Virus Bioinformatics

Nikolaos Pappas, Utrecht University, Utrecht, The Netherlands
Simon Roux, U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA, United States
Martin Hölzer, Kevin Lamkiewicz, Florian Mock, and Manja Marz, University of Jena, Jena, Germany
Bas E Dutilh, Utrecht University, Utrecht, The Netherlands and Radboud University Medical Centre, Nijmegen, Netherlands

What is up in Virus Bioinformatics

The virosphere may contain the greatest diversity known to mankind. It has been estimated that there are 10^{31} viruses on Earth, and for billions of years their ongoing proliferation and mutation has contributed to an unparalleled genomic diversity globally. Viral mutation rates range from 10^{-8} to 10^{-6} substitutions per nucleotide per cell infection for DNA viruses and from 10^{-6} to 10^{-4} substitutions per nucleotide per cell infection for RNA viruses. The only way to efficiently analyses this biodiversity is by applying powerful computational tools to (1) identify viral sequences and their encoded functional elements, (2) predict, annotate, and compare their functions, and (3) structure the data to move from measuring to understanding. Until recently, our full understanding of viruses was based on a few hundred viruses that were isolated and could be studied in detail. With recent bioinformatic developments, thousands of new viruses can be readily discovered in all natural and host-associated biomes (see also Section "Viral Metagenomics" below). Including these naturally occurring viruses in comparative analyses opens up possibilities for *de novo* computational predictions, including about the structure and function of viral genes.

Technology and Bioinformatics Drive Discoveries

The past decades have been characterized by technological innovations that revolutionized the way we do science, ranging from the development of computers and the internet, to high-throughput measurement technologies including DNA sequencing, mass spectrometry, and imaging. New fields were built based upon these developments, including bioinformatics, machine learning, and omics. These advances have expanded the scope in all scientific fields, not least in virology. One of the most profound impacts is a new view of the virosphere that is one of an unparalleled diversity. To illustrate, the number of recognized deep viral taxonomic groups has been greatly expanded and the International Committee for Taxonomy of Viruses (ICTV) has recently approved an expansion of the resolution of the viral taxonomy to 15 ranks: realm, subrealm, kingdom, subkingdom, phylum, subphylum, class, subclass, order, suborder, family, subfamily, genus, subgenus, and species.

Bioinformatic analyses of omics and other biological datasets depend on specialized computational tools. The development of these tools begins with basic analyses that are then incrementally used to create more complex applications. Examples of basic applications include software to validate the data derived from next-generation sequencing machines, build alignments of gene or protein sequences, and perform statistical tests. Higher-level analyses may include pipelines for metagenomic analysis, genome annotation, or genotype-phenotype association. Taken together, bioinformatics is arguably one of the subdisciplines in the life sciences with the broadest applicability. When calculated as the amount of computer time allotted to computational analyses, the largest consumer in virology is the analysis of omics datasets. Omics analyses are characterized as high-throughput, untargeted, and generally quantitative, and their application opens the door to systems level analysis of viruses and their effects on their hosts. For example, comparative genomics allows thousands of viruses to be analyzed, identifying important viral genes, their functions, and their evolution; metagenomics allows viruses to be discovered and identified with high throughput; and phylogenetics and phylogenomics allow new viral taxonomic groups to be identified. Some of these applications are presented and discussed in the article below.

Tools for Diagnostics

Viral infections can form a significant burden not only for human health but also for the health livestock and plants. The direct detection of viruses in clinical and other samples include microscopy, antigen detection such as ELISA, and molecular detection of the viral genomic material by PCR. Popular molecular diagnostic techniques including qPCR or RT-qPCR also allow quantification of viral loads. While these techniques are highly sensitive for the detection of specific viruses in a sample, they can only identify viral sequences that match a pre-defined search image that matches the designed PCR primers. Thus, these established diagnostic tests frequently yield negative results when a patient presents a clinical phenotype, but no virus is detected. This can be either because an uncommon variant of a known pathogen is present in the sample, or because a novel virus is the causative agent of the disease. Notably, the difference between these two possibilities is continuous, reflecting increasing evolutionary distances along the viral phylogeny.

Bioinformatic approaches allow PCR panels to be designed that capture an increasingly diverse array of viruses, but these assays will always remain limited to detecting viruses within a known range, and cannot extrapolate to identify completely novel ones. This may be resolved by untargeted (shotgun) sequencing of isolated viruses or complete sample DNA (metagenomics). Variants of

known viruses may be detected by aligning the reads derived from the sample to the reference sequence of the known virus that was originally used for designing the primers. If enough high-quality reads span the regions where the primer sequences should anneal with the target, specialized variant detection tools can call the variant with a high degree of confidence, and new PCR primers can be designed to capture them. For example, a recent PCR-based investigation of the widespread human gut-associated bacteriophage crAssphage designed globally applicable primers by screening an alignment of sequencing reads from a range of publicly available metagenomes and identifying highly variable regions of the appropriate size (1000–1400 nucleotides) that were flanked by conserved regions which could be targeted by primers, and were present in $\geq 90\%$ of all metagenomic samples ($< 10\%$ gaps). These primers allowed a range of collaborating laboratories to independently detect crAssphage in samples from 62 different localities on six continents. Detection of completely novel viruses from metagenomic datasets is less straightforward and will be discussed below.

Genome Sequencing

Obtaining the genome sequence of viral isolates is a highly standardized approach where second (massively parallel) and third (single molecule) DNA sequencing technologies have allowed immense progress. An important first step prior to any downstream analyses based on raw sequencing data is quality control. Several metrics can be used to estimate read quality. First, the sequencing machine provides quality scores for the individual nucleotides that estimate the probability that a nucleotide was wrongly measured. These are based on the logarithmic Phred scores and range from 0 (nucleotide measured with 0% accuracy) to >40 ($>99.99\%$ accuracy). Second, several heuristics have proven useful in the identification of potentially spurious sequencing reads that may be removed from the data, including the presence of any remaining primer, index, or adapter sequences, the overrepresentation of specific nucleotide subsequences (*k*-mers), divergent GC content, and the presence of duplicates of the sequence in the sequencing dataset. There is a wide array of bioinformatic tools that have been developed to calculate these metrics and produce quality reports that summarize the results in useful graphical interfaces, such as FastQC, multiQC, and PRINSEQ (**Table 1**). Once potential issues with the data have been identified, short read sequences can be pre-processed to eliminate these sources of technical variation or errors. Typically, bases that fall under a set threshold are trimmed off along with any leftover primer or adapter sequences. Depending on the downstream application, remaining reads shorter than a length threshold are also discarded. Alternatively, dedicated tools can perform error correction on the short reads themselves.

An additional step that is specific for viral datasets is the removal of any remaining host sequences. If a genome sequence of the host is available, host-derived reads may be detected by mapping the reads against the host genome. Reads that confidently map to the host may be removed to ensure that the remaining part of the read set reflects the viral fraction of the sample. In the clinical context and for human patient samples, in particular, removal of reads mapping to the human reference is essential in order to abide to established international guidelines for safeguarding the individual's privacy. Correct identification of viral sequences that might have integrated in the host genome, as is the case with retroviruses and prophages, still remains a challenge. In these cases, viral reads may still align to the sequences of proviruses or prophages in the host genome if they are sufficiently similar, and may thus be removed from the sequencing dataset. Potential solutions include masking these proviruses in the host reference genome, or postprocessing the removed putative host reads by comparing them to another database of known viral sequences that includes the sequences of the integrated viruses. Detecting integrated viral sequences in the genome sequences of cellular hosts, and accurately determining their integration boundaries remains an ongoing bioinformatic challenge.

A typical sequencing effort targeting a viral genome with current second generation sequencing technologies will result in millions of short sequences of the order of $\sim 10^2$ nucleotides in length. Typically, these sequencing reads are generated from random fragments of the genome or genomes in the sample (hence the term "shotgun sequencing"). Because viral genomes range from $\sim 10^3$–10^5 nucleotides, for most practical purposes these short reads need to be assembled, unless a very closely related reference genome sequence is available. Sequence assembly is the process whereby short reads are combined into longer stretches of contiguous sequence (contigs). In case a single non-segmented genome is assembled, the end result optimally consists of a single string of nucleotides representing the complete genome sequence. Since most viral genomes are smaller and simpler in their structure than those of cellular organisms, assembling a full viral genome is relatively straightforward. Still, sequencing errors, repetitive and low complexity regions, and especially quasispecies diversity that results from high mutation rates may pose specific hurdles into obtaining a complete genome sequence. Several virus specific genome assemblers have been developed to address these issues including VICUNA and IVA (**Table 1**).

The latest advances in long read sequencing technologies promise high quality viral genomes. Recently, long read sequencing was shown to allow whole viral genomes to be captured in a single read, for example by direct sequencing of influenza and coronavirus genomes. Long read sequencing technologies still come with a higher error rate than their short counterparts. Hybrid approaches leveraging the advantages of long reads (ability to span low complexity and coverage regions) and short reads (low error rates) produce high quality, full viral genomes. Successful long read-based genome assemblies have been reported for the human cytomegalovirus and the pig pseudorabies virus. In principle, similar pre-processing steps apply to long as to short sequencing reads, but dedicated tools are used that take into consideration their specific limitations. Extensive quality summaries can be obtained with Poretools or nanoOK. The relatively error-prone long-read sequencing data may be corrected either without the use of additional short-read sequences (i.e., non-hybrid) or with a hybrid approach. Examples of tools performing non-hybrid error-correction include Nanocorrect and PoreSeq, while hybrid methods include Nanocorr and NaS. Similarly, non-hybrid assemblers include Canu and Miniasm, with SPAdes and Unicycler performing hybrid assemblies.

Table 1 List of selected software tools and resources for virus bioinformatics tasks

Read processing tools		
Quality check	FastQC, PRINSEQ, mulitQC	Checks read sequencing quality
	Poretools, nanoOK	Quality checks for nanopore long reads
Raw reads pre-processing	Cutadapt, Trimmomatic, BBduk	Quality trimming, artefacts removal on short reads
	Nanocorrect, PoreSeq	Non-hybrid error correction for nanopore long reads
	Nanocorr, NaS	Hybrid error correction for nanopore long reads
Genome assembly tools		
Single genomes	VICUNA	Produces population consensus genome assembly
	IVA	Assembler designed for RNA viruses
	SPAdes	Generic genome assembler
	Canu, Miniasm	Non-hybrid assemblers for nanopore long reads
	Unicycler	Hybrid assembly pipeline for nanopore long reads with the use of short reads
Metagenomes	MEGAHIT, metaSPAdes, Ray-meta, IBDA-UD	Assemblers optimized for metagenomics data
	crAss	Cross-assembly analysis of multiple metagenomes
Read mapping		
	BWA, Bowtie, BBmap	Align short read sequences to a reference
	STAR	Splice-aware aligner for RNA-seq data
	GraphMap, LAST	Align long read sequences to a reference
Gene Prediction		
	ORF Finder	Searches for open reading frames in the provided sequence
	Prodigal	A protein-coding gene prediction software tool
	VIGOR	Annotation program for small viral genomes
Similarity searches		
	BLAST	A suite of tools to find regions of similarity between DNA and protein sequences
	HHpred	Sensitive protein homology detection, function, and structure prediction
	HMMER	Homology based search
Multiple Sequence Alignment		
	MAFFT, ClustalW	Multiple sequence alignment for DNA and protein sequences
	MUSCLE	Multiple sequence alignment for protein sequences
Sequence taxonomic annotation		
	CAT, Kraken, Centrifuge, Kaiju	Assign taxonomic labels to reads or assembled contigs
Phylogenies		
	RaxML, PhyML	Inference of large phylogenetic trees
	BEAST	A software package for phylogenetic analysis with an emphasis on time-scaled trees
Taxonomy and classification		
	GRAViTy	Classification of eukaryotic viruses
	vConTACT	Classification of double stranded DNA viruses of bacteria and archaea
	VICTOR	Genome based phylogeny and classification of prokaryotic viruses
	DEmARC	Classification of viruses based on genetic divergence
RNA secondary structures		
	mfold/UNAFold	RNA secondary structure prediction
	ViennaRNA package	Suite of tools to perform RNA structures prediction and comparison
	LocARNA	Structure-guided multiple sequence alignment of RNA sequences
Transcriptomics		
	DESeq2, Sleuth	Statistical analysis of RNA-seq data
Databases		
	ViralZone	Link specific knowledge for each virus family with viral protein and genomic sequences
	Virus Variation Resource	A community portal for viral sequence data
	Virus Pathogen Database and Analysis Resource (ViPR)	An integrated repository of data and analysis tools for multiple virus families

After genome assembly, annotation is an important computational analysis that is required for interpreting the functionality of the virus in its environment. This includes a prediction of functional features such as protein-coding genes and tRNAs, and a classification step where each feature is characterized by similarity to known proteins or RNAs. Importantly, most genes predicted from novel viruses are distant from known references, and common similarity detection tools like BLAST often cannot provide relevant information in this context. Conversely, the use of more sensitive tools relying on the detection of conserved residues and the representation of protein sequence diversity as Hidden Markov Model (HMM) profiles (e.g., HMMER, HH-PRED, PSI-BLAST), is much more useful when analyzing novel virus genomes. These tools are able to detect distant homologies between distantly related viral proteins, which is often the only way to detect homology and, consequently, suggest potential functions in the light of the rapid viral sequence evolution.

RNA Secondary Structures in Viruses

RNA secondary structures play an important role in the life cycle of viruses, especially RNA viruses. These are formed either via the interaction of nucleotides located at close proximity to each other or at distances of several thousand bases (i.e., long-range RNA-RNA interactions, LRI). Local RNA structures were shown to be involved in translation initiation in Hepacivirus and Tombusvirus, while LRIs between the 5′ and 3′-UTRs of *Flaviviridae* family genomes promote replication. Moreover, a network of intra- and intersegment RNA-RNA interactions facilitates reassortment between Influenza A genomic segments from different co-infecting strains. This genomic reshuffling may have important effects, including the loss of vaccine efficacy.

Current algorithms for *in silico* prediction of RNA secondary structures mainly employ thermodynamic methods and can be applied to single sequences with software tools like mfold or its successor UNAFold. Furthermore, functional RNA secondary structures are conserved among different viral strains and species and it has been shown that this conservation is higher on the structure level than on the sequence level. Improved bioinformatic methods for in silico predictions use several different sequences and their covariances. These covariances, originating from different viruses, are used to generate structure-guided multiple sequence alignments and increase the accuracy of these predictions. Such approaches have been implemented in tools like LocaRNA. Both single and multiple sequence-based predictions can be made using various tools included in the ViennaRNA package.

The *in silico* prediction of structures is limited by some assumptions of the underlying models. First, unpaired regions of two (or more) structures that interact with each other (i.e., pseudo-knots) are usually neglected. Second, the length of the input sequence is limited. The number of all possible structures increases exponentially with the length of the RNA sequence. In other words, the longer the sequence, the less confident the *in silico* prediction. Third, interaction sites that do not follow the canonical base-pairing model are usually not included in the prediction algorithms. New specific tools and analysis pipelines are constantly being developed and existing ones improved, in order to address known challenges. *In vitro* or in vivo validation of the presence of predicted structures in viruses and their biological function is of paramount importance. To this end, the close cooperation of virologists and bioinformaticians is indispensable for achieving new knowledge in the field of secondary structures in viruses.

Viral Metagenomics

Recently, a new source of viral genomic sequences has become increasingly important. Metagenomics samples genomic material directly from the environment, allowing for the reconstruction of complete viral sequences without cultivation. Early metagenomes did not allow for the assembly of large genome fragments, mostly because of a limited capacity in sequencing depth and assembly software available at the time. Hence, most analyses focused on individual marker genes or global comparison between datasets, i.e., "all-versus-all" similarity. While providing important information on the overall genetic diversity of viruses, these gene-level analyses suffered from major limitations. Specifically, gene-based approaches can only target specific groups of viruses since no universal viral marker gene exists and are thus limited in their ability to discover novel viral diversity and draw inference at the scale of whole viral communities.

An early database independent tool for cross-metagenomic comparison at the level of sequencing reads was crAss, for cross-Assembly, an approach that exploits sequence assembly to identify shared elements in different metagenomic samples. Greater sequencing depth per sample and improved bioinformatics now enable the assembly of large genome fragments and even complete genomes from metagenomes. These genomes, termed "uncultivated virus genomes" to distinguish them from genomes obtained from virus isolates, are now becoming the primary unit of most virome analyses. The community has thus recently established a set of standards and guidelines to identify, analyze, and report these genomes of uncultivated viruses in a manuscript entitled "Minimum Information about an Uncultivated Virus Genome (MIUViG)", so that these sequences can contribute to a comprehensive mapping of viral diversity on Earth. In addition, the different tools commonly used in virus genome analysis are progressively being made available on free online data analysis platforms so that all researchers can incorporate uncultivated virus genomes into their analysis.

One of the areas that strongly depends on a thorough understanding of uncultivated viruses is viral ecology. Briefly, the first step in the analysis of viral diversity from metagenomics is to identify which of the assembled sequences are derived from viral genomes. This step is required even when processing purified viromes, where most of the data is expected to be viral, because these can still contain a substantial fraction of contaminating cellular sequences. The second requirement is to evaluate whether an assembled

contig corresponds to a fragment of a larger genome, or represents the majority and possibly the entirety of a genome sequence. This information is critical to correctly interpret these data, especially for analyses such as functional potential and taxonomic classification of the identified viruses. The current standards, comparable to the ones used for uncultivated genomes of bacteria and archaea, comprise three categories defined based on estimated genome completeness and the level of genome annotation provided: "genome fragments" are sequences representing $<90\%$ of the full genome, "high-quality draft genomes" represent $\geq 90\%$ of the genome with minimal functional annotation, and "finished genomes" are complete genomes with comprehensive annotations of the encoded functional elements. When these considerations are addressed, the relative abundance of different viral groups can then be assessed through "read mapping", i.e., the reads sequenced in a metagenome are compared to the viral genomes, and the number of reads matching each genome is interpreted as a measure of the number of copies of this genome in the initial sample. Using this approach, traditional microbial ecology approaches can be applied to assess alpha and/or beta diversity of the viral community. Moreover, species-species interaction networks can be inferred based on the correlation of viral and/or microbial groups across samples, where network nodes reflect species and edges reflect their correlated abundance or occurrence patterns across samples. Thus, these networks summarize the information in individual metagenomes that represent temporary or spatial snapshots of ecosystems. Network-based approaches have been applied in the study of viruses and the interactions with their potential hosts, but their interpretation remains challenging. Nevertheless, they can help elucidate temporal and spatial dynamics of viral diversity as well as unravel the role of viruses in important ecological processes such as the carbon cycle.

The application of metagenomic deep sequencing and *de novo* assembly for virus diagnostics allows also distantly related viruses to be identified, provided that sensitive homology detection tools are used. The unbiased, high-resolution view of the viral diversity present in a sample that is offered by metagenomics, allows the identification of viruses in patient samples directly through agnostic sequencing. Advances have been made towards the application of metagenomics outside of the research context and into the clinic, with a growing number of studies evaluating metagenomics as a tool for animal and zoonotic diseases detection. These include discoveries of new viruses associated with deaths after organ transplantations in humans, polyomaviruses associated with Merkel cell carcinomas, encephalitis-causing viruses in cattle and other difficult-to-diagnose cases. However, these studies often reveal an array of viruses in most samples, and predicting for each of them whether they impact on human health remains an unresolved issue. There is still some ground to be covered for metagenomics to be established as a routine diagnostic test. Decreasing sequencing costs, improvements in the underlying bioinformatics tools, as well as standardization of protocols and well-defined guidelines for laboratory personnel will render this option even more viable in the future.

Evolution and Phylogenetics

Phylogenetic inference based on molecular data starts with the alignment of multiple homologous sequences that allows mutations to be identified. Next, several approaches exist that transform the knowledge of the identified mutations into a phylogenetic tree, including distance-based approaches like neighbor-joining that can be rapidly calculated, and more advanced approaches that build reliable phylogenetic trees based on an evolutionary model, including maximum likelihood or Bayesian optimization. Established statistical and mathematical models of evolution and the estimation of parameters such as substitution rates, divergence times, and other population genetics patterns are incorporated in various bioinformatic software packages such as RAxML, BEAST, and PhyML that are widely used in virus bioinformatics (**Table 1**).

Phylogenetic analyses of viruses encounters several challenges. First, efforts to reconstruct the "deep phylogeny" of viruses are hampered by the inability to calculate genetic distances between the highly divergent sequences of distant families. Thus, phylogenetics and phylogenomics are most successful in the context of narrowly defined groups of related genomes, whose members share a set of core genes that allow all members to be compared in a common framework. For example, the gene encoding the RNA-dependent RNA polymerase (RdRp) is the only universal gene among RNA viruses and phylogenetic reconstructions based on this have shed light into the origins and evolution of the global RNA virome. Second, phylogenetic trees of different viral genes often yield inconsistent phylogenies due to the high frequency of genomic recombination in viruses. Conventional phylogenetic trees used for graphically representing viral phylogenies are challenged by variable evolutionary rates, lack of physical "fossil records" of viruses, confounding evolutionary relationships between viruses and their hosts, high rates of horizontal gene transfer and rampant genomic rearrangements. An alternative approach to visualizing distant relationships are genome-level networks. In this context, network nodes represent virus genomes and edges are drawn between them if they share at least one gene. Using formal analytical tools the network topology can be interrogated. Such analyses have given insights in host range of phages. Furthermore, bipartite networks may also be used to depict the links between homologous gene families and genomes.

One application of phylogenetics in virology is the study and tracking of transmission networks and epidemics. Understanding the relationships between viruses can provide us with a wealth of information about when, where and how viruses are transmitted, and in what ecological or clinical context. Phylogenetic analyses can be employed to infer phylogenetic relationships between different strains, such as building a clearer picture of a viral outbreak or for the reconstruction of the demographic history of the pathogen. Transmission networks allow the analysis of evolutionary trajectories of very recently diverged strains of the range of less than years thanks to high viral mutation rates. This enables near real-time monitoring of virus outbreaks, as was shown in the 2013–2016 West African Ebola outbreak that was monitored "live" by nanopore sequencing of 142 Ebola virus genomes from patients in Guinea. Moreover, geographic mapping of 1610 Ebola virus genomes allowed the dispersal, proliferation and decline

outbreak to be analyzed, revealing a heterogeneous and spatially dissociated epidemic consisting of different transmission clusters. In an epidemiological study of the 2013–2014 Zika virus outbreak in the Americas, molecular clock estimates suggested that its introduction into Brazil predated the 2014 World Cup soccer tournament and a canoeing event, potentially pointing to its introduction during the 2013 Confederations Cup soccer tournament. The integration of genomic, epidemiologic, and mobility data has led to the blossoming of the field of phylodynamics.

Another important application of phylogenetics in virology is taxonomy. Virus taxonomy is a field in flux. No standard automated viral taxonomy framework currently exists, but several computational tools have been developed that allow clustering of viral sequences and objective demarcation of the boundaries between taxonomic levels, such as DEmARC, vConTACT, VICTOR, and GRAViTy. In this context, the ICTV is now exploring genome-based taxonomy methods for different types of viruses, which would enable an integration of high-quality and finished metagenome-assembled virus genomes in the official taxonomy, and thus a better representation of viral diversity in the official ICTV classification. Taxonomic classification is grounded in phylogenetics, and different phylogenetic characters are suitable for distinguishing recent and ancient taxonomic groups, in accordance to their rates of evolution. The emerging consensus is that gene content methods similar to those developed in the beginning of the genomic era for cellular organisms are the method of choice for resolving ancient taxa. For defining recent taxa, alignment-based methods are appropriate for taxa with widely shared marker genes.

Virus-Host Interactions

In 1973, an early breakthrough in the understanding of co-evolution came with the definition of a law known as Red Queen, named after the Alice in Wonderland character. Applied to viruses, the law describes a co-evolutionary steady state in which hosts evade the viruses that infect them by mutating certain interaction molecules, while viruses also mutate to retain virulence. Thus, for both parties "it takes all the running [in genome sequence space] they can do, to keep in the same place". Both viruses and their hosts have evolved evolvability mechanisms that boost mutations in genomic regions containing important genes involved in virus-host interaction. For example, some bacteria and bacteriophages encode mechanisms of targeted genomic hyper-variation that diversify receptor-binding proteins (RBPs). Others encode genomic islands that contain proteins involved in cell decoration, or anti-phage defense systems that may be readily gained and lost from the genome. These and other mechanisms of accelerated evolution, combined with rapidly fluctuating selection pressures, make virus-host interaction genes among the most variable elements of bacterial and phage genomes.

Traditionally, viruses were always discovered and analyzed in the context of a host, either because the host showed symptoms of the viral infection, or because the viruses were isolated by growing them in a cell culture of their host. This has changed with the advent of metagenomics, as viral genomic sequences can be identified directly from their environment (see also Section "Viral Metagenomics" above). The direct sampling of viruses without host information complicates the interpretation of the ecological roles of viruses, including fulfillment of Koch's postulates in the case of samples from diseased organisms. Thus, host prediction is an important current challenge in understanding the role of viruses identified from metagenomics. Metagenomics may reveal sequences that are distinct from those of known viruses; thus, their hosts cannot be predicted based on similarity to viruses that have been experimentally characterized. Recent advances in machine learning hold promise for predicting virus-host interactions, including several approaches that are based on the genome sequences alone. These approaches exploit genomic signals including, for example (1) the nucleotide usage profile of the genome sequence that is adapted in viruses as a result of co-evolution with their hosts; (2) regions of sequence similarity between virus and host genomes, which could reflect integrated proviruses, horizontally transferred genes, or other mechanisms; (3) CRISPR spacers in the bacterial genome matching the genomes of bacteriophages that infected that host lineage in the past; and several other signals. Nevertheless, most virus genomes assembled from metagenomes remain without any predicted host at this point, and designing new approaches to establish these linkages remains a major computational challenge in the field.

Machine Learning as an Opportunity

As new technologies allow diverse aspects of biological systems to be measured at unprecedented scale and resolution, the rapidly increasing complexity and dimensionality of the data are increasingly challenging to interpret. Emerging analysis and processing methods based on machine learning have the ability to deal with such large, complex data sets. The main power of such approaches is their ability to identify signals and patterns in the data, enabling predictions to be made by using statistical models. Machine learning describes the process of gaining general knowledge to effectively perform a specific task by analyzing samples.

Machine learning algorithms can be divided into different types based on their strategy and the kind of problem they address, primarily including supervised and unsupervised machine algorithms. Unsupervised learning algorithms analyze high-dimensional input data for patterns without a search image. This information can be extracted by clustering the input data and can be used to determine the importance of each dimension of the input data. Common approaches including PCA and t-SNE do this by reducing the dimensionality of the input data while preserving the information. This can lead to insightful visualizations of clusters or patterns in complex data. For example, the k-means clustering algorithm is an unsupervised method that clusters the data into k different groups. This is achieved by combining the samples with the highest similarities into one group. If clear clusters

are observed, this indicates variations in the data that may need further analysis before drawing general conclusions about the complete dataset. This can help to identify potential pitfalls of an experiment represented in the data. In contrast to unsupervised learning, supervised learning approaches like random forest, gradient boosting trees and support vector machines analyze features of the input data for solutions that correspond to a pre-defined pattern represented by training data. In a genomic context, such features could include the length of the genome sequence, the amino acid distribution of the encoded proteins, the age of a sampled patient, etc. From the methodological perspective, a feature can be any property of the data that can be numerically or categorically represented. The machine learning algorithm analyses the predefined features of each sample and searches for similarities which help to solve the predefined task. The predefined task is the transformation of the input data to the desired output data. For example, given a metagenomic sequence as input, a supervised machine learning algorithm could determine whether it is derived from a virus.

A major challenge in supervised learning is to identify generalizable patterns that are predictive of new "unseen" cases. To assess the performance of such algorithms, it is good practice to split the full dataset into three parts, including (1) training data that is used to identify significant patterns; (2) validation data that is used to optimize the parameters of the machine learning algorithm; and (3) testing data that is used to assess the performance of the approach on unseen data. For fair comparison, it is important that the testing data is in no way used to tweak or optimize the procedure. The amount of data used for training, validation, and testing may differ, but could represent e.g., 80:10:10 ratio, where it is important that the data points represent independent measurements. This may be especially challenging when predicting virus-host interactions (see Section "Virus-Host Interactions" above), because the viruses with known hosts that are present in the database are highly skewed for a few well-described groups. If left unaccounted for, this database bias leads to inflated performance statistics for virological machine learning predictors.

Recently, representation learning methods have gained a lot of attention due to their extraordinary ability to represent complex information in statistical models. Approaches including deep learning and transfer learning have the advantage that very complex data can be analyzed and processed, with only a minimal requirement for the user to define features. Such approaches are ideal for analyzing big omics scale datasets, but require large amounts of training data and concomitantly heavy computing power, orders of magnitude more than the classical approaches. To summarize, machine learning enables us to analyze, understand, and evaluate the huge amounts of data that are becoming available through technological innovations.

Host Transcriptomics

Viruses only come to life after infecting their cellular host. Understanding the host response is of utmost importance for the investigation of viral infections. Once the virus enters the host system (either a living organism or a virus-responsive cell line), highly specific regulators identify the threat and then stimulate the expression of a cascade of genes. For example, in the case of an RNA virus infection, antiviral type I interferons (IFN-alpha/beta) bind to their receptors, thus activating specific transcription factors and promoting the expression of several IFN-stimulated genes (ISGs) with antiviral and immunomodulatory activity.

Today, transcriptomics (RNA-Seq) is widely used to study the host response to viral infections. To this end, total RNA extracted from, e.g., uninfected (mock) and virus-infected host cells can be reverse transcribed into cDNA (complementary DNA), fragmented, and sequenced. Short-read sequencing technologies generate a high-resolution expression profile consisting of millions of short sequencing reads that represents the composition and relative amount of RNA molecules, together making up the transcriptome. Different computational approaches are combined to build bioinformatics pipelines to process and analyze these short-read data comprehensively. In the case of an available host (or closely related) reference genome, the reads can be mapped back to identify their origin and thus the origin of their corresponding RNA expression.

If no reference genome is available, RNA-Seq data can also be assembled *de novo* and subsequently characterized, to identify differentially expressed genes. A promising avenue in transcriptomics are long-read sequencing technologies such as offered by PacBio Iso-Seq and Oxford Nanopore technologies. The latter has already been used for sequencing and investigation of full-length host and virus transcripts in their native RNA form.

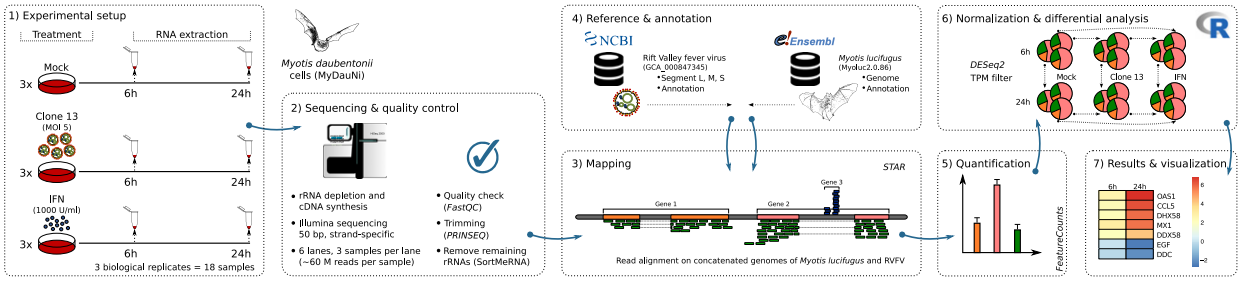

Fig. 1 Overview of an RNA-Seq bioinformatics pipeline for the identification of differentially expressed genes based on a study of an IFN-treated and virus-infected (RVFV Clone 13) bat cell line (*Myotis daubentonii*). Samples were taken in triplicates at two time points and sequenced with Illumina. Reference genomes for the virus and a close relative bat species were obtained from NCBI and Ensembl, respectively.

A gene that is more strongly transcribed during viral infection statistically yields a higher number of short reads after sequencing. Therefore, RNA-Seq not only allows the identification of transcribed host genes but also provides a quantitative value. The number of reads of the same gene derived from different conditions (e.g., mock versus infected) can be compared to identify differentially expressed genes. After normalization (taking into account different sequencing library depths and/or gene lengths), fold changes and their significance are calculated for each gene and between replicated conditions. To this end, tools such as DESeq2 and Sleuth, provided as R packages, are used to statistically evaluate the quality checked, mapped, and quantified RNA-Seq data. **Fig. 1** shows an exemplary RNA-Seq bioinformatics pipeline for the calculation of differentially expressed genes between mock, IFN-treated, and virus-infected cell lines of a microbat species (*Myotis daubentonii*) at two different times after treatment. Since no genome of this bat species was publicly available, the genome of a close relative (*M. lucifugus*) served as a reference for mapping and quantification.

Conclusions

Computers are not only indispensable to analyze data in virology, but also to store and distribute the large volumes of data generated in a reproducible way. Efforts into making the unprecedented amounts of data Findable, Accessible, Interoperable and Reusable (FAIR) should also be applicable in the field of virology. Several resources are currently available that are dedicated to viral specific sequence information and their associated metadata (**Table 1**). Currently, most analyses require a reference database, be it sequence based similarity searches for the identification of viruses, functional annotation of protein sequences or genome based phylogeny and classification. Making data publicly available in databases is important, not only as part of their general interest. As new methodologies are being developed, the available data can be further mined or reanalyzed to extract new information. A prominent example is the discovery of a highly abundant phage in the gut of humans, crAssphage; another is the recent suggestion that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. The timely deposition of genomic data and their availability to the public domain is also crucial in the epidemiological context where this information can be used for continuous surveillance, designing effective diagnostics, vaccines and antibody-based therapies.

Bioinformatics opens up a vast range of possibilities for new analyses and interpretations of viruses. While computational predictions always need to be validated by relevant in vitro experimental follow-up, the unprecedented availability of big omics datasets in the public domain already allow bioinformaticians to perform many initial validations *in silico*. These best practices can be used to estimate the accuracy of diverse bioinformatics tools, providing an important focus for wet laboratory experiments and saving valuable time and resources. Thus, bioinformatics has already become an integral and transformative component of virus research, much like techniques such as culturing, microscopy, and molecular biology have done in the past.

Further Reading

De Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., Dutilh, B.E., 2018. Molecular and evolutionary determinants of bacteriophage host range. Trends in Microbiology 27, 51–63.

Edwards, R.A., McNair, K., Faust, K., Raes, J., Dutilh, B.E., 2016. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiology Reviews 40, 258–272.

Hall, R.J., Draper, J.L., Nielsen, F.G.G., Dutilh, B.E., 2015. Beyond research: A öprimer for considerations on using viral metagenomics in the field and clinic. Frontiers in Microbiology 6, 224.

Hölzer, M., Marz, M., 2016. Differential transcriptional responses to Ebola and Marburg virus infection in bat and human cells. Scientific Reports 6, 34589.

Hölzer, M., Marz, M., 2017. Software dedicated to virus sequence analysis: "Bioinformatics goes viral". Advances in Virus Research 99, 233–257.

Jaafar, Z.A., Kieft, J.S., 2019. Viral RNA structure-based strategies to manipulate translation. Nature Reviews in Microbiology 17, 110–123.

Marz, M., Beerenwinkel, N., Drosten, C., *et al*., 2014. Challenges in RNA virus bioinformatics. Bioinformatics 30, 1793–1799.

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., Koopmans, M.P.G., 2018. Overview of virus metagenomic classification methods and their biological applications. Frontiers in Microbiology 9, 749.

Roux, S., Adriaensenss, E.M., Dutilh, B.E., *et al*., 2019. Minimum information about an uncultivated virus genome (MIUViG). Nature Biotechnology 37, 29–37.

Roux, S., Emerson, J.B., Eloe-Fadrosh, E.A., Sullivan, M.B., 2017. Benchmarking viromics: An *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ 5, e3817.

Siddell, S.G., Walker, P.J., Lefkowitz, E.J., *et al*., 2018. Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). Archives of Virology 164, 943–946.

Simmonds, P., Adams, M.J., Benkö, M., *et al*., 2017. Virus taxonomy in the age of metagenomics. Nature Reviews Microbiology 15, 161–168.

Wolf, Y.I., Kazlauskas, D., Iranzo, J., *et al*., 2018. Origins and evolution of the global RNA virome. mBio 9, e02329-18.

Relevant Websites

http://metavir-meb.univ-bpclermont.fr/
Analysis of viromes.
MetaVir.

https://www.cyverse.org/
CyVerse: Home.

https://img.jgi.doe.gov/cgi-bin/vr/main.cgi
Ecosystems.

https://talk.ictvonline.org/
International Committee on Taxonomy of Viruses (ICTV).
http://dmk-brain.ecn.uiowa.edu/pVOGs/
The pVOGs Database.
Prokaryotic Virus Orthologous Groups.
https://www.ncbi.nlm.nih.gov/genome/viruses/
Viral Genomes.
NCBI.
NIH.
https://viralzone.expasy.org/
ViralZone root.
http://viromes.org/
Viromes.
http://kbase.us/
Welcome to KBase Predictive Biology | KBase.