domains, UTR regions are not drawn to scale. Some species introduce additional reading frames that code for slightly different proteins (JEV, WNV) but retain the other proteins. Also note that cleavage of the polyprotein already occurs during translation. Figure adapted from Simmonds et al.[50]

and contains viral glycoproteins. The envelope protein E mediates viral attachment and entry into host cells and is the main target of neutralizing antibodies. Some FVs are known to produce different proteins via ribosome frameshifting leads to alternative forms of NS4b in WNV or to a C-terminally extended form of NS1 in JEV. Some insect-specific FVs translate an additional protein called *fifo* from an alternative reading frame including stretches of NS2a and NS2b.

The non-structural proteins are involved in viral replication and evasion of the host immune response. NS1 and NS2a have been shown to play an important role in formation of the viral replication complex even though their exact mechanism of action remains still unclear. NS4a and NS4b are also involved in viral replication, their mechanisms are thought to support replication in multiple ways, through maintaining of replication complex structure and organization and modulation of replicase activity. NS3 is a highly modular and multifunctional protein with protease, helicase, and nucleoside triphosphatase domains, performing tasks such as cleavage of the viral polyprotein (with NS2b as co-factor), unwinding of the viral dsRNA during replication and modification of the 5' end of the RNA for capping. It has been shown to form a complex with NS5, which is the virus-specific RNA-dependent RNA polymerase and essential for genome replication.

2.5.3 *Flavivirus UTRs*

The flaviviral Untranslated Regions (UTRs) contain a remarkable set of RNA secondary structures that mediate a number of specialized biological functions, among those replication and translation of the virus genome. In all FV species, the 5'UTR is of uniform length around 100 nts, while the size of the 3'UTR varies considerably depending on the species between 400-700 nts.

Flaviviruses have evolved to heavily utilize the functionalities provided by non-coding RNAs in their UTRs. Strikingly, the majority of flaviviral ncRNAs are defined by a remarkable conservation of RNA structure, but not nucleotide sequence. The UTR architecture furthermore critically influences the production of subgenomic flaviviral RNA (sfRNA), which are an integral part of the virus life cycle and determine its pathogenicity (Section 2.5.4).

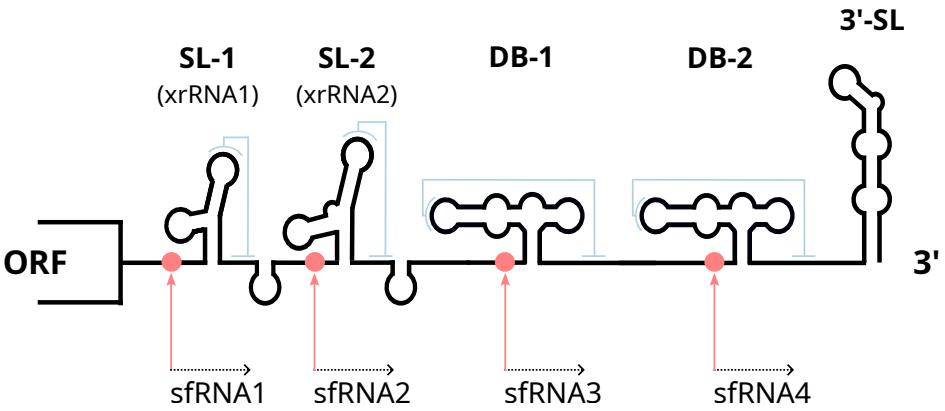


Figure 8: Schematic representation of the ZIKV 3'UTR architecture and sfRNAs, which is representative for most Mosquito Borne Flaviviruses. Secondary structure cartoons represent structural elements, blue arrows indicate pseudoknots. Red arrows indicate Xrn1 stalling sites, black dotted arrows indicate remaining sfRNAs. Structures SL-1 & SL-2 constitute UTR domain 1, DB-1 & DB-2 constitute domain 2. the 3'-terminal SL element represents domain 3.

5'UTR

The relatively short 5'UTR lies upstream of the polyprotein gene and can be divided into two parts containing the structure elements SLA and SLB. Both elements are separated by a stretch of Uridine bases and can be found in all Flaviviruses. Not part of the 5'UTR, but directly adjacent in the coding region lies the capsid-coding region hairpin (cHP) structure element, which is crucial for stable translation initiation and start codon recognition[51].

SLA (70 nts) is conserved over the entire genus *Flavivirus*. It acts as a promoter for genome synthesis, likely by recruiting the virus RNA polymerase NS5 to the viral genome. The SLB element (30 nts) forms a complex domain, together with the downstream AUG region (DAR), the cHP element and the flaviviral 5' cyclization sequence (5'CS), which have been shown to play a vital role in viral RNA replication by establishing long-range RNA-RNA interactions with other elements in the 3'UTR, leading to the cyclization of the viral genome[52].

Flavivirus 5'UTRs are of similar size in all flaviviruses, but can diverge significantly from each other in terms of actual nucleotide sequence. Strikingly, while 5'UTR sequences between TBFVs and MBFVs strongly diverge, the secondary structure is strongly conserved over all genera[53,54].

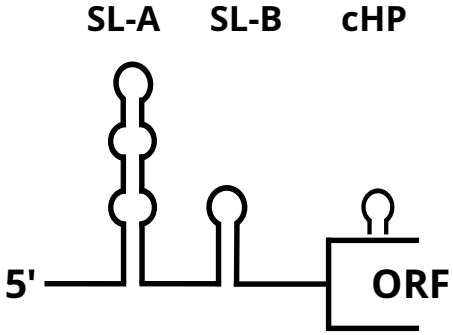


Figure 9: Schematic representation of the flaviviral 5'UTR architecture. The essential structure element cHP is located inside the reading frame of the virus polyprotein.

3'UTR

The flaviviral 3'UTR varies in length and structure between all flaviviruses and varies considerably between TBFVs, ISFVs and MBFVs. The UTR can largely be divided into three general domains that each serve a different purpose.

Domain one is directly adjacent to the polyprotein gene and shelters a variable number of Stem-Loop (SL) elements. The copy number and primary sequence of the SL elements is highly divergent between FV species, especially between TBFVs and ISFVs. Nevertheless, SL elements are highly conserved on a structural level and mediate the special ability of flaviviruses to withstand genome degradation by the host exonuclease Xrn1. SL elements are therefore often referred to as exonuclease-resistant (xrRNA) elements.

Domain 2 of the 3'UTR strongly diverges between the various sub-groups of flaviviruses. Mosquito-borne FVs typically contain one (Zika Virus (ZIKV),Yellow Fever Virus (YFV)) or two (JEV,DENV) dumbbell elements, referred to as DB1 and DB2 respectively. DB elements are thought to be essential for translation, but also have been suspected to function as additional Xrn1 stalling sites, mainly active in mosquitoes[55]. Other FV species such as TBFVs and ISFVs present a significantly different structure of domain 2. Most TBFVs do not express any forms of DB elements, but were shown to shelter another family of Y-shaped elements[49] in combination with other weakly characterized structures. Likewise, ISFVs do not express any DB elements, but show an even more complex and incoherent organization of domain 2.

Domain 3 is highly conserved over the entire genus. It contains the structurally conserved large 3'Stem Loop (3SL) structure that is indispensable for genome replication[56]. Additionally, domain 3 shelters the 3'Cyclization sequence (3'CS) and the 3'Downstream/Upstream AUG regions (3'DAR, 3'UAR) which are indispensable for circularizing of the flaviviral genome during replication[57].

2.5.4 *Subgenomic Flaviviral RNA and xrRNAs*

In vivo flavivirus genomes are constantly being degraded by the 5'-3' exonuclease Xrn1, which is part of the regular human RNA metabolism machinery and digests any uncapped monophosphorylated mRNA. The presence of Xrn1-resistant structure elements in the 3'UTRs effectively stalls Xrn1 and thereby results in the incomplete degradation of the FV genomes which accumulate as short (500 nt) fragments, called subgenomic flaviviral RNA (sfRNA). Production of sfRNA is assumed to be an inherent ability of al Flaviviruses and is best understood in MBFVs where it leads to the production of 4 distinct sfRNA species (sfRNA 1-4), corresponding to the 4 exoribonuclease stalling elements (Figure 8).

Abundance levels of sfRNAs significantly vary depending on the host of the virus. For DENV2 it has been shown that while in human cells sfRNA1 dominates the ensemble, the same genotype produces shorter sfRNA species (sfRNA2-3) in mosquitoes[58], implying a species-dependent degree of Xrn1-resistance conveyed by each individual xrRNA element. It is presumed that the apparent redundancy of Xrn1-stalling elements actually serves the virus to quickly adapt to new species during its life cycle[59], by being able to rapidly change one of the structure elements but simultaneously retaining its fitness in other hosts[60].

sfRNA have been shown to interfere with the host cell by actively counteracting innate immune evasion. In DENV, sfRNA has been shown to disturb the translation of interferon-induced mRNAs[61], which are one of the cellular responses to viral

infections. Other experiments have gathered evidence that sfRNAs might limit RNAi efficiency by acting as a decoy for Dicer2[62]. While not entirely understood, it has been also shown that in WNV, sfRNA length and abundance in mosquitoes strongly influences secretion of new virus particles in the salivary glands and as a consequence transmission efficiency[63]. In other cases, sfRNA has been shown to severely drive the evolution of diseases, when during a DENV outbreak in Puerto Rico, a novel DENV variant with higher sfRNA expression rates replaced the endemic strain[64].

The multiple roles of sfRNAs well illustrate the importance of xrRNA structure in the flaviviral life cycle. Not only does the sfRNA help in evasion of innate immune reactions, it also offers the FVs an elegant options to adapt to multiple hosts simultaneously[65], impressively shown by the dual adaptation of DENV to both humans and mosquitos[55].

2.6 CHIKUNGUNYA VIRUS

Chikungunya Virus (CHIKV) is a virus of the genus *Alphavirus* from the Togaviridae family, a family of single stranded RNA viruses that are spread by mosquito vectors from the *aedes* family. Notable members of the Alphavirus genus are O'nyong-nyong virus (ONV), Venezuelan Equine Encephalitis virus (VEEV) or CHIKV. Infections with Alphaviruses frequently manifest through an acute phase of febrile illness and arthralgia which is often followed by a chronic phase of arthralgia that can persist in CHIKV for months or even years[66,67], leaving affected individuals in crippling pain and prolonged disability.

Since the time of its first description in 1953 in Tanzania[68], CHIKV has spread from its endemic areas in Western Africa (Senegal) to all continents, including environments that were previously considered inhospitable to the virus and its vectors. This rapid spread is enabled by a continuous adaptation of CHIKV to new ecosystems and vectors resulting in frequent epidemic-like outbreaks, many of which have occurred over the past years (2004 India; 2006 La Reunion; 2014 South America).

Due to its aggressive spread, CHIKV is considered an important 'emerging Disease'[69] that poses a significant threat and therefore is of particular interest to public health agencies in their efforts to survey and predict CHIKV outbreaks in a timely and reliable manner.

2.6.1 *CHIKV life cycle*

Originally, CHIKV was endemic to the tropical regions of Africa and Southeast Asia, where it is known to persist in a sylvan transmission cycle, switching between nonhuman primate hosts and mosquitoes[66]. Occasionally, CHIKV is thought to be introduced into urban areas, thereby periodically causing outbreak events. Outbreaks are characterized by an *urban* transmission cycle, where transmission occurs primarily human-mosquito-human, primarily mediated by Aedes aegypti as vector species. Still, while urban transmission cycles account for the majority of outbreaks, there is evidence that some outbreaks in Asia are due to enzoonotic transmission[70]. While insect bites account for the overwhelming majority of CHIKV infections, rare cases of maternal-fetal transmissions have been observed, resulting in elevated rates of child mortality[71].

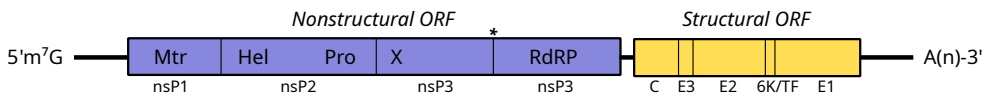


Figure 10: Schematic of a typical genome organization of viruses from the Alphavirus genus, like CHIKV. Figure adapted from Chen et al.[74].

2.6.2 *Chikungunya Virus genome*

The CHIKV genome (Figure 10 consists of one single strand of positive sense RNA of approximately 11 kb length. The genome is capped at the 5' end and polyadenylated at the 3' terminus. The protein coding genes are encoded in two large open reading frames spanning over the majority of the genome that code for two polyproteins. Polyprotein 1 codes for 4 nonstructural proteins (nsP1-4) that form the replicase complex[72], among which nsP4 carries the RNA-dependent RNA polymerase domain. Polyprotein 2 codes for 5 structural proteins (Capsid, E3, E2, 6k/TF, and E1), which form the virus capsid and its membrane domains. Both ORFs are separated by a small stretch of noncoding nucleotides, which include a promotor element specific for ORF 2[73]. The promotor element overlaps with the 3'-end of ORF2.

2.6.3 *CHIKV UTRs*

Unlike other RNA viruses, CHIKV has not been conclusively demonstrated to heavily utilize the noncoding UTR regions. Length of the 3'UTR varies considerably depending on CHIKV lineage and lies usually between 500 and 900 nucleotides. The UTR consists of several copies of repeated structure elements (RSEs) of unclear function, and one structurally conserved element (CSE), which is necessary for efficient genome replication[75]. In addition, all lineages exhibit a Y-shaped structure element of unknown function. Variation in the CHIKV 3'UTRs is predominantly introduced by the change in copy number of the repeat elements[76].

2.6.4 *CHIKV lineages*

The spatial segregation of virus strains caused a substantial diversification of the CHIKV genotype, which is accompanied by a drastically altered disease spectrum in humans. Newly evolved virulent capabilities were first encountered in the 2005 outbreak at La Réunion[67], and have become more frequent since then.

In the global circulation, a number of dominant CHIKV genotypes have been observed. They are organized in four distinct lineages, which refer to the geographic location of their (previous) circulation and are the Asian Urban lineage (AUL), Indian Ocean lineage (IOL), East-Central-South African lineage (ECSA) and West African lineage (WA). Currently, all lineages can be encountered around the globe and co-occurrence of multiple lineages in the same region is a common scenario. Recently, the IOL lineage caused outbreaks in Europa, despite being endemic to India and islands in the indian ocean. The ECSA lineage has been shown to be responsible for a major outbreak of CHIKV in Haiti[77,78] and new data from related virus isolates raise the question whether ECSA is to be considered a coherent lineage at all.

3 METHODS

This chapter introduces the most relevant methods and strategies that are applied throughout this thesis. It focuses especially on thermodynamic RNA secondary structure prediction, Consensus structure prediction and RNA structure homology search.

3.1 RNA FOLDING

The bases of an RNA polymer have the intrinsic property to form intramolecular hydrogen bonds with complementary bases, called 'pairs', leading to the formation of stretches of paired nucleotides ('stems'), and their enclosed nucleotides ('loops'). While the overall structure of an RNA molecule assumes a complex three dimensional structure *in vitro*, termed the 'tertiary structure', the local folds established by the base pairings remain remarkably stable regardless of their position relative to the rest of the molecule. Due to its remarkable stability, the entirety of an RNA molecules stems, loops and unpaired exterior bases is of particular interest and referred to as the **secondary structure**.

While the prediction of a tertiary structure from an RNA sequence is a computationally expensive process, secondary structure prediction is comparably simpler to solve, mainly due to the reduced degrees of freedom that describe a secondary structure. Multiple approaches have been devised for secondary structure prediction, approaching the problem from diverse angles such as probabilistic formal grammars[79], thermodynamics[80,81] or machine learning[82]. This chapter will focus predominantly on thermodynamic approaches, since they are used exclusively throughout this thesis.

3.1.1 *Thermodynamic RNA folding*

Folding approaches driven by thermodynamic considerations focus on finding the secondary structure that minimizes the equilibrium free energy (ΔG) of an RNA molecule, compared to the unfolded state. The rationale behind this approach lies in the fact that while an RNA molecule can adopt many structures *in vitro* and change between them seamlessly, the ensemble at thermodynamic equilibrium will be dominated by the structure that has the lowest free energy of all, accordingly termed Minimum Free Energy (MFE) structure.

This problem can be formulated in two independent parts that consist of (i) the definition of an Energy function, that determines the ΔG of a secondary structure and its sequence and (ii) finding the secondary structure that minimizes ΔG of the molecule among all possible structures that are valid for a sequence.

Typically used energy functions approximate the molecules' equilibrium free energy by summation of individual base pair ΔG contributions which are derived from a set of empirically determined ΔG values for each type of base pair[80,83,84]. The problem of finding the MFE structure is achieved by a set of dynamic programming algorithms that are able to efficiently explore the structure space of an RNA sequence. The following sections will elaborate more in detail on the details of those approaches.

Nussinov algorithm

The **Nussinov** algorithm[85] approaches the problem of finding the optimal structure by recursively evaluating all base pairs that could be formed by an RNA molecule. While the algorithm was conceived to find the secondary structure that maximizes the number of canonical base pairs, the same structure that maximizes base pairs is not necessarily the most stable one energetically. Fortunately, the Nussinov algorithm can be easily adapted to minimize for ΔG of the molecule, by inclusion of an energy term into the recursion:

$$E_{i,j} = \min \left\{ E_{i+1,j}, \min_{k,\,(i,k)pairs} \left\{ E_{i+1,k-1} + E_{k+1,j} + \beta_{ik} \right\} \right\} \tag{3.1}$$

where $E_{i,j}$ denotes the Minimum Free Energy of a structure formed by bases i to j and $\beta_i k$ represents the energy contribution resulting from base-pairing of nucleotides i and k.

Individual free energy contributions from base pairs (β_{ik}) are crucial for prediction of a correct secondary structure. Initial approaches differentiated only between contributions of individual types of base pairs, e.g. only the canonical base pairs, thereby neglecting the complex spatial topology an RNA molecule assumes *in vitro* and the resulting effects on the molecules free energy. A prominent example of such an interaction is the phenomenon of base stacking, where not the hydrogen bonds forming the base pair, but the aromatic components of nearby nucleotides greatly stabilize an RNA helix.

Zuker algorithm

In order to allow for a more granular free energy minimization, the **Zuker** algorithm[81] expands on the concepts laid out by the Nussinov algorithm and introduces additional recursion schemes and energy terms to the free energy minimization algorithm. It is able to differentiate between various types of loops (i.e. topologies) in an RNA structure, while the main goal remains the calculation of the structure with the Minimum Free Energy E_{ij}:

$$\begin{aligned}
E_{ij} &= \min \left\{ E_{i+1,j}, \min_{i \le k \le j} \left(C_{ik} + E_{k+1,j} \right) \right\} \\
C_{ij} &= \min \left\{ \mathcal{H}_{ij}, \min_{i \le k \le l \le j} \left(C_{kl} + \mathcal{I}\left(i,j;k,l\right) \right), \min_{i \le u \le j} \left(M_{i+1,u} + M^1_{u+1,j-i} + a \right) \right\} \\
M_{ij} &= \min \left\{ \min_{i \le u \le j} \left((u-i+1)c + C_{u+1,j} + b \right) \min_{i \le u \le j} \left(M_{i,u} + C_{u+1,j} + b \right), M_{i,j-1} + c \right\} \\
M^1_{ij} &= \min \left\{ M^1_{i,j-1} + c, C_{ij} + b \right\}
\end{aligned} \tag{3.2}$$

where individual recursions handle the case that single basepairs (C_{ik}) or multi-loops (M_{ij}) are formed, while the free energy contributions of hairpin loops ($\mathcal{H}_{ij}$) and interior loops ($\mathcal{I}_{ij}$) are not further decomposed but looked up in a table of empirically derived values. Implementations of the Zuker algorithm typically derive the free energy parameters from the nearest-neighbor dataset, maintained by Turner and Mathews[84].

3.1.2 *Consensus structure prediction*

Many types of RNA molecules are under selection pressure to conserve their secondary structure. Indications of conservation of structure results are typically the absence of any change of paired bases or, more nuanced, the simultaneous change of both bases involved in a pair, also termed compensatory mutation. While the pressure for structure conservation cannot be inferred from a single RNA sequence, evaluating variation and co-variation in a set of homologous or otherwise evolutionary related RNA molecules can provide evidence on positions that are likely to be involved in base pairings.

Taking (co-)variation into account, allows for the prediction of a *consensus structure* which represents the optimal secondary structure compatible with a set of related sequences. Consensus structure prediction is realized by addition of a covariation-dependent energy term into the classical Zuker algorithm. Formally, implementations like `RNAalifold`[86] extend the Zuker algorithm by quantifying covariation in an alignment of RNA sequences and assigning energy bonuses to individual base pairs. The covariation-dependent energy term $\gamma(i,j)$ is directly included as an additive term into the recursion C_{ij}:

$$C_{i,j} = \beta\gamma(i,j) + \min\left\{\mathcal{H}_{ij}, \min_{i\le k\le l\le j}\left(C_{kl} + \mathcal{I}(i,j;k,l)\right), \min_{i\le u\le j}\left(M_{i+1,u} + M^1_{u+1,j-i} + a\right)\right\} \tag{3.3}$$

where β is an empirically determined scaling factor.

The covariation itself is determined by pairwise evaluation of all basepairs $\alpha_i\alpha_j$ and $\beta_i\beta_j$ of alignment A via the help of RIBOSUM[87] like scoring matrices. These matrices quantify the likelihood of individual base pair substitutions and were created using curated alignments of tRNA sequences:

$$\gamma'(i,j) = \frac{1}{2}\sum_{\alpha,\beta\in A} xR(\alpha_i\alpha_j;\beta_i\beta_j) \tag{3.4}$$

where x is an empirically determined scaling factor.

In addition to evaluation of the covariance contribution, `RNAalifold` penalizes base pairs that would include gaps. If the number of sequences that are incompatible with a basepair is above a threshold, the base pair will be forbidden, otherwise a positive energy contribution is added if one or more bases α are not in the set of allowed base pairs B:

$$\gamma(i,j) = \gamma'(i,j) + \delta\sum_{\alpha\in A}\begin{cases}0 & \text{if } (\alpha_i,\alpha_j)\in B\\ 0.25 & \text{if } \alpha_i \wedge \alpha_j \text{ are gaps}\\ 1 & \text{otherwise}\end{cases} \tag{3.5}$$

where δ is an empirically determined scaling factor.

Since covariance contributions are exclusively evaluated on the basis of the aligned sequences, it is of importance to note that the accuracy of a consensus structure prediction is substantially influenced by the quality of the input alignment. Especially for alignments of distantly related and dissimilar sequences, common alignment algorithms such as `clustalo`[88] or `MAFFT`[89] will produce heavily gapped and thus uninformative alignments. Consensus structures predicted on the basis of such alignments will be of low accuracy, even if the sequences share a clearly conserved secondary structure. It is therefore imperative to carefully consider the choice of the alignment algorithm.

3.2 HOMOLOGY SEARCH

The search for homologous RNA sequences is one of the central topics in this thesis. Search algorithms often infer homologous sequences by looking for local sequence similarity, with BLAST[90] being a famous example of a similarity based search algorithm. The evolution of structured RNA on the other hand, is subject to constraints that arise from the secondary structure itself. Functionality in an RNA molecule is often determined by its secondary structure, often allowing significant changes in sequence provided the structure remains intact. Structurally homologous RNA sequences can be therefore vastly dissimilar from each other in terms of sequence, posing a challenge for traditional homology search algorithms.

The phenomenon of structural homology can be frequently observed in fast evolving species like viruses. Fortunately there are algorithms that can alleviate the problem of structural homology search and homology search in dissimilar sequences. The remainder of this section will focus on explanation of Hidden Markov Models and Covariance models, as well as on formal grammars, which represent their theoretical foundation.

3.2.1 *Formal Grammars*

Formal grammars are a way of describing and specifying the structure of a languages. A formal grammar consists of a set of *production rules* that define how to construct valid strings of symbols. Symbols are grouped into two distinct classes called *terminal* and *non-terminal* symbols, where only terminal symbols actually appear in the observed strings while non-terminals can be transformed into a new string of terminals or non-terminals. More specifically, production rules specify how nonterminals can be replaced by terminals and other nonterminals in order to generate a valid string in the language.

While formal grammars were originally conceived in the context of linguistics, their principles can be applied to numerous other topics beyond the study of natural language. For example, when modeling a sentence a terminal would represent a word, but terminals could also represent amino acids when modeling a protein sequence, or nucleotides when modeling DNA or RNA sequences.

Chomsky[91] introduced four types of grammars which each have a distinct set of restrictions on their production rules. They are organized in a hierarchical fashion known as the *Chomsky hierarchy*. In the following section, W denotes any nonterminal, α denotes any string of terminals and β any string of terminals and/or nonterminals:

- **Regular grammars** are the most restricted grammars and only allow production rules in the form of $W \rightarrow \alpha W$ and $W \rightarrow \alpha$, its valid strings can be parsed using a finite state automaton. Notable applications of regular grammars are *Regular Expressions* or *HMMs*.

- **Context-free grammars** allow production rules of the form $W \rightarrow \beta$, its valid strings can be parsed using a push-down automaton. Notable applications of context-free grammars are *Covariance Models*.

- **Context-sensitive grammars** allow production rules in the form of $\alpha_1 W \alpha_2 \rightarrow \alpha_1 \beta \alpha_2$. Its valid strings can be parsed by a linear bounded automaton.

- **Unrestricted Grammars** impose no limitations on production rules. Valid strings of an unrestricted grammar can be parsed using a turing machine.

The production rules of all types of grammars can furthermore have probabilities associated with them. The grammar is then referred to as **stochastic** (or probabilistic), whereas a grammar without probabilities is also referred to as **deterministic**. In contrary to deterministic grammars, where each non-terminal symbol has one unique production rule, stochastic grammars allow for multiple ways to expand a non-terminal symbol. Probabilities are usually learned from data the grammar models, for example the position-specific nucleotide distribution for modeling of DNA sequences can be learned from an alignment of homologous sequences.

From a practical perspective, only stochastic regular grammars and stochastic context-free grammars are applied on a large scale. While there are efficient algorithms for parsing strings of regular grammars and context-free grammars, no general polynomial-time algorithm for parsing of context sensitive grammars has been described. Likewise, there is no general algorithm that is guaranteed to parse an Unrestricted grammar in less than infinite time.

3.2.2 *Hidden Markov Models*

Hidden Markov Models[92] (HMMs) are probabilistic models that are built on the concept of a Markov process. The classical Markov process describes transitions between states with the restriction that transitions must be memory-less, i.e. any state must only be dependent on its previous state. Hidden Markov Models expand on the Markov process concept by differentiating between states that produce visible output and states that do not emit an observable, thus are *hidden* from the outside. Hidden states allow a more complex modeling of the processes that shape the underlying data and have been successfully applied to a wide range of problems such as speech recognition[92], sequence alignment[93] or classification of discharge patterns in electrical engineering[94].

From a formal perspective HMMs are an example of a stochastic regular grammar where non-terminal states constitute the hidden states, while terminal states produce the observables. This means that also the efficient algorithms that exist for parsing of regular grammars can be applied to HMMs. Most notably the *Viterbi* algorithm[95] computes the most likely path through the HMM, while the *Forward* algorithm[96] calculates the probability of an observation given an HMM.

HMMs excel at modeling data that can be reconciled with the Markov property, i.e. data where individual states are uncorrelated and depend just on one underlying state. Problems arise when there are dependencies between remote states, as typically seen in RNA structure analysis. In RNA, many positions in a nucleotide are involved in base pairs, which result in long-range pairwise correlation between states. So while one state (residue) might be any nucleotide, there will be a state in an arbitrary distance that needs to preserve complementarity to the first one. Such correlations cannot be modeled (or "remembered") by any HMM or any other model based on a regular grammar, but can be addressed by higher-order grammars in the Chomsky hierarchy.

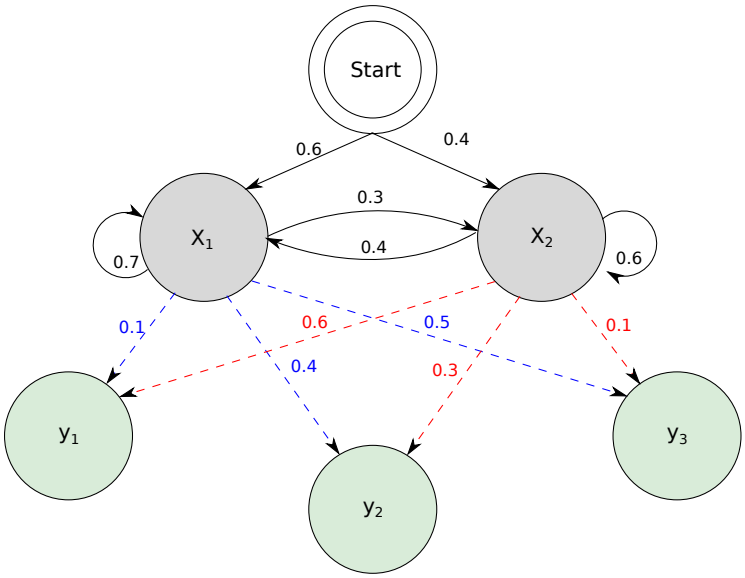


Figure 11: Example Hidden Markov Model with two hidden (grey) and three visible (green) states. Arrows indicate transitions between hidden states (solid lines). Emissions are indicated by dashed arrows between hidden and emission states. Each transition/emission has an associated probability. Figure adapted from Wikipedia.[97]

3.2.3 *Covariance Models*

The concept of a Covariance Model (CM) expands the ideas of HMMs by introducing the possibility to model dependencies between remote states. CMs are stochastic context-free grammars (SCFG) and were specifically designed to simultaneously model RNA sequence and RNA secondary structure in homology search[98]. Especially production rules of the type $W \longrightarrow \alpha W \alpha$ are ideally suited for the modeling of base-paired structures.

In analogy to HMMs, a different set of algorithms exist for efficient parsing of valid strings of a CM. The *CYK* algorithm produces the most probable sequence of a CM, while the *Inside* algorithm[99] calculates the probability of a sequence given a CM. While achieving the same goal, it is of importance that both the CYK and Inside algorithm

While a CM can therefore find homologies that remain hidden to HMMs, it comes at the price of steeply increased computational cost. Both the CYK and the Forward algorithm have a run time complexity of $O(L^3M^3)$ and memory consumption of $O(L^2M)$, where L is the length of the string and M the number of CM states, rendering them prohibitively costly for large scale applications such as genome-wide screens. This problem is often remedied by pre-filtering with cheaper methods such as HMMs before further evaluation of CMs.

3.2.4 *Infernal*

The `infernal` package (INFErence of RNa ALignment) is the central implementation of CMs and provides a collection of tools that are essential to the use of CMs.[101].

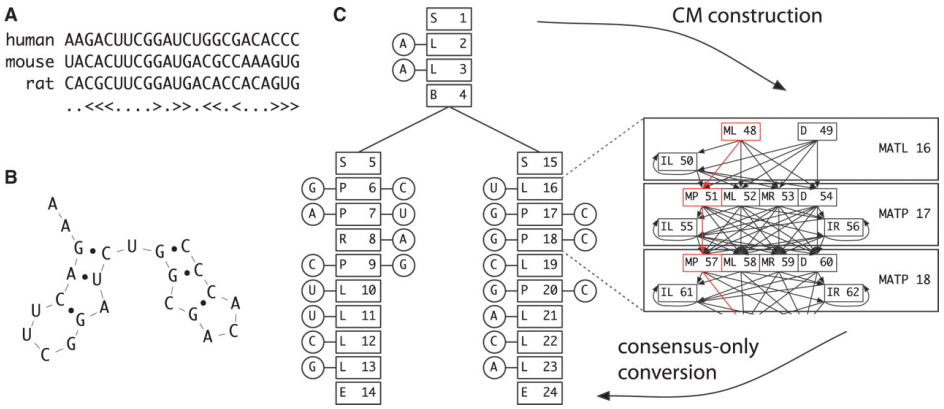


Figure 12: Exemplary Covariance Model for a simple alignment (**A**) and its associated secondary structure (**B**). The parse tree (**C**) depicts the parse tree of the CM given the consensus structure of the alignment. The right portion shows in more detail a full section of the CM with insert and delete states. Figure taken from Kolbe et al.[100]

cmbuild

`cmbuild` builds a covariance model from multiple sequence alignment and its consensus secondary structure. Additionally, it builds a profile HMM from the same alignment, which can be used downstream in fast pre-filtering for putative CM hits. It is important to note that `cmbuild` does not predict any consensus secondary structure, but accepts output of `RNAlifold`.

cmcalibrate

In order to quantify the likelihood of a match between CM and query sequence, the CM needs to undergo a calibration step with `cmcalibrate`. The process of calibration consists of repeated generation of random sequences, aligning against the CM and tabulating the bit-score. After evaluation of a sufficient number of sequences, an exponential distribution is fitted on the histogram of bit-scores. The distribution is used for E-value estimation when performing searches with the calibrated CM. The E-value[102] quantifies the likelihood that the match would occur by chance in an equivalently sized search space and is defined as:

$$E = \frac{N}{2^{S'}} \tag{3.6}$$

where E is the E-value, N is the search space size and S' is the normalized bit score. The original bit-score S is directly provided by `infernal` and can be converted using:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \tag{3.7}$$

Where S is the raw bit-score, and K and λ are scaling factors for the search space and scoring system.

cmsearch

`cmsearch` is the component of the infernal package which performs the homology search of calibrated CM s against a sequence database. Since matching a CM is

a computationally intensive task, due to the CYK algorithm scaling with $O(n^3)$, `cmsearch` offers a series of pre-filters, such as pre-filtering putative for CM matches with the help of the HMM created during the `cmbuild` step. While pre-filtering comes at the cost of sensitivity, it can speed up the search process drastically. The output of `cmsearch` contains the hit nucleotide sequences and coordinates, as well as the bit-score and its corresponding E-value.

3.3 EVALUATING RNA STRUCTURE

The main focus of this thesis is to identify functional RNA structures through computational means. Prediction of secondary structure from an RNA sequence is a well-established task, however, in most cases a secondary structure alone is not sufficient to deduce any biological function. In this thesis, functionality is broadly defined as any biological activity that is mediated by the RNA structure. Although functionality is usually identified *de novo* through experimental methods, there are algorithms that can estimate the functional relevance of an RNA structure. The research presented in this thesis heavily relies on these algorithms, and the following section will provide more details on the two most significant methods.

3.3.1 *RNAz*

The `RNAz` algorithm[103,104] was devised for genome-wide screens for functionally active non-coding RNAs. Provided an alignment of RNA sequences, RNAz can classify whether it encodes an RNA structure that resembles already known functional RNA structures. At the heart of RNAz is a Support Vector Machine (SVM) that classifies the alignment on the basis of three distinct parameters, its (i) Average minimum free energy z-score, (ii) Structure Conservation Index (SCI) and (iii) the normalized Shannon Entropy of the Alignment.

The Average z-score $\overline{z}$ of each structure x of alignment $\mathcal{A}$ quantifies how the average free energy of an alignment relates to what can on average be expected from sequences with similar dinucleotide content. It is defined as

$$\overline{z} = \frac{1}{N} \sum_{x \in \mathcal{A}} \frac{E_x - \mu_x}{\sigma_x} \tag{3.8}$$

where E_x is the minimum free energy of structure x, and parameters μ and σ denote the average energy and standard deviation of sequences with similar GC content. Parameters μ and σ are estimated through regression via a separate SVM that was trained on a large number of simulated and folded RNA sequences. Since functional RNA structures are usually more stable than chance would dictate, low ($\leq$)0 z-scores alone are already indicative of biologically relevant structures and are therefore also useful for interpretation beyond the context of RNAz.

The Structure Conservation Index *SCI* for alignment $\mathcal{A}$ relates the average consensus folding free energy $E_{consensus}$ to the individual minimum free energies of the individual sequences of the alignment:

$$SCI = \frac{E_{consensus}}{\frac{1}{N} \sum_{x \in \mathcal{A}} E_x} \tag{3.9}$$

The SCI is primarily intended as a measure on how coherently the individual sequences of an alignment support the overall consensus structure. Regardless the fact

that similar free energy values do not necessarily imply structural similarity, the SCI has proven to be a powerful metric for the assessment of structural homogeneity of an alignment, since similar structures will tend to have similar minimum free energies.

The Normalized Shannon Entropy H of the alignment is a term intended to quantify the size and sequence variation of an alignment. The concept is taken directly from the "Shannon Entropy" as described in information theory[105] and applied to each column of the alignment:

$$H = -\frac{1}{L}\sum_{i}^{L}\sum_{\alpha\in\sum} p_{\alpha}^{i}\log 2p_{\alpha}^{i} \tag{3.10}$$

where p_{α}^{i} is the probability of character α in alignment column i. Inclusion of the Normalized Shannon entropy allows the classification of alignments of an arbitrary size, as the term combines both sequence variation and the number of sequences into one measure that is scaled by the number of sequences.

All three parameters are provided as features to the RNAz Support Vector Machine, which classifies the alignment into either "structural RNA" or "other" alongside a probability estimate. Training of the SVM classifier is based on release 9.1 of the *Rfam* database[106], a comprehensive collection of manually curated RNA families. Even though the algorithm has a False Discovery Rate of roughly 50%[104], the SVM classificator is still sufficiently able to discriminate alignments with thermodynamically stable and/or structurally conserved RNA sequences from unstructured regions.

In the context of this thesis, `RNAz` is used as a means to assess the relevance of RNA structure in an alignment. While it cannot interpret or infer the biological functionality of an RNA structure, a high RNAz class probability indicates strong similarity to already known functional RNAs and warrants for further investigation of a candidate RNA structure.

3.3.2 *R-Scape*

The tool `R-Scape`[107,108] is intended to identify statistically significant covariation between columns in an alignment that might be indicative for conserved secondary structure. The use of covariation in secondary structure prediction has already been discussed in the context of the `RNAalifold` algorithm, which used a Mutual-Information (MI) derived metric in order to quantify covariation and translate it into thermodynamic parameters. `R-scape` focuses on evaluation and statistical testing of the covariation itself, rather than consensus structure prediction thereby providing additional insight into the relevance of an RNA secondary structure.

Methodically, `R-scape` evaluates the significance of covaration between columns i and j of an alignment. While it offers a broad range of metrics, by default the G-Test[109] is employed. The G-Test is defined as

$$\mathrm{GT}(i,j) = 2\sum_{a,b}\mathrm{Obs}_{ij}^{ab}\log\frac{\mathrm{Obs}_{ij}^{ab}}{\mathrm{Exp}_{ij}^{ab}} \tag{3.11}$$

where Obs_{ij}^{ab} is the observed count of a:b pairs in columns i,j (excluding gaps) and

$$\mathrm{Exp}_{ij}^{ab} = N_{ij}p_{i}^{a}p_{j}^{b} \tag{3.12}$$

is the expected frequency of pair a:b under the assumption that columns i and j are independent. p_a^i are the marginal frequencies of a residues in column i (averaged to all other positions).

In order to assess the statistical significance of a base pairs, expressed through an E-value, R-scape derives an empirical distribution of covariation scores by evaluation of covariation on simulated alignments. The simulation procedure estimates a maximum-likelihood tree for each input alignment and then infers a maximum parsimony assignment at each alignment column i. The simulation procedure therefore maintains the phylogenetic relationship between the alignment sequences and evolves columns independently from each other. Additionally, the covariation potential for each column is estimated[108] in order to exclude columns with little possible covariation from the final evaluation.

R-scape evaluates the covariation support for all base pairs in a structure if a consensus structure is provided as input with the alignment. If no consensus structure is supplied, R-scape will evaluate all possible base pairs and then generate its own structure prediction using Tornado, a SCFG-based algorithm. The prediction will be constrained by base pairs with significant covariation support by default.

4 GOALS OF THIS THESIS

Viruses represent a unique chance and challenge for structural RNA bioinformatics. In sharp contrast to other domains of life, especially RNA viruses exhibit high rates of evolution and sequence change, but are simultaneously under pressure to conserve any functional RNA structure. The plasticity of the RNA structure and its ability to accommodate for sequence changes therefore provides an ideal testing ground for RNA virus evolution.

The main interest that pervades this thesis, is to study the marks evolution has left on viruses through examination of their RNA structure. The subsequently presented goals specify how this thesis intends to add knowledge to the general question.

GOAL 1: DETERMINE THE IMPACT OF RNA STRUCTURE IN VIRUSES

One of the goals of this thesis was to investigate the relevance of RNA structure in the viral world in the context of different virus species and their unique ecological niches. Furthermore, in order to understand the general role of RNA in viruses, investigations should not only focus on individual virus species, but also more general taxonomic groups of viruses, possibly identifying general trends that shape RNA structure in viruses. Evolution of viral RNA structure should be studied in species where it was established to be indispensable for survival and vice versa.

GOAL 2: DEVELOP METHODS FOR STUDY OF RNA STRUCTURE IN VIRUSES

RNA viruses have a rate of evolution surpassing most other organisms, including most other viruses. This is best seen in the available sequence data for many viruses, which is characterized by sequence divergence. Meanwhile, algorithms frequently used in virus bioinformatics originate from the general ecosystem of tools used for comparative sequence analysis, where coherent sequence data is more widespread. As a result, investigation of viruses is often challenging, a fact that is further exacerbated by the very nature functional RNA structure, which does not necessarily require conservation of nucleotide sequence.

This thesis aimed at exploring and promoting computational approaches that address the challenges of RNA structure discovery in viruses. It should demonstrate the power of stochastic modeling approaches such as covariance models when applied to problems in virus bioinformatics. Likewise, the utility of RNA families for the description of viral structure elements should be demonstrated to a broader audience.

GOAL 3: CHARACTERIZE NEW RNA ELEMENTS IN FLAVIVIRUSES

Flaviviruses depend on RNA structure mediated functionality for survival to a degree where the evolution of parts of their genome architecture is at large determined and shaped through structure elements. Despite the importance of RNA structure, only a handful of elements have been investigated systematically and verified in selected Flavivirus (FV) species, while a large portion of FV structure elements are based on speculation or anecdotal evidence.

It was a concrete aim of this thesis to organize and advance the knowledge on FV structure elements. In order to tackle this complex matter, the problem needed to be approached from different angles: Existing elements needed to be verified on an algorithmic basis, using experimental evidence where available. A comprehensive picture of the occurrence and distribution of existing elements needed to be obtained from *all* FV species, many of which had not yet been investigated. Better understanding of already known elements furthermore allowed the discovery of new elements and enabled the structural mapping of all structure elements in FV 3'UTRs.

The combined knowledge gain yielded detailed qualitative or quantitative understanding of the architecture of FV 3'UTRs, the dominant evolutionary forces shaping its development, and the evolutionary trajectory of Flaviviruses. Additionally, the investigation also provided new evidence on which functionalities of the structure elements are mediated purely by structure and which are dependent on sequence.

INDIVIDUAL PROJECTS

The work that is presented in this thesis is distributed over three distinct projects which independently contribute to the main goals.

4.0.1 *Project Chikungunya*

Chapter 5 presents a comprehensive examination of the evolutionary history of the Chikungunya Virus (CHIKV) genome, based on recent data. The findings from the analysis provided crucial insights into the assignment of CHIKV isolates into epidemiologically significant lineages, but furthermore revealed novel details on the evolution of the 3'UTR structure. This study represents a study that focuses on a single, clinically relevant virus species that has received limited attention on the importance of RNA structure. The most important project-level goals for this project included:

1. Creation of an updated phylogeny for CHIKV, based on a comprehensive and well-curated data set
2. Revision of existing CHIKV lineages on the basis of an updated phylogeny
3. Development of a protocol for *in silico* RNA structure discovery in CHIKV 3'UTRs
4. Identification of universal and lineage-specific conserved RNA structure elements in CHIKV

Contribution to Thesis Goals

In the scope of this thesis, this project investigated the (structural) evolution of a single virus species that lacks any immediately obvious pressure for the conservation of secondary structure. The project had the added challenge that for the examined virus species there exists limited prior evidence for any significant RNA structure, which could otherwise serve as a reference point for an investigation. In the line of the same investigation, the project explored how a viruses phylogeny and lineage classifications can be reconciled with each other in order to draw a coherent picture of its evolution.

4.0.2 *Project Flaviviruses*

Chapter 6 presented an in-depth study of the structural 3'UTR architecture of all member species of the *genus* Flavivirus. In contrast to the previous project, this study extended the scope of the investigation from one well defined species to a full *genus* of viruses, a taxonomic level where sequence similarity between species can be expected to diverge. While a large divergence in sequence should make structure conservation equally unlikely, several FV species are known to depend on structural RNA elements for survival, offering the opportunity for a more granular analysis. The most important project-level goals for this project included:

1. Methods Development for RNA structure screening in divergent sequences
2. Curation of a large number of FV genomes
3. Identification of known structure elements in previously unstudied FV species
4. Search for novel structure elements in FVs

Contribution to Thesis Goals

In the scope of this thesis, this project explored a scenario where virus evolution is primarily driven by the conservation of RNA structure, rather than sequence content. It evaluated how homologous RNA structures evolve across a large group of related virus species, when the structure-mediated functionality is crucial for survival and presents an example of how RNA functionality can drive the evolution of an entire genus. Additionally, the projects explored methodological difficulties associated with analysis of RNA structure in highly divergent sequences,

4.0.3 *Project RNA structure conservation in Viruses*

Chapter 7 presents an exploratory study that assessed the relevance of RNA structure in the entire virome. The size and nature of this study entailed that it could not focus on the identification and characterization of single RNA structures in individual virus species or groups, but rather focused on the identification and quantification of evolutionary trends that have shaped RNA structure in all viruses. The most important project-level goals for this project included:

1. Definition of appropriate metrics for quantification of RNA "Structuredness"
2. Assessment of RNA structuredness in viral *all* mRNAs
3. Identification of trends in RNA structuredness among all virus families

Contribution to Thesis Goals

In the scope of this thesis, this project assessed on a large scale the occurrence and significance of RNA structure over an entire taxonomic kingdom. Due to the scale of the investigation, rather than reconstructing individual evolutionary histories of individual viruses, the study identified large-scale evolutionary trends that have shaped virus families. The project provided an appropriate context to the other two projects, which focused on RNA structure evolution in smaller taxonomic groups.

5 RNA ARCHITECTURE OF CHIKUNGUNYA VIRUS LINEAGES

The following chapter presents the results of the work conducted on the phylogeny and Untranslated Region (UTR) architecture of the Chikungunya Virus. The work was published in *Viruses. 2019 Sep; 11(9): 798*, A reprint of the article is provided as part of the chapter.

5.1 MOTIVATION

While the Chikungunya Virus (CHIKV) has been extensively studied experimentally due to its high relevance to virologists and epidemiologists, comparably little rigorous computational research has been conducted regarding its phylogeny or genome organization. Most prominently, the experimentally-driven focus becomes evident when considering the state of the CHIKV phylogeny and its further division into lineages, which are inconsistent and hard to reconcile with the genomic reality of CHIKV. Furthermore, there exists surprisingly little work regarding the genomic differences that discriminate individual CHIKV lineages from each other.

In addition, the current status quo was established on relatively sparse data and does not consider nor make use of the majority of currently known CHIKV genome sequences. A cornerstone of this project was therefore to build a curated and as comprehensive as possible collection of CHIKV genomes, that would serve as the basis for all investigations. Experience has shown that, having access to an increased amount of data will allow for fine-grained phylogenetic analyses via deep alignments and accurate secondary structure predictions by applying consensus folding.

5.2 ADDED VALUE

The results published in the article provide a granular analysis of the genomics of the Chikungunya virus and its lineages. Through the utilization of all data that was publicly available at the time of publication, experimentalists can rely on a comprehensive and detailed overview on CHIKV genomics, lineage composition and a transparent picture of isolate-lineage association. Especially the insight into lineage composition and dynamics can be especially relevant for the understanding future outbreaks of CHIKV. While the work addresses the often vaguely defined problem of virus isolate to lineage assignment by an approach of reconciliation of phylogeny and epidemiological data, the problem cannot be regarded as solved and further improvements should be encouraged.

We furthermore systematically investigated the UTRs of CHIKV for any signal of conserved RNA structure. The investigation surprised on one hand as it showed among most isolates little sequence variation that would suggest the conservation of RNA secondary structure. While the overall UTR architecture between lineages can vary and a few minor structure elements were identified, future investigations could shift the focus rather on the search for factors that interact with CHIKV UTRs in ways that are not mediated by RNA secondary structure.

5.3 REPRINT OF: UPDATED PHYLOGENY OF CHIKUNGUNYA VIRUS SUGGESTS LINEAGE-SPECIFIC RNA ARCHITECTURE

ADRIANO DE BERNARDI SCHNEIDER†, ROMAN OCHSENREITER†, REILLY HOSTAGER, IVO L. HOFACKER, DANIEL JANIES, AND MICHAEL T. WOLFINGER

Viruses, **2019**, 11, 798.
https://doi.org/10.3390/v11090798

Contributions:

ROMAN OCHSENREITER† Curated the comprehensive CHIKV data set, conducted the RNA structure analysis, built the CHIKV phylogeny and wrote the manuscript.

ADRIANO DE BERNARDI SCHNEIDER† Conceived the study, conducted the phylogeny analysis of CHIKV epidemic clades and wrote the manuscript.

REILLY HOSTAGER Contributed to the phylogeny analysis of CHIKV.

DANIEL JANIES Supervised the project and contributed to the manuscript.

IVO HOFACKER Supervised the project and contributed to the manuscript.

MICHAEL T. WOLFINGER Conceived the study, curated the dataset of Chikungunya virus UTRs and wrote the manuscript.

† *These authors contributed equally.*

viruses

MDPI

Article

Updated Phylogeny of Chikungunya Virus Suggests Lineage-Specific RNA Architecture

Adriano de Bernardi Schneider [1,†], Roman Ochsenreiter [2,†], Reilly Hostager [1], Ivo L. Hofacker [2,3], Daniel Janies [4] and Michael T. Wolfinger [2,3,*]

1 AntiViral Research Center, Department of Medicine, University of California San Diego, San Diego, CA 92103, USA
2 Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria
3 Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria
4 Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
* Correspondence: michael.wolfinger@univie.ac.at
† These authors contributed equally to this work.

Received: 10 July 2019; Accepted: 26 August 2019; Published: 29 August 2019

Abstract: Chikungunya virus (CHIKV), a mosquito-borne alphavirus of the family Togaviridae, has recently emerged in the Americas from lineages from two continents: Asia and Africa. Historically, CHIKV circulated as at least four lineages worldwide with both enzootic and epidemic transmission cycles. To understand the recent patterns of emergence and the current status of the CHIKV spread, updated analyses of the viral genetic data and metadata are needed. Here, we performed phylogenetic and comparative genomics screens of CHIKV genomes, taking advantage of the public availability of many recently sequenced isolates. Based on these new data and analyses, we derive a revised phylogeny from nucleotide sequences in coding regions. Using this phylogeny, we uncover the presence of several distinct lineages in Africa that were previously considered a single one. In parallel, we performed thermodynamic modeling of CHIKV untranslated regions (UTRs), which revealed evolutionarily conserved structured and unstructured RNA elements in the 3'UTR. We provide evidence for duplication events in recently emerged American isolates of the Asian CHIKV lineage and propose the existence of a flexible 3'UTR architecture among different CHIKV lineages.

Keywords: RNA virus; evolution; epidemics; phylogeography; secondary structure

1. Introduction

Phylogeographic inquiries can reveal viral transmission patterns and pathogenesis of viral diseases. Dramatic (re)emergences of arthropod-borne viruses (arboviruses), such as Chikungunya (CHIKV), West Nile, Zika, and Dengue, have increased in frequency and intensity, thus raising public health concerns [1]. Prolific expansion of arboviruses in countries where the viruses have previously not been recorded has drawn public attention. These trends are considered to be a function of climate change effects on the geographic range in which the vector can complete its life cycle and transmit CHIKV [2,3]. Here, we investigate the RNA architecture of CHIKV untranslated regions (UTRs) that underlies their ability to adapt and transmit in new environments and provide an updated large-scale phylogeny summarizing the geographic spread of CHIKV lineages.

CHIKV is a mosquito-borne virus that belongs to the Alphavirus genus, which encompasses mainly arboviruses, and is believed to have a marine origin [4,5]. The presence of CHIKV in sylvatic mosquitoes, such as *Aedes africanus*, *Aedes furcifer*, and *A. taylori*, and non-human primates in Uganda and Tanzania suggests that the virus originated in Central or East Africa [6,7].

The first isolation of CHIKV was in 1952 in the Newala district of Tanganyika, a British territory that is now part of Tanzania [8,9]. CHIKV has caused multiple outbreaks in all continents except Antarctica, most recently in Europe and Central and South America. In total, CHIKV has been present in over 100 countries [6,10–12].

CHIKV infects multiple human tissues, causing a febrile illness which is typically characterized by myalgia, polyarthralgia, polyarthritis, and rashes [13]. Several vaccine candidates are currently in preclinical or clinical development [14].

The virus had not been associated with life-threatening symptoms before an outbreak in 2005 on La Réunion island off the East African coast [15], which demonstrated a drastic expansion of the disease spectrum. In La Réunion, and after, higher morbidity has been associated with CHIKV infection, including neurological issues, such as visual and hearing loss, paralysis, Guillain Barré Syndrome (GBS) and renal complications [13,16,17]. Recent mutations that change protein structure have been attributed to increased ability to establish infection (infectivity) and outbreaks as the virus becomes better adapted to vector species (e.g., *A. albopictus*, the only suitable agent for transmission in La Réunion island) [18]. Species of the Aedes genus, in particular *A. albopictus*, *A. aegypti*, and *A. polynesiensis* are carriers of CHIKV, and are capable of transmission to humans [19].

1.1. Chikungunya Lineages

CHIKV is a virus for which four major lineages have been proposed: Asian Urban (AUL), Indian Ocean (IOL), East, Central and South African (ECSA), and West African (WA) [20,21]. Despite these notions, the origin and the history of how CHIKV spreads inside of Africa remains elusive. This gap in our understanding is due to undersampling and confounding factors such as co-circulating Dengue and Yellow Fever viruses. The co-circulation of viruses diminishes the value of serology test which are not species specific. While the AUL has historically been constrained in Southeast Asia, there have been recent outbreaks in the Americas [18]. The IOL, circulating in India and surrounding regions, has been mainly restricted to cause outbreaks in South and Southeastern Asia, although it has caused recent outbreaks in Europe. The WA lineage is entirely isolated.

The ECSA lineage has a unique history as it gave rise to the IOL as well as has also been identified as the origin of recent outbreaks in Brazil and Haiti [12,22,23]. The vast territory in Africa, the lack of sampling, and increasing diversity of ECSA lineage foster doubt on whether ECSA is a single lineage.

1.2. Conserved RNA Structure as Evolutionary Trait

The CHIKV genome consists of a single-stranded, (+)-sense RNA of approximately 11.8 kb, encoding nine proteins, which are divided into structural and nonstructural proteins [24]. The genome is organized as two open reading frames (ORFs), flanked by structured UTRs [25].

Populations of RNA viruses exhibit large genetic variability. This variation allows populations to have variants that are suited to changing conditions and immune escape. Viral variation results mainly from two mechanisms, mutation and recombination. Mutation occurs during replication by mis-introduction of nucleotides by error-prone RNA-dependent RNA polymerases that lack proof-reading capacity. Recombination occurs when co-circulating lineages co-infect a host and mixed segments are incorporated into the descendent viruses [26].

The role of functional RNA elements in pathogenesis is well established in flaviviruses [27,28], where they are typically conserved evolutionarily in UTRs [29–31]. In CHIKV, patterns of coupled historic recombination and mutation events are encoded in the 3'UTRs of different lineages [32]. The resulting variants have highly divergent 3'UTRs that vary in length. Characteristics of these 3'UTR variants are variable number of direct repeats (DRs) at the sequence level [2,33]. Distinctive CHIKV lineages also differ in their 3'UTR architecture, in particular in the copy number and arrangement of sequence repeats [34].

Evolutionary conservation in RNA viruses is typically maintained in secondary structure and manifested in rich base pair-conserving covariation patterns [35]. Relatively little is known about

structured RNAs in alphaviruses, both in coding and non-coding regions. Coding regions evolve under strong selection pressure to maintain protein function. Nevertheless, it has become apparent that structure within coding regions can also have biological function, as seen in the HIV RRE element [36]. In the context of alphaviruses, Kutchko et al. [37] performed RNA structure probing experiments, focusing on Sindbis virus (SINV) and Venezuelan equine encephalitis virus (VEEV), but could not find evidence for broad conservation of functional RNA structures.

The knowledge of structured RNAs in alphavirus UTRs, and CHIKV in particular, is limited. This limitation results from the fact that earlier studies focus on lineage-specific sequence repeats and could not reliably establish a relationship between all observed repeat patterns and RNA secondary structure [32,34].

A large number of genomic sequences and metadata have been deposited recently to public repositories. These data have raised the possibility of revising the lineages of CHIKV as a whole, reevaluating the historical global spread of the major epidemic clades and evaluating the RNA structural patterns among different clades. We also set out to structurally characterize the UTRs of all known CHIKV lineages, and found evidence for lineage-specific structured and unstructured repeat elements. We hypothesize such RNA elements to be involved in viral replication, vector adaptation and specificity, host factor binding, and pathogenicity.

2. Materials and Methods

2.1. Taxon Sampling

We obtained viral genome data from the National Center for Biotechnology Information (NCBI) `Genbank` database [38]. We downloaded 598 whole genome sequences of CHIKV. We compiled all geographic metadata available in these genome records to create data sets for the analyses described below. We removed eight sequences from the data set due to missing geographic location metadata or designation as a vaccine sequence. The metadata associated with geographic location was associated with the United Nations geoscheme for consistency. Therefore, "Central Africa" is labeled "Middle Africa".

2.2. Phylogenetic Tree Search

Two complete ORFs of CHIKV were extracted from the complete viral sequences for the phylogenetic analysis. Multiple sequence alignments of the individual ORFs were computed using `MAFFT` [39] under default settings and visualized in `AliView` [40]. In order to examine the possibility of recombinant sequences, we used RDP and methods therein (RDP, GENECONV, Bootscan, MaxChi, Chimaera, SiScan, and PhylPro) [41].

We conducted phylogenetic tree searches with partitioning to separate the two ORFs. We used tree search method under the optimality criterion of maximum likelihood with substitution model testing as implemented in `IQ-Tree` [42].

We executed a 1000-replicate ultrafast bootstrap analysis of clade frequencies [43] and a SH-aLRT measure of support [44]. The ultrafast bootstrap analysis is a variation of the traditional bootstrap using heuristics and constraints to speed the search for optimal tree topologies. The SH-aLRT is an approximation of the likelihood ratio, a direct measure of how much the evidence supports the hypothesis.

According to recently published phylogenies of CHIKV [20,22,45], viruses from the Western African clade form a sister group to all other CHIKV clades. Given these results and the absence of co-circulating Western African strains with any other CHIKV lineage, we selected this clade as the outgroup for the phylogeny of CHIKV. de Bernardi Schneider [45] observed Western Africa as a monophyletic clade sister to all other clades by utilizing an external outgroup (O'nyong nyong virus), thus the selection of this group should not yield an inherent bias on the tree topology at the same time that it results in a cleaner molecular sequence alignment given the outgroup sequences belong to

the same virus species as the ingroup. We performed the same procedure for a subset containing the 111 complete 3'UTR sequences from the 590 whole genome sequence data set.

The final trees with metadata annotation were rendered using custom Python scripts from the `ETE3` framework [46].

2.3. Global Spread Of Virus

We extracted the list of taxa of the three major epidemic clades, Asian Urban Americas (AUL-Am), ECSA-IOL, and ECSA-Middle African South American (ECSA-MASA), from the phylogenetic tree generated from all CHIKV sequences. We created a csv table including all locations from the dataset with corresponding latitude and longitude extracted from getlatlong.net. We performed individual alignments for each epidemic clade on concatenated ORFs and ran these on an R script pipeline (described here). Our script performs five steps:

1. Tree search under the neighbor-joining algorithm with 100 bootstrap replicates [47,48].
2. Collapsing of branches with a bootstrap value under 80% [49].
3. Replacing tree tip names with metadata information (i.e., geographic location).
4. Parsimony ancestry reconstruction using the *asr_max_parsimony* function from Castor R Package [50,51] to build an edge list (i.e., origin to destination geographic state transitions).
5. Render the geographic state transitions reconstructed on the previous step on a map using geographic coordinates for each location [52,53].

In order to evaluate the historical spread of CHIKV using a parsimonious ancestral reconstruction, the tree must be rooted to infer event directionality. Thus, we selected a root for each clade based on the complete CHIKV phylogeny. For each individual clade, the sister taxon to all other taxa was selected. When a clade had multiple sister taxa, we randomly selected one as the root. We selected HM045788.1 (India) as the outgroup for AUL-Am, KF283986.1 (Comoros) for ECSA-IOL and KX262996.1 (Cameroon) for ECSA-MASA.

This script is an adaptation from `StrainHub` [50], a web-based application to reconstruct pathogen transmission networks, to visualize the spread of the individual lineages/epidemic clades. The scripts and data sets along with a detailed explanation of the functions and packages used for the analyses described here are available online via the Supplementary Materials GitHub repository.

2.4. Conserved RNA Structures in CHIKV 3'UTR

Evolutionary conservation of functional RNAs is typically achieved by nature at the level of secondary structures. Comparative genomics approaches aim at detecting these conserved, structured RNAs as patterns of covariation in genome sequence alignments, typically constructed from phylogenetically narrow subgroups. To this end, Covariance Models (CMs) are statistical models of RNA structure formation that extend traditional Hidden Markov Models (HMMs) by simultaneously representing sequence and secondary structure information [54]. CMs provide a straightforward approach to find homologous RNAs in large RNA sequence databases.

All CHIKV 3'UTRs were processed and curated from full genomes by custom Perl scripts based on the `ViennaNGS` toolbox [55]. Structural RNA alignments were computed with `mlocarna` [56], and consensus structures were generated with `RNAalifold` [57]. Locally stable secondary structures were computed using `RNALalifold` from the `ViennaRNA` package [58]. Homologies of both sequences and secondary structures were inferred by HMMs and CMs as implemented in the `Infernal` package [59] as outlined previously [35]. All secondary structure plots were rendered using the `RNAplot` utility [58].

Viruses **2019**, *11*, 798

3. Results

3.1. Phylogeny of Chikungunya Indicates Multiple Lineages In Africa

No significant recombination events (p-value of < 0.05) were detected in the aligned data sets, therefore we kept all 590 sequences in the data sets. The phylogenetic tree search performed with `IQ-Tree` resulted in a best-scoring heuristic maximum likelihood tree for CHIKV with 590 genomic sequences (Figure 1). Deep ancestral branches that define the lineages had high (>90) SH-aLRT support/ultrafast bootstrap values. More recent branches had lower SH-aLRT support/ultrafast bootstrap values. The phylogenetic hypothesis for CHIKV with branch values is available online as Supplementary Materials dataset SD1 via GitHub.

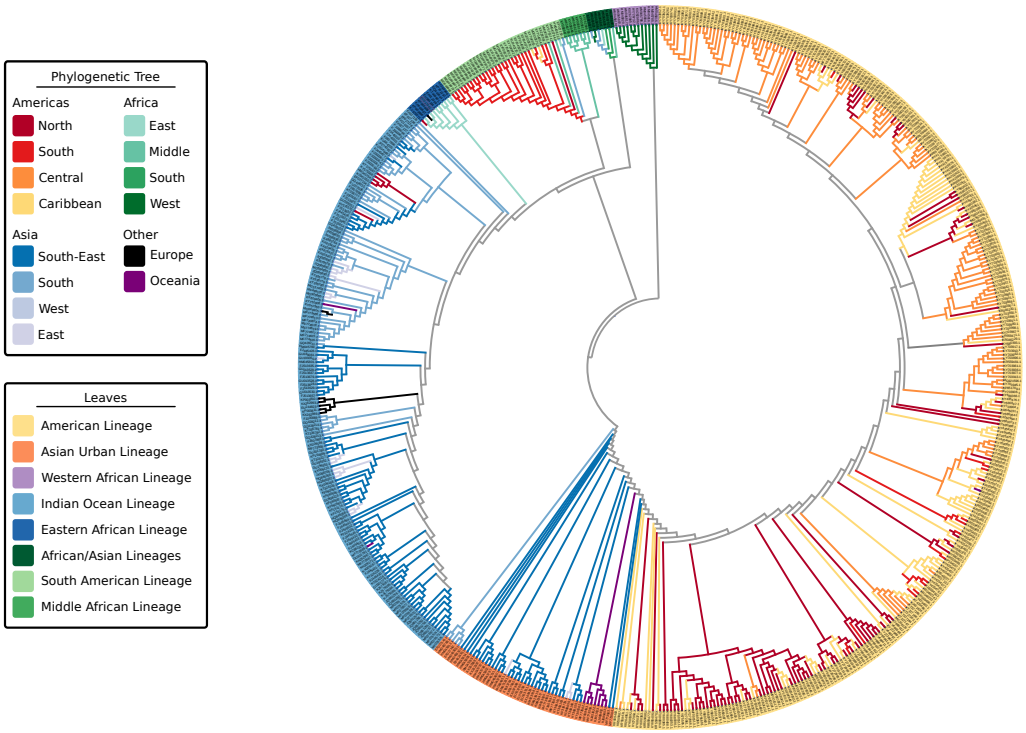


Figure 1. Maximum-likelihood phylogenetic tree of 590 Chikungunya virus genomic sequences highlighting the geographic locations (tree branches) and lineages (tree leaves). Tree was computed using `MAFFT` alignment of individual ORFs and partitioned tree search with model test using IQ-Tree. Leaves are colored by lineage association. Branches of the tree are colored by geographic location of a sample. Colors of ancestral nodes and branches do not imply geographic origin and therefore are depicted in gray.

The four viral strains grouped as expected according to reports in the literature [20,22]; however, the increased number of taxa in the tree allowed us to better investigate relationships of clades and infer source-sink events (Figure 2). In addition to the four major lineages—AUL, ECSA, IOL, and WA—our decision to split the concept of ECSA into its three regions, Eastern Africa, Central (Middle) Africa, and Southern Africa, rose from the presence of distinct clades in this large-scale phylogeny. The epidemic strain from the ECSA-MASA clade was first found in Middle Africa and later identified in South America. The AUL lineage gave rise to an extensively nested American lineage which spans North, South, and Central America, as well as the Caribbean.

Thus, we extrapolate the names for these discrete lineages based on geographic location: Indian Ocean, Eastern African, South American, Middle African, African/Asian, Asian Urban, American, and Western African (denoted as "leaves," Figure 1). Some of these lineages belong

to the three major epidemic clades: AUL-Am (Asian Urban + American lineages), ECSA-MASA (Middle Africa/South American Lineage), and ECSA-IOL (Eastern Africa + Indian Ocean Lineage). In this paper, we distinguish the individual lineages, which are delimited by geographic region (e.g., IOL and WA), from epidemic clades, which may encompass more than a single lineage (e.g., AUL, ECSA-IOL, and ECSA-MASA). Individual limited outbreaks (e.g., IOL in Europe) are not considered new lineages as the outbreak did not generate a new endemic region for the CHIKV lineage (Table 1).

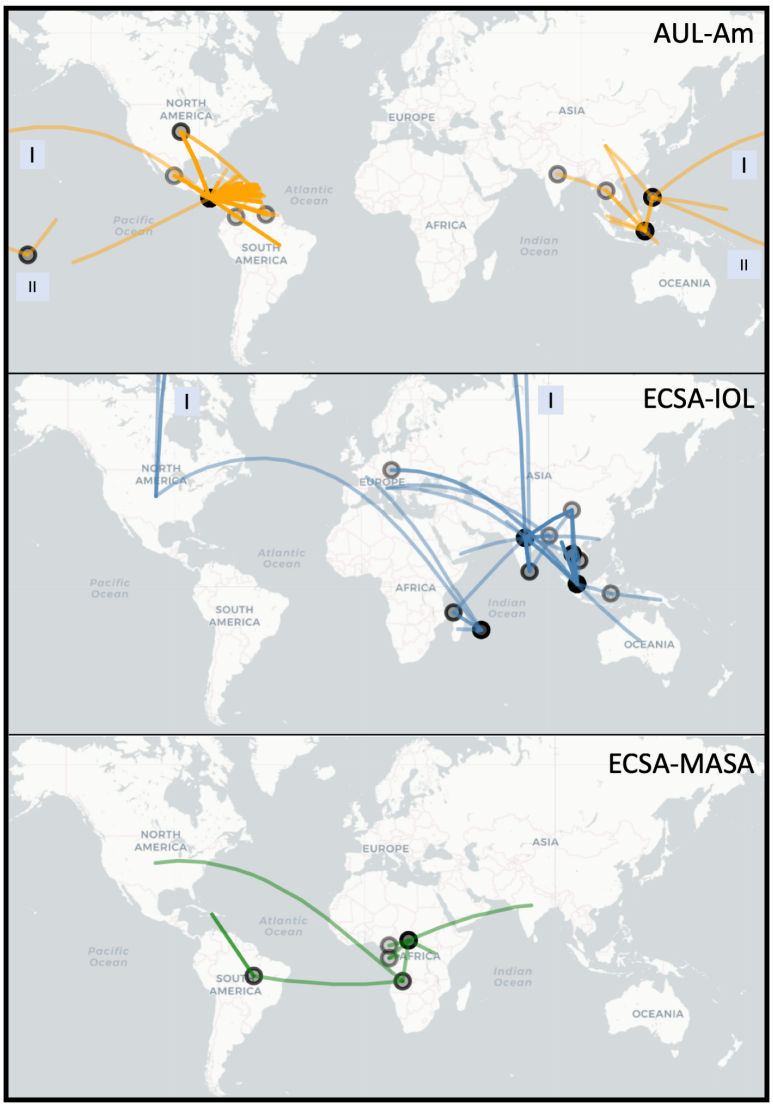


Figure 2. Worldwide spread of major epidemic Chikungunya clades: Asian Urban American (AUL-Am), ECSA Indian Ocean Lineage (ECSA-IOL) and ECSA Middle Africa South America (ECSA-MASA). This figure is rendered using the R package Leaflet with the underlying data based on parsimony ancestral reconstruction of metadata for place of isolation on the tree resulting from neighbor-joining tree search with collapsed branches under 80% bootstrap threshold (see Materials and Methods for more details). The circles on the map represent locations from which a transmission event originates. The color transparency represents the number of times a transition between characters occur on a location (node) or between locations (edge) (darker lines = more transitions). Boxes labeled "I" and "II" indicate that line crossed the International Date Line (e.g., AUL) or the shortest physical path between the two points (for mapping purposes only) was through the North Pole (e.g., ECSA-IOL).

Viruses **2019**, *11*, 798

Table 1. Comparison between literature and proposed nomenclature of Chikungunya lineages and epidemic clades based on the updated phylogeny.

Previous Work	**This Study**	
Lineages	Lineages	Epidemic Clades
Indian Ocean (IOL)	Indian Ocean	ECSA-IOL
East Central and South African (ECSA)	Eastern African	
	South American	ECSA-MASA
	Middle African	
	African/Asian	N/A
Asian Urban (AUL)	Asian Urban	Asian Urban American (AUL-Am)
	American	
West African (WA)	Western African	Western African (WA)

3.2. Spread of Major Chikungunya Epidemic Lineages

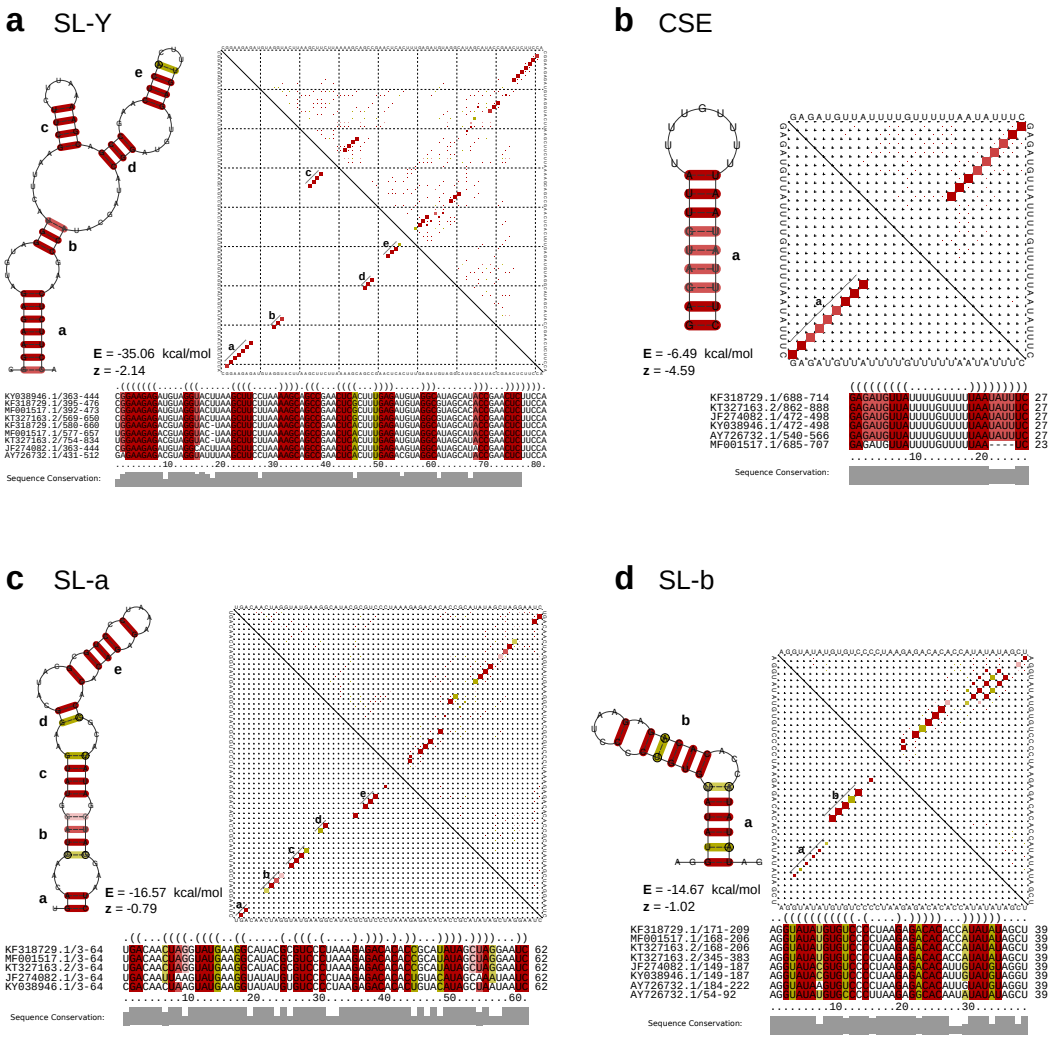
Our reconstruction of the spread of the three distinct epidemic clades is in agreement with the current literature (Figure 2). The Asian Urban lineage has spread from Southeastern Asia to the Pacific Islands and Central America, with traveler cases reaching the Continental United States (AUL-Am clade) [7,60]. A strain from Eastern Africa spread to India and is currently endemic in South and Southeast Asia (ECSA-IOL clade) [61]. More recently, a lineage from Middle Africa was introduced in South America (ECSA-MASA clade) [62,63].

3.3. Lineage-Specific 3'UTR Organization

Composition and length of the full 3'UTRs analyzed in this study differ substantially between lineages and range from 510 to 930 nucleotides. Earlier studies reported cascades of repeated sequence elements (RSEs) [33] and a structurally conserved sequence element (CSE) required for replication [2] in CHIKV 3'UTR. Chen et al. [34] proposed distinct direct repeat elements 1 and 2 (DR1 and DR2, respectively), together with a structured Y-shaped element, which all appear in lineage-specific duplication patterns. Filomatori et al. [32] adopted this classification in a recent study, however they were not able to see in silico secondary structure signals in DR sequences except for a Y-shaped element. Here, we follow up on previous knowledge, using the full set of presently available sequence data (i.e., 196 complete and truncated CHIKV 3'UTRs), and conclusively infer hitherto unresolved structural properties of all conserved elements.

From a multiple sequence alignment of all available 3'UTRs from all lineages, we manually characterized five highly conserved regions into four structured and one non-structured elements (Figure 3). We inferred sequence and structural homology from searches in the complete UTR set of all lineages, thus characterizing all instances of each conserved element. We used structural alignment and consensus structure prediction to find the conserved structures of the four structured elements.

We characterized two stem–loop (SL) elements, termed SL-a and SL-b, the latter being conserved among all lineages (Figure 4). While SL-a is only conserved among the ECSA-derived and AUL lineages, comparison of the two SL elements indicates that SL-b is likely a shortened variant of SL-a and comprised of its innermost hairpin (Figure 3c,d). Interestingly, this characteristic can also be observed in O'Nyong-Nyong virus (ONNV), which shares common ancestry with CHIKV. The genomic coordinates of both elements approximately correspond to the previously reported DRs, with DR1a spanning over the downstream portion of SL-a and DR1b being consistent with SL-b. The Y-shaped element (SL-Y) is conserved among all lineages but varies in copy number between ECSA-derived/WA (one copy) and AUL (two copies), and shows strong sequence conservation (Figure 3a). The previously described CSE [64] folds into a short hairpin loop and is found upstream of the poly-A tail in all CHKIV isolates (Figure 3b).



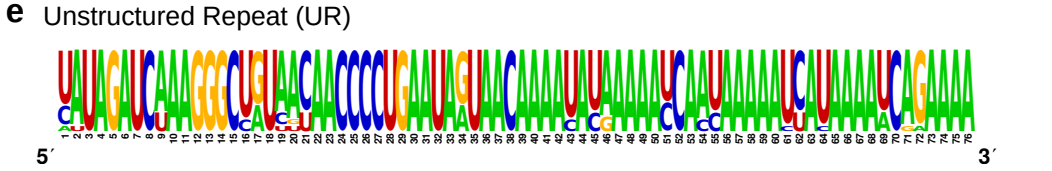


Figure 3. (**a**–**d**) Secondary structure cartoons, structural alignments and base-pair probability plots (dot plots) of the structured CHIKV 3'UTR elements SL-Y, CSE, SL-a, and SL-b. Colored squares in the dot plot report a pairing of two bases, with full squares corresponding to 100% pairing probability. The base pairings pattern of the corresponding most stable, minimum free energy structures are shown in the lower triangles, while the upper triangles show the base pairing probabilities considering all possible alternative, suboptimal structures. Colors in structure drawings, alignments and dot plots indicate how many different types of base combinations are found for a particular base pair (red: 1; yellow: 2). All structure plots show the consensus structures as calculated by `RNAalifold` [57]. For each structure, the free energy (E) and the `RNAz` [65] z-score (z) are provided. Negative z-scores indicate a structure is more stable than it would be expected for other RNAs of the same dinucleotide content. (**e**) Sequence logo of the intrinsically unstructured repeat element UR, computed from all HMM-derived homologous sequences within the data set. Structure cartoons are not shown since the overall base pairing probability is essentially zero for this sequence.

Viruses **2019**, *11*, 798

The structured elements, SL-a, SL-b, and SL-Y, are separated by one or two repeats of an intrinsically unstructured repeat (UR) element (Figure 3e). The total copy number of UR elements varies from three to five between lineages (Figure 4). This sequence repeat element lacks structure, as marked by its incapability to fold into any stable RNA secondary structure. Both single sequence predictions of individual repeats, as well as consensus structure prediction using an alignment of all instances did not yield a stable minimum free energy structure, or alternative base pairing probabilities.

Annotation of all available 3′UTR sequences with CMs and HMMs revealed that different CHIKV lineages have evolved distinct patterns of structured and unstructured functional RNA elements. This lineage-specificity of potentially functional RNAs has led to a varied 3′UTR architecture among CHIKV strains. This is particularly pronounced in the AUL and its descendants, which show a characteristic 3′UTR organization. Several geographically diverse isolates of the AUL-Am (also referenced as Caribbean lineage (AUL-Cbn)) appear to have undergone a partial duplication event in the 3′UTR [66]. As a result, an additional copy of the nonstructural sequence repeat UR and the structural element SL-b are found immediately downstream to the first copy of SL-b and UR (Figure 4).

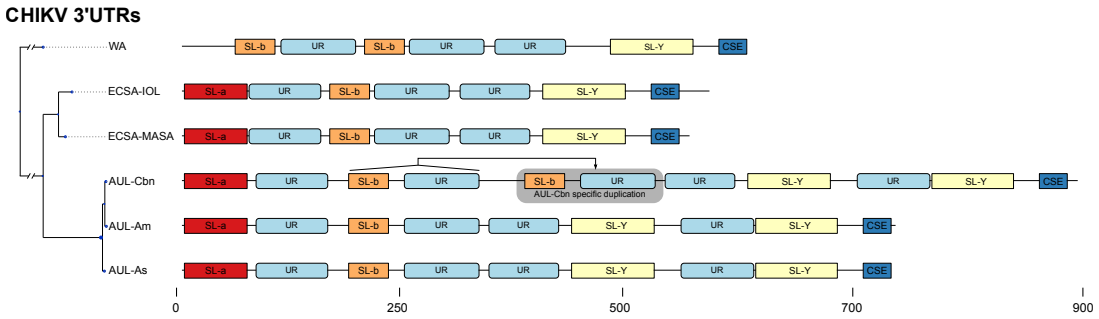


Figure 4. Schematic of the 3′UTR architecture of all major CHIKV strains. The phylogeny on the left is derived from a maximum-likelihood tree based on a genomic alignment of six representative sequences. For each strain, the sequence with the longest UTR was selected as a representative (WA: AY726732.1, ECSA-IOL: JF274082.1, ECSA-MASA: KY038946.1, AUL-Cbn: KT327163.2 (Mexico), AUL-Am: MF001517.1 (USA), AUL-Asia: KF318729.1 (China). All boxes correspond to exact coordinates of annotated structural/nonstructural elements. A scale bar is provided for the nucleotide coordinates.

4. Discussion

Studies on CHIKV phylogenetics have provided phylogenomic means of understanding disease spread in the context of correlating disease epidemiology. The phylogenetic analyses we present are in agreement with the current literature, although the increased number of sequences presented provide insights on the hidden history of ECSA. Interestingly, the phylogeny based on ORFs is in agreement with the phylogeny based on complete UTRs, despite the intuitive idea that different evolutionary forces should act on coding and non-coding regions.

4.1. Chikungunya Virus Nomenclature

It is critical to emphasize the significance of the definition of "lineage" and "epidemic clade" when discussing the evolution of a virus. Much has been proposed as to how species should be delineated. In 1977, Wiley [67] proposed that the definition of species should be that "A species is a single lineage of ancestral descendant populations of organisms which maintains its identity from other such lineages and which has its own evolutionary tendencies and historical fate." Peterson [68] criticizes the current system of viral taxonomy, mentioning that one criterion alone should dominate viral classification: the evolutionary independence of evolving lineages.

Assessing each CHIKV taxon as an individual, and bearing in mind the concept of phylogenetic independence, we cannot conclusively state that CHIKV sequences isolated in the Americas are

divergent enough from ancestral sequences in Africa and Asia to be considered two new American lineages. This logic, however, is limited when considering the similar phylogenetic structure of IOL rising from Eastern African taxa. This non-uniformity represents the difficulty in qualifying lineages based on subjective interpretation.

Further, IOL and AUL have region-specific derived clades circulating newly in endemic regions. Peterson [68] warns against the use of endemic regions to classify species. We acknowledge the flaw in solely using geography as a naming mechanism as the increased movement between continents due to airport passenger can substantially increase the number of co-circulating lineages named based on geography.

In order to remain consistent with the current literature we kept the naming scheme based on geography, but defined lineage as the pool of individual taxa found in a clade that are not isolated geographically. We define epidemic clade as a monophyletic group that may encompass one or more lineages regardless of geography.

4.2. Lineage Diversity in Chikungunya Virus

Multiple waves of CHIKV infection have occurred over the past two decades, with outbreaks traced from Eastern and Middle Africa to Southeastern Asia (e.g., Thailand and Malaysia), Southern Asia (India) and the Americas [20]. Clades of CHIKV observed in the analyses are consistent with those found in the literature with an important exception. Similar to White et al. [23], who alluded to the presence of ECSA subgroups, we observed that ECSA has multiple co-circulating lineages which gave rise to the current outbreaks in Southern and Southeastern Asia and the Americas. With the large amount of new sequences we analyzed, what was once called ECSA, East Central South African lineage, is no longer a single lineage because it is not monophyletic. Figure 1 demonstrates that Middle African taxa gave rise to the outbreak in South America, while Eastern African strains gave origin to IOL. IOL has remained nearly exclusively on the Asian continent, with exception of a few African and European outbreaks. AUL, a sister clade to all ECSA derived lineages, has reached the American continent with extensive nesting diversity (Figure 1). In Africa, AUL is known to have emerged in 2004 and spread explosively in the Comoros islands and incited a severe epidemic in Kenya [18]. In the Americas, strains from AUL are observed as coming from the US to Central America (Figure 2), and given the absence of locally acquired cases in the US, the indication of an ancestral sequence originating a transmission event to Central America could be due to lack of sampling the ancestral strain sequence and the close similarity of it with strain sequences found in Central America.

Asian Urban and IOL are the main lineages of CHIKV in Asia [69,70]. Moreover, Asian Urban has since emerged in Central America in 2013 and South America in 2014 [71] forming the AUL-Am clade. The Americas faced a devastating outbreak in 2013–2014 after the AUL lineage was introduced to the Caribbean [72], causing nearly 3,000,000 estimated cases. In 2014, the ECSA lineage was detected in Brazil as well [71]. Based on our phylogeny, the additional sister group lineage to both ECSA-IOL and ECSA-MASA, which encompasses Southern African isolates and is related to a few Asian isolates (of the African/Asian lineage) has no major outbreaks reported.

We propose that the ECSA genotype is too reductive in light of the rapid expansion and increased taxonomic sampling, and should be separated into their respective regions of the African continent: Eastern, Central, and Southern. Nevertheless, it is important to raise the attention to the lack of sampling in Africa which obscures the Chikungunya virus dynamics between the distinct regions within Africa. The current but limited amount of African sequences utilized in this study indicate that there may be multiple co-circulating lineages and these results should serve as justification for the importance of understanding the virus dynamics in Africa. As more isolates are sequenced, a more nuanced history of CHIKV evolution will emerge which will aid in our understanding of the dynamics of these outbreaks.

4.3. Lineage-Specific 3'UTR Architecture

Earlier work in alphavirus 3'UTRs reported a varied architecture of repeat sequence elements (RSEs) [2,33,73], which appear to influence virus replication in mosquito cells without affecting infectivity in mammalian cells [34]. Interestingly, molecular evolution experiments with engineered CHIKV 3'UTRs in mosquito and mammalian cells suggest that improved fitness was acquired mainly due to mutations in protein coding regions [74]. Still, the specific biological functions of individual 3'UTR RSEs have not been studied in detail. The comparative genomics approach employed in this study confirms the previously reported repeat organization in CHIKV 3'UTR sequences. It is now possible to conclusively associate repeat sequences with distinct RNA secondary structures that are conserved among all lineages. Importantly, the organization of all CHIKV 3'UTRs follows a common pattern characterized by an alternating arrangement of structured, i.e., SL-a, SL-b, and SL-Y, and unstructured regions. The latter are defined by a lack of any RNA structure, as predicted by single sequence and consensus folding algorithms. This suggests that these regions of the CHIKV genome exist in a predominantly unpaired state, rendering them highly accessible.

We hypothesize that such intrinsically unstructured repeat regions act as structural insulators that separate functional secondary structures from each other. Unstructured stretches can mediate the formation of structure in adjacent regions by prohibiting base pairs that are in conflict with the element structure. Likewise, single-stranded, unstructured RNAs can serve as easily accessible protein binding sites. In particular, the enrichment of `A` nucleotides in the unstructured region not only indicates the lack of secondary structure, but might also facilitate the sponging of RNA-binding proteins [75]. Earlier studies suggested an involvement of alphavirus 3'UTRs with cellular HuR and La proteins [76,77]. While HuR binding in CHIKV appears impaired [78], stabilization of viral RNA through host factors is an evolutionary trait to counteract viral genome degradation by host nuclease activity [25]. Conversely, other RNA-binding proteins, such as the Musashi family of translational regulators, have been shown to efficiently interact with viral 3'UTRs [31,79]. The canonical Musashi binding motif, `UAG`, is present at multiple loci within the unstructured repeat of all CHIKV 3'UTR sequences, supporting the biological role of intrinsic lack of structure. Besides protein-aided regulation, RNA interference is a crucial mechanism that controls viral replication in mosquitoes. Host-encoded microRNAs (miRNAs) were shown to interact with the 3'UTRs of Eastern equine encephalitis virus (EEEV) [80] and CHIKV [81,82]. The miRNA binding site reported by Dubey et al. [81] overlaps with the structured Y-shaped element (SL-Y) characterized here. Whether or not the CHIKV genome encodes for endogenous viral miRNAs, as observed in flaviviruses [83], remains to be established.

The structural aspects of alphavirus RNA is understudied when compared to other arboviruses. While repeat motifs were shown to be a crucial architectural pattern in alphavirus UTRs, many studies focused on the role and nature of these sequence repeats, without asking how far RNA structure might be the evolutionary driving force behind UTR evolution. A recent study employing SHAPE structure probing (selective 2'-hydroxyl acylation analyzed by primer extension) experiments could not find evidence for large-scale structural conservation of RNA within alphaviruses [37]. Work in the 5'UTRs of VEEV [84] and SINV [85] could associate RNA structure elements with replication efficiency and antiviral response [86]. We identified structured elements in the 5'UTR of CHIKV, however did not observe strong covariation patterns (data not shown). Likewise, the structured elements SL-a and SL-b in the 3'UTR exhibit only a few covariations. This observed lack of covariation might be due to the high degree of sequence conservation in the 3'UTR. Alternatively, it could also be a consequence of a sampling bias, as many samples in our dataset originate from individual outbreaks and therefore only represent a fraction of the CHIKV population. In contrast to other arboviruses, such as flaviviruses carrying exoribonuclease-resistant RNAs (xrRNAs) [35], the lack of covariations makes it difficult to assert to what extent secondary structure conservation is a driving force of CHIKV UTR evolution. This is underlined by a weak support of structure-associated z-scores for elements SL-a and SL-b (Figure 3c,d). Contrary, the SL-Y and CSE elements have very stable structures as indicated by strongly negative z-scores. One of the most striking features of the CHIKV 3'UTR is presence of multiple

copies of the unstructured UR region separating structured regions. Our findings indicate that the evolutionary forces acting on CHIKV 3'UTR manifest as a process that can rapidly rearrange structured and unstructured RNA elements, as observed in a recently emerged partial duplication of CHIKV strains isolated from the Caribbean [66]. Consequently, our predictions encourage further experimental validation to assess the structural and functional role of these proposed elements. In particular, it would be highly informative to elucidate whether or not these elements mediate the production of standalone non-coding RNAs. Additionally, experiments could reveal whether these conserved elements act in cis or trans.

To obtain a more comprehensive picture of the 3'UTR phylogeny in the context of diverged CHIKV lineages, we computed a separate phylogeny based exclusively on 3'UTR sequences (Figure S1). Although coding and non-coding parts of the viral genome are subject to different evolutionary pressures, mapping lineage association onto this tree reproduces the topology of the full CDS tree (Figure 1). This is to some extent surprising, as the nature of evolutionary forces acting on both entities is supposed to be different, best illustrated by the requirement to maintain a reading frame. This means that the 3'UTR evolution is not disentangled from CDS evolution in CHIKV, as expected intuitively. On a broader scale, our data provide evidence that CHIKV CDS and 3'UTRs show coherent evolutionary patterns, with leads to the hypothesis that the 3'UTR might represent an evolutionary scaffold that can rapidly replace functional elements, as required for adaptation to environmental change, e.g., by means of recombination [32]. Still, CHIKV 3'UTRs do not exhibit the high degree of plasticity known from other arboviruses with functional 3'UTRs, such as flaviviruses [28,35].

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/9/798/s1, Figure S1: Maximum likelihood tree of CHIKV 3'UTRs. Scripts and data sets are available via GitHub at https://github.com/abschneider/paper-chikungunya-phylogeny.

Author Contributions: M.T.W. and A.d.B.S. conceived the study; A.d.B.S., R.O., R.H., and M.T.W. analyzed the data; D.J. supported the phylogenetic analysis; I.L.H. supported the RNA structure analysis; All authors contributed to writing the paper and approved of the submitted version.

Funding: This work was supported by the National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (grant number AI135992), the Doktoratskolleg RNA Biology at University of Vienna, and the Austrian Science Fund (SFB F43, I-1303).

Acknowledgments: We thank Adiba Hassan, MPH, for the valuable discussions during the development of this work. We acknowledge the support of the Department of Bioinformatics and Genomics, The College of Computing and Informatics, and the Graduate School of the University of North Carolina at Charlotte. Open Access Funding by the University of Vienna.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gould, E.; Pettersson, J.; Higgs, S.; Charrel, R.; de Lamballerie, X. Emerging arboviruses: Why today? *ONE Health* **2017**, *4*, 1–13.
2. Pfeffer, M.; Kinney, R.M.; Kaaden, O.R. The Alphavirus 3'-Nontranslated Region: Size Heterogeneity and Arrangement of Repeated Sequence Elements. *Virology* **1998**, *240*, 100–108.
3. Campbell, L.P.; Luther, C.; Moo-Llanes, D.; Ramsey, J.M.; Danis-Lozano, R.; Peterson, A.T. Climate change influences on global distributions of dengue and chikungunya virus vectors. *Philos. Trans. R. Soc. B* **2015**, *370*, 20140135.
4. Powers, A.M.; Brault, A.C.; Shirako, Y.; Strauss, E.G.; Kang, W.; Strauss, J.H.; Weaver, S.C. Evolutionary Relationships and Systematics of the Alphaviruses. *J. Virol.* **2001**, *75*, 10118–10131.
5. Forrester, N.; Palacios, G.; Tesh, R.; Savji, N.; Guzman, H.; Sherman, M.; Weaver, S.; Lipkin, W. Genome-scale phylogeny of the alphavirus genus suggests a marine origin. *J. Virol.* **2012**, *86*, 2729–2738.
6. Zeller, H.; Van Bortel, W.; Sudre, B. Chikungunya: Its history in Africa and Asia and its spread to new regions in 2013–2014. *J. Infect. Dis.* **2016**, *214*, S436–S440.

7. Chen, R.; Puri, V.; Fedorova, N.; Lin, D.; Hari, K.L.; Jain, R.; Rodas, J.D.; Das, S.R.; Shabman, R.S.; Weaver, S.C. Comprehensive genome scale phylogenetic study provides new insights on the global expansion of Chikungunya virus. *J. Virol.* **2016**, *90*, 10600–10611.
8. Ross, R. The Newala epidemic: III. The virus: Isolation, pathogenic properties and relationship to the epidemic. *Epidemiol. Infect.* **1956**, *54*, 177–191.
9. Mason, P.; Haddow, A. An epidemic of virus disease in Southern Province, Tanganyika Territory, in 1952–1953: An additional note on Chikungunya virus isolations and serum antibodies. *Trans. R. Soc. Trop. Med. H* **1957**, *51*, 238–240.
10. Rezza, G.; Nicoletti, L.; Angelini, R.; Romi, R.; Finarelli, A.; Panning, M.; Cordioli, P.; Fortuna, C.; Boros, S.; Magurano, F.; et al. Infection with Chikungunya virus in Italy: An outbreak in a temperate region. *Lancet* **2007**, *370*, 1840–1846.
11. Van Bortel, W.; Dorleans, F.; Rosine, J.; Blateau, A.; Rousset, D.; Matheus, S.; Leparc-Goffart, I.; Flusin, O.; Prat, C.; Cesaire, R.; et al. Chikungunya outbreak in the Caribbean region, December 2013 to March 2014, and the significance for Europe. *Eurosurveillance* **2014**, *19*, 20759.
12. Teixeira, M.G.; Andrade, A.M.; Maria da Conceição, N.C.; Castro, J.S.; Oliveira, F.L.; Goes, C.S.; Maia, M.; Santana, E.B.; Nunes, B.T.; Vasconcelos, P.F. East/Central/South African genotype Chikungunya virus, Brazil, 2014. *Emerg. Infect. Dis.* **2015**, *21*, 906.
13. Weaver, S.C.; Lecuit, M. Chikungunya virus and the global spread of a mosquito-borne disease. *N. Engl. J. Med.* **2015**, *372*, 1231–1239.
14. Rezza, G.; Weaver, S.C. Chikungunya as a paradigm for emerging viral diseases: Evaluating disease impact and hurdles to vaccine development. *PLoS Negl. Trop. D* **2019**, *13*, e0006919.
15. Borgherini, G.; Poubeau, P.; Staikowsky, F.; Lory, M.; Moullec, N.L.; Becquart, J.P.; Wengling, C.; Michault, A.; Paganin, F. Outbreak of Chikungunya on Reunion Island: Early clinical and laboratory features in 157 adult patients. *Clin. Infect. Dis.* **2007**, *44*, 1401–1407.
16. Pulmanausahakul, R.; Roytrakul, S.; Auewarakul, P.; Smith, D.R. Chikungunya in Southeast Asia: Understanding the emergence and finding solutions. *Int. J. Infect. Dis.* **2011**, *15*, e671–e676.
17. Suhrbier, A.; La Linn, M. Clinical and pathologic aspects of arthritis due to Ross River virus and other alphaviruses. *Curr. Opin. Rheumatol.* **2004**, *16*, 374–379.
18. Wahid, B.; Ali, A.; Rafique, S.; Idrees, M. Global expansion of chikungunya virus: Mapping the 64-year history. *Int. J. Infect. Dis.* **2017**, *58*, 69–76.
19. Diallo, M.; Thonnon, J.; Traore-Lamizana, M.; Fontenille, D. Vectors of Chikungunya virus in Senegal: Current data and transmission cycles. *Am. J. Trop. Med. Hyg.* **1999**, *60*, 281–286.
20. Weaver, S.C.; Forrester, N.L. Chikungunya: Evolutionary history and recent epidemic spread. *Antivir. Res.* **2015**, *120*, 32–39.
21. Langsjoen, R.M.; Haller, S.L.; Roy, C.J.; Vinet-Oliphant, H.; Bergren, N.A.; Erasmus, J.H.; Livengood, J.A.; Powell, T.D.; Weaver, S.C.; Rossi, S.L. Chikungunya virus strains show lineage-specific variations in virulence and cross-protective ability in murine and nonhuman primate models. *mBio* **2018**, *9*, e02449-17.
22. Volk, S.M.; Chen, R.; Tsetsarkin, K.A.; Adams, A.P.; Garcia, T.I.; Sall, A.A.; Nasar, F.; Schuh, A.J.; Holmes, E.C.; Higgs, S.; et al. Genome-Scale Phylogenetic Analyses of Chikungunya Virus Reveal Independent Emergences of Recent Epidemics and Various Evolutionary Rates. *J. Virol.* **2010**, *84*, 6497–6504.
23. White, S.K.; Mavian, C.; Salemi, M.; Morris, J.G., Jr.; Elbadry, M.A.; Okech, B.A.; Lednicky, J.A.; Dunford, J.C. A new "American" subgroup of African-lineage Chikungunya virus detected in and isolated from mosquitoes collected in Haiti, 2016. *PLoS ONE* **2018**, *13*, e0196857.
24. Li, X.F.; Jiang, T.; Deng, Y.Q.; Zhao, H.; Yu, X.D.; Ye, Q.; Wang, H.J.; Zhu, S.Y.; Zhang, F.C.; Qin, E.D.; et al. Complete genome sequence of a Chikungunya virus isolated in Guangdong, China. *J. Virol.* **2012**, *86*, 8904–8905.
25. Hyde, J.L.; Chen, R.; Trobaugh, D.W.; Diamond, M.S.; Weaver, S.C.; Klimstra, W.B.; Wilusz, J. The 5′ and 3′ ends of alphavirus RNAs–non-coding is not non-functional. *Virus Res.* **2015**, *206*, 99–107.
26. Casal, P.E.; Chouhy, D.; Bolatti, E.M.; Perez, G.R.; Stella, E.J.; Giri, A.A. Evidence for homologous recombination in Chikungunya Virus. *Mol. Phylogenet. Evol.* **2015**, *85*, 68–75.
27. Roby, J.A.; Pijlman, G.P.; Wilusz, J.; Khromykh, A.A. Noncoding subgenomic flavivirus RNA: Multiple functions in West Nile virus pathogenesis and modulation of host responses. *Viruses* **2014**, *6*, 404–427.

28. Villordo, S.M.; Carballeda, J.M.; Filomatori, C.V.; Gamarnik, A.V. RNA Structure Duplications and Flavivirus Host Adaptation. *Trends Microbiol.* **2016**, *24*, 270–283.
29. Chapman, E.G.; Moon, S.L.; Wilusz, J.; Kieft, J.S. RNA structures that resist degradation by Xrn1 produce a pathogenic Dengue virus RNA. *eLife* **2014**, *3*, e01892.
30. MacFadden, A.; O'Donoghue, Z.; Silva, P.A.; Chapman, E.G.; Olsthoorn, R.C.; Sterken, M.G.; Pijlman, G.P.; Bredenbeek, P.J.; Kieft, J.S. Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. *Nat. Commun.* **2018**, *9*, 119.
31. de Bernardi Schneider, A.; Wolfinger, M.T. Musashi binding elements in Zika and related Flavivirus 3′ UTRs: A comparative study *in silico*. *Sci. Rep.* **2019**, *9*, 6911.
32. Filomatori, C.V.; Bardossy, E.S.; Merwaiss, F.; Suzuki, Y.; Henrion, A.; Saleh, M.C.; Alvarez, D.E. RNA recombination at Chikungunya virus 3'UTR as an evolutionary mechanism that provides adaptability. *PLoS Pathog.* **2019**, *15*, e1007706.
33. Ou, J.H.; Trent, D.W.; Strauss, J.H. The 3′-non-coding regions of alpharivus RNAs contain repeating sequences. *J. Mol. Biol.* **1982**, *156*, 719–730.
34. Chen, R.; Wang, E.; Tsetsarkin, K.A.; Weaver, S.C. Chikungunya virus 3′ untranslated region: Adaptation to mosquitoes and a population bottleneck as major evolutionary forces. *PLoS Pathog.* **2013**, *9*, e1003591.
35. Ochsenreiter, R.; Hofacker, I.L.; Wolfinger, M.T. Functional RNA Structures in the 3′UTR of Tick-Borne, Insect-Specific and No-Known-Vector Flaviviruses. *Viruses* **2019**, *11*, 298.
36. Fernandes, J.; Jayaraman, B.; Frankel, A. The HIV-1 Rev response element: An RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biol.* **2012**, *9*, 6–11.
37. Kutchko, K.M.; Madden, E.A.; Morrison, C.; Plante, K.S.; Sanders, W.; Vincent, H.A.; Cruz Cisneros, M.C.; Long, K.M.; Moorman, N.J.; Heise, M.T.; et al. Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Res.* **2018**, *46*, 3657–3670.
38. Benson, D.A.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2013** *42*, D32–D37.
39. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.
40. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **2014**, *30*, 3276–3278.
41. Martin, D.P.; Murrell, B.; Golden, M.; Khoosal, A.; Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **2015**, *1*. doi:10.1093/ve/vev003.
42. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2014**, *32*, 268–274.
43. Minh, B.Q.; Nguyen, M.A.T.; von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **2013**, *30*, 1188–1195.
44. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321.
45. de Bernardi Schneider, A. Arboviruses: The Hidden Path of an Imminent Threat. Ph.D. Thesis, The University of North Carolina at Charlotte, Charlotte, NC, USA, 2018.
46. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638.
47. Schliep, K.P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **2010**, *27*, 592–593.
48. Charif, D.; Lobry, J.R. SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural Approaches to Sequence Evolution*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 207–232.
49. Paradis, E.; Claude, J.; Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **2004**, *20*, 289–290.
50. de Bernardi Schneider, A.; Ford, C.T.; Hostager, R.; Williams, J.; Cioce, M.; Çatalyürek, Ü.V.; Wertheim, J.O.; Janies, D. StrainHub: A phylogenetic tool to construct pathogen transmission networks. *bioRxiv* **2019**. doi:10.1101/650283.
51. Louca, S.; Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **2017**, *34*, 1053–1055.

52. Graul, C. LeafletR: Interactive Web-Maps Based on the Leaflet JavaScript Library. R-package version 0.4-0. 2016. Available online: https://github.com/chgrl/leafletR (accessed on 28 August 2019).
53. Hijmans, R.J. Geosphere: Spherical Trigonometry. R Package Version 1.5-5. 2016. Available online: https://cran.r-project.org/package=geosphere (accessed on 28 August 2019).
54. Eddy, S.R.; Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **1994**, *22*, 2079–2088.
55. Wolfinger, M.T.; Fallmann, J.; Eggenhofer, F.; Amman, F. ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines. *F1000Research* **2015**, *4:50*, doi:10.12688/f1000research.6157.2.
56. Will, S.; Reiche, K.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.* **2007**, *3*, e65.
57. Bernhart, S.H.; Hofacker, I.L.; Will, S.; Gruber, A.R.; Stadler, P.F. RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinform.* **2008**, *9*, 474.
58. Lorenz, R.; Bernhart, S.H.; Zu Siederdissen, C.H.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithm. Mol. Biol.* **2011**, *6*, 26.
59. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold Faster RNA Homology Searches. *Bioinformatics* **2013**, *29*, 2933–2935, doi:10.1093/bioinformatics/btt509.
60. Lindsey, N.P.; Staples, J.E.; Fischer, M. Chikungunya virus disease among travelers—United States, 2014–2016. *Am. J. Trop. Med. Hyg.* **2018**, *98*, 192–197.
61. Hapuarachchi, H.; Bandara, K.; Sumanadasa, S.; Hapugoda, M.; Lai, Y.L.; Lee, K.S.; Tan, L.K.; Lin, R.T.; Ng, L.F.; Bucht, G.; et al. Re-emergence of Chikungunya virus in South-east Asia: virological evidence from Sri Lanka and Singapore. *J. Gen. Virol.* **2010**, *91*, 1067–1076.
62. da Costa, A.C.; Thézé, J.; Komninakis, S.C.V.; Sanz-Duro, R.L.; Felinto, M.R.L.; Moura, L.C.C.; de Oliveira Barroso, I.M.; Santos, L.E.C.; de Lemos Nunes, M.A.; Moura, A.A.; et al. Spread of Chikungunya virus east/central/south African genotype in northeast Brazil. *Emerg. Infect. Dis.* **2017**, *23*, 1742.
63. Souza, T.M.A.; Azeredo, E.L.; Badolato-Corrêa, J.; Damasco, P.V.; Santos, C.; Petitinga-Paiva, F.; Nunes, P.C.G.; Barbosa, L.S.; Cipitelli, M.C.; Chouin-Carneiro, T.; et al. First report of the East-Central South African genotype of Chikungunya virus in Rio de Janeiro, Brazil. *PLoS Curr.* **2017**, *9*, doi:10.1371/currents.outbreaks.4200119978d62ccaa454599cd2735727.
64. Hardy, R.W.; Rice, C.M. Requirements at the 3′ end of the *Sindbis Virus* Genome Efficient Synthesis Minus-Strand RNA. *J. Virol.* **2005**, *79*, 4630–4639.
65. Gruber, A.R.; Findeiß, S.; Washietl, S.; Hofacker, I.L.; Stadler, P.F. RNAz 2.0: Improved noncoding RNA detection. In *Biocomp 2010*; World Scientific: London, UK, 2010; pp. 69–79.
66. Stapleford, K.A.; Moratorio, G.; Henningsson, R.; Chen, R.; Matheus, S.; Enfissi, A.; Weissglas-Volkov, D.; Isakov, O.; Blanc, H.; Mounce, B.C.; et al. Whole-Genome Sequencing Analysis from the Chikungunya Virus Caribbean Outbreak Reveals Novel Evolutionary Genomic Elements. *PLoS Negl. Trop. D* **2016**, *10*, e0004402.
67. Wiley, E.O. The evolutionary species concept reconsidered. *Syst. Zool* **1978**, *27*, 17–26.
68. Peterson, A.T. Defining viral species: Making taxonomy useful. *Virol. J.* **2014**, *11*, 131.
69. Yang, C.F.; Su, C.L.; Hsu, T.C.; Chang, S.F.; Lin, C.C.; Huang, J.C.; Shu, P.Y. Imported Chikungunya Virus Strains, Taiwan, 2006–2014. *Emerg. Infect. Dis.* **2016**, *22*, 1981.
70. Maha, M.S.; Susilarini, N.K.; Nur Ika Hariastuti, S. Chikungunya virus mutation, Indonesia, 2011. *Emerg. Infect. Dis.* **2015**, *21*, 379.
71. Nunes, M.R.T.; Faria, N.R.; de Vasconcelos, J.M.; Golding, N.; Kraemer, M.U.; de Oliveira, L.F.; da Silva Azevedo, R.d.S.; da Silva, D.E.A.; da Silva, E.V.P.; da Silva, S.P.; et al. Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med.* **2015**, *13*, 102.
72. Yactayo, S.; Staples, J.E.; Millot, V.; Cibrelus, L.; Ramon-Pardo, P. Epidemiology of Chikungunya in the Americas. *J. Infect. Dis.* **2016**, *214*, S441–S445.
73. Strauss, J.H.; Strauss, E.G. The Alphaviruses: Gene Expression, Replication, and Evolution. *Microbiol. Mol. Biol. R* **1994**, *58*, 491–562.
74. Morley, V.J.; Noval, M.G.; Chen, R.; Weaver, S.C.; Vignuzzi, M.; Stapleford, K.A.; Turner, P.E. Chikungunya virus evolution following a large 3′ UTR deletion results in host-specific molecular changes in protein-coding regions. *Virus Evol.* **2018**, *4*, vey012.
75. Charley, P.A.; Wilusz, J. Sponging of cellular proteins by viral RNAs. *Curr. Opin. Virol.* **2014**, *9*, 14–18.
76. Pardigon, N.; Strauss, J.H. Mosquito homolog of the La autoantigen binds to Sindbis virus RNA. *J. Virol.* **1996**, *70*, 1173–1181.

77. Sokoloski, K.J.; Dickson, A.M.; Chaskey, E.L.; Garneau, N.L.; Wilusz, C.J.; Wilusz, J. Sindbis Virus Usurps the Cellular HuR Protein to Stabilize its Transcripts and Promote Productive Infections in Mmammalian and Mmosquito Cells. *Cell Host Microbe* **2010**, *8*, 196–207.
78. Dickson, A.M.; Anderson, J.R.; Barnhart, M.D.; Sokoloski, K.J.; Oko, L.; Opyrchal, M.; Galanis, E.; Wilusz, C.J.; Morrison, T.E.; Wilusz, J. Dephosphorylation of HuR protein during alphavirus infection is Associated with HuR relocalization to the cytoplasm. *J. Biol. Chem.* **2012**, *287*, 36229–36238.
79. Chavali, P.L.; Stojic, L.; Meredith, L.W.; Joseph, N.; Nahorski, M.S.; Sanford, T.J.; Sweeney, T.R.; Krishna, B.A.; Hosmillo, M.; Firth, A.E.; et al. Neurodevelopmental protein Musashi 1 interacts with the Zika genome and promotes viral replication. *Science* **2017**, *357*, 83–88.
80. Trobaugh, D.W.; Gardner, C.L.; Sun, C.; Haddow, A.D.; Wang, E.; Chapnik, E.; Mildner, A.; Weaver, S.C.; Ryman, K.D.; Klimstra, W.B. RNA viruses can hijack vertebrate microRNAs to suppress innate immunity. *Nature* **2014**, *506*, 245.
81. Dubey, S.K.; Shrinet, J.; Sunil, S. Aedes aegypti microRNA, miR-2944b-5p interacts with 3'UTR of Chikungunya virus and cellular target vps-13 to regulate viral replication. *PLoS Negl. Trop. D* **2019**, *13*, e0007429.
82. Yen, P.S.; Chen, C.H.; Sreenu, V.; Kohl, A.; Failloux, A.B. Assessing the Potential Interactions between Cellular miRNA and Arboviral Genomic RNA in the Yellow Fever Mosquito, *Aedes Aegypti*. *Viruses* **2019**, *11*, 540.
83. Hussain, M.; Torres, S.; Schnettler, E.; Funk, A.; Grundhoff, A.; Pijlman, G.P.; Khromykh, A.A.; Asgari, S. West Nile virus encodes a microRNA-like small RNA in the 3' untranslated region which up-regulates GATA4 mRNA and facilitates virus replication in mosquito cells. *Nucleic Acids Res.* **2012**, *40*, 2210–2223.
84. Kulasegaran-Shylini, R.; Atasheva, S.; Gorenstein, D.G.; Frolov, I. Structural and Functional Elements of the Promoter Encoded by the 5' Untranslated Region of the Venezuelan Equine Encephalitis Virus Genome. *J. Virol.* **2009**, *83*, 8327–8339.
85. Fayzulin, R.; Frolov, I. Changes of the Ssecondary Sstructure of the 5' end of the Sindbis Virus Genome Inhibit Virus Growth in Mosquito Cells and Lead to Accumulation of Adaptive Mutations. *J. Virol.* **2004**, *78*, 4953–4964.
86. Hyde, J.L.; Gardner, C.L.; Kimura, T.; White, J.P.; Liu, G.; Trobaugh, D.W.; Huang, C.; Tonelli, M.; Paessler, S.; Takeda, K.; et al. A viral RNA structural element alters host recognition of nonself RNA. *Science* **2014**, *343*, 783–787.

6 FUNCTIONAL STRUCTURES IN THE UTRS OF FVS

The following chapter presents the work that was conducted as part of a systematic investigation of Flavivirus UTRs for functional elements, published in the article *Viruses. 2019 Sep; 11(9): 798.* A reprint of the article is provided as part of the chapter.

6.1 MOVIATION

There has been a long standing interest in the viruses of the entire *genus* Flavivirus (FV) and their RNA structure elements, but typically only the human-pathogenic members species are studied extensively due to their RNA elements that convey resistance against exoribonucleases. Despite the special relevance for human disease, RNA structure is suspected to also play a crucial role in all FVs, most notably in the context of switching between reservoir hosts, vectors and humans. Still, many FVs remain uninvestigated, especially since novel Flaviviruses are discovered at an increasingly rapid pace.

Given this steady increase in fully sequenced FV genomes, the introduction of comparative methodologies like covariance models and consensus folding could substantially help to better understand the weakly studied FV species and reveal new insights about the RNA structure dynamics of the Untranslated Region (UTR) of the well studied members of the *genus*. Especially given the fact that FV RNA structure elements seem to rely rather on conservation of structure than sequence, makes them an ideal candidate for advanced structure homology search via Covariance Model (CM), an approach that, somewhat surprisingly, had not been applied before to FVs. At the same time, a series of experimental works reported nucleotide-level resolution data on several exonuclease resistant elements in Tick Born Encephalitis Virus, which represented an ideal opportunity to seed a CM based investigation.

6.2 ADDED VALUE

This study provides a comprehensive and systematic analysis of structural elements in all currently known Flavivirus 3'UTRs. Many known elements could be confirmed and novel instances were detected in previously unstudied FVs. Additionally, novel conserved elements were characterized, exhibiting a remarkable conservation of secondary structure but not nucleotide sequence. This further adds evidence to the hypothesis, that RNA structure is indeed *the* defining characteristic causing the exoribonuclease stalling capabilites.

The published work in addition demonstrates the power of comparative methods when applied to problems in virus bioinformatics. In particular, a combination of methods as Structural Alignments, Consensus Folding and homology search by CMs, is demonstrated to yield significant insights into the structural architecture of UTRs and should be considered as a standard approach for similar investigations. Furthermore, the study emphasizes the importance of high quality experimental data, which served as an enabler for the entire study and demonstrates the synergies between experimental and theoretical work.

6.3 reprint of: functional structures in the 3'utr of tick-borne, insect-specific and no-known-vector flaviviruses

Roman Ochsenreiter, Ivo L. Hofacker and Michael T. Wolfinger

Viruses, **2019**, 11(3), 298.
https://doi.org/10.3390/v11030298

Contributions:

Roman Ochsenreiter Conceived the study, developed the methods, curated the dataset, analyzed the data and wrote the manuscript.

Ivo L. Hofacker Supervised the project and wrote the manuscript.

Michael T. Wolfinger Conceived the study, supervised the project, curated the dataset and wrote the manuscript.

viruses

MDPI

Article

Functional RNA Structures in the 3′UTR of Tick-Borne, Insect-Specific and No-Known-Vector Flaviviruses

Roman Ochsenreiter [1], Ivo L. Hofacker [1,2] and Michael T. Wolfinger [1,2,*]

[1] Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria; romanoch@tbi.univie.ac.at (R.O.); ivo@tbi.univie.ac.at (I.L.H.)

[2] Research Group BCB, Faculty of Computer Science, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria

* Correspondence: michael.wolfinger@univie.ac.at

Received: 1 March 2019; Accepted: 20 March 2019; Published: 24 March 2019

Abstract: Untranslated regions (UTRs) of flaviviruses contain a large number of RNA structural elements involved in mediating the viral life cycle, including cyclisation, replication, and encapsidation. Here we report on a comparative genomics approach to characterize evolutionarily conserved RNAs in the 3′UTR of tick-borne, insect-specific and no-known-vector flaviviruses in silico. Our data support the wide distribution of previously experimentally characterized exoribonuclease resistant RNAs (xrRNAs) within tick-borne and no-known-vector flaviviruses and provide evidence for the existence of a cascade of duplicated RNA structures within insect-specific flaviviruses. On a broader scale, our findings indicate that viral 3′UTRs represent a flexible scaffold for evolution to come up with novel xrRNAs.

Keywords: flavivirus; non-coding RNA; secondary structure

1. Introduction

Flaviviruses are small, single-stranded positive-sense RNA viruses that are typically transmitted between arthropod vectors and vertebrate hosts. They are endemic in tropic and sub-tropic regions and represent a global health threat, although humans are considered dead end hosts in many cases.

The genus *Flavivirus* within the *Flaviviridae* family comprises more than 70 species, which are organized into four groups, each with a specific host association: Mosquito-borne flaviviruses (MBFVs) and tick-borne flaviviruses (TBFVs) spread between vertebrate (mammals and birds) and invertebrate (mosquitoes and ticks) hosts, whereas insect-specific flaviviruses (ISFVs) replicate specifically in mosquitoes and no-known-vector flaviviruses (NKVs) have only been found in rodents and bats, respectively. This natural host-range-based classification is in good agreement with sequence-based phylogenetic clustering, mainly because all flaviviruses share a common genome organization [1]. Conversely, epidemiology, disease association [2] and transmission cycles [3] are fundamentally different among different flavivirus groups.

Emerging and re-emerging MBFVs such as Dengue virus (DENV), Japanese encephalitis virus (JEV), West Nile virus (WNV), Yellow fever virus (YFV) or Zika virus (ZIKV) are the causative agents of large-scale outbreaks that result in millions of human and veterinary infections every year [4]. Likewise, tick-borne encephalitis virus (TBEV), Powassan virus (POWV) and other members of the tick-borne serocomplex are neuropathogenic agents that cause a large number of infections every year, resulting in a massive incidence increase since the 1970ies [5]. Consequently, much research effort has gone into studying MBFV and TBFV biology, biochemistry and phylogeny [6]. The two remaining groups, ISFVs and NKVs, however, have received limited attention in the research community, mainly

because they are generally not associated with human or veterinary disease and therefore are still underrepresented in the literature. The phylogenetic relationship among the four ecological flavivirus groups is shown in Figure 1. Table A1 lists all viral species considered in the present study.

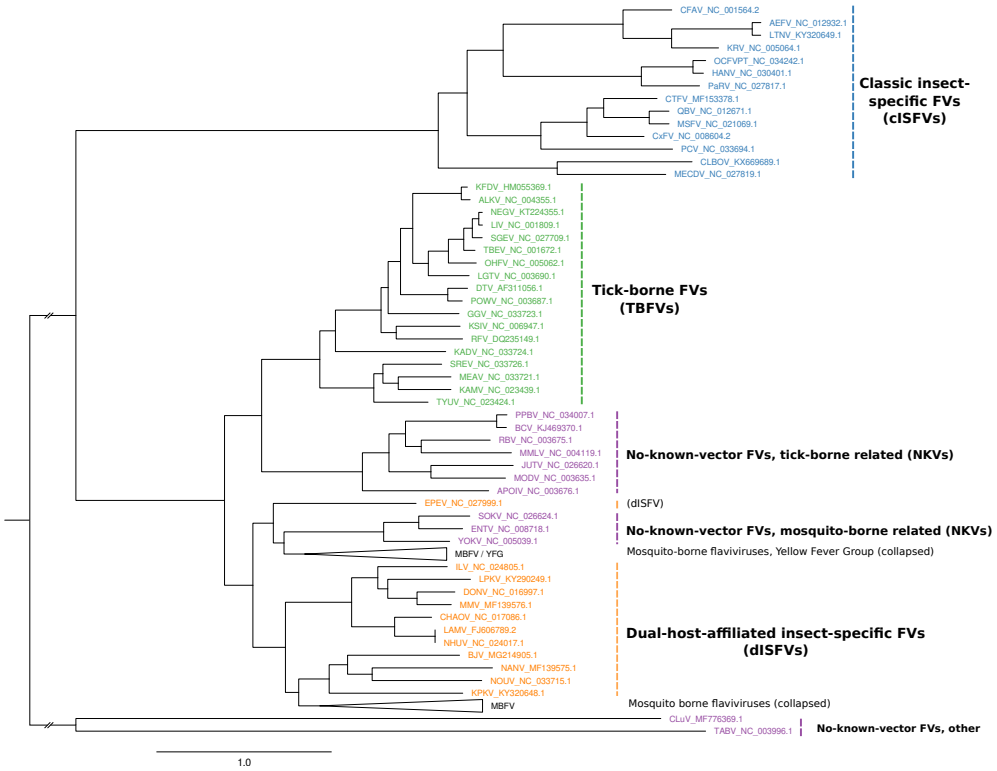


Figure 1. Maximum-likelihood phylogenetic tree of the genus *Flavivirus*, highlighting the major groups ISFVs (blue), dISFVs(orange), TBFVs (green), and NKVs (magenta). The MBFV Yellow Fever virus group (YFG) and the main MBFV branch were not covered in this study and are both collapsed. The tree has been computed from a `MAFFT` alignment of complete polyprotein amino acid sequences with `iq-tree`. Figure rendered with `FigTree`.

TBFVs form a monophyletic group consisting of a single serocomplex, although pathology and clinical manifestations vary among different viruses. They comprise more than a dozen of recognized species and separate into three groups: Mammalian tick-borne flaviviruses (M-TBFV), seabird tick-borne flaviviruses (S-TBFV) and the Kadam virus group. See [7] for a comprehensive review.

ISFVs naturally infect hematophagous *Diptera* and are typically divided into two groups [8]: Classical insect-specific flaviviruses (cISFVs) naturally infect mosquitoes and excursively replicate in mosquito cells in vitro. They form a phylogenetically distinct clade among known flaviviruses, appearing at the root of the MBFV, TBFV and NKV branches. The cISFV group separates into two clades, one associated with *Aedes* spp. mosquitoes and the other associated with *Culex* spp. mosquitoes, respectively [9]. They lack the ability to infect vertebrates and to replicate in vertebrate cell lines and have not been in the research spotlight until very recently. The second group is comprised of arbovirus-related or dual-host affiliated insect-specific flaviviruses (dISFVs), which represent a non-monophyletic group which is phylogenetically and antigenically related to mosquito/vertebrate flaviviruses, although they do not appear to infect vertebrate cells [10]. Insect-specific viruses play a crucial role in the mosquito microbiome and have been shown to modulate the replication of other arboviruses [11]. In this line, they are currently considered as biological control agents and vaccine platforms [12].

NKVs represent an ecologically and phylogenetically diverse set of viruses which have been isolated exclusively from vertebrates (mainly bats and rodents), without evidence for transmission by arthropod vectors. They form a non-monophyletic group among flaviviruses and are typically divided into bat- (B-NKV) and rodent-associated (R-NKV) groups, see Table A1. B-NKVs can be further separated into Entebbe virus group, which is phylogenetically closer to MBFVs, and Rio Bravo virus group, which is a sister clade to TBFVs. Species in the R-NKV group form the Modoc virus group, which is phylogenetically close to the B-NKV Rio Bravo group [13]. While NKVs are poorly characterized they represent a valuable resource to study evolutionary traits related to host-switch capacity mediated by conserved genomic elements.

Conserved RNA Structures Mediate Pathogenesis

Conserved RNA structures in the untranslated regions (UTRs) of RNA viruses are of particular interest because they mediate the viral life cycle by promoting or enhancing replication, as proposed for elements in both 5′UTRs [14] and 3′UTRs [15–18]. Mosquito/vertebrate viruses must operate efficiently in vectors and hosts, phylogenetically distinct organisms with different cellular machineries. This requires a high level of flexibility of viral regulatory elements to evade various antiviral response strategies while assuring proper replication conditions required for maintaining a stable quasispecies population. To achieve this resilience in host adaptation, RNA duplication strategies have been proposed as an evolutionary trait for MBFVs [19]. Tandem RNA structures within DENV 3′UTR are under different selective pressures in alternating hosts, suggesting the idea that duplicated RNA structures differentially evolved to accommodate specific functions in the two hosts [20]. Likewise, there is evidence for evolutionary pressure on maintaining the primary sequence of parts of duplicated RNA elements, as recently shown for flaviviral dumbbell (DB) elements in the context of finding a biophysical model for explaining a possible route for ZIKV-induced neurotropism [21].

Viral RNA genomes are different from procaryotic and eucaryotic mRNA. In addition to coding for and regulating the viral machinery, viral genomic RNA (gRNA) exhibits functional regions that act upon different stages of the viral life cycle. The 10-12kB flaviviral gRNA is capped, but non-polyadenylated and encodes a single open reading frame (ORF). Upon translation, a polyprotein is produced, which is then cleaved by viral and cellular enzymes into structural and non-structural proteins [22]. The ORF is flanked by highly structured untranslated regions (UTRs), which contain evolutionary conserved RNA elements that are crucially related to regulation of the viral life cycle, thereby inducing processes such as genome circularization, viral replication and packaging [23–25].

Upon flavivirus infection, accumulation of both gRNA as well as viral long non-coding RNAs (lncRNAs) is observed. These lncRNAs, which have been referred to as subgenomic flaviviral RNAs (sfRNAs) [26] are stable decay intermediates produced by exploiting the host's mRNA degradation machinery [27] and are associated with viral replication, pathogenesis and cytopathicity [28,29]. The production of sfRNA is induced by partial degradation of viral gRNA by the 5′-3′ exoribonuclease Xrn1, an enzyme associated with the cell's RNA turnover machinery [30,31]. Mechanistically, sfRNAs are generated by stalling Xrn1 at conserved structural elements in the viral 3′UTR, termed xrRNA (exoribonuclease-resistant RNA elements). These structures efficiently stall Xrn1 from progressing through from the 5′ direction, thus protecting the downstream RNA from degradation, while pass-through in the 3′-5′ direction, as required for viral RNA-dependent RNA-polymerase is still possible [32]. In particular, different types of stem-loop (SL) and dumbbell (DB) elements found in many MBFVs and TBFVs have been related to quantitative protection of downstream virus RNA against exoribonuclease degradation [33].

Xrn1 stalling results in dysregulation of cellular function with the aim of promoting viral infections. In this regard, functions of sfRNA in modulating cellular mRNA decay and RNAi pathways [34] as well as modulating anti-viral interferon response [35,36] have been reported.

The genomic architecture of flaviviruses has been extensively studied to understand the molecular principles required for sfRNA production. Chemical and enzymatic probing methods [37], together

with x-ray crystallography revealed the 3′UTR structure of the MBFVs WNV [38], YFV [39], DENV [40], Murray Valley encephalitis virus (MVEV) [41], ZIKV [42] and recently different species of the TBFV and NKV groups [33], highlighting the possibility that exoribonuclease resistance might be a pervasive mechanism of the viral world. Interestingly, several conserved RNA structural elements in viral 3′UTRs have been predicted in our group [43–47], some of which have later been attributed to xrRNA functionality [26]. To further expand the set of potential xrRNAs, we report here on a comparative genomics survey aimed at characterization of evolutionary conserved RNA structures in flavivirus 3′UTRs, focusing on TBFVs and the hitherto understudied groups of ISFVs and NKVs. A detailed study on the evolutionary traits of conserved RNAs in MBFV 3′UTRs will be published elsewhere.

2. Materials and Methods

Viral genome data for the present study were obtained from the public National Center for Biotechnology Information `(NCBI) refseq` (https://www.ncbi.nlm.nih.gov/refseq/) and `genbank` (https://www.ncbi.nlm.nih.gov/genbank/) databases on 28 May 2018. We downloaded all complete viral genomes under taxonomy ID 11051 (genus *Flavivirus*) and filtered for TBFV, ISFV and NKV species listed in Table A1. Whenever `refseq` annotation was not available for a species, we selected the longest complete genome from the `genbank` set as representative sequence. In total, the data set is comprised of 86 ISFV, 275 TBFV and 27 NKV isolates, respectively. The number of isolates with available 3′UTR sequence data per species varies between 1 and 167.

2.1. Phylogeny Reconstruction

The polyprotein/coding sequence (CDS) regions of most flaviviruses can be aligned consistently, however, UTRs typically show large variance both in length and sequence composition, rendering them ill-suited for phylogeny reconstruction. A phylogeny of all members of the genus *Flavivirus* (Figure 1) was therefore reconstructed via a multiple sequence alignment (MSA) of the nucleotide sequences of the CDS regions only. The MSA was computed with `MAFFT`[48] and subsequent maximum-likelihood tree reconstruction was performed using `iq-tree` [49] using the GTR+F+R7 substitution model.

2.2. Structural Homology Search with Covariance Models

The present study is centered around structural homology of RNA elements among phylogenetically narrow subgroups. A straightforward approach to finding novel homologous RNA structures is to search RNA sequence databases with Covariance Models (CMs), i.e., statistical models of RNA structure that extend classic Hidden-Markov-Models (HMMs) to simultaneously represent sequence and secondary structure. CMs, as implemented in the `infernal` package [50] allow for rapid screening of large RNA sequence databases to find even weakly conserved sequence-only or structurally homologous RNAs. We have recently applied this approach to identify novel telomerase RNAs in *Saccharomycetes* [51].

Here, structural multiple sequence alignments of the viral 3′UTR sequences were generated with `locARNA` [52] and CMs were built for known or experimentally verified xrRNAs [33]. All 3′UTR sequences were then screened and novel candidate sequences were added to perform iterative refinement until convergence. Weak sequence conservation of putative xrRNA elements resulted in initially fragile results, indicating that `infernal` default parameters are typically not optimal. Adjusting parameters, in particular disabling both heuristic filtering and local end detection, however, allowed our CMs to find homologs with strongly conserved secondary structures in presence of large sequence deviation from the original sequence the CM was built from. Likewise, `cmsearch` E-values turned out unsuitable for assessing hit quality in case of major sequence divergence. We therefore employed a cutoff approach, requiring a hit to form at least 75% of all base pairs listed in the CM in order to be considered significant.

2.3. De novo Discovery of Conserved RNA Elements

Beside characterization of RNAs with homology to known structurally conserved elements, we aimed at identifying novel elements, considering both thermodynamic stability and sequence covariation as evolutionary traits. In this line, `locARNA`-generated structural alignments of full UTR sequences were cut manually into blocks corresponding to conserved secondary structures. Alternatively, we employed `RNALalifold` from the `ViennaRNA` package [53] to compute locally stable secondary structures for aligned UTR sequences. A CM was built for each structure and searched against all flavivirus 3′UTRs, keeping only CMs that scored well multiple times per UTR. The rationale here is that the occurrence of multiple copies hints towards a possible functional role of a structural element, given that the ability of two or more independently evolving sequences to form a common structure is unlikely.

The above approach is implemented as a set of custom Perl and Python scripts for semi-automatic characterization and annotation of conserved RNAs in viral UTR sequences. Internally, these scripts build on the `ViennaRNA` scripting language interface for thermodynamics calculations, the `ViennaNGS` [54] suite for extraction of genomic loci, the `RNAaliSplit` package [55] for splitting alignments into subparts with common consensus structures (i.e., common structures formed by all individual sequences), `R2R` [56] for visualization, and the `ETE3` framework [57] for tree annotation and visualization.

3. Results

Several flaviviruses have previously been studied in great detail, yielding a varied landscape of repeated RNA sequence and structure elements within the 3′UTRs of these viruses, which are likely to have evolved from numerous duplications [19,58,59]. Many of these studies relied on single sequence predictions, which resulted in a good understanding of both structure and genomic position of conserved elements in individual species. A unified picture of homologous RNAs within the 3′UTRs of flaviviruses, however, has not been available.

The comparative approach applied in the present study outperforms single sequence predictions by considering consensus structures formed by all sequences. This allows us not only to confirm previously described RNA structures but also to elucidate hitherto unrecognized tandem repeats in many species. In this line our results can help in understanding the complex evolution of flavivirus 3′UTRs.

3.1. Construction of Seed Alignments

Based on recent experimental evidence for the existence of xrRNAs in TBFVs, ISFVs and NKVs, and previously characterized conserved RNA elements in flaviviral 3′UTRs, we built seed alignments for initial CMs, which were then refined iteratively, i.e., subjected to multiple rounds of screening and incorporation of best hits into the CM. Likewise, candidate structures from `RNALalifold` calculations were used as seeds for identification of conserved RNA structures. Figure 2 shows an overview of refined consensus structures for each ecologic group of flaviviruses analyzed here.

The four ecologic groups of flaviviruses show a varied 3′UTR architecture, however, the terminal 3′ stem-loop structure (3′SL, also referred to as 3′ long stable hairpin, 3′LSH) has been shown to be associated with panhandle-formation during virus replication and is therefore present in the terminal region of all flaviviruses [14]. The element is listed in Rfam as RF00185 (Flavivirus 3′UTR cis-acting replication element, Flavi_CRE) and we could use it to consistently identify terminal regions within 3′UTRs. Absence of this element from a UTR sequence is indicative of incomplete or truncated sequence data. The underlying sequences generally form a stable stem-loop structure upon structural alignment and single sequence folding. We built individual 3′SL seed alignments and CMs for each ecologic group, termed T.3SL, N.3SL and I.3SL, respectively.

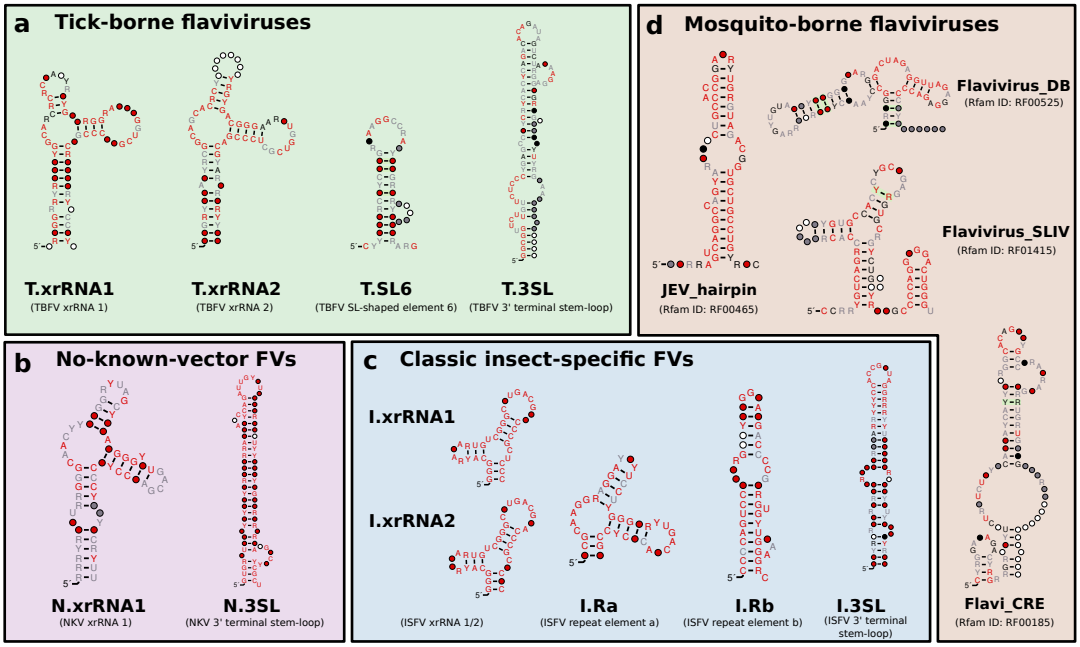


Figure 2. Overview of consensus structures of all CMs used for the annotation of flavivirus 3′UTRs. TBFV, ISFV, and NKV elements were refined from published experimental data (T.xrRNA1/2, I.xrRNA1/2, N.xrRNA) or identified computationally (T.SL6, I.Ra, I.Rb as well as all 3′-terminal stem-loop structures). MBFV elements were obtained from Rfam. Throughout this paper, all CMs are referred to by the name written in bold. References to xrRNA-like structures refer to the generalized xrRNA CM (Section 3.6).

3.2. Tick-Borne Flaviviruses

MacFadden et al. [33] suggested two different exoribonuclease-resistant structures in TBFV 3′UTRs. We used the proposed sequences from TBEV, POWV, Karshi virus (KSIV), Langat virus (LGTV), Louping ill virus (LIV), Omsk hemorrhagic fever virus (OHFV) and Alkhumra hemorrhagic fever virus (ALKV) as templates for a set of initial structural alignments and CMs. These models were then employed to search for high confidence hits within all TBFV 3′UTRs to construct seed alignments of the two exoribonuclease resistant structures in TBFVs, termed T.xrRNA1 and T.xrRNA2. These models allowed us to construct highly specific CMs for both TBFV xrRNAs, which were subsequently used to annotate xrRNA instances in already studied and previously unstudied TBFV species (Figure 3a).

The full structural alignment of the 3′UTRs of selected tick-borne species moreover suggests a short stem-loop element in several species, which is characterized by high sequence heterogeneity but heavily conserved structure supported by multiple covariations. Evidence for this element, termed stem-loop 6 (SL6), has been reported earlier for at least TBEV, LGTV and OHFV [58,59]. We kept this nomenclature and identified the exact position in each TBFV 3′UTR (Figure 3d).

Our data further shows that both TBFV xrRNA CMs (Figure 2a), as well as NKV xrRNA CMs (Figure 2b and Section 3.5) consistently yield plausible hits with a high degree of structure conservation immediately upstream of the strongly conserved terminal stem-loop element. Existence of a Y-shaped element (termed Y1) and putative similarity to NKVs has been proposed earlier based on single sequence structure predictions [58]. Structural `locARNA` alignment and subsequent `RNAalifold` consensus structure prediction indicates strong secondary structure conservation with frequent structure-conserving sequence covariations. Taken together, this suggests good evidence that respective regions in TBFVs harbor a putatively structured and functional xrRNA-like RNA (Y1, Figure 3d).

Viruses **2019**, *11*, 298

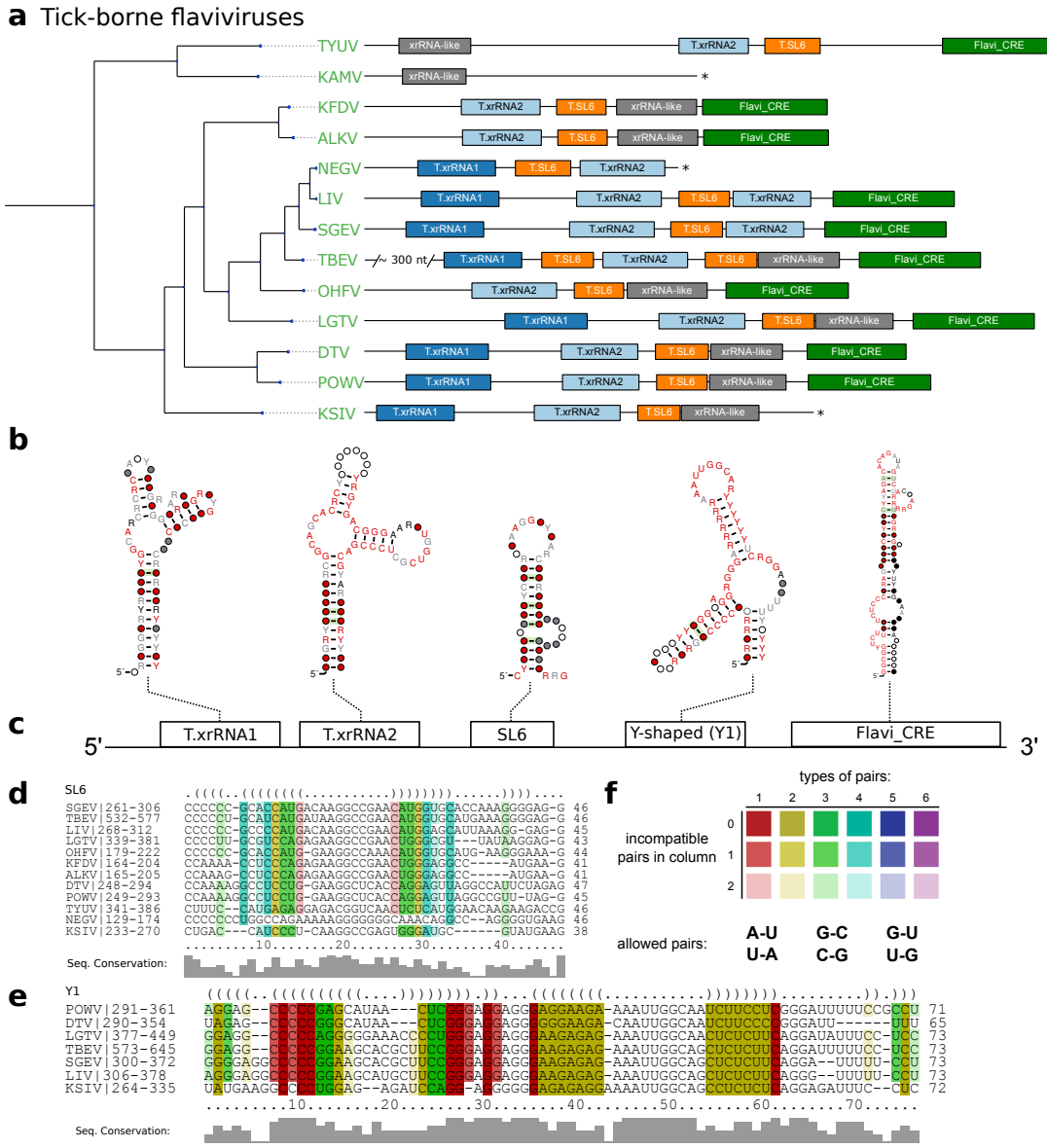


Figure 3. (**a**) Annotated 3′UTRs of TBFVs. The phylogenetic tree on the left has been computed from complete coding sequence nucleotide alignments and corresponds to the TBFV subtree in Figure 1. For each species with available 3′UTR sequence a sketch of the 3′UTR is drawn to scale next to the leaves of the tree. Colored boxes represent conserved RNA structural elements. Identifiers within the boxes indicate the CM which was used to infer homology at this position. Asterisks indicate incomplete 3′UTR sequences. Species without available 3′UTR are not shown. (**b**) Consensus structure plots of CM hits as calculated by *mlocarna*. (**c**) Schematic depiction of the common structural architecture of TBFV 3′UTRs. (**d**,**e**) Structural alignments of elements SL6 and Y1. (**f**) `RNAalifold` coloring scheme for paired columns in alignments. Colors indicate the number of basepair combinations found in pair of columns. Fainter colors indicate that some sequences cannot form a base pair.

Despite the differences in length and sequence composition, the 3′UTRs of most species in the TBFV group share a common architecture. Similar to MBFV SL-elements [19], two copies of xrRNAs are found in almost every species of this ecologic group, generally succeeded by one instance of SL6 and Y1. Likewise, the terminal 3′ stem loop is conserved in all TBFVs and can be reliably annotated by both our CM, T.3SL, and the Rfam Flavi_CRE model, which is used in Figure 3. Among all

investigated species, only ALKV, OHFV and Kyasanur forest disease virus (KFDV) do not have a copy of xrRNA1, indicating that these viruses may have previously lost this element. Conversely, the two seabird-associated TBFVs with available 3′UTR data, Tyuleniy virus (TYUV) and Kama virus (KAMV) do not fit into this general scheme. Likewise, we were not able to annotate additional homologous or conserved structures with any CM used in this screen in the variable region of the 3′UTR of TBEV [5], despite the substantially longer UTR (+300 nts).

3.3. Classic Insect-Specific Flaviviruses

Classic insect-specific flaviviruses present diverged 3′UTR architectures, which likely result from the association of different species to *Aedes* spp. and *Culex* spp. vectors, respectively, which are also reflected by clade separation in the ISFV phylogenetic tree. Previous studies employed single sequence predictions to propose a varied set of homologous RNA structures in combination with an unusually large number of duplicated sequence signals [59]. Recent experimental evidence, however, suggests the presence of xrRNAs that have a similar fold to those known from MBFVs in cISFVs. Consequently, we set out to independently characterize conserved RNA elements for different subclades.

3.3.1. Exoribonuclease-Resistant RNAs in *Aedes*-Associated cISFVs

MacFadden et al. [33] used SHAPE structure probing to report the presence of two exoribonuclease-resistant RNAs in Cell fusing agent virus (CFAV), and provided evidence for a duplicated set of homologous structures in Aedes flavivirus (AEFV) and Kamiti river virus (KRV). We constructed initial alignments from the reported sequences in this clade in *Aedes* spp. associated viruses, resulting in two seed alignments, termed I.xrRNA1 and I.xrRNA2 (Figure 2c). For both elements, seed CMs were iteratively built from structural `locARNA` alignments. Minor manual adjustments to the alignments were required here, since the predicted consensus structures diverged slightly from the published SHAPE-guided prediction. Both models were then employed to search for additional high confidence hits within other isolates of CFAV, AEFV and KRV which were subsequently added to the seed alignments.

Screening the entire set of flavivirus 3′UTRs revealed that both ISFV xrRNA elements, I.xrRNA1 and I.xrRNA2, are only found in CFAV, AEFV and KRV, i.e., species initially used for the construction of the respective CMs. Furthermore, also in terms of pure structural conservation, no reliable hits in any other ISFV species could be obtained with any of these CMs (Figure 4a). This suggests that both ISFV xrRNAs may represent a specialized class of xrRNA elements only present in CFAV, AEFV and KRV. The 3′UTR of KRV is unique among all known flaviviruses because it harbors an additional copy of the terminal 3′ stem-loop element 600 nts upstream of the actual 3′-terminus, supporting previous reports that the KRV 3′UTR has undergone a full duplication during its evolution [60].

3.3.2. Conserved Structures in *Culex*-Associated cISFVs

The second distinct clade of cISFVs includes Culex flavivirus (CxFV), Quang Binh virus (QBV), Mosquito flavivirus (MSFV), Palm Creek virus (PCV), Culex theileri flavivirus (CTFV) as well as a few other species with only partial genome sequence availability [9] and is associated with *Culex* spp. vectors. An interesting observation in this clade is that no other CM from any of the four ecologic flavivirus groups shows a hit, not even with remote sequence or structure conservation. We therefore set out to produce a a high quality structural alignment of the complete 3′UTRs of CxFV, QBV and MSFV. Consensus structure folding of the full alignment revealed each species to harbor 3–4 repeats of two highly conserved elements supported by multiple co-varying base pairs (Figure 4b,c). We termed these "Repeat element a/b", respectively (Ra and Rb). Both elements, while strongly conserving their folds, show highly variable loop regions as well as weak sequence conservation in the case of the Ra element. Structure conservation and occurrence in multiple copies, as typically seen with other exoribonuclease-stalling elements, hints towards possible functional importance. These results

Viruses **2019**, *11*, 298

complement earlier reports of sequence repeats in the 3′UTR of CxFV [61] with the identification of a quadruplicated pair of conserved structures.

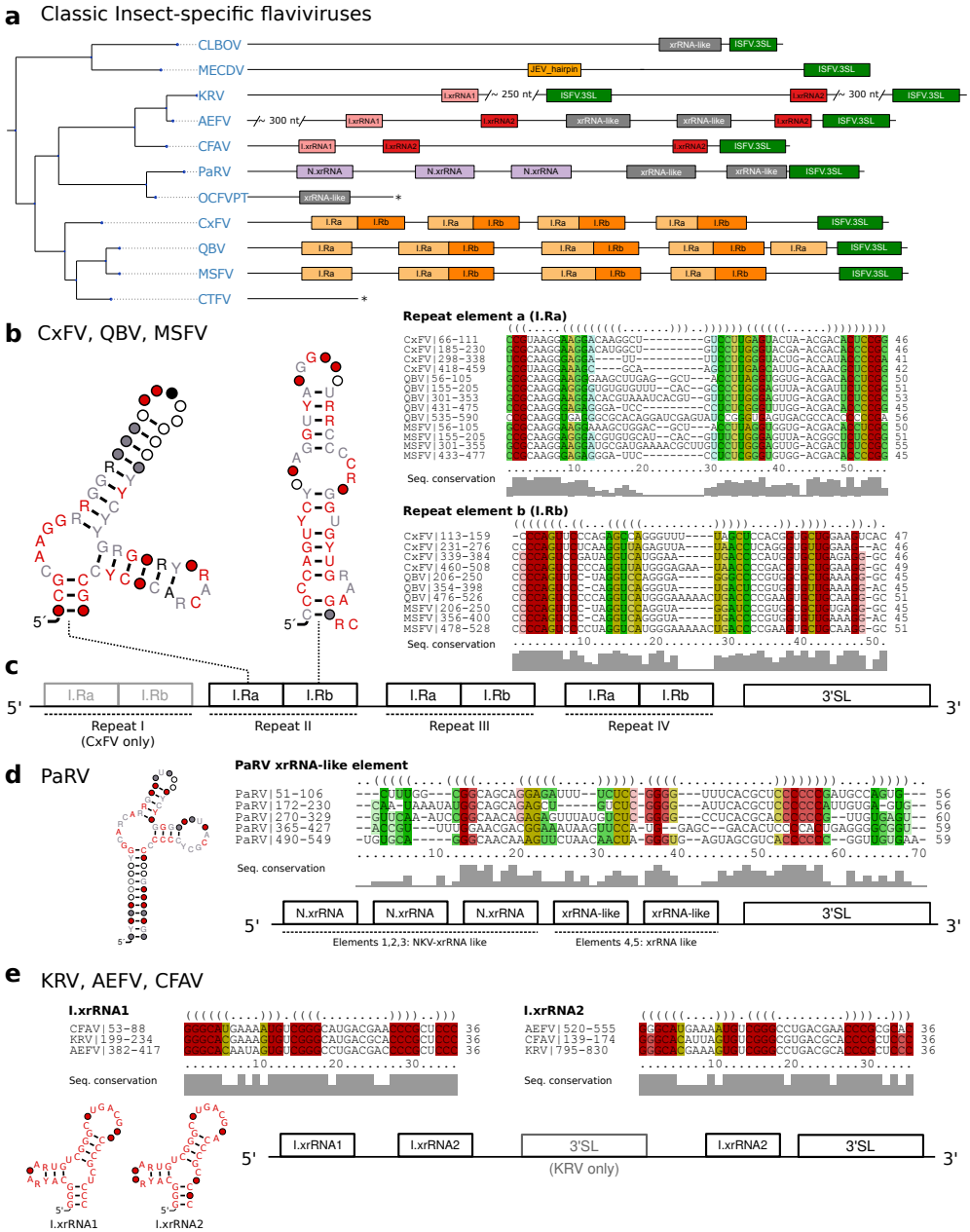


Figure 4. (**a**) Annotated Tree of cISFV 3′UTRs. Asterisks denote incomplete 3′UTR sequences. Species without available 3′UTR are not shown. (**b**) Consensus secondary structure plots and structural alignments of CM hits of Repeat a/b elements in CxFV, QBV, and MSFV. (**c**) Schematic of the common architecture of CxFV, QBV, and MSFV. Element I-IV refers to the respective repeat of elements. (**d**) Consensus structure plot and structural alignment of all CM hits of xrRNA-like elements in PaRV. (**e**) Proposed 3′UTR architecture of KRV, AEFV, and CFAV with consensus structure plots and structural alignments of I.xrRNA1 and I.xrRNA2.

3.3.3. Diverged 3′UTR Architecture in Many cISFVs

Interestingly, a screen of all available CMs in Parramatta River virus (PaRV) revealed five xrRNA-like elements (Figure 4a), with elements 1–3 sharing sequence and structure properties with NKV xrRNAs (Section 3.5), while elements 4 and 5 only conserve N.xrRNAs structure. All five hits can be structurally aligned into a consistent consensus structure (Figure 4d), despite the overall weak sequence consensus.

Conversely, the 3′UTRs of Calbertado virus (CLBOV) and Mercadeo virus (MECDV) appear structurally different from the other cISFVs. A general lack of characteristic CM hits lets these species appear more like an outgroup among cISFVs. In particular, we could only find significant hits for the omnipresent terminal 3′stem-loop structure, a putative xrRNA-like element in CLBOV and a single instance of a structure homologous to the Rfam model RF00465 (Japanese encephalitis virus hairpin structure) in MECDV. Still, limited availability of 3′UTR sequence data renders the characterization of conserved elements and interpretation difficult here.

Our data suggests that the 3′UTRs of cISFVs, in contrast to TBFVs (Section 3.2 and dISFVs (Section 3.4), do not appear to have a consistent architectural organization. In agreement with the cISFV phylogenetic subtree (Figure 1) we constitute three diverged groups with common 3′UTR organization that conform to their respective sub-clades: (i) CFAV-AEFV-KRV, each with two instances of xrRNAs, (ii) CxFV-QBV-MSFV with 3–4 copies of I.Ra/I.Rb elements and (iii) PaRV with 4–5 copies of xrRNA-like structures. Although no full 3′UTR sequences are available for the phylogenetically closest relatives of PaRV, HANV and OCFVPT, an xrRNA-like element in the small available fragment (syntenic to PaRV UTR) of OCFVPT 3′UTR suggests that both viruses might be organized in a similar manner, as supported by earlier reports that these viruses should be classified within the same species [1]. For CLBOV and MECDV, no clear pattern of conserved elements can be identified with our CMs. Both viruses either employ an entirely different class of elements or might not require capability for exoribonuclease stalling at all. The only element shared universally among all cISFVs is the 3′-terminal stem-loop, although cISFVs seem to diverge from other flaviviruses here, indicated by the inability of Rfam model RF00185 (Flavivirus CRE) to reliably annotate any cISFV 3′UTR.

3.4. Dual-Host Affiliated Insect-Specific Flaviviruses

Isolated almost exclusively from mosquitoes, dISFVs do not seem to infect vertebrate cells, despite their phylogenetic proximity to MBFVs (Figure 1). This association is reflected by good hits of the Rfam covariance models RF00525 (Flavivirus DB element) and RF00465 (Japanese encephalitis virus hairpin structure) in all dISFV isolates studied here (Figure 5). Interestingly, we could not find evidence for any sequences or structures homologous to tick-borne or other insect-specific flaviviruses. In this line, our data is in good agreement with the phylogenetic location of these viruses, which share ancestral roots with MBFVs [10].

An unusual species within this group is Ecuador Paraiso Escondido virus (EPEV), which has been isolated from New World sandflies and has been classified as insect-specific virus. EPEV phylogenetically appears at the root of the Entebbe bat virus group (ENTVG), a clade comprised of the three NKVs Entebbe bat virus (ENTV), Sokoluk virus (SOKV) and Yokose virus (YOKV). While all of these viruses contain homologs of conserved stem-loop (SL) and dumbbell (DB) elements found in MBFVs, ENTVG species may have lost their vector dependence [1].

Viruses **2019**, *11*, 298

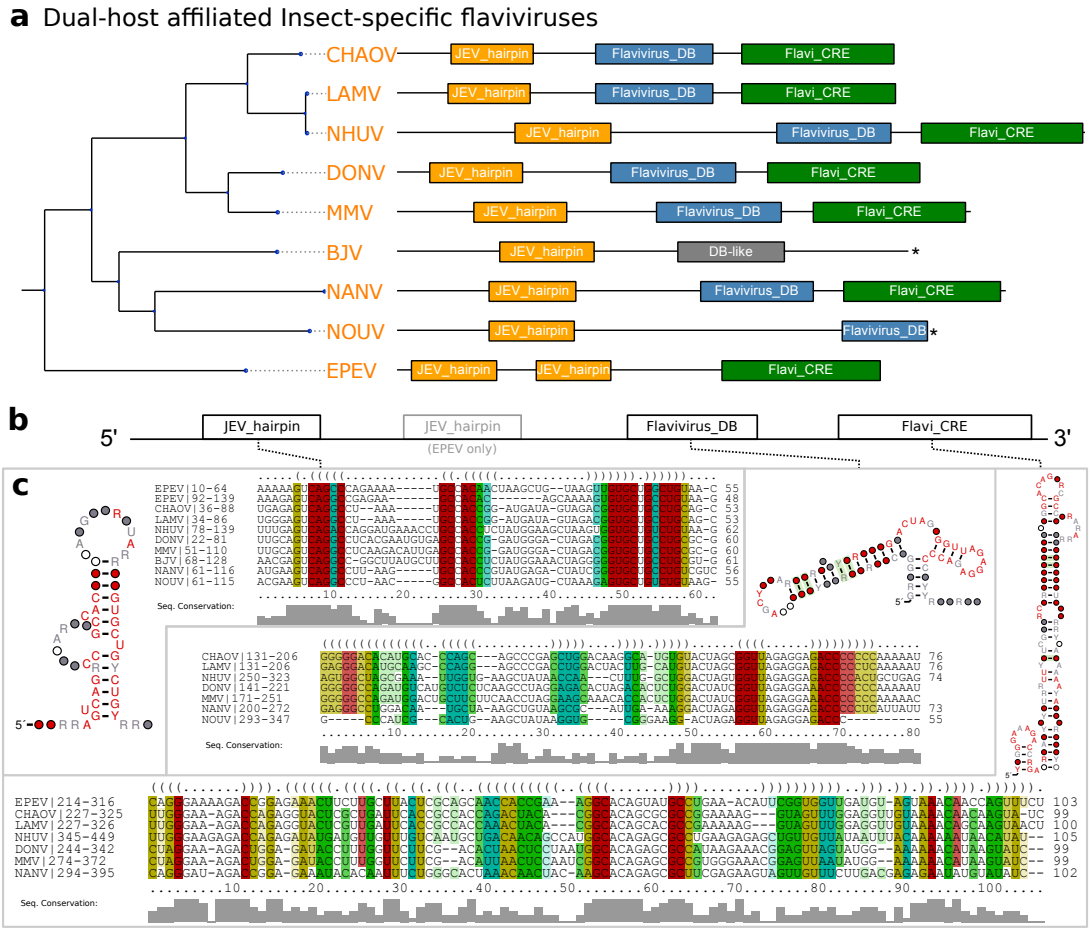


Figure 5. (**a**) Annotated Tree of dISFV 3′UTRs. Asterisks denote incomplete 3′UTR sequences. Species without available 3′UTR are not shown. (**b**) Schematic Architecture of the dISFV 3′UTR. (**c**) Structural alignments and consensus structure plots of dISFV elements.

3.5. No-Known-Vector Flaviviruses

Rather than forming a monophyletic group, the no-known-vector flaviviruses can be separated into two distinct lineages, which are closely related to either TBFVs or MBFVs, respectively (Figure 1). Two additional NKVs, Tamana bat virus (TABV) and Cyclopterus lumpus virus (CLuV), are phylogenetically distant and serve as an outgroup to all flaviviruses. In analogy to the procedure outlined above for TBFVs (Section 3.2) and ISFVs (Section 3.3), we built a CM for experimentally verified xrRNAs in tick-borne related NKVs, termed N.xrRNA.

We found multiple hits of this CM at various loci within the 3′UTRs of tick-borne related NKVs, indicating that these species, in contrast to TBFVs, do not conserve a common 3′UTR architecture (Figure 6a). Surprisingly, we could identify several high-quality hits of the Rfam model RF00525 (Flavivirus DB element), an element typically found in MBFVs, in Rio Bravo virus (RBV), Montana myotis leukoencephalitis virus (MMLV) and Modoc virus (MODV). This is in so far remarkable as there is no evidence for conservation of this element in TBFVs, which phylogenetically cluster with this clade of NKVs. This element might have been introduced by an ancestral recombination event. Alternatively, conservation of an MBFV element in NKVs might be indicative of an association with an unknown vector, in agreement with the hypothesis that vector specificity is mediated by characteristic 3′UTR elements [19].

Conversely, there seems to be no generally conserved 3′UTR architecture among members of the mosquito-borne related NKVs (Figure 6c). While sequence data has not been available for Sokoluk virus (SOKV), we could annotate typical MBFV elements in the next relatives Entebbe bat virus (ENTV) and Yokose virus (YOKV), as proposed previously [32].

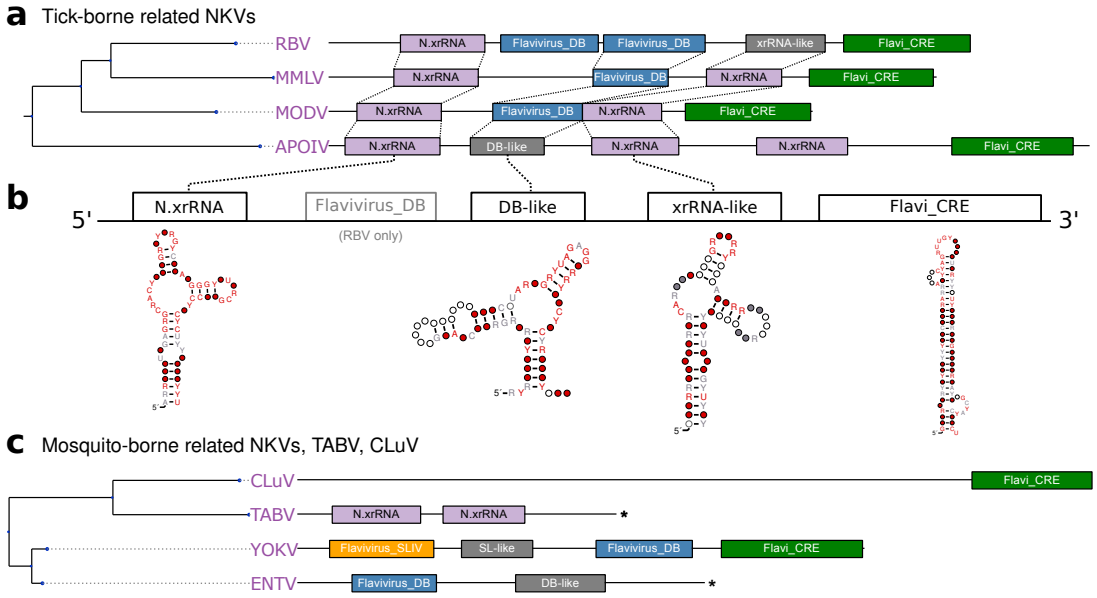


Figure 6. (**a**,**c**) Annotated 3′UTRs of NKVs. Asterisks denote incomplete 3′UTR sequences. (**b**) Schematic of TBFV-associated NKV-FV UTR architecture with consensus structures of NKV structure elements.

3.6. A Generalized xrRNA Structure

Earlier work suggested that xrRNAs from TBFVs and tick-borne related NKVs fall into a more general structural class of xrRNAs [33]. Following this line of reasoning, we investigated whether all high confidence hits obtained with our TBFV and NKV CMs could be assembled into one coherent CM that conserves the xrRNA-typical fold. A further advantage of a generalized CM would be higher sensitivity, allowing for identification of common features and eventually lead to annotation of previously unannotated xrRNAs.

Structural alignment and consensus structure prediction revealed all high confidence hits to fold into a common secondary structure (Figure 7a,b). While most of the consensus structure is characterized by low sequence conservation, stem 3 (S3) and loop 1 (L1) show medium to high degree of sequence conservation. The length of all stems is well conserved, although both major loop regions L2 and L3 show large fluctuations, with the length of L3 being de facto constant and L2 showing a high degree of flexibility.

We further investigated whether any high confidence hits from I.xrRNA1/2 in cISFVs could be aligned to the generalized xrRNA model. Although both cISFV xrRNAs (Figure 2c) bear some similarity to the generalized model, in particular to S3 and L3, we were not able to build a common alignment or consensus structure. Despite seemingly similar shape, our data also suggests that cISFV xrRNAs form a separate xrRNA subclass, unrelated to MBFV xrRNAs. In particular, we could not obtain hits of Rfam CMs (which can be seen as representatives of MBFV elements) in cISFVs, nor could we confirm any hits of cISFV-specific elements in MBFVs.

In addition to learning xrRNA features, a more generalized CM enabled us to detect xrRNA-like structures (indicated as such in all annotation plots) that could not be found previously.

Viruses **2019**, *11*, 298

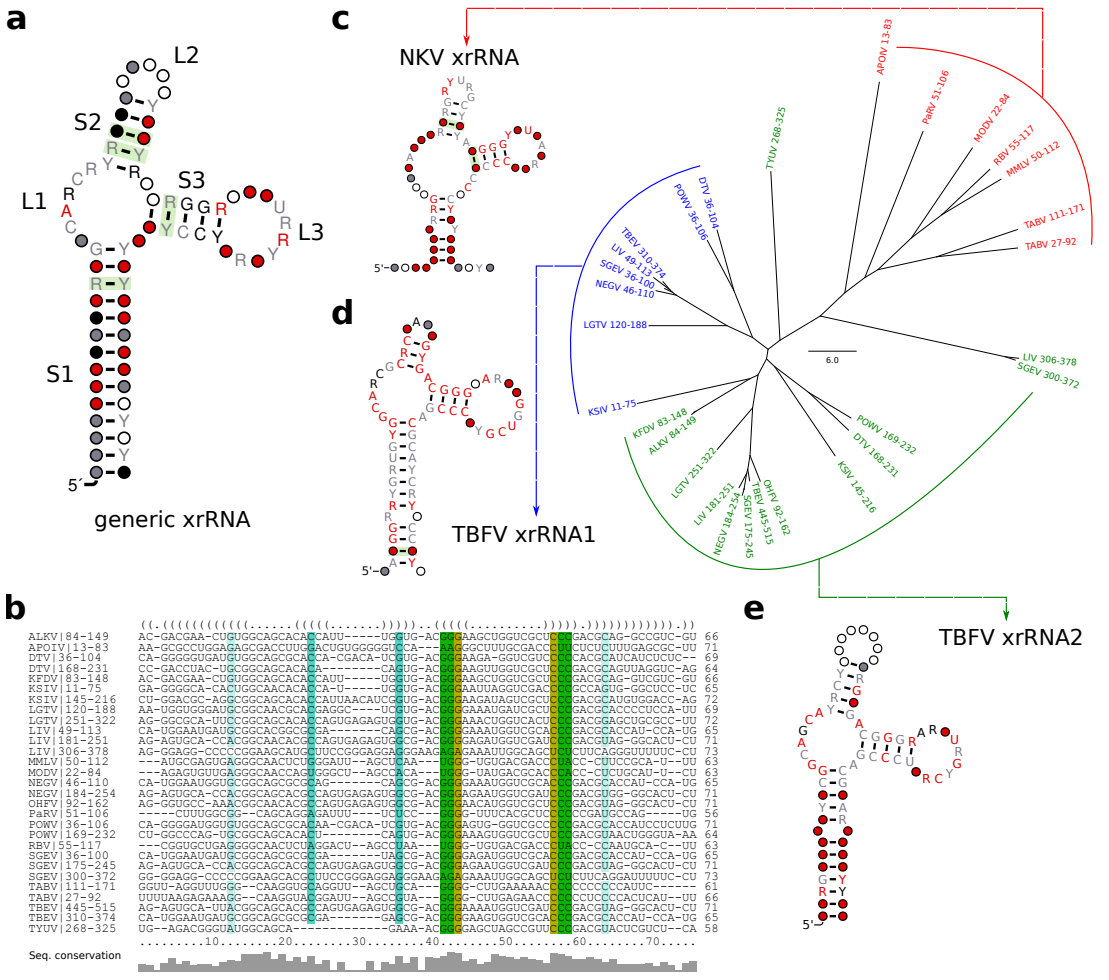


Figure 7. Generalized structure of all high confidence (cmsearch evalue $< 10^{-5}$) hits of T.xrRNA1, T.xrRNA2 and N.xrRNA. (**a**) Consensus structure prediction and (**b**) structural alignment of all high confidence hits. (**c**–**e**) Neighbor-joining tree of all high confidence hits of N.xrRNA (**c**), T.xrRNA1 (**d**), and T.xrRNA2 (**e**). Leaves are grouped and colored by the CM used for annotation, coordinates correspond to the position in the respective 3′UTR. For each group a separate structural alignment was computed, the consensus structures are shown.

4. Discussion

Mediated gRNA decay in the form of exoribonuclease resistance seems to be a pervasive strategy employed by viruses to circumvent host immune responses. Evidence of sfRNA production following incomplete Xrn1 degradation has not only been observed in different members of the *Flavivirus* genus [62], but also in other species of the *Flaviviridae* family, however, with major differences in xrRNA structure and sfRNA characteristics. While MBFV produce a 300-500nt sfRNA that corresponds to degradation products of the gRNA 3′UTR, hepaciviruses and pestiviruses produce a long subgenomic RNA whose 5′ end is located within the first 130nt of the viral gRNA [63].

Moreover, recent studies have identified xrRNA functionality in several phylogenetically distant RNA viruses, such as animal-infecting, segmented viruses of the *Bunyaviridae* and *Arenavividae* [64] families, as well as plant-infecting viruses of the *Tombusuviridae* and *Luteoviridae* families [65,66]. The interesting question whether exoribonucleases other than Xrn1 would be blocked as well has recently been answered. MacFadden et al. [33] could show that both RNAse J1 and Dxo1 are stalled by MBFV xrRNAs, thereby demonstrating the general nature of this structure-induced blocking

mechanism. These novel findings, together with previous knowledge of Xrn1 stalling in segmented plant viruses [67,68] provide evidence for a convergent evolution scenario where xrRNAs depend on a specific folded RNA structure and form a distinct class of functional RNAs.

Repeated RNA elements appear to be a hallmark of flavivirus 3′UTR architecture. While there seems to be a plethora of conserved structure classes, our data emphasizes the consistent trend that these elements typically do not occur as single copies. Rather, duplicate or even multiple occurrences of these elements hint towards functional relevance. This is further underlined by the evolutionary conservation of both patterns and elements among different species. Since, exoribonuclease stalling is presumably never perfect, it makes sense that viruses might employ multiple copies of such elements.

In this contribution we set out to identify homologs of known exoribonuclease stalling elements and novel conserved structures. To this end, we computationally characterized homologs of experimentally verified xrRNA in tick-borne and no-known vector viruses that seem to form a coherent class of RNA structures with capability to stall exoribonucleases (T.xrRNA1/2, N.xrRNA). Likewise, we identified another class of xrRNAs in classic insect-specific flaviviruses (I.xrRNA1/2) which appears to be only distantly related to the former class. In the same line, we predicted a set of novel conserved elements in cISFVs that appear in quadruples and do not coincide with other insect-specific elements (I.Ra,I.Rb).

While we did not focus on studying the evolutionary history of these groups of elements in detail, our data suggests that many elements share ancestral roots. This is supported by the observation that at least the tick-borne, no-known-vector and *Aedes* spp. associated xrRNAs fold into a similar Y-shaped substructure, although the exact fold varies significantly among individual species.

We compiled a set of covariance models that can be used for rapid screening assays in the identification and characterization of novel flaviviruses. All models are available from GitHub via https://github.com/mtw/ITNFV-Data.

A major problem is the limited availability of diverse 3′UTR sequence data for many viruses analyzed here, particularly within cISFVs. Many novel ISFVs have previously been discovered, but 3′UTR sequences were only available for a subset of them. Future studies are required to shed more light on the evolutionary history of 3′UTR evolution in flaviviruses.

Author Contributions: M.T.W. conceived the study. R.O. and M.T.W. analyzed the data, characterized conserved RNA elements and performed phylogenetic studies. All authors contributed to writing the paper and approved of the submitted version.

Funding: This work was funded in part by the Doktoratskolleg RNA Biology at Univ. Vienna and the Austrian Science Fund (SFB F43, I-1303).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Data Set Analyzed in This Study

Table A1. Viral genomes considered in this study. Flaviviruses are categorized into the groups tick-borne flaviviruses (TBFV), insect-specific flaviviruses (ISFV) and no-known-vector flaviviruses (NKV). The length of the 3′UTR is listed for each isolate. † Representative accession number from the `refseq` database. Whenever a `refseq` genome was not available, the isolate with the longest 3′UTR was selected as representative species. [a] Mammalian TBFVs. [b] Seabird TBFVs. [c] Classic ISFVs. [d] Dual-host affiliated ISFVs. [e] Rodent-associated NKVs. [f] Bat-associated NKVs. [N/A] 3′UTR partial or not available in the `refseq` data set.

Group	Accession Number †	Acronym	Scientific Name	3′UTR Length (nt)	Isolates
ISFV [c]	NC_012932.1	AEFV	Aedes flavivirus	942	3
ISFV [c]	NC_001564.2	CFAV	Cell fusing agent virus	553	3
ISFV [c]	KX669689.1	CLBOV	Calbertado virus	546	8
ISFV [c]	MF153378.1	CTFV	Culex theileri flavivirus	112	1

Table A1. *Cont.*

Group	Accession Number †	Acronym	Scientific Name	3′UTR Length (nt)	Isolates
ISFV [c]	NC_008604.2	CxFV	Culex flavivirus	654	13
ISFV [c]	NC_030401.1	HANV	Hanko virus	*N/A*	*N/A*
ISFV [c]	NC_005064.1	KRV	Kamiti River virus	1208	3
ISFV [c]	NC_027819.1	MECDV	Mercadeo virus	638	3
ISFV [c]	NC_021069.1	MSFV	Mosquito flavivirus	674	6
ISFV [c]	NC_034242.1	OCFVPT	Ochlerotatus caspius flavivirus	148	2
ISFV [c]	NC_027817.1	PaRV	Parramatta River virus	629	2
ISFV [c]	NC_033694.1	PCV	Palm Creek virus	*N/A*	*N/A*
ISFV [c]	NC_012671.1	QBV	Quang Binh virus	673	2
ISFV [d]	MG214905.1	BJV	Barkedji virus	335	1
ISFV [d]	NC_017086.1	CHAOV	Chaoyang virus	326	2
ISFV [d]	NC_016997.1	DONV	Donggang virus	343	2
ISFV [d]	NC_027999.1	EPEV	Paraiso Escondido virus	316	2
ISFV [d]	NC_024805.1	ILV	Ilomantsi virus	*N/A*	*N/A*
ISFV [d]	KY320648.1	KPKV	Kampung Karu virus	*N/A*	*N/A*
ISFV [d]	FJ606789.2	LAMV	Lammi virus	326	1
ISFV [d]	KY290249.1	LPKV	Long Pine Key virus	*N/A*	*N/A*
ISFV [d]	KY320649.1	LTNV	La Tina virus	*N/A*	*N/A*
ISFV [d]	MF139576.1	MMV	Marisma mosquito virus	376	1
ISFV [d]	MF139575.1	NANV	Nanay virus	399	1
ISFV [d]	NC_024017.1	NHUV	Nhumirim virus	451	1
ISFV [d]	NC_033715.1	NOUV	Nounane virus	347	3
NKV [e]	NC_003676.1	APOIV	Apoi virus	576	1
NKV	KJ469370.1	BCV	Batu Cave virus	*N/A*	*N/A*
NKV	MF776369.1	CLuV	Cyclopterus lumpus virus	601	1
NKV [f]	NC_008718.1	ENTV	Entebbe bat virus	308	3
NKV [e]	NC_026620.1	JUTV	Jutiapa virus	*N/A*	*N/A*
NKV [f]	NC_004119.1	MMLV	Montana myotis leukoencephalitis virus	460	1
NKV [e]	NC_003635.1	MODV	Modoc virus	366	1
NKV [f]	NC_034007.1	PPBV	Phnom Penh bat virus	*N/A*	*N/A*
NKV [f]	NC_003675.1	RBV	Rio Bravo virus	486	2
NKV [f]	NC_026624.1	SOKV	Sokoluk virus	*N/A*	*N/A*
NKV	NC_003996.1	TABV	Tamana bat virus	241	1
NKV [f]	NC_005039.1	YOKV	Yokose virus	429	1
TBFV [a]	NC_004355.1	ALKV	Alkhumra hemorrhagic fever virus	393	21
TBFV [a]	AF311056.1	DTV	Deer tick virus	459	1
TBFV [a]	NC_033723.1	GGV	Gadgets Gully virus	*N/A*	*N/A*
TBFV	NC_033724.1	KADV	Kadam virus	*N/A*	*N/A*
TBFV [b]	NC_023439.1	KAMV	Kama virus	282	2
TBFV [a]	HM055369.1	KFDV	Kyasanur forest disease virus	392	6
TBFV [a]	NC_006947.1	KSIV	Karshi virus	381	3
TBFV [a]	NC_003690.1	LGTV	Langat virus	568	5
TBFV [a]	NC_001809.1	LIV	Louping ill virus	500	5
TBFV [b]	NC_033721.1	MEAV	Meaban virus	*N/A*	*N/A*
TBFV [a]	KT224355.1	NEGV	Negishi virus	266	1
TBFV [a]	NC_005062.1	OHFV	Omsk hemorrhagic fever virus	410	4
TBFV [a]	NC_003687.1	POWV	Powassan virus	480	23
TBFV	DQ235149.1	RFV	Royal Farm virus	*N/A*	*N/A*
TBFV [a]	NC_027709.1	SGEV	Spanish goat encephalitis virus	493	2
TBFV [b]	NC_033726.1	SREV	Saumarez Reef virus	*N/A*	*N/A*
TBFV [a]	NC_001672.1	TBEV	Tick-borne encephalitis virus	764	167
TBFV [b]	NC_023424.1	TYUV	Tyuleniy virus	591	3

References

1. Kuno, G.; Chang, G.J.J.; Tsuchiya, K.R.; Karabatsos, N.; Cropp, C.B. Phylogeny of the Genus *Flavivirus*. *J. Virol.* **1998**, *72*, 73–83.
2. Gaunt, M.W.; Sall, A.A.; de Lamballerie, X.; Falconar, A.K.; Dzhivanian, T.I.; Gould, E.A. Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. *J. Gen. Virol.* **2001**, *82*, 1867–1876. [CrossRef]

3. Vasilakis, N.; Weaver, S.C. Flavivirus transmission focusing on Zika. *Curr. Opin. Virol.* **2017**, *22*, 30–35. [CrossRef]
4. Barrows, N.J.; Campos, R.K.; Liao, K.C.; Prasanth, K.R.; Soto-Acosta, R.; Yeh, S.C.; Schott-Lerner, G.; Pompon, J.; Sessions, O.M.; Bradrick, S.S.; et al. Biochemistry and Molecular Biology of Flaviviruses. *Chem. Rev.* **2018**, *118*, 4448–4482. [CrossRef]
5. Kellman, E.; Offerdahl, D.; Melik, W.; Bloom, M. Viral Determinants of Virulence in Tick-Borne Flaviviruses. *Viruses* **2018**, *10*, 329. [CrossRef] [PubMed]
6. Braack, L.; de Almeida, A.P.G.; Cornel, A.J.; Swanepoel, R.; De Jager, C. Mosquito-borne arboviruses of African origin: Review of key viruses and vectors. *Parasite Vector* **2018**, *11*, 29. [CrossRef] [PubMed]
7. Gritsun, T.; Lashkevich, V.; Gould, E. Tick-borne encephalitis. *Antivir. Res.* **2003**, *57*, 129–146. [CrossRef]
8. Guzman, H.; Contreras-Gutierrez, M.A.; da Rosa, A.P.T.; Nunes, M.R.; Cardoso, J.F.; Popov, V.L.; Young, K.I.; Savit, C.; Wood, T.G.; Widen, S.G.; et al. Characterization of three new insect-specific flaviviruses: Their relationship to the mosquito-borne flavivirus pathogens. *Am. J. Trop. Med. Hyg.* **2018**, *98*, 410–419. [CrossRef] [PubMed]
9. Blitvich, B.; Firth, A. Insect-Specific Flaviviruses: A Systematic Review of Their Discovery, Host Range, Mode of Transmission, Superinfection Exclusion Potential and Genomic Organization. *Viruses* **2015**, *7*, 1927–1959. [CrossRef] [PubMed]
10. Alkan, C.; Zapata, S.; Bichaud, L.; Moureau, G.; Lemey, P.; Firth, A.E.; Gritsun, T.S.; Gould, E.A.; de Lamballerie, X.; Depaquit, J.; et al. Ecuador Paraiso Escondido virus, a new flavivirus isolated from New World sandflies in Ecuador is the first representative of a novel clade in the genus *Flavivirus*. *J. Virol.* **2015**, *89*, 11773–11785. [CrossRef] [PubMed]
11. Halbach, R.; Junglen, S.; van Rij, R.P. Mosquito-specific and mosquito-borne viruses: Evolution, infection, and host defense. *Curr. Opin. Insect. Sci.* **2017**, *22*, 16–27. [CrossRef] [PubMed]
12. Bolling, B.; Weaver, S.; Tesh, R.; Vasilakis, N. Insect-Specific Virus Discovery: Significance for the Arbovirus Community. *Viruses* **2015**, *7*, 4911–4928. [CrossRef]
13. Blitvich, B.; Firth, A. A Review of Flaviviruses that Have No Known Arthropod Vector. *Viruses* **2017**, *9*, 154. [CrossRef] [PubMed]
14. Filomatori, C.V.; Lodeiro, M.F.; Alvarez, D.E.; Samsa, M.M.; Pietrasanta, L.; Gamarnik, A.V. A 5′ RNA element promotes dengue virus RNA synthesis on a circular genome. *Gene Dev.* **2006**, *20*, 2238–2249. [CrossRef] [PubMed]
15. Alvarez, D.E.; Ezcurra, A.L.D.L.; Fucito, S.; Gamarnik, A.V. Role of RNA structures present at the 3′ UTR of dengue virus on translation, RNA synthesis, and viral replication. *Virology* **2005**, *339*, 200–212. [CrossRef] [PubMed]
16. Manzano, M.; Reichert, E.D.; Polo, S.; Falgout, B.; Kasprzak, W.; Shapiro, B.A.; Padmanabhan, R. Identification of Cis-Acting Elements in the 3′-Untranslated Region of the Dengue Virus Type 2 RNA That Modulate Translation and Replication. *J Biol. Chem.* **2011**, *286*, 22521–22534. [CrossRef]
17. Brinton, M.A.; Basu, M. Functions of the 3′ and 5′ genome RNA regions of members of the genus Flavivirus. *Virus Res.* **2015**, *206*, 108–119. [CrossRef]
18. Ng, W.; Soto-Acosta, R.; Bradrick, S.; Garcia-Blanco, M.; Ooi, E. The 5′ and 3′ untranslated regions of the flaviviral genome. *Viruses* **2017**, *9*, 137. [CrossRef] [PubMed]
19. Villordo, S.M.; Carballeda, J.M.; Filomatori, C.V.; Gamarnik, A.V. RNA Structure Duplications and Flavivirus Host Adaptation. *Trends Microbiol.* **2016**, *24*, 270–283. [CrossRef] [PubMed]
20. de Borba, L.; Villordo, S.M.; Marsico, F.L.; Carballeda, J.M.; Filomatori, C.V.; Gebhard, L.G.; Pallarés, H.M.; Lequime, S.; Lambrechts, L.; Vargas, I.S.; et al. RNA Structure Duplication in the Dengue Virus 3′ UTR: Redundancy or Host Specificity? *mBio* **2019**, *10*, e02506–18. [CrossRef]
21. de Bernardi Schneider, A.; Wolfinger, M.T. Musashi binding elements in Zika and related Flavivirus 3′UTRs: A comparative study in silico. *bioRxiv* **2018**. [CrossRef]
22. Rice, C.M.; Lenches, E.M.; Eddy, S.R.; Shin, S.; Sheets, R.; Strauss, J. Nucleotide sequence of yellow fever virus: Implications for flavivirus gene expression and evolution. *Science* **1985**, *229*, 726–733. [CrossRef] [PubMed]
23. Hahn, C.S.; Hahn, Y.S.; Rice, C.M.; Lee, E.; Dalgarno, L.; Strauss, E.G.; Strauss, J.H. Conserved elements in the 3′ untranslated region of flavivirus RNAs and potential cyclization sequences. *J. Mol. Biol.* **1987**, *198*, 33–41. [CrossRef]

24. Villordo, S.M.; Alvarez, D.E.; Gamarnik, A.V. A balance between circular and linear forms of the dengue virus genome is crucial for viral replication. *RNA* **2010**, *16*, 2325–2335. [CrossRef] [PubMed]
25. de Borba, L.; Villordo, S.M.; Iglesias, N.G.; Filomatori, C.V.; Gebhard, L.G.; Gamarnik, A.V. Overlapping local and long-range RNA-RNA interactions modulate dengue virus genome cyclization and replication. *J. Virol.* **2015**, *89*, 3430–3437. [CrossRef] [PubMed]
26. Pijlman, G.P.; Funk, A.; Kondratieva, N.; Leung, J.; Torres, S.; Van der Aa, L.; Liu, W.J.; Palmenberg, A.C.; Shi, P.Y.; Hall, R.A.; et al. A highly structured, nuclease-resistant, noncoding RNA produced by flaviviruses is required for pathogenicity. *Cell Host Microbe* **2008**, *4*, 579–591. [CrossRef] [PubMed]
27. Akiyama, B.M.; Eiler, D.; Kieft, J.S. Structured RNAs that evade or confound exonucleases: Function follows form. *Curr. Opin. Struct. Biol.* **2016**, *36*, 40–47. [CrossRef] [PubMed]
28. Moon, S.L.; Anderson, J.R.; Kumagai, Y.; Wilusz, C.J.; Akira, S.; Khromykh, A.A.; Wilusz, J. A noncoding RNA produced by arthropod-borne flaviviruses inhibits the cellular exoribonuclease XRN1 and alters host mRNA stability. *RNA* **2012**, *18*, 2029–2040. [CrossRef]
29. Roby, J.A.; Pijlman, G.P.; Wilusz, J.; Khromykh, A.A. Noncoding subgenomic flavivirus RNA: Multiple functions in West Nile virus pathogenesis and modulation of host responses. *Viruses* **2014**, *6*, 404–427. [CrossRef]
30. Jones, C.I.; Zabolotskaya, M.V.; Newbury, S.F. The 5′-> 3′ exoribonuclease Xrn1/Pacman and its functions in cellular processes and development. *Wiley Interdiscip. Rev. RNA* **2012**, *3*, 455–468. [CrossRef]
31. Antic, S.; Wolfinger, M.T.; Skucha, A.; Hosiner, S.; Dorner, S. General and miRNA-mediated mRNA degradation occurs on ribosome complexes in Drosophila cells. *Mol. Cell Biol.* **2015**, doi:10.1128/MCB.01346-14. [CrossRef]
32. Kieft, J.S.; Rabe, J.L.; Chapman, E.G. New hypotheses derived from the structure of a flaviviral Xrn1-resistant RNA: Conservation, folding, and host adaptation. *RNA Biol.* **2015**, *12*, 1169–1177. [CrossRef]
33. MacFadden, A.; O'Donoghue, Z.; Silva, P.A.; Chapman, E.G.; Olsthoorn, R.C.; Sterken, M.G.; Pijlman, G.P.; Bredenbeek, P.J.; Kieft, J.S. Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. *Nat. Commun.* **2018**, *9*, 119. [CrossRef]
34. Hussain, M.; Torres, S.; Schnettler, E.; Funk, A.; Grundhoff, A.; Pijlman, G.P.; Khromykh, A.A.; Asgari, S. West Nile virus encodes a microRNA-like small RNA in the 3′ untranslated region which up-regulates GATA4 mRNA and facilitates virus replication in mosquito cells. *Nucleic Acids Res.* **2012**, *40*, 2210–2223. [CrossRef] [PubMed]
35. Schuessler, A.; Funk, A.; Lazear, H.M.; Cooper, D.A.; Torres, S.; Daffis, S.; Jha, B.K.; Kumagai, Y.; Takeuchi, O.; Hertzog, P.; et al. West Nile virus noncoding subgenomic RNA contributes to viral evasion of the type I interferon-mediated antiviral response. *J. Virol.* **2012**, *86*, 5708–5718. [CrossRef] [PubMed]
36. Bidet, K.; Dadlani, D.; Garcia-Blanco, M.A. G3BP1, G3BP2 and CAPRIN1 are required for translation of interferon stimulated mRNAs and are targeted by a dengue virus non-coding RNA. *PLoS Pathog.* **2014**, *10*, e1004242. [CrossRef] [PubMed]
37. Lorenz, R.; Wolfinger, M.T.; Tanzer, A.; Hofacker, I.L. Predicting RNA Structures from Sequence and Probing Data. *Methods* **2016**, *103*, 86–98. [CrossRef] [PubMed]
38. Funk, A.; Truong, K.; Nagasaki, T.; Torres, S.; Floden, N.; Melian, E.B.; Edmonds, J.; Dong, H.; Shi, P.Y.; Khromykh, A.A. RNA structures required for production of subgenomic flavivirus RNA. *J. Virol.* **2010**, *84*, 11407–11417. [CrossRef]
39. Silva, P.A.; Pereira, C.F.; Dalebout, T.J.; Spaan, W.J.; Bredenbeek, P.J. An RNA pseudoknot is required for production of yellow fever virus subgenomic RNA by the host nuclease XRN1. *J. Virol.* **2010**, *84*, 11395–11406. [CrossRef] [PubMed]
40. Sztuba-Solinska, J.; Teramoto, T.; Rausch, J.W.; Shapiro, B.A.; Padmanabhan, R.; Le Grice, S.F. Structural complexity of Dengue virus untranslated regions: cis-acting RNA motifs and pseudoknot interactions modulating functionality of the viral genome. *Nucleic Acids Res.* **2013**, *41*, 5075–5089. [CrossRef] [PubMed]
41. Chapman, E.G.; Costantino, D.A.; Rabe, J.L.; Moon, S.L.; Wilusz, J.; Nix, J.C.; Kieft, J.S. The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production. *Science* **2014**, *344*, 307–310. [CrossRef] [PubMed]
42. Akiyama, B.M.; Laurence, H.M.; Massey, A.R.; Costantino, D.A.; Xie, X.; Yang, Y.; Shi, P.Y.; Nix, J.C.; Beckham, J.D.; Kieft, J.S. Zika virus produces noncoding RNAs using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science* **2016**, *354*, 1148–1152. [CrossRef]

43. Rauscher, S.; Flamm, C.; Mandl, C.W.; Heinz, F.X.; Stadler, P.F. Secondary structure of the 3′-noncoding region of flavivirus genomes: comparative analysis of base pairing probabilities. *RNA* **1997**, *3*, 779–791. [PubMed]
44. Hofacker, I.L.; Fekete, M.; Flamm, C.; Huynen, M.A.; Rauscher, S.; Stolorz, P.E.; Stadler, P.F. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acid Res.* **1998**, *26*, 3825–3836. [CrossRef] [PubMed]
45. Witwer, C.; Rauscher, S.; Hofacker, I.L.; Stadler, P.F. Conserved RNA secondary structures in Picornaviridae genomes. *Nucleic Acid Res.* **2001**, *29*, 5079–5089. [CrossRef]
46. Hofacker, I.L.; Stadler, P.F.; Stocsits, R.R. Conserved RNA secondary structures in viral genomes: A survey. *Bioinformatics* **2004**, *20*, 1495–1499. [CrossRef]
47. Thurner, C.; Witwer, C.; Hofacker, I.L.; Stadler, P.F. Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.* **2004**, *85*, 1113–1124. [CrossRef]
48. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
49. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2014**, *32*, 268–274. [CrossRef] [PubMed]
50. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold Faster RNA Homology Searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef]
51. Waldl, M.; Thiel, B.C.; Ochsenreiter, R.; Holzenleiter, A.; de Araujo Oliveira, J.V.; Walter, M.E.M.; Wolfinger, M.T.; Stadler, P.F. TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes. *Genes* **2018**, *9*, 372. [CrossRef] [PubMed]
52. Will, S.; Reiche, K.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comp. Biol.* **2007**, *3*, e65. [CrossRef] [PubMed]
53. Lorenz, R.; Bernhart, S.H.; Zu Siederdissen, C.H.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithm Mol. Biol.* **2011**, *6*, 26. [CrossRef] [PubMed]
54. Wolfinger, M.T.; Fallmann, J.; Eggenhofer, F.; Amman, F. ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines. *F1000Research* **2015**, *4*. [CrossRef]
55. Wolfinger, M.T. Bio::RNA::RNAaliSplit 0.09. 2019. Available online: https://github.com/mtw/Bio-RNA-RNAaliSplit (accessed on 23 March 2019).
56. Weinberg, Z.; Breaker, R.R. R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinform.* **2011**, *12*, 3. [CrossRef]
57. Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **2016**, *33*, 1635–1638. [CrossRef] [PubMed]
58. Gritsun, T.; Gould, E. Origin and Evolution of 3′UTR of Flaviviruses: Long Direct Repeats as A Basis for the Formation of Secondary Structures and Their Significance for Virus Transmission. *Adv. Virus Res.* **2006**, *69*, 203–248.
59. Gritsun, D.J.; Jones, I.M.; Gould, E.A.; Gritsun, T.S. Molecular Archaeology of Flaviviridae Untranslated Regions: Duplicated RNA Structures in the Replication Enhancer of Flaviviruses and Pestiviruses Emerged via Convergent Evolution. *PLoS ONE* **2014**, *9*, e92056. [CrossRef]
60. Gritsun, T.; Gould, E. The 3′ untranslated regions of Kamiti River virus and Cell fusing agent virus originated by self-duplication. *J. Gen. Virol.* **2006**, *87*, 2615–2619. [CrossRef]
61. Hoshino, K.; Isawa, H.; Tsuda, Y.; Yano, K.; Sasaki, T.; Yuda, M.; Takasaki, T.; Kobayashi, M.; Sawabe, K. Genetic characterization of a new insect flavivirus isolated from *Culex pipiens* mosquito in Japan. *Virology* **2007**, *359*, 405–414. [CrossRef]
62. Clarke, B.; Roby, J.; Slonchak, A.; Khromykh, A. Functional non-coding RNAs derived from the flavivirus 3′ untranslated region. *Virus Res.* **2015**, *206*, 53–61. [CrossRef]
63. Moon, S.L.; Blackinton, J.G.; Anderson, J.R.; Dozier, M.K.; Dodd, B.J.; Keene, J.D.; Wilusz, C.J.; Bradrick, S.S.; Wilusz, J. X1 Stalling in the 5′UTR of Hepatitis C Virus and Bovine Viral Diarrhea Virus Is Associated with Dysregulated Host mRNA Stability. *PLoS Pathog.* **2015**, *11*, e1004708. [CrossRef]
64. Charley, P.A.; Wilusz, C.J.; Wilusz, J. Identification of Phlebovirus and Arenavirus RNA sequences that stall and repress the exoribonuclease XRN1. *J. Biol. Chem.* **2018**, *293*, 285–295. [CrossRef] [PubMed]

65. Steckelberg, A.L.; Akiyama, B.M.; Costantino, D.A.; Sit, T.L.; Nix, J.C.; Kieft, J.S. A folded viral noncoding RNA blocks host cell exoribonucleases through a conformationally dynamic RNA structure. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 6404–6409. [CrossRef] [PubMed]
66. Steckelberg, A.L.; Vicens, Q.; Kieft, J.S. Exoribonuclease-Resistant RNAs Exist within both Coding and Noncoding Subgenomic RNAs. *mBio* **2018**, *9*, e02461-18. [CrossRef] [PubMed]
67. Iwakawa, H.o.; Mizumoto, H.; Nagano, H.; Imoto, Y.; Takigawa, K.; Sarawaneeyaruk, S.; Kaido, M.; Mise, K.; Okuno, T. A viral noncoding RNA generated by cis-element-mediated protection against 5′- 3′ RNA decay represses both cap-independent and cap-dependent translation. *J. Virol.* **2008**, *82*, 10162–10174. [CrossRef] [PubMed]
68. Flobinus, A.; Hleibieh, K.; Klein, E.; Ratti, C.; Bouzoubaa, S.; Gilmer, D. A viral noncoding RNA complements a weakened viral RNA silencing suppressor and promotes efficient systemic host infection. *Viruses* **2016**, *8*, 272. [CrossRef] [PubMed]

7 CONSERVED SECONDARY STRUCTURES IN VIRAL RNAS

The following chapter presents the work that was conducted as part of a systematic investigation of RNA structure over all Viral Ortholog Groups, published in the article *Viruses 2019, 11, 401*. A reprint of the article is provided as part of the chapter.

7.1 MOTIVATION

The general question that motivated this project was to investigate the degree of RNA structure conservation over a large number of viral species. While in strains of an individual virus or within viruses of the same *genus*, conserved RNA structure elements have been characterized (e.g. see chapters 5 and 6), it is largely unexplored how much structure conservation can be found between viruses that are only related to each other by higher orders of taxonomy, i.e. for viruses that have a very distant common ancestor.

Since common ancestry in viruses is challenging to define, the study focuses on the mRNA sequences of viral proteins that were grouped together on the basis of orthology, thereby implying a minimum degree of common ancestry. This approach was seeded by the then recent release of the *Viral Ortholog Groups* (VOGs), a comprehensive and high quality data source which provided high quality orthology relations for most viral proteins.

7.2 ADDED VALUE

While the study does not give one conclusive answer on the general relevance of RNA structure in viruses, which did not surprise due to its large scale, it highlights multiple examples of potentially relevant secondary structures across many virus species. It furthermore highlights individual genera or even families of viruses that show unexpected structure conservation. Since the study describes structure conservation on a high level and not on the basis of individual structures, the findings should be regarded more as a broad overview of the virosphere. It is plausible to assume that the presented data can be used to guide future studies that focus on investigation of specific viruses for conserved RNA structures.

To facilitate accessibility and analysis of the presented data, it was made available as a web server in the form of the *RNA Structure in Viruses* (RNASIV) database. This database enables a user to explore the data set in detail and will hopefully help in better understanding the RNA structures found in individual viruses.

7.3 REPRINT OF: CONSERVED SECONDARY STRUCTURES IN VIRAL MRNAS

MICHAEL KIENING, ROMAN OCHSENREITER, HANS-JÖRG HELLINGER, THOMAS RATTEI, IVO L. HOFACKER AND DIMITRIJ FRISHMAN

Viruses **2019**, 11, 401.
https://doi.org/10.3390/v11050401

Contributions:

MICHAEL KIENING Developed and Implemented the methods, Implemented the RNASIV-database, analyzed the data and wrote the manuscript.

ROMAN OCHSENREITER Developed methods and contributed to the manuscript.

HANS-JÖRG HELLINGER Provided the latest release of VOGs for analysis.

THOMAS RATTEI Contributed to the manuscript.

IVO L. HOFACKER Supervised the project and contributed to the manuscript.

DIMITRIJ FRISHMAN Conceived the study, supervised the project and wrote the manuscript.

viruses

MDPI

Article

Conserved Secondary Structures in Viral mRNAs

Michael Kiening [1], Roman Ochsenreiter [2], Hans-Jörg Hellinger [4], Thomas Rattei [4], Ivo Hofacker [2,3] and Dmitrij Frishman [1,5,*]

1 Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Maximus-von-Imhof-Forum 3, D-85354 Freising, Germany; m.kiening@wzw.tum.de
2 Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währingerstr. 29, 1090 Vienna, Austria; romanoch@tbi.univie.ac.at (R.O.); ivo@tbi.univie.ac.at (I.H.)
3 Department of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, Währingerstrasse 17, 1090 Vienna, Austria
4 Division of Computational Systems Biology, Department of Microbiology and Ecosystem Science, University of Vienna, Althanstraße 14, 1090 Vienna, Austria; joerg.hellinger@univie.ac.at (H.-J.H.); thomas.rattei@univie.ac.at (T.R.)
5 St. Petersburg State Polytechnic University, St. Petersburg 195251, Russia
* Correspondence: d.frishman@wzw.tum.de

Received: 30 March 2019; Accepted: 26 April 2019; Published: 29 April 2019

Abstract: RNA secondary structure in untranslated and protein coding regions has been shown to play an important role in regulatory processes and the viral replication cycle. While structures in non-coding regions have been investigated extensively, a thorough overview of the structural repertoire of protein coding mRNAs, especially for viruses, is lacking. Secondary structure prediction of large molecules, such as long mRNAs remains a challenging task, as the contingent of structures a sequence can theoretically fold into grows exponentially with sequence length. We applied a structure prediction pipeline to Viral Orthologous Groups that first identifies the local boundaries of potentially structured regions and subsequently predicts their functional importance. Using this procedure, the orthologous groups were split into structurally homogenous subgroups, which we call subVOGs. This is the first compilation of potentially functional conserved RNA structures in viral coding regions, covering the complete RefSeq viral database. We were able to recover structural elements from previous studies and discovered a variety of novel structured regions. The subVOGs are available through our web resource RNASIV (RNA structure in viruses).

Keywords: mRNA structure; structure database; secondary structure; viral mRNA; subVOG; structurally related; RNA structure; structurally homogenous; structurally related; mRNA families

1. Introduction

Secondary structures formed in single-stranded mRNA molecules through complementary self-interactions, both in the untranslated (UTR) and coding (CDS) regions of mRNAs, have been implicated in a variety of regulatory functions [1]. For example, riboswitches modulate gene expression through conformational changes in response to various stimuli [2]. Translation initiation, elongation, and termination as well as translation efficiency depend on higher order mRNA secondary structures in non-coding regions [3,4]. CDS hairpins have also been suggested to play a role in the regulation of translation [5], in particular by causing ribosomal stalling and modulating translational efficiency [6]. The relationship between mRNA structure in the CDS and gene expression has been demonstrated both computationally and experimentally [7–11]. In particular, reduced mRNA stability near the start codon has been observed in a wide range of species, probably as a mechanism to facilitate ribosome

binding or start codon recognition by initiator-tRNA [12]. Structured elements within CDS directly influence mRNA abundance [13]. Computational studies show that native mRNAs have lower folding energies and are thus more stable than codon-randomized ones [5]. The three mRNA functional domains—5'UTR, CDS, and 3'UTR—form largely independent folding units, while base pairing across domain borders is rare [14]. The ability of viruses to persist in their host in a genus-specific manner is influenced by the interplay between local structural motifs and genome-scale ordered RNA structures (GORS) [15], which impose additional restraints on the RNA sequence space. Evolutionarily conserved local secondary structures have been identified in CDSs [16] and shown to be functional [17]. An indirect indication of the global importance of RNA structures in the coding regions comes from the recent study of Fricke et al. who identified selection favoring specific pairing patterns between synonymous codons within RNA hairpins [18].

Increasing evidence suggests that secondary structural elements in the CDSs of viral RNAs also constitute a previously underappreciated, evolutionarily conserved level of functional organization of viruses. A large number of conserved secondary structural motifs were computationally identified in the Flavivirus genomes [19–21], predicted to restrain sequence variability [22] and experimentally shown to regulate important biological processes, such as replication and infection [21]. Multiple secondary structures were described in the coding regions of the (+) sense RNA of the Influenza A virus [23]. Another example is a secondary structural element within the coding region of the Dengue virus type 2, which is essential for its replication [24]. More recently, using a comparative genomics approach, Goz and Tuller identified a large number of potentially functionally important regions in the coding regions of Dengue viruses, in which the RNA folding strength is conserved independently of sequence conservation and compositional bias [25]. Specific regions in the HIV structural genes were reported to be under strong selection for stable secondary structures [26]. Recent research shows that mechanisms of translational control by RNA structures can be shared between viruses and cellular organisms [27].

Given the important role played by RNA structures in shaping the evolutionary dynamics of viruses and modulating their interaction with the host, a large-scale investigation of RNA motifs in viruses would be warranted. However, there are two major challenges that need to be addressed before embarking on such an investigation. First, accurate structure prediction for long RNA molecules, such as mRNAs, is generally out of reach for the existing computational methods. Second, conserved stem-loop structures can only be derived from a collection of high-quality alignments of orthologous viral transcripts, which are difficult to obtain, given the rapid pace of viral evolution and the ensuing poor sequence conservation, even between closely related species.

Here, we propose a computational approach to explore the RNA structurome of the viral coding regions, in which local structure predictions are applied to VOG (Viral Orthologous Groups, http://vogdb.org), the first comprehensive collection of orthologous groups derived for all viral proteins contained in the RefSeq [28] database. We utilize RNALalifold [29] to scan long input sequences for locally optimal secondary structures. The identified structural boundaries are more accurate than those derived from using a sliding window of fixed length. Functional importance of structured regions is assessed by RNAz [30]. We present a novel database, RNASIV (RNA structure in viruses; http://rnasiv.bio.wzw.tum.de), which contains the largest currently available collection of predicted RNA structures in viruses. It provides access to 201,708 viral mRNA sequences clustered into 42,293 structurally homogenous groups and is intended to become a useful tool for exploring structure–function relationships in virus families.

2. Materials and Methods

2.1. Viral Orthologous Groups (VOGs)

All genome sequences and their annotations were retrieved from the RefSeq viral database release 79 [31] and grouped into phages and non-phages, based on the available taxonomic information.

Assemblies containing inconsistently annotated or completely unannotated polyproteins were identified based on the manually curated information provided by ViralZone [32] and excluded from consideration. Phage and non-phage protein sequences were clustered into phage and non-phage preVOGs, using the NCBI's COG software package with all default settings.

For all phage and non-phage preVOGs, multiple sequence alignments were constructed with Clustal Omega v1.2.4 [33] and used to build HMM-profiles using HMMer 3 [34]. The profiles were subsequently aligned against each other, using HHalign from the HHsuite toolkit [35]. The number of aligned HMM columns was used as an alignment score. All scores for alignments with HHalign probability >85, HHalign *e*-Value < 10^{-5}, and more than 70% of aligned columns between the query and the match HMM were stored as an all-against-all matrix. This matrix was clustered into 21,200 VOGs, using the MCL (Markov Clustering) method [36]. Based on the manual inspection of the homogeneity of the protein function descriptions in the resulting clusters, we selected the inflation value of 2.0 for the MCL clustering. For all VOG member proteins, we determined the closest homolog in the UniProt database [37] from BLAST [38] hits with *E*-values better than 10^{-5} and a minimal query coverage of 90%. Functional descriptions of VOGs were automatically derived based on the most frequent protein description found in the UniProt entries or, if not available, in the RefSeq annotation [31]. The complete VOG dataset, which was used in this study, and supplementary files are available for download at http://vogdb.org.

2.2. Mapping VOG Sequences to Specific Hosts

We used Virus-Host DB [39] to assign host information to VOG proteins. For 20757 VOGs, we were able to map all contained sequences to a specific host, while 428 VOGs contain proteins from at least one viral species for which we could not find host annotation. Most VOGs include viruses infecting hosts from only one domain of life, i.e., bacteria (~72%), eukaryotes (~22%), or archaea (4%), while only 2% of VOGs are taxonomically mixed (Figure 1). Only 12 VOGs contain viruses that infect hosts from all three domains of life. The VOG sizes range from 15 proteins of 12 distinct species, up to 265 proteins belonging to 261 different species (on average, 104 proteins from 95 different species). These VOGs mostly harbor highly conserved core enzymes of double-stranded DNA viruses, such as kinases, ligases, methylases, helicases, hydrolases, and synthases [40]. The other two VOGs additionally contain proteins from viruses belonging to the order of Caudovirales, which belong to the bacteriophages, which are not classified as double-stranded DNA viruses, according to the NCBI taxonomy. We excluded from consideration 15 VOGs containing satellite viruses infecting other viruses.

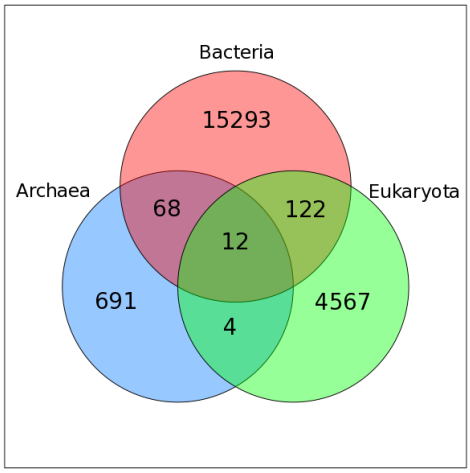


Figure 1. Venn diagram showing the taxonomy of the host organisms within all viral orthologous groups (VOGs). Only those VOGs are included for which host annotation for all viruses is available in the Virus-Host DB.

Viruses **2019**, *11*, 401

2.3. Distance Trees of VOG Proteins

Expectedly, we found that RNA structure conservation within VOGs decreases with increasing VOG size. Most VOGs (66%) consist of at least three sequences (size distribution shown in Figure 2) and can therefore potentially be split into smaller groups containing structures that are not conserved across the entire VOG. We therefore utilized distance trees derived by the neighbor-joining algorithm [41] to identify structurally homogeneous subsets of VOGs (subVOGs). All-against-all pairwise alignments of protein sequences were calculated using Clustal Omega and then converted to the nucleotide alphabet. The distance matrices were derived from pairwise sequence identity values, and the trees were created from the matrices using neighbor joining, as implemented in the BioPerl toolkit [42]. The inner nodes of the sequence trees represent possible subVOG candidates, potentially containing structurally homogenous sequences.

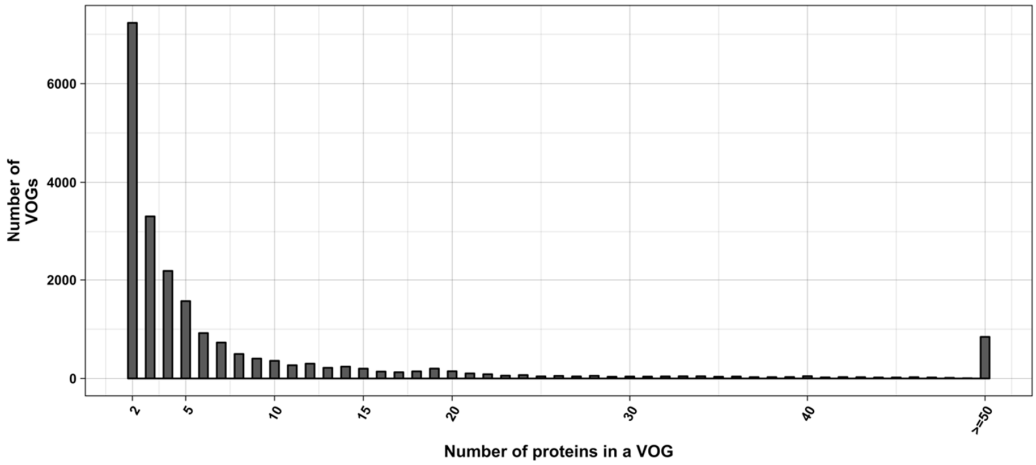


Figure 2. Distribution of VOG sizes.

2.4. Structure Prediction and subVOG Assignment

In order to assess the amount of structural RNA conservation present in subVOG candidates, multiple sequence alignments (MSAs) of proteins were calculated for each inner node of the distance trees and converted to the nucleotide alphabet. The RefSeq nucleotide and protein sequences were obtained from the VOGDB. We then employed RNALalifold from the ViennaRNA package [29], with default parameters, to determine the boundaries of locally stable structures within each MSA, and realigned these local regions using mLocARNA [43]. MLocARNA produces structure-guided multiple sequence alignments, using an adapted version of the Sankoff algorithm. The significance and conservation of the found structures was assessed with RNAz [30]. This procedure is simpler and arguably more accurate than the usual approach of applying RNAz to the entire MSA within a sliding window. RNAz classifies fragments of an MSA pre-selected by RNALalifold as containing or not containing a functional RNA secondary structural element. Realignment with mLocARNA significantly increases the precision of RNAz [30]. As no sequence of a potential subVOG can be regarded as a reference sequence, the option "no reference" was used for the subsequent RNAz analysis. The RNAz method uses the RNAfold algorithm from the ViennaRNA package to calculate secondary structures and the corresponding minimum free energy (MFE) for each individual RNA sequence in the alignment. In addition, for each aligned sequence set, RNAz calculates a consensus secondary structure and its MFE using the RNAalifold algorithm. RNAz assumes that conserved and thermodynamically stable structures are functional, in which case it outputs "RNA". Otherwise, it outputs "OTHER". For this purpose, a class probability value, combining all information on an input alignment is calculated. We used a stringent threshold of 0.9 (default 0.5) for the class probability value, which is recommended for finding high confidence structures [30]. Subsequently, the trees

were scanned for subtrees containing at least one conserved structural element, that is, predicted to be functional, and the largest subtrees were designated as structurally homogenous subVOGs. We found that sequences that are only distantly related according to the neighbor-joining tree may still share conserved RNA structures. In order to account for structure-level relationships between sequences, we built covariance models for all conserved structures found within subVOGs, using the tool cmbuild from the infernal package [44], and used them to search against all sequences in the entire VOG database.

2.5. mRNA Stability

Following Tuller et al. [45] and Faure et al. [46], we employed RNAfold to calculate the folding energy of the most and the least stable 30-nucleotide segment of mRNAs (ΔGmin and ΔGmax, respectively), as well as the average folding energy of all possible 30 nucleotide segments (ΔGmean). Faure et al. investigated the effect of mRNA stability on the translation rate and protein folding. During translation, the ribosome sequentially unfolds parts of the mRNA. These parts are typically 30 nucleotides long, which explains the choice of segment length in Faure et al. As this procedure does not take into account the actual boundaries of local structures, but rather limits all structures to the size of 30 nucleotides, we additionally calculated the three energy values for all local optimal structures found with RNALfold.

2.6. mRNA Structures and Protein Function

We investigated the relationship between protein function, described in terms of gene ontology (GO) annotation [47], and mRNA structures. Instead of using the global folding energy for classifying mRNAs as highly or lowly structured [48], we considered structural coverage—the portion of an mRNA covered by functional and conserved structures. GO terms for all VOG proteins were downloaded using QuickGO [49], where available. Based on the Evidence & Conclusion Ontology (ECO) evidence codes [50], two separate datasets were created: (i) Proteins annotated by manually or experimentally derived GO terms (ECO evidence codes: ECO:0000352, ECO:0000269), and (ii) proteins annotated by GO terms with any evidence codes. To find out whether mRNAs of proteins with certain functions tend to harbor more or fewer structures, we pooled together functionally similar GO terms with the average structural coverage of their corresponding mRNAs, using Revigo [51]. Revigo uses a semantic similarity measure to group similar GO terms together, which results in a concise list of distinct functions. To perform this analysis, we calculated the average structural coverage of all subVOG mRNAs with available GO annotation. For the experimental dataset we allowed a coverage value to be associated with a GO term if more than 50% of the sequences in a particular subVOG were annotated with this term. Within the dataset based on all evidence codes, we only allowed GO terms shared by all sequences of a subVOG. We only used mRNAs that were clustered into a subVOG. For sequences that were not part of any subVOG, we did not find conserved structures, although this does not necessarily mean that the mRNA did not contain functional structures. The distributions of standard deviations of the structural coverage values were compared within the actual and randomly generated Revigo clusters. Randomization was performed 1000 times by preserving the size of the clusters and filling them with randomly chosen GO terms.

3. Results

3.1. Overview of the Study

A graphical overview of the study is given in Figure 3. In a first step, we created distance trees for all protein sequences contained in each VOG, using the neighbor joining method, as described in Materials and Methods. All sequences of the inner nodes of each tree, representing potential subVOGs, were multiply aligned, converted to the nucleotide alphabet and processed with RNALalifold to obtain all potentially conserved local optimal structures. Each part of the alignment covering a potential

structure was then realigned with the structure-guided alignment method mLocARNA and checked for functionality using RNAz. The use of structure-guided alignments as input for RNAz improves the performance, compared to pure sequence-based alignments [30]. The tree nodes containing the most sequences that yielded conserved structures were taken as final subVOGs. For all obtained subVOG structures, we computed covariance models that could be used to search for similar structures in future research.

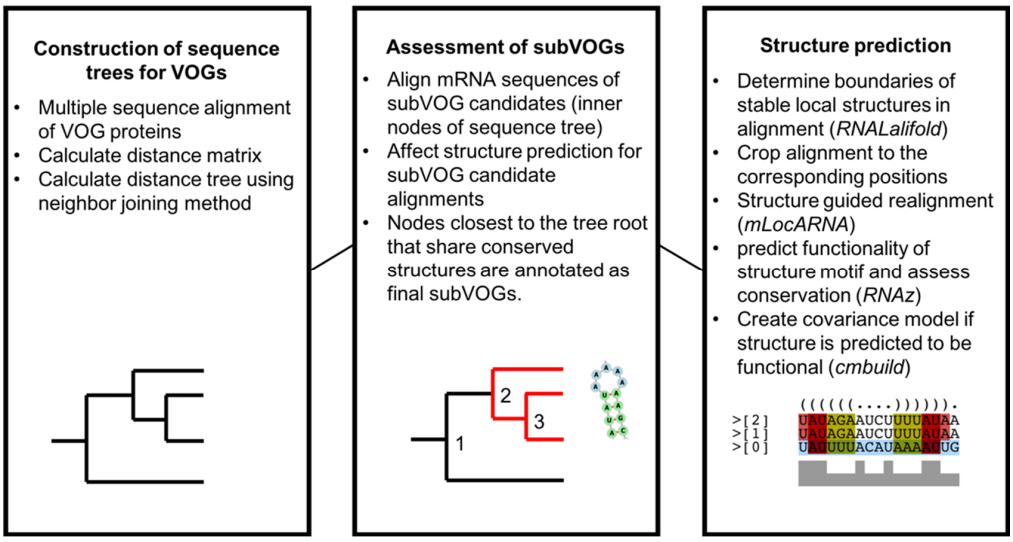


Figure 3. Overview of the analysis of conserved RNA structures in VOGs.

3.2. Structure Conservation in VOGs

The current release of the VOG database, derived from the RefSeq release 77, contains 21,200 VOGs, composed of 251,796 proteins from 6252 phages and eukaryotic viruses (Figure S1). Protein sequences in each VOG were aligned by Clustal Omega, converted to the nucleotide alphabet, and used as input for RNA structure prediction by RNALalifold. As seen in Figure 4, the number of local optimal structures conserved within entire VOGs decreases quickly with the number of aligned sequences, which may in part be the consequence of poor multiple alignment quality in large sets of sequences. Indeed, we found that proteins in smaller VOGs tend to be more closely related (Figure S2). To exclude structures found due to sequence conservation only, the potential functionality of structures was verified with RNAz. However, even those VOGs that only consist of a few sequences do not always contain conserved structures. There are 7232 VOGs with exactly two sequences, and for 1237 of these, we could not find any conserved structures. The remaining 5995 VOGs of size two had an average structural coverage of approximately 25% (Figure 5a). Out of the 13,968 VOGs with more than two sequences, 7238 VOGs were predicted to contain RNA structures conserved across the entire VOG, with an average structural coverage of approximately 18% (Figure 5b). These contain between 3 and 96 sequences, with an average of 6. On average, VOGs contain sequences from three different genera, mostly belonging to the same taxonomic family and thus also to the same order (Figure 6a–c). The 25 most diverse VOGs contain sequences from three different orders and up to 19 taxonomic families. On average, a VOG contains mRNAs from viruses that infect hosts from four different genera, belonging to three different taxonomic families and two orders. The VOG with the highest host diversity corresponds to 209 different host genera from 114 families and 64 orders.

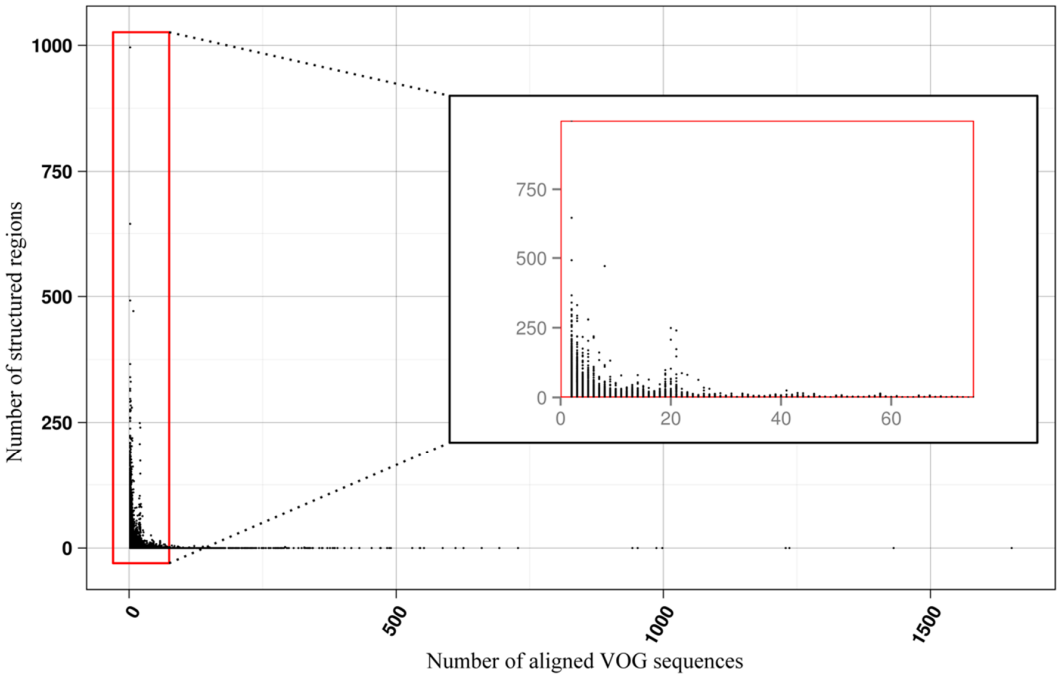


Figure 4. Number of local RNA structures as a function of VOG size.

(a)

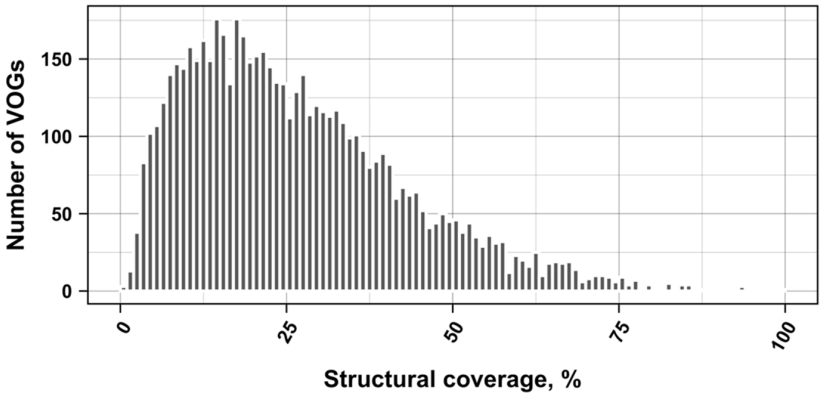
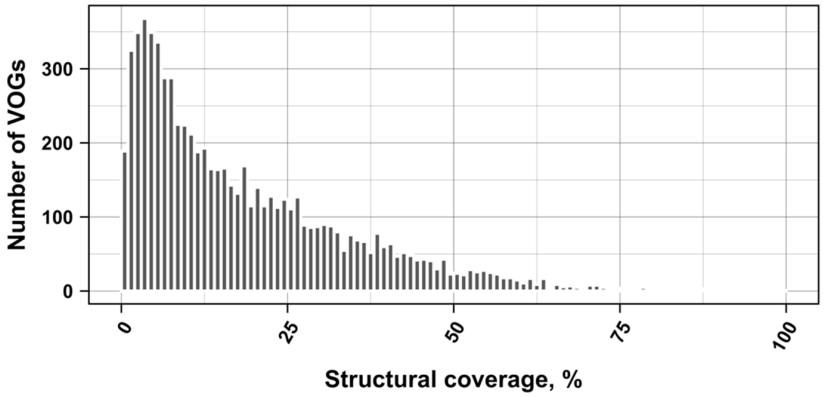


Figure 5. *Cont.*

(b)



(c)

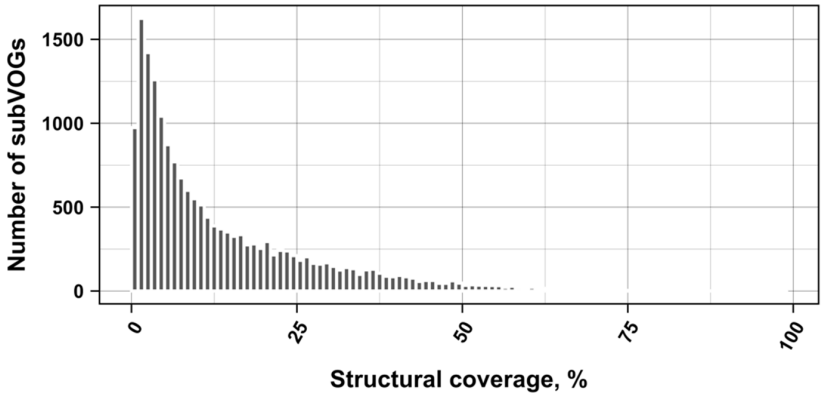


Figure 5. Coverage of VOG alignments by local optimal RNA structures. (**a**) VOGs with two sequences. (**b**) VOGs with more than two sequences, in which structures are conserved across all sequences. (**c**) subVOGs. VOGs that did not contain conserved structures, even after splitting into subVOGs, are not shown.

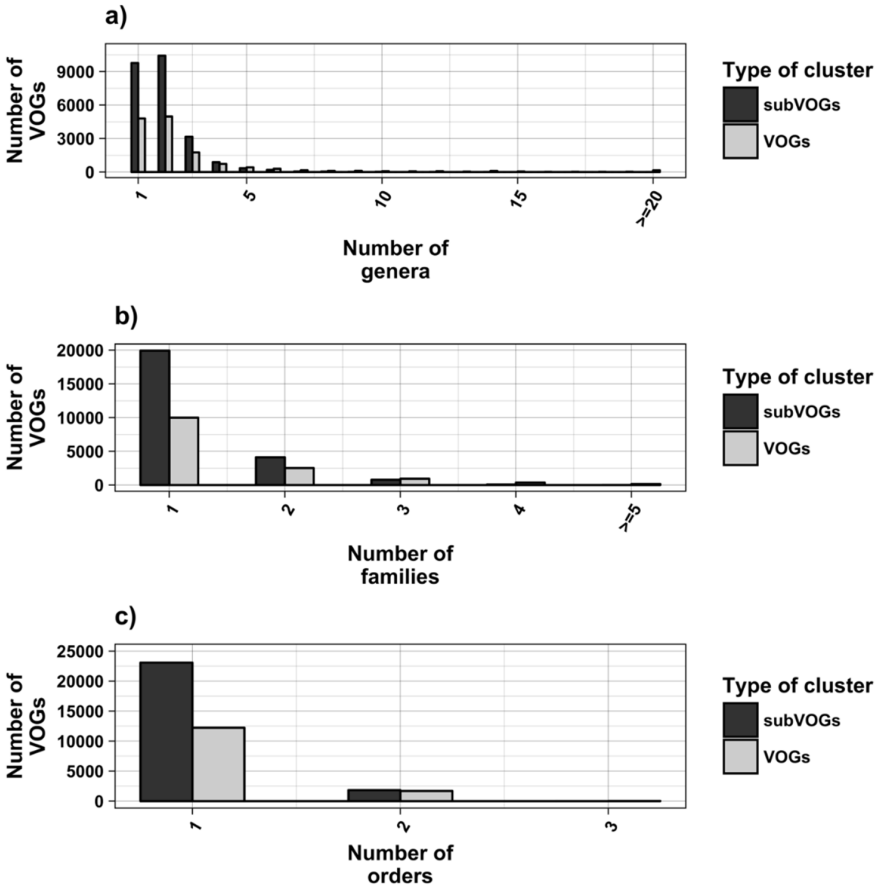


Figure 6. Taxonomic distribution of proteins in VOGs (with more than two sequences) and subVOGs.

3.3. Structure Conservation in subVOGs

We attempted to subdivide 6730 VOGs with more than two sequences and without conserved structures into structurally homogeneous subsets, which we call subVOGs, using phylogenetic trees derived by the neighbor-joining method. This procedure resulted in 17,678 subVOGs with an average structural coverage of approximately 13% (Figure 5c). The average number of genera per subVOG is 2 and the most diverse of them contains sequences from three orders and 14 families. A subVOG contains on average sequences that infect two different host genera, and the most diverse subVOG infects hosts of 42 different genera, belonging to 33 families and 20 different orders (Figure 7a–c). Thus, unsurprisingly, subVOGs, which constitute subsets of full VOGs with increased structural homogeneity, exhibit a reduced taxonomic spread, both of the viruses they contain and their hosts. A large fraction of subVOGs (63%) contains sequences from more than one genus and 21% contain sequences from more than one family. The structural coverage of subVOGs, i.e., the fraction of alignment positions that are located within conserved RNA structures, decreases with increasing taxonomic diversity of the viruses and their hosts (Figure 8). An example that demonstrates the reduction of taxonomic spread between a VOG and its corresponding subVOGs is given in Figure 9. Here, the VOG 00052, which contains 20 proteins from 12 different virus species belonging to 4 different taxonomic families, was split into four structurally homogenous subVOGs. Two of the four subVOGs consist of mRNAs belonging to the genus Avipoxvirus from the family Poxviridae, the third subVOG contains sequences from the family Mimiviridae, and the fourth subVOG consists of two mRNAs belonging to viruses from two different taxonomic families, the Ascoviridae and the Iridoviridae. For two mRNAs, we could not find structures conserved in any of the other VOG members and they are therefore not part of any subVOG.

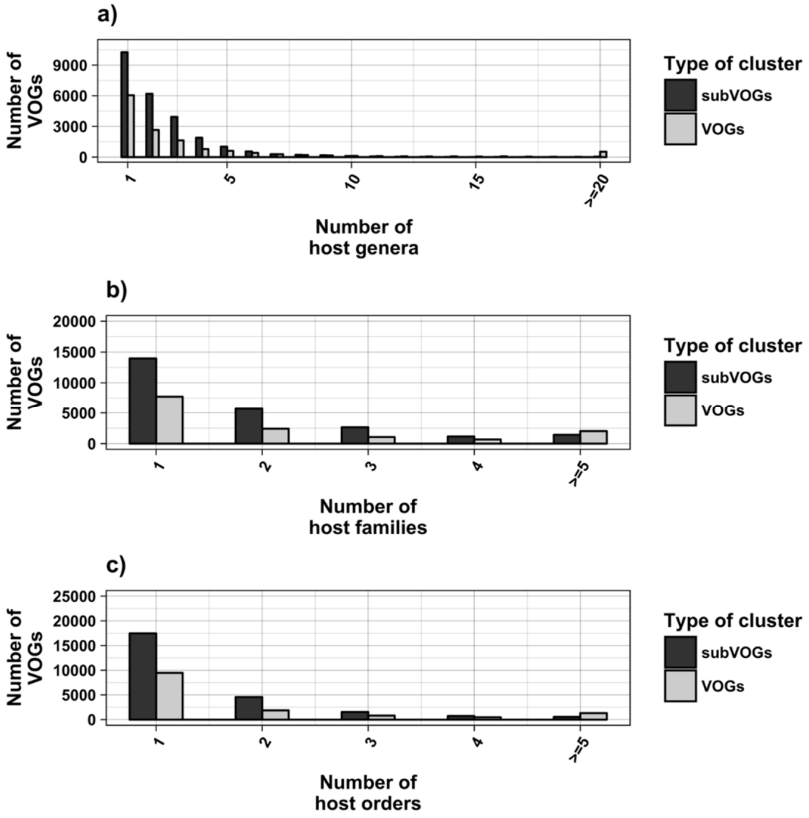


Figure 7. Taxonomic distribution of hosts in VOGs (with more than two sequences) and subVOGs.

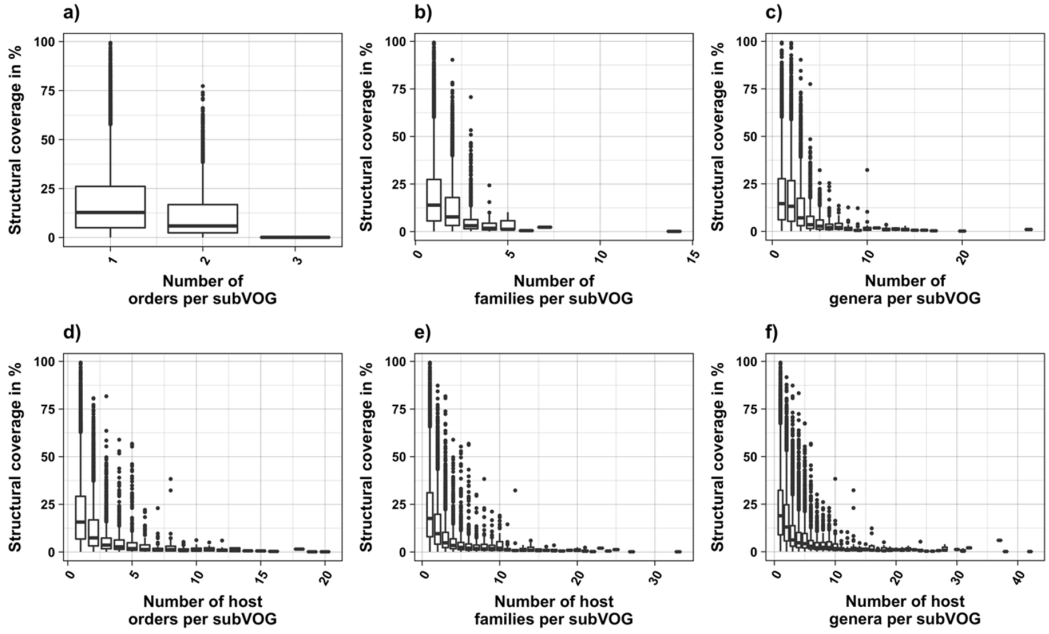


Figure 8. Structural coverage as a function of the taxonomic variety of subVOGs and their host organisms.

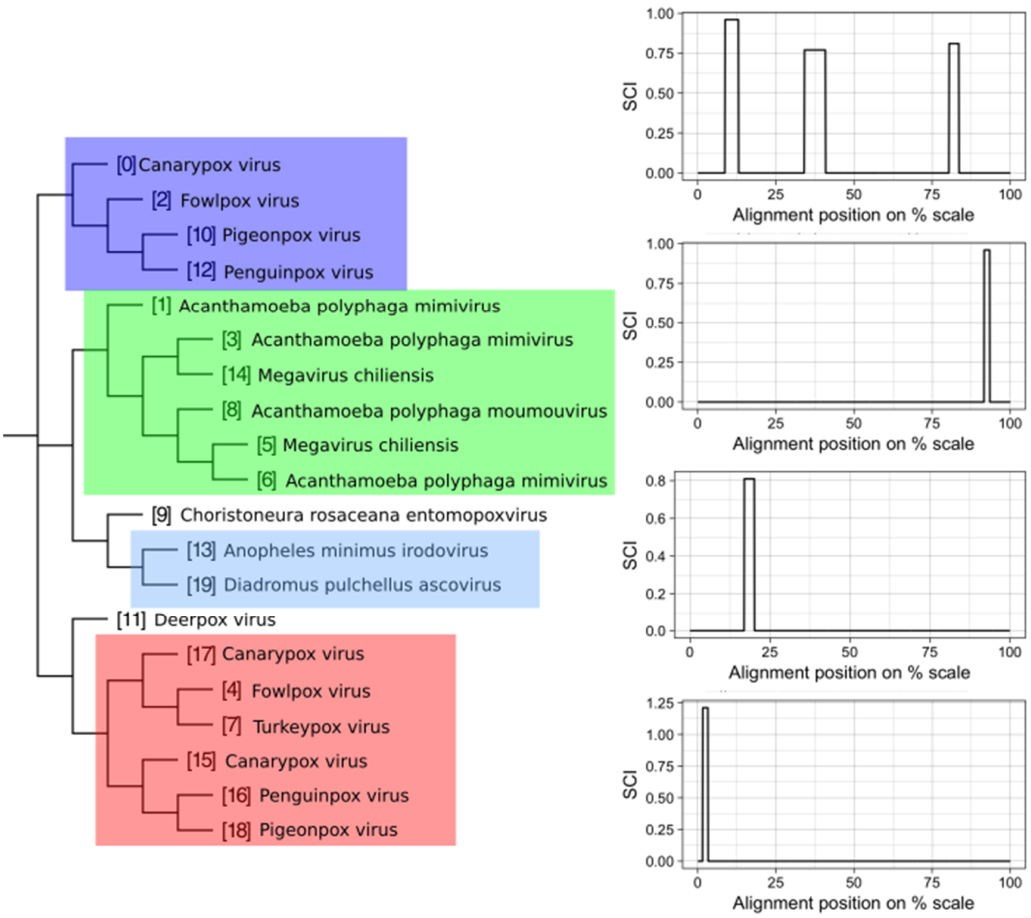


Figure 9. Example of a VOG split into structurally homogenous subVOGs. Shown is the VOG 00052 containing 20 mRNAs, encoding for Kila-N domain proteins, from 12 virus species. On the left, the neighbor-joining tree based on the pairwise sequence identity between the protein sequences is shown. Colored boxes indicate subVOGs, within which conserved structures were predicted. The tree nodes outside colored boxes did not yield any conserved structures. On the right, the structure conservation index (SCI) (black line for each subVOG alignment) is plotted against the alignment position on the percentage scale. Plots are ordered according to the subVOG position in the tree.

As an example, Figure 10 shows the subVOG 1 of VOG11160, which contains two mRNAs encoding the matrix protein 1 from the Influenza A virus (H3N2) and the Influenza B virus. There are three RNA structural motifs described in the literature for the Influenza A mRNA. Nucleotides 105 to 192 form either a multibranch structure, according to Moss et al. [23] and Jiang et al. [52], or a double hairpin structure, proposed by Jiang et al. [52]. Two consecutive stem-loop structures are formed from position 682 to 744, according to Moss et al. [23]. Despite the sequences' dissimilarity between Influenza A and B, both motifs are partly conserved, according to our RNAz analysis of the corresponding subVOG (Figure 10). Our analysis supports the second hairpin loop from the double hairpin structure, described by Jiang et al. (Figure 10a–c). From the second motif, proposed by Moss et al., we also found that the second hairpin structure was partly conserved (Figure 10d–e). The consensus structure of the first motif has a high structure conservation index (SCI) of 0.78, although the part of the alignment covering the structure has a low pairwise identity of 29%. The second motif has an SCI of 0.58 and a pairwise identity of 32%. Our analysis also revealed three further conserved stem-loop structures—position 346 to 369, 438 to 483, and 654 to 674, with SCIs and mPIDs of 0.81 and 29%, 0.66 and 48%, and 0.65 and 33%, respectively.

A recent study of secondary structures in alphaviruses by Kutchko et al. revealed that Sindbis virus mRNAs harbor many functional structures, but they are poorly conserved in the closely related Venezuelan equine encephalitis virus [53]. The corresponding subVOG containing mRNAs coding for the non-structural protein 1 includes orthologous mRNAs from 12 further alphaviruses. We identified three short structures that are conserved in all of the contained species and overlap with the functional structures described by Kutchko et al., while all other structures reported by Kutchko et al. are indeed poorly conserved in further Alphavirus species.

An example of a subVOG in which structures are conserved across mRNAs from different taxonomic families is given in Figure 11. Shown is a subVOG containing proteins from two mosaic viruses (Maracuja mosaic virus, Tobacco mosaic virus), the Bell pepper mottle virus, and the Odontoglossum ringspot virus (Figure 11a,b). The proteins are classified as replicases and RNA polymerases. The subVOG contains overall 15 locally conserved structured regions. Figure 11 shows the region covering alignment positions 4766 to 4815. The alignment covering this structure has an mPID of 72% and the structures are conserved with an SCI of 0.9.

Overall, we subdivided 21,200 VOGs containing, on average, 11 proteins (233,380 in total) into a total of 42,293 subVOGs, containing, on average, five mRNAs (201,708 in total) and three structured regions (147,087 in total). The VOGs with more than two sequences that had to be split up contain, on average, four subVOGs.

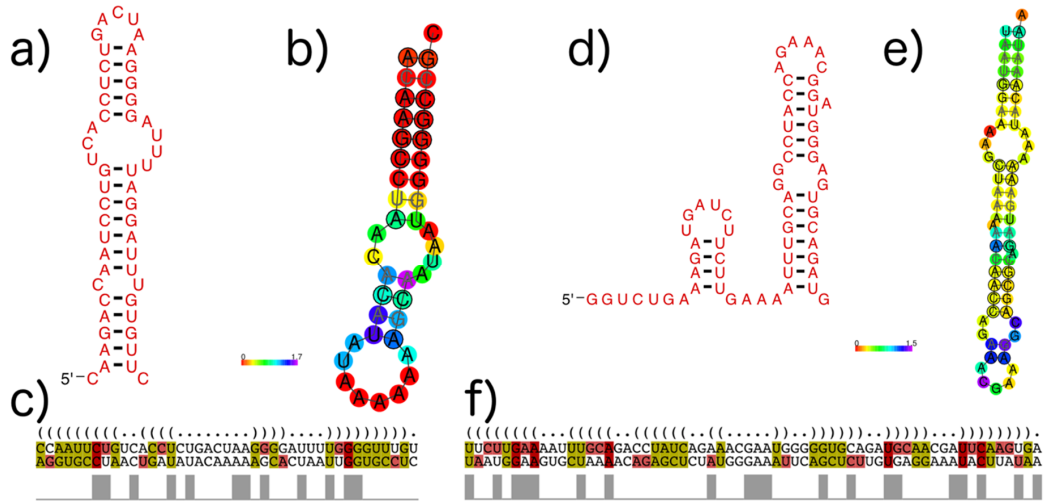


Figure 10. Structures found in Influenza A and B mRNAs encoding the matrix protein (VOG11160). Colors in MSA pictures encode compensatory mutations supporting the consensus structure. Red marks pairs with no sequence variation; ochre, green, turquoise, blue, and violet mark pairs with 2, 3, 4, 5, and 6 different types of pairs, respectively. (**a**) The second of the two consecutive stem loops of the structure proposed by Jiang et al. [52], covering positions 147–192, visualized with R2R [54]; (**b**) The predicted conserved consensus structure for nucleotides 148–188 supports the second hairpin loop of the model of Jiang et al., shown in (**a**). Colors encode the positional entropy; (**c**) Structure-guided alignment and dot bracket structure notation for the consensus structure shown in (**a**). The upper sequence corresponds to Influenza A and the lower sequence to Influenza B; (**d**) Shown are two consecutive hairpin loops for nucleotide positions 682 to 744, proposed by Moss et al. [23], visualized with R2R; (**e**) The predicted conserved structure for nucleotides 697–758 partly supports the model shown in (**e**). Colors encode the positional entropy; (**f**) Structure-guided alignment and dot bracket notation for the consensus structure shown in (**e**). The upper sequence corresponds to Influenza A and the lower sequence to Influenza B.

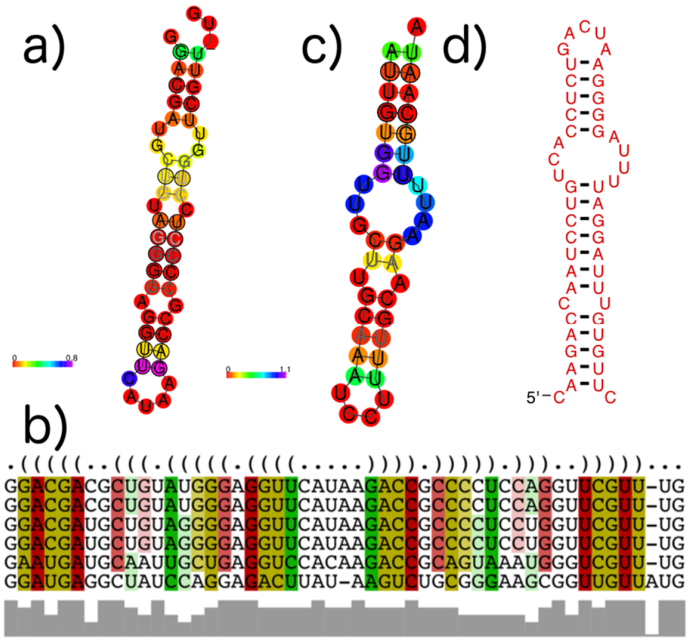


Figure 11. Example structures that were identified within subVOGs. (**a**) Structural annotation of the subVOG 30, belonging to VOG00029, which contains six mRNAs encoding a replicase protein of different Tobamovirus species. Consensus structure visualized by RNAalifold. Colors encode the positional entropy; (**b**) Structure-guided MSA and consensus structure in dot bracket notation corresponding to consensus structure shown in (**a**). Colors encode compensatory mutations supporting the consensus structure. Red marks pairs with no sequence variation; ochre, green, turquoise, blue, and violet mark pairs with 2, 3, 4, 5, and 6 different types of pairs, respectively; (**c**) Consensus structure of subVOG 64 from VOG00003, which contains four mRNAs coding for a p28-like protein of different alphabaculoviruses; (**d**) Structure found in a Heliothis virescens ascovirus 3e, by covariance model search of the structure shown in (**c**), using cmsearch in the entire sequence space of all VOGs.

3.4. subVOG Covariance Models

We built covariance models for all structures found within subVOGs and, using cmsearch, found that in many cases, structures are conserved between different subVOGs and even between different VOGs. In most cases, this was due to a shared sequence domain. For example, the subVOG 64 from VOG00003 harbors four mRNA sequences from different nucleopolyhedroviruses, belonging to the family Baculoviridae. This subVOG was predicted to contain four conserved structures. One of these structures is a highly conserved stem-loop structure (Figure 11c). This structure can also be found in an mRNA of Heliothis virescens ascovirus 3e, belonging to the family Ascoviridae, which is part of VOG01276 (Figure 11d). The two structures are highly conserved with an SCI close to 1, although they are part of different VOGs and belong to mRNAs of different virus families. The alignment of the corresponding proteins revealed that these sequences share a common domain, but the sequence similarity is below the inclusion threshold of the VOG pipeline (Figure S3).

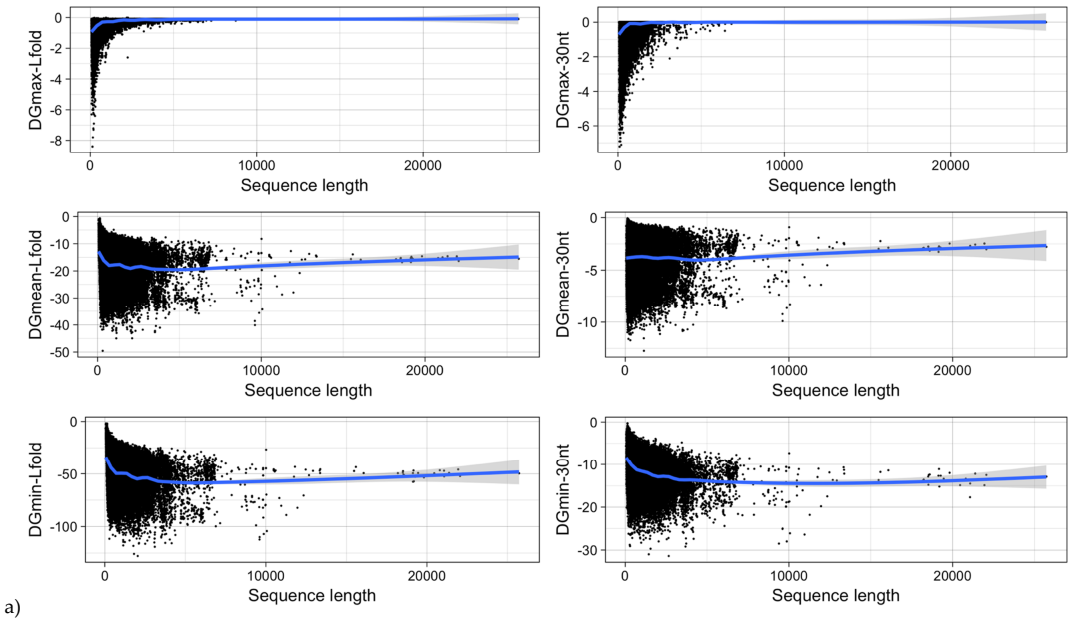
3.5. mRNA Stability and Length

It was shown for a number of eukaryotic and prokaryotic organisms that longer mRNAs exhibit more stable RNA structures, which allows for more efficient control of co-translational protein folding [45,46]. In our dataset of viral mRNA sequences, we also found a correlation between the free energy of the most stable 30-nucleotide segment of an mRNA (ΔGmin) and mRNA length (Pearson correlation coefficient −0.27; from here on referred to as Pearson's *r*), but no correlation between the average energy of all possible 30-nucleotide windows (ΔGmean) and mRNA length (Table 1,

Figure 12a). We additionally calculated the free energy of the most and least stable local optimal segment found by RNALalifold as well as the mean energy of all found RNALalifold segments, and obtained Pearson's r values of −0.25, −0.07, and 0.29 respectively. The Pearson's r of folding energy and GC content lies between −0.5 for ΔGmax and −0.94 for ΔGmean (Table 1, Figure 12b). The number of bases that are within functional structures is positively correlated with the alignment length of subVOGS (Pearson's r 0.40, p-value < 2.2^{-16}), while this correlation becomes negative when considering the percentage of bases within structures (structural coverage) instead of the absolute value (Pearson's r −0.27, p-value < 2.2^{-16}) (Figure 13). In other words, longer mRNAs harbor more or longer structured regions, but at the same time, the percentage of positions in functional structures decreases with increasing length. The only explanation for this effect that we can think of is that there is a certain number of structured elements needed for regulatory functions, which is largely independent of the mRNA length. As expected (see Figure 8), there is a weak but significant negative correlation (Pearson's r −0.23, p-value < 2.2^{-16}) between structural coverage and the number of sequences in the MSA, with more taxonomically diverse alignments containing fewer conserved structures.

Table 1. Pearson correlation between alignment length or GC-content and the minimum (ΔGmin), maximum (ΔGmax), or mean (ΔGmean) folding energy of either all possible 30-nucleotide long-sequence windows or all local optimal structures found with RNALfold, of all mRNAs in our data set. *P*-values are given in parentheses.

Type of ΔG	**Pearson Correlation Coefficient**	
	ΔG vs. Sequence Length	**ΔG vs. GC-Content**
ΔGmin	−0.27 (<2.2^{-16})	−0.73 (<2.2^{-16})
ΔGmean	0.004 (0.1655)	−0.94 (<2.2^{-16})
ΔGmax	0.17 (<2.2^{-16})	−0.50 (<2.2^{-16})
ΔGmin (RNALfold)	−0.24 (<2.2^{-16})	−0.86 (<2.2^{-16})
ΔGmean (RNALfold)	−0.16 (<2.2^{-16})	−0.86 (<2.2^{-16})
ΔGmax (RNALfold)	0.29 (<2.2^{-16})	−0.07 (<2.2^{-16})



a)

Figure 12. *Cont.*

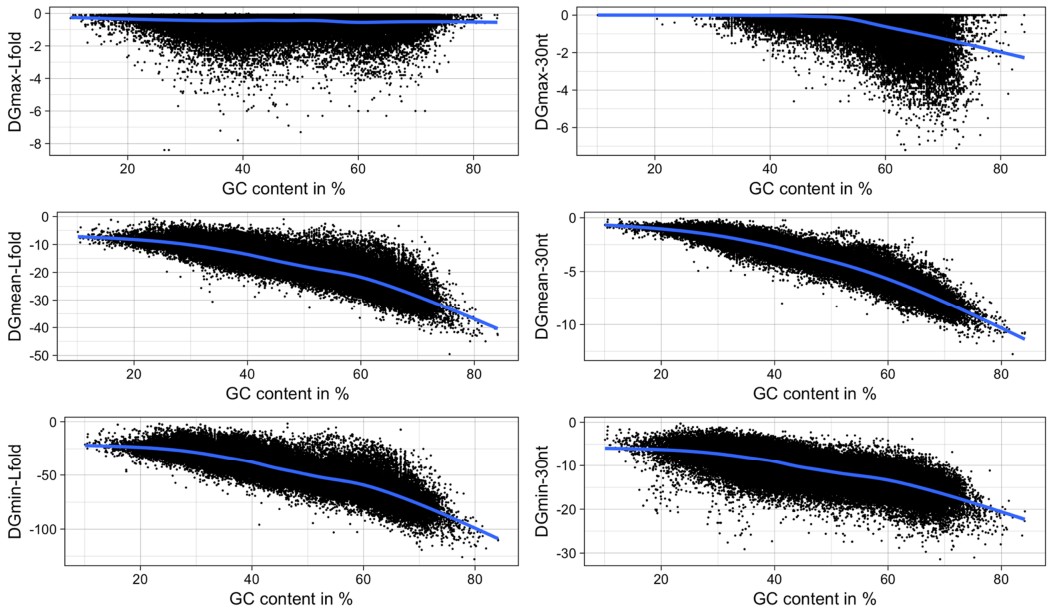


Figure 12. MRNA folding energy as a function of (**a**) sequence length and (**b**) GC-content. DGmin: Minimum folding energy of either all possible 30-nucleotide windows of a sequence or all found local optimal structures using RNALfold. DGmean and DGmax: Mean and maximum of all windows, respectively.

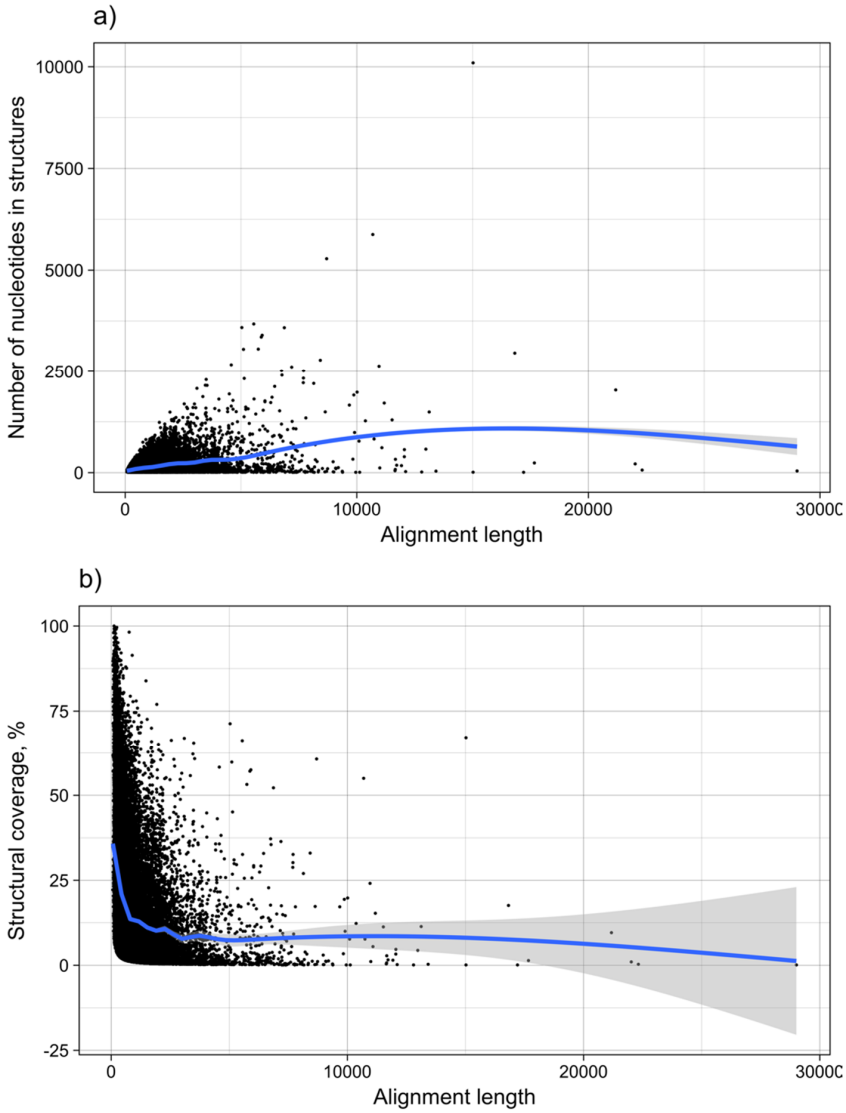


Figure 13. MRNA structure as a function of length. The graph shows the dependence of (**a**) the number of nucleotides within structures predicted to be functional, and (**b**) the structural coverage of the mRNAs in %, from the total length of mRNAs. Each point corresponds to one subVOG.

3.6. mRNA Structures and Protein Function

We analyzed the relationship between protein function and mRNA structure in viral subVOGs by comparing RNA structural coverage with gene ontology (GO) annotation. Using the QuickGO database, we identified a total of 814 VOG proteins that are manually or experimentally annotated (according to ECO evidence codes, as described in Materials and Methods) with GO terms, of which 727 are part of a subVOG, and thus harbor conserved structures according to our analysis. (For the sake of completeness, we also performed the same analysis for all GO annotated proteins, without regard for the annotation evidence codes, see Table S2). For each individual GO term, we only considered the structural coverage of mRNA sequences if that term was assigned to more than 50% of the proteins in a given subVOG. This resulted in 106 GO terms from the biological process sub-ontology and 17 terms from the molecular function sub-ontology. Note that no GO terms from the cellular component sub-ontology satisfied the criteria explained above.

Using Revigo, we derived 70 functionally similar groups of GO terms, with 57 belonging to the biological process ontology and 13 to the function sub-ontology (Table S1). The resulting GO term groups were subdivided into three categories, according to the average structural coverage of the corresponding subVOGs: Low structural coverage (up to 10%), medium structural coverage (up to 20%), and high structural coverage (more than 20%). We found that the standard deviation of the structural coverage values within the Revigo clusters was significantly smaller (Wilcoxon test p-value 1.068^{-10}), compared to randomized clusters (Figure 14). In other words, our findings suggest that mRNAs encoding the proteins with coherent functions tend to exhibit a similar structural coverage.

These findings are in line with the previous study by Vandivier et al. who found that transcripts in *Arabidopsis thaliana* with similar levels of secondary structure in their untranslated and coding regions tend to encode functionally similar proteins [48]. Likewise, Wang et al. also identified GO terms associated with highly or lowly folded mRNAs in yeast [55]. Four of the GO terms associated with highly structured mRNAs, according to Wang et al. (regulation of translation, posttranscriptional regulation of gene expression, regulation of cellular protein metabolic process, and cellular nitrogen compound biosynthetic process), correspond to highly structured viral mRNAs in our data. At the same time, none of the GO terms corresponding to lowly structured yeast mRNAs according to Wang et al. were enriched in our results. On the other hand, Fan Li et al. found that *Arabidopsis thaliana* mRNAs related to "regulation of transcription" were structurally unstable [56], while we found that mRNAs encoding the proteins related to "viral transcription" do harbor conserved RNA structures. We also found virus-specific trends not previously observed for cellular proteins, such as the high structure of viral mRNAs coding for proteins that regulate replication and transcription, suppression by viruses of host translation, or modulation by viruses of host process (Table S1). It has been reported that mRNA folding strength influences the efficiency of gene expression and that mRNAs encoding abundant proteins generally tend to be more structured [57]. In the future, once RNA-seq data for a sufficient number of viral genes becomes available, it will be interesting to investigate whether functional coherence between mRNAs with similar structural coverage is, at least in part, caused by similar expression levels.

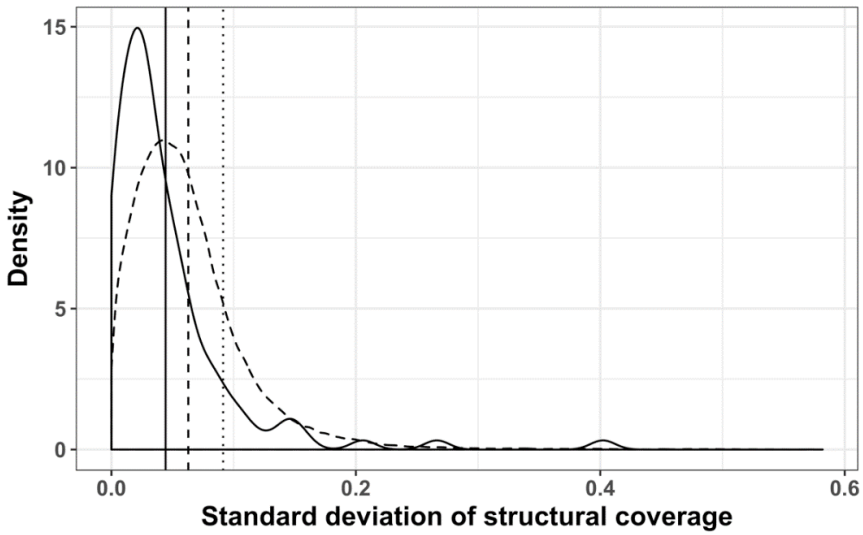


Figure 14. Distribution of standard deviations of mRNA structural coverage, mapped to GO-terms: Clustered with Revigo (solid line); randomized Revigo clusters (dashed line); not clustered (dotted line); vertical lines represent the mean of the corresponding dataset.

3.7. subVOG Online Resource

Structurally homogenous subVOGs can be accessed online (http://rnasiv.bio.wzw.tum.de) through two entry points: "Browse by VOG" and "Browse by taxonomy". The first option is a list of all VOGs, together with the consensus description of their constituent proteins. The list can be filtered with a keyword search and links to the corresponding subVOGs of each VOG are provided. The second option is an expandable taxonomic tree, based on the NCBI taxonomy [58], which allows navigation to the viral species of interest. For each species, mRNA sequences are provided, if available, interlinked to the corresponding subVOGs. Tree nodes containing only mRNAs that are not part of any subVOG are colored grey. Each subVOG contains at least two sequences that share at least one structural element predicted to be functional. If a species of interest is not contained in the subVOG database, the taxonomy tree makes it possible to find the taxonomically closest species. Web pages describing individual subVOGs contain four parts:

(i) General information, i.e., number of mRNAs in the subVOG, the number of proteins and species in the parent VOG, as well as a consensus functional description;
(ii) Information on conserved structures among the subVOG sequences. A plot outlining the SCI for each column of the subVOG MSA gives a brief overview over the structure of the subVOG members. Also provided is a table that shows a list of all structures found, including the corresponding values of SCI, mPID, and the GC content. The consensus structure can also be visualized by Forna, and a covariance model is provided, which can be used to search for similar structures. Additionally, the RNAz results for each individual structured region can be accessed, including structure visualization, dot plots, and the local structure-guided alignments;
(iii) The global MSA for the subVOG sequences. Alignment columns colored in blue correspond to the structured regions described in the previous section. The alignment is visualized with the javascript library MSAviewer [59], which is based on Jalview [60];
(iv) The list of subVOG members, including protein names, descriptions, and taxonomy. For each protein, a link to the RefSEQ entry is provided, as well as the amino acid and nucleotide sequences. The leftmost column of the list contains a checkbox for each subVOG member, which can be used to build a subset of members and analyze the RNA structures shared by these.

4. Discussion

In this work we set out to create a possibly complete census of conserved RNA secondary structures in the coding regions of viruses and to shed light on their biological role. Using sequence comparison and structure prediction methods, we derived structurally homogenous groups of viral mRNAs from subsets of viral orthologous groups (VOGs), which we call subVOGs. We identified a total of 147,087 conserved structures in 42,293 subVOGs, which we make accessible through our database RNASIV (RNA Structures in Viruses). On average, subVOGs contain three structured regions and their structural homogeneity decreases with increasing taxonomic diversity of the viruses and their hosts. We found that 63% of all subVOGs contain mRNAs from at least two genera and 21% from more than one taxonomic family. In line with the previous studies on cellular organisms, we confirm that, in viruses, longer mRNAs tend to contain more stable structures. However, the number of structures grows only slowly with length, which implies that there is a certain minimum amount of structures required to maintain regulatory functions and control protein folding. MRNAs annotated with similar GO terms tend to have a similar structural coverage, hinting at possible commonalities in the regulatory mechanisms of functionally related proteins. It is hoped that RNASIV will be a useful resource for exploring the structure–function relationships in viral mRNAs.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/5/401/s1: Figure S1: Virus lineages included in the VOGs; Figure S2: Mean pairwise sequence identity of VOG alignments as a function of VOG size; Figure S3: Sequence Alignment of a protein from Heliothis virescens ascovirus 3e and proteins belonging to the mRNAs of subVOG 64 of VOG00003; Table S1: Clustering of GO terms of subVOG

proteins and the average structural coverage of their corresponding mRNAs; Table S2: Clustering of GO terms of subVOG proteins and the average structural coverage of their corresponding mRNAs (regardless of GO evidence codes).

Author Contributions: Conceptualization, D.F., I.H., and T.R.; methodology, D.F., R.O., and M.K.; software, M.K.; validation, D.F. and M.K.; formal analysis, D.F. and M.K.; investigation, D.F., M.K., and T.R.; resources, I.H. and T.R.; data curation, M.K. and H.-J.H.; writing—original draft preparation, D.F. and M.K.; writing—review and editing, D.F. and M.K.; visualization, D.F. and M.K.; supervision, D.F.; project administration, D.F., I.H., and T.R.; funding acquisition, D.F., I.H., and T.R.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft, grant number FR1411/10-1 and the FWF-Grant I-1303.

Acknowledgments: We are grateful to Jan Zaucha for his helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bevilacqua, P.C.; Blose, J.M. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annu. Rev. Phys. Chem.* **2008**, *59*, 79–103. [CrossRef] [PubMed]
2. Serganov, A.; Patel, D.J. Ribozymes, riboswitches and beyond: Regulation of gene expression without proteins. *Nat. Rev. Genet.* **2007**, *8*, 776–790. [CrossRef] [PubMed]
3. Gray, N.K.; Hentze, M.W. Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.* **1994**, *19*, 195–200. [CrossRef] [PubMed]
4. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **2005**, *361*, 13–37. [CrossRef]
5. Katz, L.; Burge, C.B. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.* **2003**, *13*, 2042–2051. [CrossRef]
6. Mortimer, S.A.; Kidwell, M.A.; Doudna, J.A. Insights into RNA structure and function from genome-wide studies. *Nat. Rev. Genet.* **2014**, *15*, 469–479. [CrossRef]
7. Kudla, G.; Murray, A.W.; Tollervey, D.; Plotkin, J.B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **2009**, *324*, 255–258. [CrossRef]
8. Duan, J.; Wainwright, M.S.; Comeron, J.M.; Saitou, N.; Sanders, A.R.; Gelernter, J.; Gejman, P.V. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.* **2003**, *12*, 205–216. [CrossRef]
9. Ilyinskii, P.O.; Schmidt, T.; Lukashev, D.; Meriin, A.B.; Thoidis, G.; Frishman, D.; Shneider, A.M. Importance of mRNA secondary structural elements for the expression of influenza virus genes. *OMICS* **2009**, *13*, 421–430. [CrossRef]
10. Carlini, D.B.; Chen, Y.; Stephan, W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics* **2001**, *159*, 623–633. [PubMed]
11. Nackley, A.G.; Shabalina, S.A.; Tchivileva, I.E.; Satterfield, K.; Korchynskyi, O.; Makarov, S.S.; Maixner, W.; Diatchenko, L. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* **2006**, *314*, 1930–1933. [CrossRef]
12. Gu, W.; Zhou, T.; Wilke, C.O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.* **2010**, *6*, e1000664. [CrossRef]
13. Del Campo, C.; Bartholomäus, A.; Fedyunin, I.; Ignatova, Z. Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genet.* **2015**, *11*, e1005613. [CrossRef]
14. Shabalina, S.A.; Ogurtsov, A.Y.; Spiridonov, N.A. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* **2006**, *34*, 2428–2437. [CrossRef]
15. Simmonds, P.; Tuplin, A.; Evans, D.J. Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA* **2004**, *10*, 1337–1351. [CrossRef]
16. Meyer, I.M.; Miklós, I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.* **2005**, *33*, 6338–6348. [CrossRef]

17. Olivier, C.; Poirier, G.; Gendron, P.; Boisgontier, A.; Major, F.; Chartrand, P. Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol.* **2005**, *25*, 4752–4766. [CrossRef]
18. Fricke, M.; Gerst, R.; Ibrahim, B.; Niepmann, M.; Marz, M. Global importance of RNA secondary structures in protein-coding sequences. *Bioinformatics* **2019**, *35*, 579–583. [CrossRef]
19. Thurner, C.; Witwer, C.; Hofacker, I.L.; Stadler, P.F. Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.* **2004**, *85*, 1113–1124. [CrossRef]
20. Fricke, M.; Dünnes, N.; Zayas, M.; Bartenschlager, R.; Niepmann, M.; Marz, M. Conserved RNA secondary structures and long-range interactions in hepatitis C viruses. *RNA* **2015**, *21*, 1219–1232. [CrossRef]
21. Pirakitikulr, N.; Kohlway, A.; Lindenbach, B.D.; Pyle, A.M. The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. *Mol. Cell* **2016**, *62*, 111–120. [CrossRef]
22. Simmonds, P.; Smith, D.B. Structural constraints on RNA virus evolution. *J. Virol.* **1999**, *73*, 5787–5794.
23. Moss, W.N.; Priore, S.F.; Turner, D.H. Identification of potential conserved RNA secondary structure throughout influenza A coding regions. *RNA* **2011**, *17*, 991–1011. [CrossRef]
24. Clyde, K.; Harris, E. RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. *J. Virol.* **2006**, *80*, 2170–2182. [CrossRef]
25. Goz, E.; Tuller, T. Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes. *BMC Genomics.* **2015**, *16* Suppl. 10, S4. [CrossRef]
26. Goz, E.; Tuller, T. Evidence of a Direct Evolutionary Selection for Strong Folding and Mutational Robustness Within HIV Coding Regions. *J. Comput. Biol.* **2016**, *23*, 641–650. [CrossRef]
27. Díez, J.; Jungfleisch, J. Translation control: Learning from viruses, again. *RNA Biol.* **2017**, *14*, 835–837. [CrossRef]
28. Pruitt, K.D.; Tatusova, T.; Brown, G.R.; Maglott, D.R. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **2012**, *40*, D130–D135. [CrossRef]
29. Lorenz, R.; Bernhart, S.H.; Höner Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26. [CrossRef]
30. Gruber, A.R.; Findeiß, S.; Washietl, S.; Hofacker, I.L.; Stadler, P.F. RNAz 2.0: Improved noncoding RNA detection. *Pac. Symp Biocomput* **2010**, 69–79. [CrossRef]
31. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
32. Hulo, C.; de Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: A knowledge resource to understand virus diversity. *Nucleic Acids Res.* **2011**, *39*, D576–D582. [CrossRef]
33. Sievers, F.; Higgins, D.G. Clustal omega. *Curr. Protoc. Bioinformatics* **2014**, *48*, 3.13.1–3.13.16.
34. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **2009**, *23*, 205–211.
35. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2011**, *9*, 173–175. [CrossRef]
36. Enright, A.J.; Van Dongen, S.; Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584. [CrossRef]
37. UniProt Consortium UniProt: A hub for protein information. *Nucleic Acids Res.* **2015**, *43*, D204–D212. [CrossRef]
38. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
39. Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H. Linking Virus Genomes with Host Taxonomy. *Viruses* **2016**, *8*, 66. [CrossRef]
40. Kazlauskas, D.; Krupovic, M.; Venclovas, Č. The logic of DNA replication in double-stranded DNA viruses: Insights from global analysis of viral genomes. *Nucleic Acids Res.* **2016**, *44*, 4551–4564. [CrossRef]
41. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
42. Stajich, J.E.; Block, D.; Boulez, K.; Brenner, S.E.; Chervitz, S.A.; Dagdigian, C.; Fuellen, G.; Gilbert, J.G.R.; Korf, I.; Lapp, H.; et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **2002**, *12*, 1611–1618. [CrossRef]

43. Will, S.; Joshi, T.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* **2012**, *18*, 900–914. [CrossRef]
44. Cui, X.; Lu, Z.; Wang, S.; Jing-Yan Wang, J.; Gao, X. CMsearch: Simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics* **2016**, *32*, i332–i340. [CrossRef]
45. Tuller, T.; Veksler-Lublinsky, I.; Gazit, N.; Kupiec, M.; Ruppin, E.; Ziv-Ukelson, M. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* **2011**, *12*, R110. [CrossRef]
46. Faure, G.; Ogurtsov, A.Y.; Shabalina, S.A.; Koonin, E.V. Role of mRNA structure in the control of protein folding. *Nucleic Acids Res.* **2016**, *44*, 10898–10911. [CrossRef]
47. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
48. Vandivier, L.; Li, F.; Zheng, Q.; Willmann, M.; Chen, Y.; Gregory, B. Arabidopsis mRNA secondary structure correlates with protein function and domains. *Plant. Signal. Behav.* **2013**, *8*, e24301. [CrossRef]
49. Binns, D.; Dimmer, E.; Huntley, R.; Barrell, D.; O'Donovan, C.; Apweiler, R. QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics* **2009**, *25*, 3045–3046. [CrossRef]
50. Giglio, M.; Tauber, R.; Nadendla, S.; Munro, J.; Olley, D.; Ball, S.; Mitraka, E.; Schriml, L.M.; Gaudet, P.; Hobbs, E.T.; et al. ECO, the Evidence & Conclusion Ontology: Community standard for evidence information. *Nucleic Acids Res.* **2019**, *47*, D1186–D1194.
51. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **2011**, *6*, e21800. [CrossRef] [PubMed]
52. Jiang, T.; Nogales, A.; Baker, S.F.; Martinez-Sobrido, L.; Turner, D.H. Mutations Designed by Ensemble Defect to Misfold Conserved RNA Structures of Influenza A Segments 7 and 8 Affect Splicing and Attenuate Viral Replication in Cell Culture. *PLoS ONE* **2016**, *11*, e0156906. [CrossRef]
53. Kutchko, K.M.; Madden, E.A.; Morrison, C.; Plante, K.S.; Sanders, W.; Vincent, H.A.; Cruz Cisneros, M.C.; Long, K.M.; Moorman, N.J.; Heise, M.T.; et al. Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Res.* **2018**, *46*, 3657–3670. [CrossRef]
54. Weinberg, Z.; Breaker, R.R. R2R–software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics* **2011**, *12*, 3. [CrossRef] [PubMed]
55. Wang, X.; Li, P.; Gutenkunst, R.N. Systematic Effects Of mRNA Secondary Structure On Gene Expression And Molecular Function In Budding Yeast. *BioRxiv* **2017**. [CrossRef]
56. Li, F.; Zheng, Q.; Vandivier, L.E.; Willmann, M.R.; Chen, Y.; Gregory, B.D. Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant. Cell* **2012**, *24*, 4346–4359. [CrossRef]
57. Zur, H.; Tuller, T. Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*. *EMBO Rep.* **2012**, *13*, 272–277. [CrossRef]
58. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **2012**, *40*, D136–D143. [CrossRef]
59. Yachdav, G.; Wilzbach, S.; Rauscher, B.; Sheridan, R.; Sillitoe, I.; Procter, J.; Lewis, S.E.; Rost, B.; Goldberg, T. MSAViewer: Interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* **2016**, *32*, 3501–3503. [CrossRef]
60. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*, 1189–1191. [CrossRef]

8 SUMMARY AND CONCLUSION

The overarching goal of this thesis was to study the evolution of RNA structure in viruses by examination of significant RNA structures in a diverse variety of contexts. Three studies were conducted, of which each examined an aspect of RNA structure in different viruses and on a different level of granularity. All studies explored also the methodological difficulties associated with characterization of RNA structure in viruses in different scopes.

8.1 CENTRAL FINDINGS

8.1.1 *Insights into dynamics of viral 3'UTR architecture*

It could be demonstrated in many examples that all species of the genus Flavivirus possess diversely structured 3'UTRs, many of which harbor elements that are likely of exoribonuclease-stalling function. Previously established models of structural elements, predominantly from the exoribonuclease-stalling family, were greatly improved with the help of experimental data from several critical hallmark Flavivirus species. The improved models enabled the discovery of structurally similar elements in a large number of understudied Flaviviruses, predominantly from the Tick-borne, Insect-specific, and No-Known-Vector Flaviviruses. Other highly group-specific elements were identified and shown to be of special relevance to the viruses on the basis of covariation and structure conservation, even though no experimental evidence is available. Importantly, rather than simply adding new elements to the vast and sometimes contradictory collection of elements already described in the literature[110,111], newly discovered or refined elements and their denominations were reconciled with the literature where possible. Further relevance of the identified elements is emphasized by the fact that in the meantime, many elements have been included in the Rfam database[42].

The analysis of 3'UTR secondary structure in Flaviviruses furthermore indirectly provided significant insight into the mechanisms that are responsible for host-tropism of different Flavivirus species. While all major host-tropic groups of Flaviviruses strongly diverge from each other in the 3'UTR architecture, species within a group exhibit strong similarities. Differences within groups often manifest via structure conserving sequence changes or variations in copy numbers of elements, while inter-group differences are characterized by large changes in UTR sequence that give rise to new elements that are often but not always similar in secondary structure. The fact that 3'UTRs within host-specific groups always tend to have similar structure elements strongly support the hypothesis that flaviviral 3'UTR structure is a critical contributor to the viruses ability to survive the passage through different hosts[59], during its life cycle. The connection between RNA structure and host-tropism additionally allowed to re-interpret the family of No-Known-Vector Flaviviruses, since instances of elements specific to Tick-borne or Mosquito-borne FVs could be found within the 3'UTRs of NKVs, suggesting a novel classification of individual viruses from this clearly non-monophyletic group. The approach of rapidly classifying future NKVs via 3'UTR structure therefore seems promising and should be considered in future investigations.

In sharp contrast to Flaviviruses, the Chikungunya Virus (of the Alphaviruses) exhibits a substantially different strategy regarding 3'UTR architecture. Only limited evidence of functionally relevant or conserved RNA structure could be uncovered in all lineages of CHIKV. There is limited amount of variation between different lineages of CHIKV, though the nature of variation mainly consists of duplication events. Though 3'UTR sequences are not uniform, sequence changes hardly occur and rarely provide any evidence of covariation that might be used to infer RNA structure. In fact, rather than using the non-coding regions as a ground for experimenting, CHIKV is likely experiencing evolutionary pressure to conserve UTR sequences in order to preserve sequence motifs acting as protein binding sites. Interestingly, while in Flaviviruses secondary structure is characteristic for individual viruses to a point that it allows for the precise assignment of individual specimens to viruses or lineages, structure diversity in CHIKV seems to be largely insignificant and is rather a coincidental side effect of virus evolution.

The Comparison of Flaviviruses and Alphaviruses showcases two examples of drastically different approaches of viruses towards the utilization of RNA structure in an 3'UTR architecture. Even though both families are arthropod-borne RNA viruses that share similar life cycles, Flaviviruses seem to embrace RNA structure as a versatile tool for host adaption and interference that can be mediated by conservation of RNA structure, while Alphaviruses seem to utilize non-coding RNA in a more traditional way via RNA binding proteins which is mediated via conservation of RNA sequence. Comparison of both virus groups furthermore aptly reinforces the vast diversity of the viral world, especially given the fact that such drastic differences can already be found between viruses that are quite closely related in the scope of the large-scale virus taxonomy.

8.1.2 *Methodological Advances*

One of the goals of this thesis was to promote novel or more sophisticated ways of *in silico* RNA structure examination in the field of virology. While computational methods for secondary structure prediction have been applied in the past to viruses, many approaches were rather simplistic from the methodological side and strongly focused on the prediction of structures from single sequences. While single-sequence approaches may give reasonable accurate predictions, comparative approaches further improve results by opening up the possibility to take a novel layer of information which is the evolutionary relation of sequences into account. Especially regarding RNA structure, covariation harvested from multiple sequence alignments can be a decisive indicator for the occurrence of secondary (and tertiary) structure and enables predictions which lie far beyond the capabilities of any single-sequence approach.

This thesis successfully demonstrates how to move beyond single-sequence predictions by applying a broad arsenal of comparative methods to problems in virus bioinformatics. Especially the application of thermodynamic consensus structure prediction and construction of Covariance Models, which was (surprisingly) not applied before, is demonstrated to accurately model but also detect viral RNA structure elements. Especially when used in combination with methods that guarantee a high quality of sequence data and sophisticated procedures for curation of seed alignments for viral RNA families, Covariance models are demonstrated to have the power to generate detailed insight into structure homology not only between different viruses but also between individual elements within one virus. The utility

of CMs could be further demonstrated to be especially of value in virus families or parts of the virus genome that are characterized by a high sequence divergence, that would normally obscure relationships from analysis

At the same time it is demonstrated that major improvements come not alone from using Covariance Models, but from being used in combination with other methods. Furthermore, the CM building process is still frequently dependent on manual intervention, which strongly restricts its capabilities to be rolled out to a large scale use. In addition, the procedures necessary to build, calibrate and search through CMs with sufficient accuracy can reach prohibitive resource requirements in terms of computing time and memory. Still, for comparably sized data sets the sensitivity gains that can be obtained by at least partly including a CM step in analysis pipelines is significant, and should not be ruled out, especially in scenarios where homology between viruses cannot be easily established on the basis of sequence content alone in a straightforward manner.

A critical learning from the investigations performed as part of this thesis is that the availability of experimental evidence for RNA structure massively synergizes with this type of investigations. Especially the analyses regarding the elements in Flaviviruses were rendered possible to be conduct with confidence only by the availability of high-resolution structure probing data, which allowed to build high confidence seed models. It is important to note that even a handful experimentally verified SHAPE-structures already immensely enhance and facilitate the model building process. Since structure model building is an inherently iterative process, the possibility to start from an experimentally verified core profoundly impacts the final result. Therefore, it is a general suggestion for any future in silico investigation involving RNA Covariance models or RNA families, to also focus on the acquisition or utilization of experimental evidence, where possible.

8.2 OUTLOOK

While the utility of comparative approaches, especially CM-based approaches, could be demonstrated to succeed in selected and virus families with varying degrees of diversity, the imminent next steps would ideally see the same approaches being applied to a substantially wider and diverse set of viruses. While RNA viruses are obviously the immediate target for RNA structure discovery, also the massive group of DNA viruses has been shown to utilize RNA structure and should therefore be investigated. Though, several additional obstacles need to be overcome in order to reliably achieve any amount of scaling. Seed alignment construction and the accompanying quality assessment procedures were in this work only achieved by semi-automated procedures which are well suited for the examination of a handful of RNA families, but cannot be scaled to entire virus families or even to the entire virome. Approaches for automated model building have been published, such as `CMfinder`[112] for motif finding or `RNAlien`[113] for unsupervised model building, none of them are directly suited to deal with the challenges found in viruses, such as large sequence diversity or the dissolution of traditional taxonomic relations. In addition, both tools are computationally expensive, which puts a further obstacle for any large scale investigations in the entire virome. Since significant speedups in the core CM algorithms (Inside-Outside, CYK) are unlikely to arrive due to the algorithms complexity, improvements could rather focus on automating and improving the process that decides on which CMs to build, thereby curbing the search space. While

it might seem attractive to circumvent the problem by addition of computational capacity, it is practically counteracted by the simultaneously happening exponential growth of novel and fully sequenced virus genomes. The presented work chose to reduce the number of candidate CMs by either utilization of experimental data, which is not realistically available at large scale, or by pre-filtering with other tools for ncRNA evaluation such as `R-Scape` and `RNAz`. While those tools have proven to be of significant use, they are not without bias themselves and undoubtedly introduce some amount of bias into the study. It is therefore suggested that future approaches should investigate ways that either replace those tools or at least include further factors in deciding whether a structure is relevant and can be further funneled into the computationally expensive stels of CM building.

Furthermore, while the presented work successfully showed how analyses can utilize comparative approaches in order to move beyond predictions from single sequences for better RNA structure models, the possibilities for further improvements are far from being exhausted. The work in this thesis focused primarily on improvements to the structure prediction process, which is a crucial but not the only factor impacting the quality and robustness of a RNA structure model. As indicated at various points in this thesis, important upstream processes such as multiple sequence alignment significantly impact the quality of a consensus structure prediction. Recent advances in the area of bi-alignments[114,115] have shown the problems that surface with traditional alignment algorithms when structural and sequence evolution are not concordant any more and presented novel algorithms that are suited to appropriately align such sequences. Alignments procured in this way might be a promising enhancement for the construction of RNA families for viral sequences and should therefore be considered in further similar investigations in viruses.

The application of screening methods for viral RNA structures at a large large scale could provide invaluable insights on the evolution of RNA structures in the entire biosphere. Recently, large-scale genomic analyses have fundamentally rewritten our perception of the viral world through a combination of virus gene homology inference and network analysis[116]. Similar approaches could be imagined by using noncoding RNA structures rather than genes as the elements that are shared by different viruses. Such structures might exist in plenty in viruses but have evaded any means of detection until now, due to sequence divergence obscuring any amount of structure conservation beyond recognition. Such an investigation might uncover hitherto unseen relationships between viruses, or if relationship seems implausible, uncover novel cases of convergent evolution in RNA structure, that indicate possible functionality. Even the prospect of building a structure-based phylogeny, based on common structures between viruses seems tempting, especially imagining examples where it could or cannot recapitulate a sequence- or gene-based phylogeny.

[1] E. V. Koonin, V. V. Dolja, M. Krupovic, J. H. Kuhn: Viruses Defined by the Position of the Virosphere within the Replicator Space, *Microbiology and Molecular Biology Reviews*, **85**, e00193 (2021).

[2] J. Kuhn, V. Dolja, M. Krupovic, E. Adriaenssens, F. Di Serio, B. Dutilh, R. Flores, B. Harrach, A. Mushegian, B. Owens, J. Randles, L. Rubino, S. Sabanadzovic, P. Simmonds, A. Varsani, F. Zerbini, E. Koonin: *Expand, amend, and emend the International Code of Virus Classification and Nomenclature (ICVCN; "the Code") and the Statutes to clearly define the remit of the ICTV*, ICTV (2021).

[3] C. H. Wigington, D. Sonderegger, C. P. D. Brussaard, A. Buchan, J. F. Finke, J. A. Fuhrman, J. T. Lennon, M. Middelboe, C. A. Suttle, C. Stock, W. H. Wilson, K. E. Wommack, S. W. Wilhelm, J. S. Weitz: Re-examination of the relationship between marine virus and microbial cell abundances, *Nature Microbiology*, **1**, 1 (2016).

[4] C. A. Suttle: Viruses in the sea, *Nature*, **437**, 356 (2005).

[5] L. P. P. Braga, S. M. Soucy, D. E. Amgarten, A. M. da Silva, J. C. Setubal: Bacterial Diversification in the Light of the Interactions with Phages: The Genetic Symbionts and Their Role in Ecological Speciation, *Frontiers in Ecology and Evolution*, **6** (2018).

[6] H. Liu, Y. Fu, D. Jiang, G. Li, J. Xie, J. Cheng, Y. Peng, S. A. Ghabrial, X. Yi: Widespread Horizontal Gene Transfer from Double-Stranded RNA Viruses to Eukaryotic Nuclear Genomes, *Journal of Virology*, **84**, 11876 (2010).

[7] P. N. NELSON, P. HOOLEY, D. RODEN, H. DAVARI EJTEHADI, P. RYLANCE, P. WARREN, J. MARTIN, P. G. MURRAY: Human endogenous retroviruses: transposable elements with potential?, *Clinical and Experimental Immunology*, **138**, 1 (2004).

[8] R. Belshaw, V. Pereira, A. Katzourakis, G. Talbot, J. Pačes, A. Burt, M. Tristem: Long-term reinfection of the human genome by endogenous retroviruses, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4894 (2004).

[9] D. M. Ozata, I. Gainetdinov, A. Zoch, D. O'Carroll, P. D. Zamore: PIWI-interacting RNAs: small RNAs with big functions, *Nature Reviews Genetics*, **20**, 89 (2019).

[10] F. X. Abad, R. M. Pintó, A. Bosch: Survival of enteric viruses on environmental fomites., *Applied and Environmental Microbiology*, **60**, 3704 (1994).

[11] E. V. Koonin, A. S. Novozhilov: Origin and Evolution of the Universal Genetic Code, *Annual Review of Genetics*, **51**, 45 (2017).

[12] E. V. Koonin, T. G. Senkevich, V. V. Dolja: The ancient Virus World and evolution of cells, *Biology Direct*, **1**, 29 (2006).

[13] Wikimedia: T4virussa.gif (620×664), `https://upload.wikimedia.org/wikipedia/commons/2/26/T4VirusSA.gif`, (Accessed on 02/21/2023) (2023).

[14] Wikimedia: Enterobacteria_phage_t2_transmission_electron_micrograph.jpg (2021×2021), `https://upload.wikimedia.org/wikipedia/commons/9/98/Enterobacteria_phage_T2_transmission_electron_micrograph.jpg`, (Accessed on 02/21/2023) (2023).

[15] H. J. Gross, H. Domdey, C. Lossow, P. Jank, M. Raba, H. Alberty, H. L. Sänger: Nucleotide sequence and secondary structure of potato spindle tuber viroid, *Nature*, **273**, 203 (1978).

[16] T. SANO: Progress in 50 years of viroid research—Molecular structure, pathogenicity, and host adaptation, *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, **97**, 371 (2021).

[17] P. Colson, N. Yutin, S. A. Shabalina, C. Robert, G. Fournous, B. La Scola, D. Raoult, E. V. Koonin: Viruses with More Than 1,000 Genes: Mamavirus, a New Acanthamoeba polyphagamimivirus Strain, and Reannotation of Mimivirus Genes, *Genome Biology and Evolution*, **3**, 737 (2011).

[18] E. V. Koonin: Comparative genomics, minimal gene-sets and the last universal common ancestor, *Nature Reviews Microbiology*, **1**, 127 (2003).

[19] H. Ruska: Versuch zu einer Ordnung der Virusarten, *Archives of Virology*, **2**, 480 (1943).

[20] P. D. Cooper: A Chemical Basis for the Classification of Animal Viruses, *Nature*, **190**, 302 (1961).

[21] A. Lwoff: Principles of Classification and Nomenclature of Viruses, *Nature*, **215**, 13 (1967).

[22] PDB: Rcsb pdb - 5ire: The cryo-em structure of zika virus, `https://www.rcsb.org/structure/5IRE`, (Accessed on 02/23/2023) (2016).

[23] P. David Goodsell: Pdb-101: Molecule of the month: Tobacco mosaic virus, `https://pdb101.rcsb.org/motm/109`, (Accessed on 02/23/2023) (2009).

[24] P. Dr. Victor Padilla-Sanchez: Bacteriophage lambda structural model at atomic resolution, `https://zenodo.org/record/5110221`, (Accessed on 02/23/2023) (2021).

[25] D. Baltimore: Expression of animal virus genomes, *Bacteriological Reviews*, **35**, 235 (1971).

[26] ViralZone: Baltimore classification ~ viralzone, `https://viralzone.expasy.org/254`, (Accessed on 02/21/2023) (2023).

[27] A. E. Gorbalenya: Increasing the number of available ranks in virus taxonomy from five to ten and adopting the Baltimore classes as taxa at the basal rank, *Archives of Virology*, **163**, 2933 (2018).

[28] S. G. Siddell, P. J. Walker, E. J. Lefkowitz, A. R. Mushegian, M. J. Adams, B. E. Dutilh, A. E. Gorbalenya, B. Harrach, R. L. Harrison, S. Junglen, N. J. Knowles, A. M. Kropinski, M. Krupovic, J. H. Kuhn, M. Nibert, L. Rubino, S. Sabanadzovic, H. Sanfaçon, P. Simmonds, A. Varsani, F. M. Zerbini, A. J. Davison: Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018), *Archives of Virology*, **164**, 943 (2019).

[29] E. V. Koonin, V. V. Dolja, M. Krupovic, A. Varsani, Y. I. Wolf, N. Yutin, F. M. Zerbini, J. H. Kuhn: Global Organization and Proposed Megataxonomy of the Virus World, *Microbiology and Molecular Biology Reviews : MMBR*, **84** (2020).

[30] J. H. Kuhn: Virus Taxonomy, in D. H. Bamford, M. Zuckerman (editors), *Encyclopedia of Virology (Fourth Edition)*, 28–37, Academic Press, Oxford (2021).

[31] P. Simmonds, M. J. Adams, M. Benkő, M. Breitbart, J. R. Brister, E. B. Carstens, A. J. Davison, E. Delwart, A. E. Gorbalenya, B. Harrach, R. Hull, A. M. Q. King, E. V. Koonin, M. Krupovic, J. H. Kuhn, E. J. Lefkowitz, M. L. Nibert, R. Orton, M. J. Roossinck, S. Sabanadzovic, M. B. Sullivan, C. A. Suttle, R. B. Tesh, R. A. van der Vlugt, A. Varsani, F. M. Zerbini: Virus taxonomy in the age of metagenomics, *Nature Reviews Microbiology*, **15**, 161 (2017).

[32] ICTV: About virus taxonomic classification | ictv, `https://ictv.global/taxonomy/about`, (Accessed on 02/22/2023) (2023).

[33] Z. A. Jaafar, J. S. Kieft: Viral RNA structure-based strategies to manipulate translation, *Nature Reviews Microbiology*, **17**, 110 (2019).

[34] A. MacFadden, Z. O'Donoghue, P. A. G. C. Silva, E. G. Chapman, R. C. Olsthoorn, M. G. Sterken, G. P. Pijlman, P. J. Bredenbeek, J. S. Kieft: Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs, *Nature Communications*, **9**, 119 (2018).

[35] P. Simmonds, D. B. Smith: Structural Constraints on RNA Virus Evolution, *Journal of Virology*, **73**, 5787 (1999).

[36] W. Fontana: Modelling 'evo-devo' with RNA, *BioEssays*, **24**, 1164 (2002).

[37] J. A. G. M. d. Visser, J. Hermisson, G. P. Wagner, L. A. Meyers, H. Bagheri-Chaichian, J. L. Blanchard, L. Chao, J. M. Cheverud, S. F. Elena, W. Fontana, G. Gibson, T. F. Hansen, D. Krakauer, R. C. Lewontin, C. Ofria, S. H. Rice, G. von Dassow, A. Wagner, M. C. Whitlock: Perspective: Evolution and Detection of Genetic Robustness, *Evolution*, **57**, 1959 (2003).

[38] B. Houck-Loomis, M. A. Durney, C. Salguero, N. Shankar, J. M. Nagle, S. P. Goff, V. M. D'Souza: An equilibrium-dependent retroviral mRNA switch regulates translational recoding, *Nature*, **480**, 561 (2011).

[39] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S. R. Eddy: Rfam: an RNA family database, *Nucleic Acids Research*, **31**, 439 (2003).

[40] P. P. Gardner, J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, R. D. Finn, E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, A. Bateman: Rfam: Wikipedia, clans and the "decimal" release, *Nucleic Acids Research*, **39**, D141 (2011).

[41] T. R. Cech, J. A. Steitz: The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones, *Cell*, **157**, 77 (2014).

[42] I. Kalvari, E. P. Nawrocki, N. Ontiveros-Palacios, J. Argasinska, K. Lamkiewicz, M. Marz, S. Griffiths-Jones, C. Toffano-Nioche, D. Gautheret, Z. Weinberg, E. Rivas, S. R. Eddy, R. Finn, A. Bateman, A. I. Petrov: Rfam 14: expanded coverage of metagenomic, viral and microRNA families, *Nucleic Acids Research*, **49**, D192 (2021).

[43] B. J. Blitvich, A. E. Firth: A Review of Flaviviruses that Have No Known Arthropod Vector, *Viruses*, **9**, 154 (2017).

[44] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, S. I. Hay: The global distribution and burden of dengue, *Nature*, **496**, 504 (2013).

[45] H. Weissenböck, J. Kolodziejek, A. Url, H. Lussy, B. Rebel-Bauder, N. Nowotny: Emergence of Usutu virus, an African Mosquito-Borne Flavivirus of the Japanese Encephalitis Virus Group, Central Europe - Volume 8, Number 7—July 2002 - Emerging Infectious Diseases journal - CDC, *CDC Reports* (2002).

[46] T. V. B. d. Araújo, L. C. Rodrigues, R. A. d. A. Ximenes, D. d. B. Miranda-Filho, U. R. Montarroyos, A. P. L. d. Melo, S. Valongueiro, M. d. F. P. M. d. Albuquerque, W. V. Souza, C. Braga, S. P. B. Filho, M. T. Cordeiro, E. Vazquez, D. D. C. S. Cruz, C. M. P. Henriques, L. C. A. Bezerra, P. M. d. S. Castanha, R. Dhalia, E. T. A. Marques-Júnior, C. M. T. Martelli: Association between Zika virus infection and microcephaly in Brazil, January to May, 2016: preliminary report of a case-control study, *The Lancet Infectious Diseases*, **16**, 1356 (2016).

[47] N. Vasilakis, S. C. Weaver: Flavivirus transmission focusing on Zika, *Current opinion in virology*, **22**, 30 (2017).

[48] B. J. Blitvich, A. E. Firth: Insect-Specific Flaviviruses: A Systematic Review of Their Discovery, Host Range, Mode of Transmission, Superinfection Exclusion Potential and Genomic Organization, *Viruses*, **7**, 1927 (2015).

[49] R. Ochsenreiter, I. L. Hofacker, M. T. Wolfinger: Functional RNA Structures in the 3′UTR of Tick-Borne, Insect-Specific and No-Known-Vector Flaviviruses, *Viruses*, **11**, 298 (2019).

[50] P. Simmonds, P. Becher, J. Bukh, E. A. Gould, G. Meyers, T. Monath, S. Muerhoff, A. Pletnev, R. Rico-Hesse, D. B. Smith, J. T. Stapleton, ICTV Report Consortium: ICTV Virus Taxonomy Profile: Flaviviridae, *Journal of General Virology*, **98**, 2 (2017).

[51] L. Yu, M. Nomaguchi, R. Padmanabhan, L. Markoff: Specific requirements for elements of the 5' and 3' terminal regions in flavivirus RNA synthesis and viral replication, *Virology*, **374**, 170 (2008).

[52] D. E. Alvarez, M. F. Lodeiro, S. J. Ludueña, L. I. Pietrasanta, A. V. Gamarnik: Long-range RNA-RNA interactions circularize the dengue virus genome, *Journal of Virology*, **79**, 6631 (2005).

[53] R. M. Kofler, V. M. Hoenninger, C. Thurner, C. W. Mandl: Functional Analysis of the Tick-Borne Encephalitis Virus Cyclization ElementsIndicates Major Differences between Mosquito-Borne and Tick-Borne Flaviviruses, *Journal of Virology*, **80**, 4099 (2006).

[54] H. Rouha, V. M. Hoenninger, C. Thurner, C. W. Mandl: Mutational analysis of three predicted 5′-proximal stem-loop structures in the genome of tick-borne encephalitis virus indicates different roles in RNA replication and translation, *Virology*, **417**, 79 (2011).

[55] C. V. Filomatori, J. M. Carballeda, S. M. Villordo, S. Aguirre, H. M. Pallarés, A. M. Maestre, I. Sánchez-Vargas, C. D. Blair, C. Fabri, M. A. Morales, A. Fernandez-Sesma, A. V. Gamarnik: Dengue virus genomic variation associated with mosquito adaptation defines the pattern of viral non-coding RNAs and fitness in human cells, *PLOS Pathogens*, **13**, e1006265 (2017).

[56] S. M. Villordo, A. V. Gamarnik: Differential RNA Sequence Requirement for Dengue Virus Replication in Mosquito and Mammalian Cells, *Journal of Virology*, **87**, 9365 (2013).

[57] C. V. Filomatori, M. F. Lodeiro, D. E. Alvarez, M. M. Samsa, L. Pietrasanta, A. V. Gamarnik: A 5′ RNA element promotes dengue virus RNA synthesis on a circular genome, *Genes & Development*, **20**, 2238 (2006).

[58] K. Bidet, M. A. Garcia-Blanco: Flaviviral RNAs: weapons and targets in the war between virus and host, *Biochemical Journal*, **462**, 215 (2014).

[59] S. M. Villordo, C. V. Filomatori, I. Sánchez-Vargas, C. D. Blair, A. V. Gamarnik: Dengue Virus RNA Structure Specialization Facilitates Host Adaptation, *PLOS Pathogens*, **11**, e1004604 (2015).

[60] S. M. Villordo, J. M. Carballeda, C. V. Filomatori, A. V. Gamarnik: RNA Structure Duplications and Flavivirus Host Adaptation, *Trends in Microbiology*, **24**, 270 (2016).

[61] K. Bidet, D. Dadlani, M. A. Garcia-Blanco: G3BP1, G3BP2 and CAPRIN1 Are Required for Translation of Interferon Stimulated mRNAs and Are Targeted by a Dengue Virus Non-coding RNA, *PLOS Pathogens*, **10**, e1004242 (2014).

[62] S. L. Moon, B. J. T. Dodd, D. E. Brackney, C. J. Wilusz, G. D. Ebel, J. Wilusz: Flavivirus sfRNA suppresses antiviral RNA interference in cultured cells and mosquitoes and directly interacts with the RNAi machinery, *Virology*, **485**, 322 (2015).

[63] G. P. Göertz, J. J. Fros, P. Miesen, C. B. F. Vogels, M. L. van der Bent, C. Geertsema, C. J. M. Koenraadt, R. P. van Rij, M. M. van Oers, G. P. Pijlman: Noncoding Subgenomic Flavivirus RNA Is Processed by the Mosquito RNA Interference Machinery and Determines West Nile Virus Transmission by Culex pipiens Mosquitoes, *Journal of Virology*, **90**, 10145 (2016).

[64] G. Manokaran, E. Finol, C. Wang, J. Gunaratne, J. Bahl, E. Z. Ong, H. C. Tan, O. M. Sessions, A. M. Ward, D. J. Gubler, E. Harris, M. A. Garcia-Blanco, E. E. Ooi: Dengue subgenomic RNA binds TRIM25 to inhibit interferon expression for epidemiological fitness, *Science (New York, N.Y.)*, **350**, 217 (2015).

[65] A. Schuessler, A. Funk, H. M. Lazear, D. A. Cooper, S. Torres, S. Daffis, B. K. Jha, Y. Kumagai, O. Takeuchi, P. Hertzog, R. Silverman, S. Akira, D. J. Barton, M. S. Diamond, A. A. Khromykh: West Nile virus noncoding subgenomic RNA contributes to viral evasion of the type I interferon-mediated antiviral response, *Journal of Virology*, **86**, 5708 (2012).

[66] S.-D. Thiberville, V. Boisson, J. Gaudart, F. Simon, A. Flahault, X. de Lamballerie: Chikungunya Fever: A Clinical and Virological Investigation of Outpatients on Reunion Island, South-West Indian Ocean, *PLoS Neglected Tropical Diseases*, **7** (2013).

[67] E. Javelle, A. Ribera, I. Degasne, B.-A. Gaüzère, C. Marimoutou, F. Simon: Specific Management of Post-Chikungunya Rheumatic Disorders: A Retrospective Study of 159 Cases in Reunion Island from 2006-2012, *PLoS Neglected Tropical Diseases*, **9** (2015).

[68] M. C. Robinson: An epidemic of virus disease in Southern Province, Tanganyika territory, in 1952–1953. I. Clinical Features, *Transactions of The Royal Society of Tropical Medicine and Hygiene*, **49**, 28 (1955).

[69] G. Rezza, S. C. Weaver: Chikungunya as a paradigm for emerging viral diseases: Evaluating disease impact and hurdles to vaccine development, *PLoS Neglected Tropical Diseases*, **13** (2019).

[70] S. Inoue, K. Morita, R. Matias, J. Tuplano, R. Resuello, J. Candelario, D. Cruz, C. Mapua, F. Hasebe, A. Igarashi, F. Natividad: Distribution of three arbovirus antibodies among monkeys (Macaca fascicularis) in the Philippines, *Journal of Medical Primatology*, **32**, 89 (2003).

[71] P. Gérardin, G. Barau, A. Michault, M. Bintner, H. Randrianaivo, G. Choker, Y. Lenglet, Y. Touret, A. Bouveret, P. Grivard, K. Le Roux, S. Blanc, I. Schuffenecker, T. Couderc, F. Arenzana-Seisdedos, M. Lecuit, P.-Y. Robillard: Multidisciplinary prospective study of mother-to-child chikungunya virus infections on the island of La Réunion, *PLoS medicine*, **5**, e60 (2008).

[72] L. A. Silva, T. S. Dermody: Chikungunya virus: epidemiology, replication, disease mechanisms, and prospective intervention strategies, *The Journal of Clinical Investigation*, **127**, 737 (2017).

[73] A. C. Holmes, K. Basore, D. H. Fremont, M. S. Diamond: A molecular understanding of alphavirus entry, *PLOS Pathogens*, **16**, e1008876 (2020).

[74] R. Chen, S. Mukhopadhyay, A. Merits, B. Bolling, F. Nasar, L. L. Coffey, A. Powers, S. C. Weaver, I. R. Consortium: ICTV Virus Taxonomy Profile: Togaviridae, *Journal of General Virology*, **99**, 761 (2018).

[75] M. Pfeffer, R. M. Kinney, O. R. Kaaden: The alphavirus 3'-nontranslated region: size heterogeneity and arrangement of repeated sequence elements, *Virology*, **240**, 100 (1998).

[76] R. Chen, E. Wang, K. A. Tsetsarkin, S. C. Weaver: Chikungunya Virus 3′ Untranslated Region: Adaptation to Mosquitoes and a Population Bottleneck as Major Evolutionary Forces, *PLOS Pathogens*, **9**, e1003591 (2013).

[77] S. K. White, C. Mavian, M. Salemi, J. G. Morris, M. A. Elbadry, B. A. Okech, J. A. Lednicky, J. C. Dunford: A new "American" subgroup of African-lineage Chikungunya virus detected in and isolated from mosquitoes collected in Haiti, 2016, *PloS One*, **13**, e0196857 (2018).

[78] S. M. Volk, R. Chen, K. A. Tsetsarkin, A. P. Adams, T. I. Garcia, A. A. Sall, F. Nasar, A. J. Schuh, E. C. Holmes, S. Higgs, P. D. Maharaj, A. C. Brault, S. C. Weaver: Genome-Scale Phylogenetic Analyses of Chikungunya Virus Reveal Independent Emergences of Recent Epidemics and Various Evolutionary Rates, *Journal of Virology*, **84**, 6497 (2010).

[79] B. Knudsen, J. Hein: RNA secondary structure prediction using stochastic context-free grammars and evolutionary history., *Bioinformatics*, **15**, 446 (1999).

[80] I. Tinoco, O. C. Uhlenbeck, M. D. Levine: Estimation of Secondary Structure in Ribonucleic Acids, *Nature*, **230**, 362 (1971).

[81] M. Zuker, P. Stiegler: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information., *Nucleic Acids Research*, **9**, 133 (1981).

[82] L. Fu, Y. Cao, J. Wu, Q. Peng, Q. Nie, X. Xie: UFold: fast and accurate RNA secondary structure prediction with deep learning, *Nucleic Acids Research*, **50**, e14 (2022).

[83] T. Xia, J. J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, D. H. Turner: Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson–Crick Base Pairs, *Biochemistry*, **37**, 14719 (1998).

[84] D. H. Turner, D. H. Mathews: NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure, *Nucleic Acids Research*, **38**, D280 (2010).

[85] R. Nussinov, G. Pieczenik, J. R. Griggs, D. J. Kleitman: Algorithms for Loop Matchings, *SIAM Journal on Applied Mathematics*, **35**, 68 (1978).

[86] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, P. F. Stadler: RNAalifold: improved consensus structure prediction for RNA alignments, *BMC Bioinformatics*, **9**, 474 (2008).

[87] R. J. Klein, S. R. Eddy: RSEARCH: finding homologs of single structured RNA sequences, *BMC bioinformatics*, **4**, 44 (2003).

[88] F. Madeira, M. Pearce, A. R. N. Tivey, P. Basutkar, J. Lee, O. Edbali, N. Madhusoodanan, A. Kolesnikov, R. Lopez: Search and sequence analysis tools services from EMBL-EBI in 2022, *Nucleic acids research*, gkac240 (2022).

[89] K. Katoh, J. Rozewicki, K. D. Yamada: MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization, *Briefings in Bioinformatics*, **20**, 1160 (2019).

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

[91] N. Chomsky: Three models for the description of language, *IRE Transactions on Information Theory*, **2**, 113 (1956).

[92] L. Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77**, 257 (1989).

[93] S. R. Eddy: Hidden Markov models, *Current Opinion in Structural Biology*, **6**, 361 (1996).

[94] L. Satish, B. Gururaj: Use of hidden Markov models for partial discharge pattern classification, *IEEE Transactions on Electrical Insulation*, **28**, 172 (1993).

[95] A. Viterbi: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, **13**, 260 (1967).

[96] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1 edition (1998).

[97] Hidden markov model - wikipedia, `https://en.wikipedia.org/wiki/Hidden_Markov_model`, (Accessed on 03/09/2023) (2009).

[98] S. R. Eddy, R. Durbin: RNA sequence analysis using covariance models, *Nucleic Acids Research*, **22**, 2079 (1994).

[99] K. Lari, S. J. Young: The estimation of stochastic context-free grammars using the Inside-Outside algorithm, *Computer Speech & Language*, **4**, 35 (1990).

[100] D. L. Kolbe, S. R. Eddy: Fast filtering for RNA homology search, *Bioinformatics*, **27**, 3102 (2011).

[101] E. P. Nawrocki, S. R. Eddy: Infernal 1.1: 100-fold faster RNA homology searches, *Bioinformatics (Oxford, England)*, **29**, 2933 (2013).

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

[103] S. Washietl, I. L. Hofacker, P. F. Stadler: Fast and reliable prediction of noncoding RNAs, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2454 (2005).

[104] A. R. Gruber, S. FINDEIss, S. Washietl, I. L. Hofacker, P. F. Stadler: Rnaz 2.0: improved noncoding rna detection, in *Biocomputing 2010*, 69–79, WORLD SCIENTIFIC (2009).

[105] C. E. Shannon: A mathematical theory of communication, *The Bell System Technical Journal*, **27**, 379 (1948).

[106] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, A. Bateman: Rfam: updates to the RNA families database, *Nucleic Acids Research*, **37**, D136 (2009).

[107] E. Rivas, J. Clements, S. R. Eddy: A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs, *Nature Methods*, **14**, 45 (2017).

[108] E. Rivas, J. Clements, S. R. Eddy: Estimating the power of sequence covariation for detecting conserved RNA structure, *Bioinformatics*, **36**, 3072 (2020).

[109] B. Woolf: The Log Likelihood Ratio Test (the G-Test), *Annals of Human Genetics*, **21**, 397 (1957).

[110] D. J. Gritsun, I. M. Jones, E. A. Gould, T. S. Gritsun: Molecular Archaeology of Flaviviridae Untranslated Regions: Duplicated RNA Structures in the Replication Enhancer of Flaviviruses and Pestiviruses Emerged via Convergent Evolution, *PLOS ONE*, **9**, e92056 (2014).

[111] T. S. Gritsun, P. A. Nuttall, E. A. Gould: Tick-borne flaviviruses, *Advances in Virus Research*, **61**, 317 (2003).

[112] Z. Yao, Z. Weinberg, W. L. Ruzzo: CMfinder—a covariance model based RNA motif finding algorithm, *Bioinformatics*, **22**, 445 (2006).

[113] F. Eggenhofer, I. L. Hofacker, C. Höner zu Siederdissen: RNAlien – Unsupervised RNA family model construction, *Nucleic Acids Research*, **44**, 8433 (2016).

[114] M. Waldl, S. Will, M. T. Wolfinger, I. L. Hofacker, P. F. Stadler: Bi-alignments as Models of Incongruent Evolution of RNA Sequence and Secondary Structure, in P. Cazzaniga, D. Besozzi, I. Merelli, L. Manzoni (editors), *Computational Intelligence Methods for Bioinformatics and Biostatistics*, Lecture Notes in Computer Science, 159–170, Springer International Publishing, Cham (2020).

[115] P. F. Stadler, S. Will: Bi-alignments with affine gaps costs, *Algorithms for Molecular Biology*, **17**, 10 (2022).

[116] J. Iranzo, E. V. Koonin, D. Prangishvili, M. Krupovic: Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements, *Journal of Virology*, **90**, 11043 (2016).

[117] A. de Bernardi Schneider, R. Ochsenreiter, R. Hostager, I. L. Hofacker, D. Janies, M. T. Wolfinger: Updated Phylogeny of Chikungunya Virus Suggests Lineage-Specific RNA Architecture, *Viruses*, **11**, 798 (2019).

[118] M. Kiening, R. Ochsenreiter, H.-J. Hellinger, T. Rattei, I. Hofacker, D. Frishman: Conserved Secondary Structures in Viral mRNAs, *Viruses*, **11**, 401 (2019).

[119] B. C. Thiel, R. Ochsenreiter, V. P. Gadekar, A. Tanzer, I. L. Hofacker: RNA Structure Elements Conserved between Mouse and 59 Other Vertebrates, *Genes*, **9**, 392 (2018).

[120] M. Waldl, B. C. Thiel, R. Ochsenreiter, A. Holzenleiter, J. V. De Araujo Oliveira, M. E. M. T. Walter, M. T. Wolfinger, P. F. Stadler: TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes, *Genes*, **9**, 372 (2018).

[121] M. Hölzer, V. Krähling, F. Amman, E. Barth, S. H. Bernhart, V. A. O. Carmelo, M. Collatz, G. Doose, F. Eggenhofer, J. Ewald, J. Fallmann, L. M. Feldhahn, M. Fricke, J. Gebauer, A. J. Gruber, F. Hufsky, H. Indrischek, S. Kanton, J. Linde, N. Mostajo, R. Ochsenreiter, K. Riege, L. Rivarola-Duarte, A. H. Sahyoun, S. J. Saunders, S. E. Seemann, A. Tanzer, B. Vogel, S. Wehner, M. T. Wolfinger, R. Backofen, J. Gorodkin, I. Grosse, I. Hofacker, S. Hoffmann, C. Kaleta, P. F. Stadler, S. Becker, M. Marz: Differential transcriptional responses to Ebola and Marburg virus infection in bat and human cells, *Scientific Reports*, **6**, 34589 (2016).

[122] M. T. Wolfinger, R. Ochsenreiter, I. L. Hofacker: Functional RNA Structures in the 3′ UTR of Mosquito-Borne Flaviviruses, in *Virus Bioinformatics*, Chapman and Hall/CRC (2021), num Pages: 36.

LIST OF PUBLICATIONS

Publications in bold are included in this thesis.

- **R. Ochsenreiter, I. L. Hofacker, M. T. Wolfinger: Functional RNA Structures in the 3′UTR of Tick-Borne, Insect-Specific and No-Known-Vector Flaviviruses, *Viruses*, 11, 298 (2019)**
- **A. de Bernardi Schneider, R. Ochsenreiter, R. Hostager, I. L. Hofacker, D. Janies, M. T. Wolfinger: Updated Phylogeny of Chikungunya Virus Suggests Lineage-Specific RNA Architecture, *Viruses*, 11, 798 (2019)**
- **M. Kiening, R. Ochsenreiter, H.-J. Hellinger, T. Rattei, I. Hofacker, D. Frishman: Conserved Secondary Structures in Viral mRNAs, *Viruses*, 11, 401 (2019)**
- B. C. Thiel, R. Ochsenreiter, V. P. Gadekar, A. Tanzer, I. L. Hofacker: RNA Structure Elements Conserved between Mouse and 59 Other Vertebrates, *Genes*, **9**, 392 (2018)
- M. Waldl, B. C. Thiel, R. Ochsenreiter, A. Holzenleiter, J. V. De Araujo Oliveira, M. E. M. T. Walter, M. T. Wolfinger, P. F. Stadler: TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes, *Genes*, **9**, 372 (2018)
- M. Hölzer, V. Krähling, F. Amman, E. Barth, S. H. Bernhart, V. A. O. Carmelo, M. Collatz, G. Doose, F. Eggenhofer, J. Ewald, J. Fallmann, L. M. Feldhahn, M. Fricke, J. Gebauer, A. J. Gruber, F. Hufsky, H. Indrischek, S. Kanton, J. Linde, N. Mostajo, R. Ochsenreiter, K. Riege, L. Rivarola-Duarte, A. H. Sahyoun, S. J. Saunders, S. E. Seemann, A. Tanzer, B. Vogel, S. Wehner, M. T. Wolfinger, R. Backofen, J. Gorodkin, I. Grosse, I. Hofacker, S. Hoffmann, C. Kaleta, P. F. Stadler, S. Becker, M. Marz: Differential transcriptional responses to Ebola and Marburg virus infection in bat and human cells, *Scientific Reports*, **6**, 34589 (2016)
- M. T. Wolfinger, R. Ochsenreiter, I. L. Hofacker: Functional RNA Structures in the 3′ UTR of Mosquito-Borne Flaviviruses, in *Virus Bioinformatics*, Chapman and Hall/CRC (2021), num Pages: 36

ACKNOWLEDGMENTS

This thesis marks the end of a journey. At this point I would like to thank all people who have influenced and shaped me while coming to this point.

Ivo, I would like to express my biggest gtatitude first and foremost to you, who made my journey possible. The past years have been an invaluable formative time in my life, and you played an instrumental role. I want to thank you for being my supervisor, teacher and mentor, but also for being a friend and source of a myriad of equally interesting and educational conversations. More than all, I want to thank you for showing me the value innate to rigorous, analytical thinking and not settling for low-hanging fruits.

Mtw, I want to thank you for our extremely productive cooperations. Working with you proved not only to be very fruitful, but also very pleasant - I am sure that sifting through dubious RNA structures in exotic Viruses was never as hilarious before. I am also grateful for your guidance and traction when it came to finishing our projects.

Xtof, through you I could significantly broaden my horizon with numerous diverse and exotic topics I would not have discovered on my own. I learned the true meaning of ""keeping an open mind" through your inexhaustible drive to discover new things that are (or seem to be) out of the box.

Django aka Stefan Hammer, thank you for your friendship, compassion and so many other things during the years we shared an office. Thank you for providing so many splendid opportunities for venting frustration by shouting at each other - I would not want to share a room with anyone else!

Maria Waldl and **Bernhard Thiel**, thank you for numerous interesting discussions, not only through the projects we worked on. Thank you Maria especially for always sharing your knowledge and your inexhaustible efforts in keeping the institute spirit up (especially at lunchtime!).

Ronny Lorenz, thank you for everything related to the RNA package. You provided for us the software backbone most computational scientists can only dream of. Thank you also for answering so many of my (silly) questions regarding the package and RNA folding!

Richard Neuböck, thank you for providing the TBI infrastructure. I don't know how, but you managed to single-handedly maintain a comprehensive IT and HPC infrastructure in what can just be described as an herculean effort. Live long and Prosper!

Judith Ivansits and **Amy**, thank you for a smooth administration of the institute and a healthy social atmosphere.

Michael Gastegger, thank you for being my opera and Konzerthaus ally, where we enjoyed Wagner and Mozart for hours just to get to the Würstl in the break(s). There never have been better opportunities to discuss science than during three hours of queueing for five hours of *Parsifal*.

Finally my biggest gratitude belongs to my family, my parents **Wilhelm** and **Maria** and my brother **Julian** for supporting me through my whole life. The final and most important thank you goes to you, **Dr.** ***Vroni***: Thank you for your love through all parts of the journey, I could not have made it without your support.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne Zuhilfenahme weiterer als der aufgeführten Quellen angefertigt habe. Alle wörtlich oder sinngemäß übernommenen Textstellen anderer Verfasser wurden als solche gekennzeichnet.

Unterschrift Wien, 2023