Virus Research 251 (2018) 86–90

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Review

A new era of virus bioinformatics

Bashar Ibrahim[a,b], Dino P. McMahon[a,c,d], Franziska Hufsky[a,b], Martin Beer[a,e], Li Deng[a,f], Philippe Le Mercier[a,g], Massimo Palmarini[h], Volker Thiel[a,i,j], Manja Marz[a,b,*]

[a] European Virus Bioinformatics Center, Jena, Germany
[b] RNA Bioinformatics and High Throughput Analysis Jena, Friedrich Schiller University Jena, Jena, Germany
[c] Host Parasite Evolution and Ecology, Institute of Biology, Free University of Berlin, Berlin, Germany
[d] Department for Materials and Environment, BAM Federal Institute for Materials Research and Testing, Berlin, Germany
[e] Institute of Diagnostic Virology, Friedrich-Loeffler-Institute, Greifswald, Germany
[f] Institute of Virology, Helmholtz Zentrum Munich, Munich, Germany
[g] Swiss-Prot Group, SIB,CMU, University of Geneva Medical School, Geneva, Switzerland
[h] MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom
[i] Federal Department of Home Affairs, Institute of Virology and Immunology, Bern and Mittelhausen, Switzerland
[j] Department of Infectious Diseases and Pathobiology, University of Bern, Bern, Switzerland

ARTICLE INFO

Keywords:
Bioinformatics
Virology
Viruses
Software

ABSTRACT

Despite the recognized excellence of virology and bioinformatics, these two communities have interacted surprisingly sporadically, aside from some pioneering work on HIV-1 and influenza. Bringing together the expertise of bioinformaticians and virologists is crucial, since very specific but fundamental computational approaches are required for virus research, particularly in an era of big data. Collaboration between virologists and bioinformaticians is necessary to improve existing analytical tools, cloud-based systems, computational resources, data sharing approaches, new diagnostic tools, and bioinformatic training. Here, we highlight current progress and discuss potential avenues for future developments in this promising era of virus bioinformatics. We end by presenting an overview of current technologies, and by outlining some of the major challenges and advantages that bioinformatics will bring to the field of virology.

1. Crosstalk between virology and bioinformatics

Viruses are the cause of a considerable burden to human and animal health (Kirk et al., 2015). In recent years, we have witnessed both the emergence of new viral diseases (e.g. MERS, SARS; see Fig. 1) and the re-emergence of known diseases in new geographical areas (e.g. Zika, Dengue and Chikungunya). The increased global risk of viral emergence is due to a variety of social, environmental and ecological factors. Climate change, deforestation, urbanization, and the unprecedented mobility of goods, people, animals and disease vectors are all elements that are facilitating the spread of viral diseases and creating potentially ideal conditions for pandemics.

The economic burden of viral diseases is enormous. The costs of all global disasters are currently estimated at 150 billion USD per year of which 30 billion USD are attributable to infectious disease outbreaks alone.[1] Viruses can also cause diseases in animals and plants. Diseases of livestock affect food security and inflict considerable economic damage. For example, annual losses due to foot-and-mouth disease are between 6.5 and 21 billion USD in endemic areas (Knight-Jones and Rushton, 2013).

Virologists have traditionally concentrated on studying viruses that cause disease in humans, animals or plants. However, there is a staggeringly large number of viruses in the biosphere (estimated to be around 10^{31}, about ten times more abundant than bacteria (Breitbart and Rohwer, 2005; Suttle, 2005; Edwards and Rohwer, 2005; Clokie et al., 2011)) and only a minuscule fraction has been identified (Mokili et al., 2012). Diverse phenomena critical to the biology of microbes have been described to be driven by viruses, especially in response to rapid environmental change (Suttle, 2005). Therefore, the view that viruses are "only" parasites is no longer valid. In the environment, viruses are able to transfer and store genetic information of their host population and influence entire biogeochemical cycles. Hence, some viruses are pathogens causing important diseases (in humans, animals or plants) but the great majority can play important roles in regulating entire ecosystems.

The field of "virology" also needs to deal with a variety of different

* Corresponding author at: European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany.
E-mail address: manja@uni-jena.de (M. Marz).

[1] http://www.who.int/csr/research-and-development/r_d_blueprint_plan_of_action.pdf.

https://doi.org/10.1016/j.virusres.2018.05.009
Received 17 April 2018; Accepted 7 May 2018
Available online 08 May 2018
0168-1702/ © 2018 Elsevier B.V. All rights reserved.

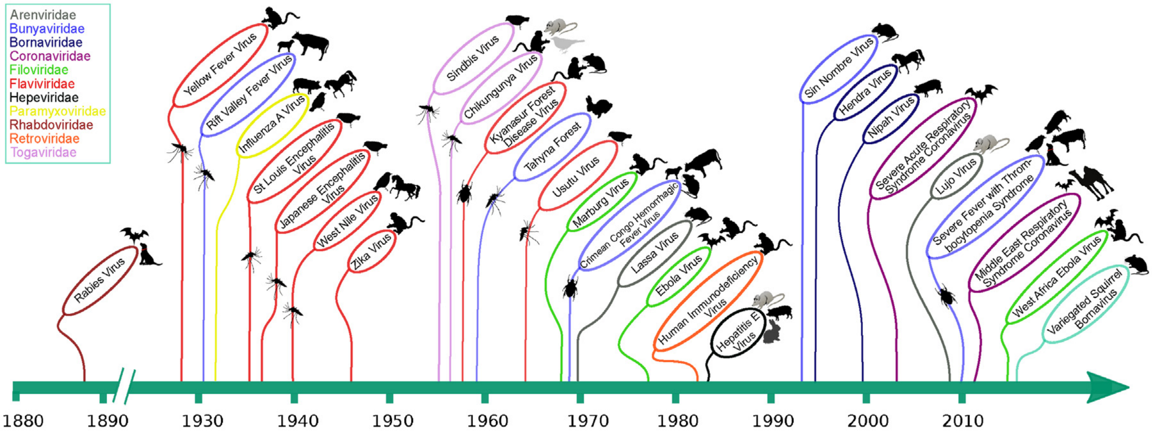

Fig. 1. Unknown/new viruses emerge all the time. Figure is an extended and redrawn version of https://www.microbiologysociety.org/publication/past-issues/zoonotic-diseases.html.

viruses with fundamentally different biological properties including their genetic organization, replication strategies, host range and host interactions (commensal, antagonists, mutualistic). Importantly, viruses evolve very rapidly and can quickly vary their genomes in response to various selective pressures including the most sophisticated control measures deployed by host (eukaryotic and prokaryotic) immune systems and/or therapeutic interventions.

There are many fundamental question in virology that need to be tackled. For example, how can we capture the full diversity of virus families in different hosts and environments? How do viruses evolve and how important is recombination in viral evolution? Is there a single common viral origin or do we find clearly independent origins? How can we identify the dynamic gene pool carried by viruses in various ecosystems? Many other questions will help us to develop strategies to control and treat viral diseases but also to understand the broader ecological role of viruses.

The power of new genome sequencing technologies, associated with new tools to handle "big data", provide unprecedented opportunities to address fundamental questions in virology. We would like to emphasize that many of the common questions raised in virology require specific bioinformatics support (Chang, 2015; Marz et al., 2014) and require the combined expertise of both bioinformaticians and virologist. This is because highly specific computational approaches are becoming absolutely necessary to address some of the key questions that we highlight here, in addition to the many other questions being addressed in virology laboratories across the globe. Approaches to tackle these important questions are discussed elsewhere (Marz et al., 2014).

2. Is bioinformatics ready to go viral?

There have been remarkably few bioinformatics communities focusing on viruses. With few exceptions, viral genomes are therefore rather poorly annotated and few computational tools and techniques have been developed specifically to analyze the idiosyncratic features of individual virus families.

Technically, the small size of viral genomes makes it possible to sequence large numbers of isolates, usually in clinical contexts, an advantage that is generally unavailable for any other living system. This flood of sequencing data in itself calls for specific methods of analysis, which so far are partially available at best. Nevertheless, the current sequencing technologies available for viral genomes pose challenges because most analysis steps are not easily automated and every method approach has its own peculiar set of technical limitations (Marz et al., 2014). However, by integrating bioinformatic methods, it could in future be possible to predict viral evolution in patients just based on individual virus population characteristics, such as whether an individual contains a low prevalence virus population with limited genetic variation. Here, the ultimate goal would be to forecast the course of a virus infection and to adjust therapeutic treatments accordingly.

Clearly there is an emerging need for an integrated workflow combining the different processing steps in viral diversity studies (Hufsky et al., 2018). Such a workflow could then assist clinicians and virologists on a daily basis to discover and characterize the underlying virus populations that are causing disease. Attempts have recently been made towards achieving this aim, some of which are listed in the following section and Table 1.

3. Virus related databases and tools

A major challenge for algorithm and software development in the big data field is the biodiversity of viruses with its coverage of multiple scales and its high complexity (Hölzer and Marz, 2017). Recently, a handful of new databases and tools have become available to virologists that will be discussed in the following section. A summary of some of these databases and tools is shown in the first column of Table 1.

Table 1
List of *selected* virus bioinformatical databases and tools. Further specific details can be found at http://evbc.uni-jena.de/tools.

Databases	De novo assembly	Secondary structure	Sequencing and annotation	Phylogenetic inference
DIGS (Database-integrated, 2017)	AV454 (Henn et al., 2012)	mfold (Zuker, 2003)	ATHLATES (Athlates, 2017)	AdaPatch (Adapatch, 2017a)
EpiFlu (Shu and McCauley, 2017)	SPAdes (Nurk et al., 2013)	LocARNA (Will et al., 2007)	GLUE (Glue, 2017)	AntiPatch (Antipatch, 2017b)
HCV (Kuiken et al., 2005)	RIEMS (Scheuch et al., 2015)	LRIscan (Fricke and Marz, 2016)	PriSM (Prism, 2017)	AntigenicTree (Antigenictree, 2017)
HIV (Druce et al., 2016)	V-FAT (Charlebois et al., 2017)	RNAalifold (Hofacker, 2007)	Tanoti (Tanoti, 2017)	
ICTV (The international, 2017)	VICUNA (Yang et al., 2012)	RNAfold (Gruber et al., 2008)		
ViPR (Pickett et al., 2012)	VrAP (Fricke et al., 2017)			
ViralZone (Hulo et al., 2011)	SOAP (Luo et al., 2012)			
VVR (Hatcher et al., 2017)				

3.1. Virus-specific databases

A few virus-specific databases exist so far for virologists (Table 1, first column), but a general database for all viruses needs to be urgently developed. ViPR database for example integrates genomes for multiple virus families belonging to the Arenaviridae, Bunyaviridae, Caliciviridae, Coronaviridae, Flaviviridae, Filoviridae, Hepeviridae, Herpesviridae, Paramyxoviridae, Picornaviridae, Poxviridae, Reoviridae, Rhabdoviridae and Togaviridae (Pickett et al., 2012). EpiFlu is currently the most complete collection of genetic sequence data of influenza viruses and related clinical and epidemiological data (Shu and McCauley, 2017). The HIV database includes genetic sequences and immunological epitope data (Druce et al., 2016) while HCV is a complete database of the Hepatitis C Virus (Kuiken et al., 2005). ViralZone provides general molecular and epidemiological information, along with virion and genome figures. Each virus or family page gives an easy access to UniProtKB/Swiss-Prot viral protein entries (Hulo et al., 2011). The virus variation resource (VVR) is a selection of web interfaces, analysis and visualization tools for virus sequence datasets (Hatcher et al., 2017).

3.2. Viral genome de novo assembly tools

There have been many tools developed for whole genome assembly (e.g. Velvet (Zerbino and Birney, 2008), ABySS (Simpson et al., 2009) or Geneious (Kearse et al., 2012)). However, these tools cannot be used for complete viral genomes, due to repetitive elements in the viral UTR regions and also a low and uneven read coverage (Peng et al., 2012). However, algorithms dedicated for single-cell sequencing, such as SPAdes (Bankevich et al., 2012) or IDBA-UD (Peng et al., 2012) work well for tested samples. Further, they outperform assembly tools such as VICUNA (Yang et al., 2012) designed for viral data. An example of contig alignment view of EBOV (Ebola) Zaire virus is shown in Fig. 2. We compare ten assembly tools based on an Illumina sequenced HuH7 cell line infected with the EBOV Zaire virus 3 h post-infection (Hölzer et al., 2016). Interestingly, this example indicates that the non-virus specific tools such as SOAPdenovo-Trans (Luo et al., 2012) perform very well in comparison to virus-related tools such as VrAP (Fricke et al., 2017).

3.3. Secondary structures in RNA viruses

The prediction of RNA structure is essential for understanding the regulatory functions performed by RNA. To date, there are many tools working on predicting RNA secondary structure. Common software tools that predict the secondary structure of an RNA are based on the calculation of the minimum free energy and can fold reliably on small local windows of up to 300 nt, like mfold (Zuker, 2003) and RNAfold (Gruber et al., 2008). Secondary structures of larger genomic segments are still computationally challenging. Viruses usually have a high mutation rate and thus provide sets of similar sequences that are ideal for large alignments and secondary structure predictions. Large genomic regions up to 800 nt can be accurately predicted based on recent tools like LocARNA (Will et al., 2007) which creates a multiple alignment based on the calculation of sequence and structure simultaneously.

3.4. Viral phylogeny

Phylogenetic trees are the most conventional graphical presentation model for viral phylogenies in the literature. This approach however faces serious challenges including variation in evolutionary rate, lack of physical “fossil records” of viruses, and confounding evolutionary relationships between viruses and their hosts. Several methods and software tools exist in the literature such as AdaPatch (Adapatch, 2017a), AntiPatch (Antipatch, 2017b) and AntigenicTree (Antigenictree, 2017). Nevertheless, phylogenetic trees still struggle to account for many virus evolutionary processes such as horizontal gene transfer, recombination, or the evolutionary relationships between viruses and their hosts. Together, there is a need for unconventional computational methods to help resolve these special aspects of virus phylogeny.

3.5. Virus evolution

Processes such genetic reassortment and recombination, exemplified by influenza viruses (Tumpey et al., 2005), but also common in nonsegmented viruses (Eden et al., 2013), can enable viruses to dramatically change their epidemiology and host range, being frequently associated with disease emergence such as in the recent case of H1N1 Swine flu (Team, 2009). Studying genome evolution is thus

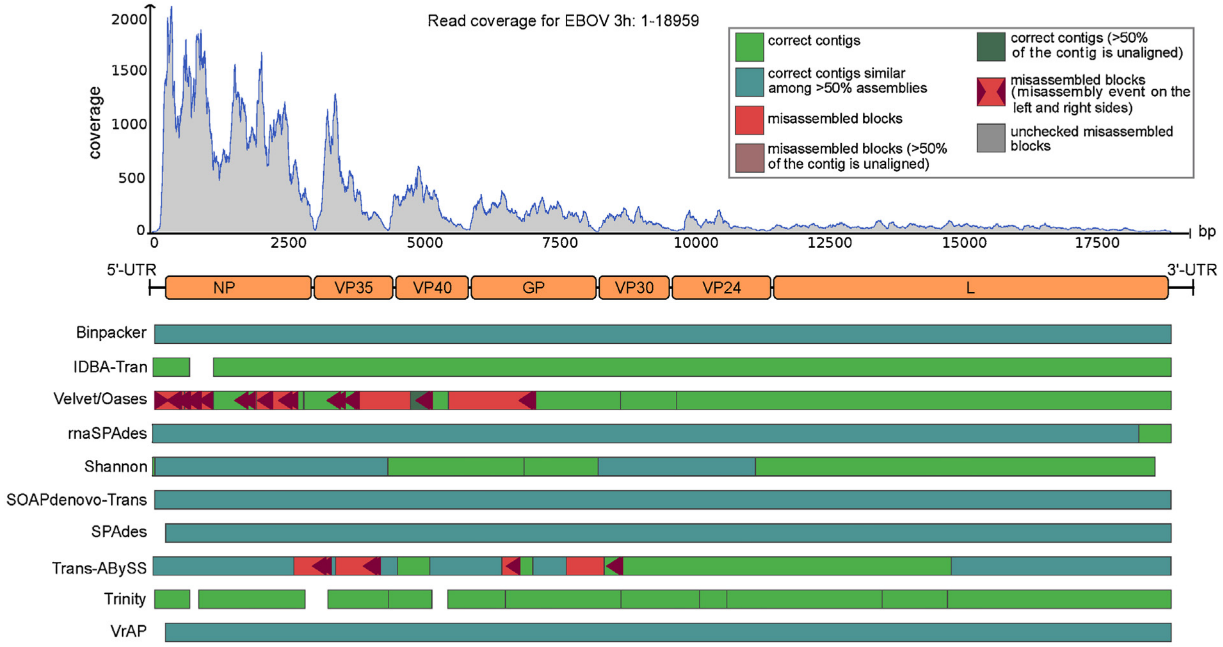

Fig. 2. Contig alignment view. Comparison of ten assembly tools based on an Illumina sequenced HuH7 cell line infected with the EBOV Zaire virus 3 h post-infection (Hölzer et al., 2016).

integral to understanding viruses and their potential to emerge in novel host species. A point to consider here is that virus evolution and host ecology are inseparable. RNA viruses in particular are among the fastest evolving biological entities with evolutionary rates of 10^3–10^{-5} substitutions per site per year (Holmes, 2008). For these viruses in contrast to their eukaryotic hosts ecological and evolutionary timescales overlap dramatically, making virus evolution integral to any study of virus biology. The extremely high mutation rates of some viruses pose particular challenges to sequencing technologies and bioinformatics, where distinguishing sequencing error from real mutations can pose problems. Novel wet-lab techniques such as Cir-seq (Acevedo et al., 2014), which links replicated reads of an original molecule through circularization and then rolling circle amplification (RCA), can lessen the impact of sequencing error. Alternatively, approaches such as BAsE-Seq (Hong et al., 2014) can produce a similar effect by transposing template-specific barcodes onto virus genomes. BAsE-Seq reduces error by assigning reads containing identical barcodes to the same read-family. This has the added feature of permitting assembly of complete haplotypes. Unfortunately, both RCA- and unique barcode-based approaches are hampered by significant limitations in yield and high sequencing waste, although a recent technique combining both methodologies could hold promise (Wang et al., 2017).

3.6. *Virus annotation and genotyping*

Genome annotation is an essential process for identifying gene locations, functions, and the coding and non-coding regions of a genome. Recently, a few virus related tools have been developed for annotation. For example, GLUE is an open-source software toolkit that can be used for the storage and interpretation of sequence data. It can be used to organize viral sequence data, even with multiple sequence alignments (MSAs). Thus, GLUE can be used as both, a resource for data and as analysis and storage platform. On the other hand, ATHLATES identifies human leukocyte antigens (HLAs) from Illumina exome sequencing data. With correctly identified HLAs one is able to determine pathogens, which can be viruses. PRiSM is a set of algorithms specifically to create primers for the amplification and sequencing of short viral genomes. The advantage of PRiSM lies in the maintaining of sample population diversity. Tanoti is a BLAST guided reference based short read aligner. It is developed for maximising alignment in highly variable next generation sequence data sets (Illumina) (Tanoti, 2017) and RotaC2.0 is automated genotyping tool which is specific for the group A rotaviruses.

4. Where to now

The future of virus bioinformatics depends on rapid specific bioinformatical software development, establishment of useful virus-specific databases and tools, and the establishment of joint interdisciplinary research projects. It also requires immediate actions, including graduate summer schools, ring trials, courses for principal investigators and annual meetings and workshops. These cannot be met by individual countries acting alone. To this end, the European Virus Bioinformatics Center (EVBC), consisting of virologists and bioinformaticians from all over Europe, was recently founded to coordinate efforts in a new era of virus bioinformatics. The EVBC hopes to fill some of the fundamental outstanding knowledge gaps facing virus research. Meanwhile, DiaMETA-net is a network of German research groups devoted to the very broad detection and characterization of pathogens (viruses, bacteria, parasites) by means of NGS.

References

Acevedo, A., Brodsky, L., Andino, R., 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505, 686–690. http://dx.doi.org/10.1038/nature12861.

Adapatch. (accessed 03.07.17). http://research.bifo.helmholtz-hzi.de/software/.

Antigenictree. (accessed 03.07.17). http://research.bifo.helmholtz-hzi.de/software/.

Antipatch. (accessed 03.07.17). https://github.com/hzi-bifo/PatchInference.

Athlates. (accessed 03.07.17). https://www.broadinstitute.org/viral-genomics/athlates.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19 (5), 455–477. http://dx.doi.org/10.1089/cmb.2012.0021.

Breitbart, M., Rohwer, F., 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13, 278–284. http://dx.doi.org/10.1016/j.tim.2005.04.003.

Chang, J., 2015. Core services: reward bioinformaticians. Nature 520, 151–152. http://dx.doi.org/10.1038/520151a.

Charlebois, P., Yang, X., Newman, R., Henn, M., Zody, M., 2017. V-FAT: A Post-assembly Pipeline for the Finishing and Annotation of Viral Genomes. (accessed 03.07.17). https://www.broadinstitute.org/viral-genomics/v-fat.

Clokie, M.R., Millard, A.D., Letarov, A.V., Heaphy, S., 2011. Phages in nature. Bacteriophage 1, 31–45. http://dx.doi.org/10.4161/bact.1.1.14942.

Database-Integrated Genome Screening (DIGS). (accessed 03.07.17). http://giffordlabcvr.github.io/DIGS-tool/.

Druce, M., Hulo, C., Masson, P., Sommer, P., Xenarios, I., Le Mercier, P., De Oliveira, T., 2016. Improving HIV proteome annotation: new features of BioAfrica HIV proteomics resource. Database: J. Biol. Databases Curation. http://dx.doi.org/10.1093/database/baw045.

Eden, J.-S., Tanaka, M.M., Boni, M.F., Rawlinson, W.D., White, P.A., 2013. Recombination within the pandemic norovirus GII.4 lineage. J. Virol. 87 (11), 6270–6282.

Edwards, R.A., Rohwer, F., 2005. Viral metagenomics. Nat. Rev. Microbiol. 3, 504–510. http://dx.doi.org/10.1038/nrmicro1163.

Fricke, M., Marz, M., 2016. Prediction of conserved long-range RNA–RNA interactions in full viral genomes. Bioinformatics (Oxf., Engl.) 32, 2928–2935. http://dx.doi.org/10.1093/bioinformatics/btw323.

Fricke, M., Zirkel, F., Drosten, C., Junglen, S., Marz, M., 2017. VrAP: full length *de novo* genome assembly of unknown RNA viruses. Nucleic Acids Res.

Glue. (accessed 03.07.17). http://tools.glue.cvr.ac.uk/#/home.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neuböck, R., Hofacker, I.L., 2008. The Vienna RNA websuite. Nucleic Acids Res. 36, W70–W74. http://dx.doi.org/10.1093/nar/gkn188.

Hölzer, M., Krähling, V., Amman, F., Barth, E., Bernhart, S.H., Carmelo, V.A.O., Collatz, M., Doose, G., Eggenhofer, F., Ewald, J., Fallmann, J., Feldhahn, L.M., Fricke, M., Gebauer, J., Gruber, A.J., Hufsky, F., Indrischek, H., Kanton, S., Linde, J., Mostajo, N., Ochsenreiter, R., Riege, K., Rivarola-Duarte, L., Sahyoun, A.H., Saunders, S.J., Seemann, S.E., Tanzer, A., Vogel, B., Wehner, S., Wolfinger, M.T., Backofen, R., Gorodkin, J., Grosse, I., Hofacker, I., Hoffmann, S., Kaleta, C., Stadler, P.F., Becker, S., Marz, M., 2016. Differential transcriptional responses to Ebola and Marburg virus infection in bat and human cells. Sci. Rep. 6, 34589. http://dx.doi.org/10.1038/srep34589.

Hölzer, M., Marz, M., 2017. Software dedicated to virus sequence analysis "bioinformatics goes viral". Adv. Virus Res. http://dx.doi.org/10.1016/bs.aivir.2017.08.004.

Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schäffer, A.A., Brister, J.R., 2017. Virus variation resource – improved response to emergent viral outbreaks. Nucleic Acids Res. 45, D482–D490. http://dx.doi.org/10.1093/nar/gkw1065.

Henn, M.R., Boutwell, C.L., Charlebois, P., Lennon, N.J., Power, K.A., Macalalad, A.R., Berlin, A.M., Malboeuf, C.M., Ryan, E.M., Gnerre, S., Zody, M.C., Erlich, R.L., Green, L.M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A.K., Dudek, T., Tully, D., Newman, R., Axten, K.L., Gladden, A.D., Battis, L., Kemper, M., Zeng, Q., Shea, T.P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Günthard, H.F., Brumme, Z.L., Brumme, C.J., Bazner, S., Rychert, J., Tinsley, J.P., Mayer, K.H., Rosenberg, E., Pereyra, F., Levin, J.Z., Young, S.K., Jessen, H., Altfeld, M., Birren, B.W., Walker, B.D., Allen, T.M., 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8, e1002529. http://dx.doi.org/10.1371/journal.ppat.1002529.

Hofacker, I.L., 2007. RNA consensus structure prediction with RNAalifold. Methods Mol. Biol. 395, 527–544.

Holmes, E.C., 2008. Evolutionary history and phylogeography of human viruses. Annu. Rev. Microbiol. 62, 307–328.

Hong, L.Z., Hong, S., Wong, H.T., Aw, P.P.K., Cheng, Y., Wilm, A., de Sessions, P.F., Lim, S.G., Nagarajan, N., Hibberd, M.L., Quake, S.R., Burkholder, W.F., 2014. BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads. Genome Biol. 15, 517. http://dx.doi.org/10.1186/PREACCEPT-6768001251451949.

Hufsky, F., Ibrahim, B., Beer, M., Deng, L., Mercier, P., McMahon, D., Palmarini, M., Thiel, V., Marz, M., 2018. Virologists – heroes need weapons. PLoS Pathog. 14 (2), e1006771.

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., Le Mercier, P., 2011. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 39 (Database issue), D576–D582. http://dx.doi.org/10.1093/nar/gkq901.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics (Oxf., Engl.) 28, 1647–1649. http://dx.doi.org/10.1093/bioinformatics/bts199.

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., Hall, A.J., Keddy, K.H., Lake, R.J., Lanata, C.F., Torgerson, P.R., Havelaar, A.H., Angulo, F.J., 2015. World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Med. 12, e1001921. http://dx.doi.org/10.1371/journal.pmed.1001921.

Knight-Jones, T., Rushton, J., 2013. The economic impacts of foot and mouth disease – what are they, how big are they and where do they occur? Prev. Vet. Med. 112 (3–4), 161–173. http://dx.doi.org/10.1016/j.prevetmed.2013.07.013.

Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics (Oxf., Engl.) 21, 379–384. http://dx.doi.org/10.1093/bioinformatics/bth485.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., Wang, J., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience 1, 18. http://dx.doi.org/10.1186/2047-217X-1-18.

Marz, M., Beerenwinkel, N., Drosten, C., Fricke, M., Frishman, D., Hofacker, I.L., Hoffmann, D., Middendorf, M., Rattei, T., Stadler, P.F., Töpfer, A., 2014. Challenges in RNA virus bioinformatics. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btu105.

Mokili, J.L., Rohwer, F., Dutilh, B.E., 2012. Metagenomics and future perspectives in virus discovery. Curr. Opin. Virol. 2, 63–77. http://dx.doi.org/10.1016/j.coviro.2011.12.004.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., McLean, J., Lasken, R., Clingenpeel, S.R., Woyke, T., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2013. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 158–170. http://dx.doi.org/10.1007/978-3-642-37195-0.

Peng, Y., Leung, H.C.M., Yiu, S., Chin, F.Y.L., 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28 (11), 1420–1428. http://dx.doi.org/10.1093/bioinformatics/bts174.

Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C.N., Dietrich, J., Klem, E.B., Scheuermann, R.H., 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 40 (Database issue), D593–D598. http://dx.doi.org/10.1093/nar/gkr859.

Prism. (accessed 03.07.17). https://www.broadinstitute.org/viral-genomics/prism.

Scheuch, M., Höper, D., Beer, M., 2015. RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. BMC Bioinf. 16, 69. http://dx.doi.org/10.1186/s12859-015-0503-6.

Shu, Y., McCauley, J., 2017. GISAID: global initiative on sharing all influenza data – from vision to reality. Euro Surveill. 22. http://dx.doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123. http://dx.doi.org/10.1101/gr.089532.108.

Suttle, C.A., 2005. Viruses in the sea. Nature 437, 356–361. http://dx.doi.org/10.1038/nature04160.

Tanoti. (accessed 03.07.17). http://www.bioinformatics.cvr.ac.uk/tanoti.php.

Team, N.S.-O.I.A.H.V.I., 2009. Emergence of a novel swine-origin influenza a (H1N1) virus in humans. N. Engl. J. Med. 360 (25), 2605–2615.

The International Committee on Taxonomy of Viruses. (accessed 03.07.17). https://talk.ictvonline.org/taxonomy/.

Tumpey, T.M., Basler, C.F., Aguilar, P.V., Zeng, H., Solórzano, A., Swayne, D.E., Cox, N.J., Katz, J.M., Taubenberger, J.K., Palese, P., et al., 2005. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science 310 (5745), 77–80.

Wang, K., Lai, S., Yang, X., Zhu, T., Lu, X., Wu, C.-I., Ruan, J., 2017. Ultrasensitive and high-efficiency screen of de novo low-frequency mutations by o2n-seq. Nat. Commun. 8, 15335.

Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F., Backofen, R., 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput. Biol. 3 (4), e65.

Yang, X., Charlebois, P., Gnerre, S., Coole, M.G., Lennon, N.J., Levin, J.Z., Qu, J., Ryan, E.M., Zody, M.C., Henn, M.R., 2012. *De novo* assembly of highly diverse viral populations. BMC Genomics 13, 475. http://dx.doi.org/10.1186/1471-2164-13-475.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using *de Bruijn* graphs. Genome Res. 18, 821–829. http://dx.doi.org/10.1101/gr.074492.107.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415.