CHAPTER 1

General Introduction

RNA virus diversity in a nutshell

One of the most fundamental characteristics of RNA viruses is genetic variation. It originates from different sources which include (i) mutation (misincorporation of nucleotides during genome copying), (ii) duplication of genome regions (e.g. genes) followed by subsequent diversification, (iii) acquisition of foreign genetic material, (iv) loss of genetic material, and (v) overprinting (opening of an alternative, possibly overlapping reading frame)[34,252]. The underlying molecular processes act on different scales and may have profoundly different impacts on the virus affected.

At the lowermost level, genetic variation in RNA viruses is observed within single-host infections where a cloud of potentially beneficial mutations is formed and maintained. The cloud sequences can be thought of representing a fitness landscape – sequences with slightly lower or higher fitness – which allows the virus to quickly respond to environmental changes during infection. What is obtained by genome sequencing is a so-called *master sequence* which basically represents a consensus over the cloud. Interestingly, when infecting a new individual with the same virus, essentially the same cloud is formed which suggests that the sequences, linked by mutational coupling, act as one entity which is targeted as a whole by evolutionary selection[43]. This prompted researchers to refer to it as *Quasispecies*[110,131,132,471]. However, others are reluctant to accept the applicability of this concept to RNA viruses[219,237]. Regardless of this debate, it is apparent that genetic variation naturally increases on higher levels, e.g. when distinct viruses diverge gradually during evolution, eventually erasing any detectable homology. The reason for this high sequence variability of RNA viruses is the extreme error rate of their RNA-dependent polymerases, which introduce roughly one mutation per 10,000 nt copied[117,121]. In contrast to their cellular counterparts with an error rate that is orders of magnitude lower, polymerases of RNA viruses lack a proofreading-repair functionality (with one possible exception, see below).

Besides mutation, the second major source of genetic variation in RNA viruses is recombination. Recombination is thought to be guided by the level of local sequence similarity of the participant nucleic acid molecules[492] and, hence, is predominantly observed between closely related viruses. As a result, a recombinant genome in which homologous parts have been exchanged between the parents is produced (homologous recombination), potentially causing several substitutions in one go with respect to either parent. A prerequisite for homologous recombination to take place is co-infection of the same cell. It is worth mentioning that, while representing a source of innovation on a small scale, homologous recombination limits sequence divergence on a large scale (evolutionary time scale) by both the propagation of beneficial substitutions throughout a population of closely related viruses and the removal of deleterious mutations[305]. Moreover, recombination may also happen with remotely related viruses or the host transcriptome, and as a result new genes or other functional elements are integrated into the recipient viral genome (non-homologous recombination). In contrast to the mechanism of recombination in the host

genome which works through breakage and rejoining of DNA strands[80,251], it is generally accepted that that of most RNA viruses is characterized by template switching of the viral polymerase during replication[6,260,329].

With increasing genetic divergence, RNA virus diversity is also observed at the level of the proteome. A hallmark of RNA viruses and viruses in general is that they do not encode ribosomes and are thus condemned to parasitize host cells for protein synthesis. However, they express various other types of proteins. The only recurrent theme here is that all *bona fide* RNA viruses encode at least one virion protein as well as a polymerase, for genome dissemination and replication, respectively. The polymerase can be of different type depending on the mechanism of mRNA production (see next section). Its core component, the palm subdomain, which is crucial for catalysis, is structurally conserved across RNA viruses[184,367]. Many but not all RNA viruses express one or several proteinases which are utilized either for cleavage of viral polyproteins or to interact with the host environment[173]. RNA virus proteinases are grouped, on the basis of sequence similarity, into the chymotrypsin-related and papain-related superfamilies and proteases of unknown type. The chymotrypsin-related superfamily is further divided into the 3C-like family (named after the picornavirus proteinase encoded in the 3C genome region) and the CP-like family (named after the alphavirus capsid protein)[175]. Another type of enzyme frequently found in the RNA virus proteome is a helicase, characterized by containing a NTP-binding sequence pattern. It provides an activity – the unwinding of dsRNA or base-paired ssRNA – that is crucial to many genetic processes like replication, expression or recombination[177]. RNA virus helicases are grouped, on the basis of sequence similarity, into three superfamilies, two of which also include cellular proteins[175,177]. Besides these widespread enzymes, various other types of proteins are employed by RNA viruses including, for instance, ribonucleases, methyltransferases, phosphatases or membrane-associated proteins[73,96,128,150,174,376]. Notably, many RNA virus-encoded proteins have so far not been characterized functionally, partially due to the lack of any apparent similarity with cellular counterparts. Owing to the limited coding capacity of RNA viruses most of their proteins are multifunctional, but, at the same time, there is a certain division of labor between the enzymes[7].

More fundamentally, RNA virus diversity is observed at the level of the genome. RNA viruses vary in employing one of three possible genome types – double stranded (dsRNA viruses), single-stranded negative sense (ssRNA- viruses) or single-stranded positive sense (ssRNA+ viruses and RNA retroviruses). The latter represents by far the most abundant genome type and outnumbers by around fivefold each of the other two[25]. What's more, RNA virus genomes show a considerable variability in size[174]. Despite being limited to relatively small sizes compared to cellular organisms and most DNA viruses, RNA virus genomes still range from approximately 2 to 32 kb with an average of about 10 kb. Larger genomes are supposed to result in a so-called *error catastrophe* caused by too many deleterious mutations that would accumulate during the error-prone copying of RNA genomes[216]. Yet, the diversity of RNA genome sizes is striking and it was proposed that the

acquisition of specific enzymes that refined the rudimentary viral replication machinery empowered some RNA viruses to undergo major evolutionary transitions accompanied with an enlargement of the genome beyond the size of the acquired gene[174].

Last but not least, there is variation among RNA viruses from an economic and medical perspective. Most RNA viruses have rather mild effects on their hosts, which makes sense from an evolutionary perspective because the virus should avoid compromising its means of existence. However, there are also numerous viral pathogens that present serious threats to health care and economy. Examples are influenza A virus, measles virus, rinderpest virus or ebolavirus (ssRNA-), as well as rotaviruses and bluetongue virus (dsRNA). The most abundant viruses with a significant impact on society, however, are those with ssRNA+ genomes and include poliovirus, foot-and-mouth disease virus, tobacco mosaic virus, hepatitis C virus, yellow fever virus, dengue fever virus, noroviruses, or severe acute respiratory syndrome (SARS) coronavirus, to name only few. The focus of this thesis lies mainly, but not exclusively, on picorna- and nidoviruses.

Picornaviruses are among the most extensively studied and best characterized viruses. The first human virus (poliovirus) and the first animal virus (foot-and-mouth disease virus) to be discovered are picornaviruses. They employ small, non-enveloped virions and a ssRNA+ genome of around 6.5-9 kb length (Fig. 1A). The genome directly acts as an mRNA to encode (with few exceptions) a single polyprotein that is cleaved into a dozen or so mature proteins[480]. Among these viral proteins are three to four virion proteins, a superfamily 3 helicase with putative activity (denoted 2C in Fig. 1A), a chymotrypsin-related proteinase (3C), a RNA-dependent RNA polymerase (RdRp; 3D), and several other proteins that are deployed mainly to interact with the host environment. Picornaviruses can cause various, mostly acute diseases in animals and humans including poliomyelitis, foot-and-mouth disease, common cold, gastroenteritis, hepatitis, meningitis, myocarditis and uveitis[130]. Picornavirus studies contributed crucially to the general understanding of various aspects of (viral and cellular) molecular biology and virus-host interactions[7].

Also nidoviruses are known to infect only animals and to employ a ssRNA+ genome. They often cause fatal diseases like SARS in humans, feline infectious peritonitis in cats or infectious bronchitis of chickens and yellow head disease of prawns in livestock. Nidoviruses differ tremendously from picornaviruses in many aspects. The genome is organized in multiple ORFs (Fig. 1B), of which the biggest two are expressed directly from the genomic RNA[58] involving a ribosomal frameshifting event[56]. The other, smaller ORFs, whose number vary between three[429] to twelve[432], are expressed from subgenomic mRNAs that are synthesized by discontinuous extension during subgenome-length minus-strand synthesis[408]. Only three enzymes are common to all nidoviruses which are a chymotrypsin-related 3C-like proteinase (3CLpro), a superfamily 1 helicase (HEL1), and a RdRp. The nidoviral 3CLpro and RdRp show detectable sequence similarity to their homologs in picornaviruses[175,184,433].

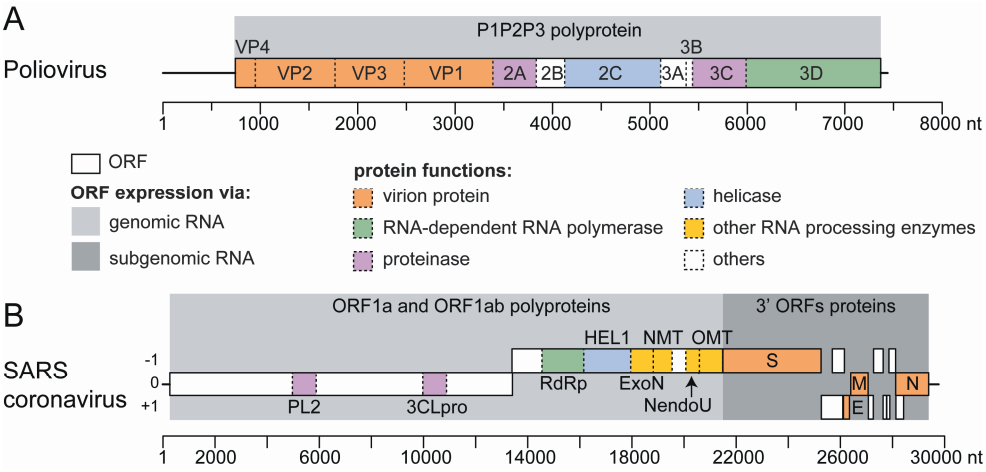

Figure 1. Genomic organization of picorna- and nidoviruses. The ssRNA+ genomes of poliovirus (A) and SARS coronavirus (B) are shown in 5'-to-3' direction and drawn to different scales. ORFs are depicted as bold rectangles and 5' and 3' untranslated regions as horizontal lines. The ORF expression mechanisms for (poly)protein production are indicated by grey background shading, and a selected set of protein functions is highlighted by coloring. An equal color does not necessarily imply sequence homology. (A) The genome of most picornaviruses expresses a single polyprotein that is cleaved into mature products by one or several viral proteinases. Picornaviruses other than poliovirus may differ in the polyprotein regions 2A and 3B and may additionally encode one or several leader proteins upstream of VP4. (B) The nidovirus genome encodes multiple, potentially overlapping ORFs in all three reading frames (-1, 0, +1). Using the 5'-proximal ORFs 1a and 1b two large polyproteins, pp1a and pp1ab, are expressed. The production of the latter involves a -1 ribosomal frameshifting event. The 3'-proximal ORFs are expressed via multiple subgenomic mRNAs to produce virion and accessory proteins. Only three enzymes - chymotrypsin-related 3C-like proteinase (3CLpro), RNA-dependent RNA polymerase (RdRp) and superfamily 1 helicase (HEL1) - are conserved across nidoviruses. Other highlighted proteins include: papain-related proteinase (PL2), exoribonuclease (ExoN), N7-methyltransferase (NMT), uridylate-specific endoribonuclease (NendoU), 2'O-methyltransferase (OMT), spike protein (S), envelope protein (E), matrix protein (M), and nucleocapsid protein (N).

Other proteins encoded by some nidoviruses have diverse functionalities including endo- and exoribonuclease, methyltransferase or phosphatase activities[174] that are rarely or never observed outside this virus group. The extraordinary genetic complexity of nidoviruses is reflected in their large genomes whose sizes are well above that of the average RNA virus genome. The smallest nidovirus genome is approximately 12.5 kb in size whereas the largest ones reach almost 32 kb which, in fact, makes them the largest RNA genomes known to date. Interestingly, nidovirus genomes of sizes above 20 kb are uniquely associated with the expression of a special enzyme, a 3'-to-5' exoribonuclease of the DEDD superfamily. This enzyme was suggested to improve the fidelity of RNA replication[432], a role which is supported by three independent lines of evidence. First, it is distantly related in sequence to a cellular proofreading enzyme. Second, the gene encoding the exoribonuclease is genetically segregated with genes that encode key enzymes of the nidovirus replication complex. And third, its functional activity was shown for mouse hepatitis

virus[124] and SARS coronavirus[123]. The extreme genome size and complex proteome composition make nidoviruses an exciting model to study the relation of genome size and mutation rate in RNA viruses[174] and, potentially, in biology in general.

Whys and wherefores of virus classification

Having in mind that tremendous diversity of RNA viruses briefly outlined in the previous section (as well as that of their DNA cousins), it comes not as a surprise that scientists aimed to structure the *Virosphere* ever since these infectious agents came to the fore in science at the end of the 19th century[30,233,301]. In order to do so, a classification has to be devised which, as a basic principle, groups together similar objects in a biologically meaningful way. A classification relies on one or several distinctive characteristics which are often referred to as *demarcation criteria*.

One such criterion is the mode of viral mRNA production. It is related to the viral genome type and defines a coarse-grained classification that consists of seven virus classes[23]. Three of them are formed by viruses with DNA genomes that produce mRNA either by direct transcription of the genome (dsDNA viruses), by involving the formation of a double-stranded intermediate (ssDNA), or by first filling their gapped genomes using reverse transcription (dsDNA with RNA intermediate). The remaining four virus classes employ RNA genomes that serve as a template for mRNA synthesis (dsRNA and ssRNA-), directly act as mRNA (ssRNA+), or require a DNA intermediate which is integrated into the host genome to subsequently undergo the cellular transcription process (retroviruses). The ssRNA+ genome type is by far the most abundant and outnumbers around fivefold each of the other types[374]. Since this classification scheme ignores most variety observed for viruses, it is suitable only for very general purposes.

The most widely used form of virus classification, as with cellular organisms, is *taxonomy*. Virus taxonomy presents a hierarchical system which comprises the ranks, from top to bottom, order, family, subfamily, genus, and species[257]. The two most crucial layers here are the family and species ranks due to different reasons. A family is used for grouping together viruses that share some rather general properties which uniquely discriminate them from other families. For each virus family there is a group of specialized virologists who propose, update or revise taxonomic entities largely independently for that particular family. These so-called *virus study groups* operate under the direction of the *International Committee on Taxonomy of Viruses* (ICTV) which, in turn, is administered by the Virology Division of the International Union of Microbiological Societies. The ICTV is the (only) official body with the legitimation to approve taxonomic assignments of viruses. According to the latest taxonomy release in 2012, 94 virus families are currently recognized by the ICTV[257]. Virus species, on the other hand, are the basic taxonomic entities[373] and imply highly similar virus phenotypes. They are defined as “a polythetic class of viruses that constitute a replicating lineage and occupy a particular ecological niche”[257]. With other words, for virus

species demarcation it is taken into consideration a broad range of characteristics, including genetic and phenotypic properties, each of which is shared by some but not necessarily all members of the species – a form of consensus decision. Currently, 2475 different virus species are recognized by the ICTV[257]. The ICTV offers an extremely powerful and flexible framework for virus classification which can accommodate any virus diversity. However, this comes with the costs of a substantial amount of scientists' labor time and a relatively lengthy decision process. Names of ICTV-defined taxa are written in italics to discriminate them from virus strains or isolates.

The expert-based ICTV classification framework might approach its logistic limitations in the near future due to the rapid increase of newly described viral genome sequences[358] (Fig.2). Consequently, ICTV study groups increasingly explore the usability of genetic data for decision making in virus taxonomy[264]. Moreover, several research groups proposed a purely genetic-based classification for a virus family or genus (see Introduction in chapter 2). They thereby exploited the fact that genome sequencing greatly outpaces all other types of virus characterization and produces high-quality data which is most suitable for quantitative analysis. A recurrent theme in these analyses is the utilization of a so-called *pairwise distance distribution* formed by genetic distances between all pairs of viruses under consideration. Often, the percentage of sequence identity is used as a measure of pairwise distance. With this sequence-based approach a classification is derived by partitioning the pairwise distance distribution into intra-rank (e.g. intra-species) and inter-rank regions. For example, all distances below a certain threshold are attributed to viruses from the same species and those above the threshold to virus pairs from different species. In doing so, it is assumed that there exist common factors that result in the separation of taxonomic ranks on a genetic level. The crucial challenge which remains is to define the number and position of distance thresholds. Preferably, this should be done in an objective manner which, however, is a goal not aimed at by current approaches.

RNA virus classification makes a difference

RNA virus classification efforts brought some insightful findings with great impact on virology. Perhaps one of the most illustrative examples is the picornavirus species *Human enterovirus C*[263] which includes the major human pathogen poliovirus (PV) as well as eleven serotypes of the rather benign C-cluster coxsackie A viruses (CCAVs). Initially, PV and CCAVs were classified as two separate species[436]. Largely due to sequence-based phylogenetic analyses, it was found that the genetic similarity of PV and CCAVs is large enough to form a single species[238,263]. In line with that are recombinants between PV and CCAVs that have been identified in the field and are mostly vaccine-dervied[21,59,390]. These findings have great implications for the Global Polio Eradication Initiative[238], a campaign to eradicate poliovirus from our planet which was initiated in 1988 but still did not succeed globally[112], and it is unclear if it ever will[331].

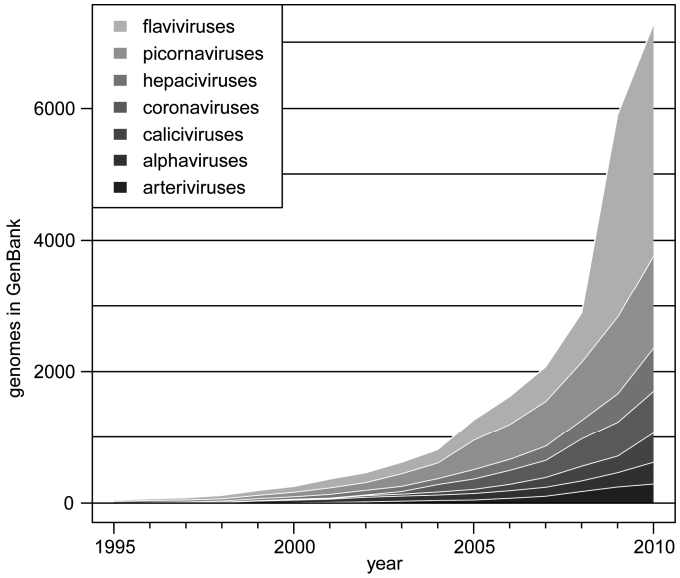

Figure 2. Non-linear increase in the number of ssRNA+ viruses with sequenced genomes. The accumulation of complete genome sequences in the VRL and PAT divisions of GenBank during the last one and a half decade is shown for selected ssRNA+ virus groups (see legend). Genome sequences have been obtained and grouped using the HAYGENS tool[423].

Another example for the potential value a virus classification can have is the species *Theilovirus*[263] that was initially thought to infect only rodents[206,343,449]. Later, phylogenetically closely related viruses have been isolated from humans and it was shown that the rodent and human lineages are genetically similar enough to be members of the same species, basing on the comparison with other picornavirus species[161,299]. This indicates that a virus species is not bound to infect a single host species, a finding supported by phylogenetic analyses of other virus families[262].

Furthermore, from a more technical perspective, a virus classification provides means for the reduction of a dataset in size in order to perform certain resource-consuming analyses. For phylogeny reconstruction (see next section), for instance, it is often necessary to chose representatives from a much larger set of viral sequences. Depending on the evolutionary scale of the analysis, virus species may present a meaningful and objective choice. In the case of picornaviruses, it allows for decreasing the dataset of available complete genomes by more than one order of magnitude from hundreds of sequences to a few dozen species. As a second effect, such a virus selection per species partially corrects for the inherent sampling bias of a dataset which can be enormous and is caused by the difference in medical or economical importance of the viruses. *Foot-and-mouth disease virus* (FMDV)[109,265] for example, one of the most economically important and best studied RNA viruses, shows the highest sampling among all picornavirus species with a few hundred

complete genomes available, whereas other species of that family, like *Seneca Valley virus* (SVV)[197], are represented only by a single sequence.

Bioinformatics meets virology

All bioinformatics analyses utilized in this thesis depend in one way or another on sequencing data which is accessible from public databases like GenBank/RefSeq[36]. In virology this type of information presents an invaluable and virtually infinite resource nowadays (Fig.2), however, one should keep in mind that it comes with certain challenges due to the nature of viruses and their elevated rate of introducing sequence innovations. Various bioinformatics techniques are available that can withstand or at least account for theses challenges, some of which have been utilized in one or several of the following chapters. They are briefly discussed below.

A commonly used tool to compare related biological sequences is a multiple sequence alignment. It is typically built by maximizing the similarity among the sequences, which involves the introduction of gaps at sequence positions that are not conserved. These gap positions predict insertion/deletion (indel) events that have happened during evolution, whereas non-gap positions infer common ancestry (homology) of the respective sequence residues. For an alignment of highly diverged sequences, like that of a typical evolutionary study of RNA viruses, it is common to partition it into conserved regions, so-called *blocks*, and poorly conserved parts[18,65]. The latter are prone to contain alignment artifacts due to an elevated substitution and/or indel rate in these regions, and, hence, may need to be discarded from subsequent analyses. Multiple sequence alignments build the foundation for many other types of bioinformatics analyses. For example, they can be converted to profiles, which are statistical models and capture for each alignment column the degree of conservation and the likelihood to observe a certain residue or gap. One type of profiles are profile HMMs[125,273,434] which operate inside a probabilistic framework and are particularly suitable for the detection of remote sequence homology. A profile HMM can be compared to other HMMs or used to search for motifs in a single sequence.

Another approach to detect or support remote homology, especially when sequence divergence erased most similarity, is structure predictions, which can be applied to both RNA and protein sequences. In the case of RNA, the folding that results in the minimal free energy is considered biologically most meaningful and desirable. A folding is defined by basepairing interactions (stems) and unpaired regions (loops), and may include pseudoknots (interaction of two or more stem-loop structures). Pseudoknots and other structural elements are often found in UTRs of RNA viruses[57,389]. For protein sequences, three basic secondary structure elements are distinguished – alpha helices, beta sheets, and loops. During prediction, each sequence residue is assigned to one of these three states, thereby taking into account physiochemical properties of the amino acids, for

instances hydrophobicity. Protein secondary structure predication can be combined with profile searches (see above) to improve the sensitivity of the latter[434].

Assuming that sequence homology is still detectable, a widely used approach to infer the relatedness of a set of sequences is phylogeny reconstruction. Typically starting from a multiple alignment, it results in an evolutionary tree that represents order and degree of sequence divergence. The tree can be in unrooted or rooted form depending on the availability of additional information, often brought by an outgroup (a sister lineage related to the sequences of interest). Importantly, certain assumptions have to be met in order to obtain meaningful results from phylogenetic analyses. It includes common ancestry of the sequences under consideration, neutrality of the vast majority of molecular changes[256], and adequate approximation of the true substitution patterns by the evolutionary model used[275]. There are several frameworks for phylogeny reconstruction, one of which is distance methods. Here, for each pair of sequences an evolutionary distance is computed and the relationships among these values are used to gradually group the sequences starting with the most closely related pair. Another popular and speedy technique is Maximum Parsimony (MP), which takes a greedy approach by seeking to find the tree that would result in the least number of substitutions throughout the phylogeny[122,152,201]. Maximum Likelihood (ML) is one of the most powerful reconstruction techniques, as being a probabilistic framework[66,145]. Under a specific substitution model, ML identifies the tree that maximizes the likelihood to observe the given set of sequences. Importantly, the distances between sequences, so-called *branch lengths*, are a parameter of the method and are thus optimized together with the branching order, the *topology*. Finally, there is Bayesian methods which are related to ML[381]. They allow for the incorporation of external knowledge, so-called *priors*. Prior knowledge could be, for instance, a known substitution rate or the monophyly of a subset of sequences. Furthermore, Bayesian methods provide confidence intervals for every parameter estimated, which represents a major advantage over the other methods. ML and Bayesian frameworks are considered more sophisticated and hence were favored throughout this thesis.

As indicated in the FMDV-SVV example above (see previous section), sampling bias is a common problem in sequence-based bioinformatics. In certain situations it may distort or even obscure signals in the data. Fortunately, such effects can be circumvented or at least diminished by using sequence weights. As a consequence, the sequences contribute unequally to the results, with most unique sequences having the highest impact. There are several approaches how to calculate these weights[157,209,472].

One of many situations where sequence weights can be of high value is regression analysis. For instance, it could be of interest to determine if there is any relation between two parameters shared by all sequences of a dataset (a naive example would be sequence length compared to G+C content). Regression analysis is a general statistical technique to model the relationship between a dependent variable and one or more independent variables. Such a relation, if any exists, can be of linear or non-linear nature and is often

referred to as *correlation*. Once a correlation has been established, regression analysis can be used to make predictions for new data points (e.g. new sequences). It is important to note that correlation does not necessarily imply causation. For example, since the middle of the last century, both atmospheric CO_2 level and obesity level increased strongly. Though, one might not infer that atmospheric CO_2 causes obesity. Instead, these two parameters could be linked by an increase in car sales.

Scientific questions

The first part of this thesis is devoted to RNA virus classification in an evolutionary context. Specifically, it is asked whether viruses of a family can be classified objectively and accurately by basing solely on genome sequences, the footprint of evolution, and whether this approach can be extended to multi-family analyses. It is further explored whether such a classification can deliver novel insight into constraints on genetic diversity in RNA viruses and what benefit for applied research in virology it can provide.

The second part is devoted to the analysis of gene and genome size change during RNA virus evolution. It is asked whether there are factors other than the fidelity of replication that limit genetic size in RNA viruses, what principal proteins are involved in the control of genetic size, and whether size change is linked to the accumulation of mutations. Moreover, it is sought to unravel biological factors that drove the emergence of the largest known RNA genomes employed by nidoviruses.