Standard finishing categories for high-throughput sequencing of viral genomes

J.T. Ladner [(1)], J.H. Kuhn [(2)] & G. Palacios [(1)*]

(1) Center for Genome Sciences, United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, MD, United States of America
(2) Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD, United States of America
*Corresponding author: gustavo.f.palacios.ctr@mail.mil

Summary

Viral genome sequencing has become the cornerstone of almost all aspects of virology. In particular, high-throughput, next-generation viral genome sequencing has become an integral part of molecular epidemiological investigations into outbreaks of viral disease, such as the recent outbreaks of Middle Eastern respiratory syndrome, Ebola virus disease and Zika virus infection. Multiple institutes have acquired the expertise and necessary infrastructure to perform such investigations, as evidenced by the accumulation of thousands of novel viral sequences over progressively shorter time periods. The authors recently proposed a nomenclature comprised of five high-throughput sequencing standard categories to describe the quality of determined viral genome sequences. These five categories (standard draft, high quality, coding complete, complete and finished) cover all levels of viral genome finishing and can be applied to sequences determined by any technology platform or assembly technique.

Keywords

Data curation – Genome assembly – High-throughput sequencing – Medical countermeasure development – Molecular epidemiology – Population genomics – Virus.

Introduction

Rapid advances in high-throughput sequencing (HTS) technologies have drastically changed the way that genome sequencing is performed by sequencing many samples simultaneously, leading to lower cost, time and personnel requirements. The implementation of these technologies has resulted in an explosion in the number of available genome sequences. The number of institutions sequencing viral genomes has also increased substantially and continues to grow (1, 2). It is crucial for the final quality of each newly determined genome sequence to be categorised using commonly agreed-upon standards so that third parties can assess and compare independently derived datsets and determine the suitability of genome sequences for a variety of applications. Such standards have been proposed for prokaryotic and eukaryotic genomes (1). Until recently, no standards have been available for small virus genomes (≤50,000 bases), which present unique challenges for genome assembly and categorisation. The authors addressed the need for standardisation by defining a set of five genome quality categories (standard draft, high quality, coding complete, complete and finished) for viruses with small genome length, along with appropriate downstream applications (3). Assignment to a given category was determined by a simple set of criteria independent from the sequencing technology or sequence assembly method for consistent categorisation across international research groups.

doi: http://dx.doi.org/10.20506/rst.35.1.2416

The dynamic and elusive viral genome

Viruses represent the greatest source of biological diversity on earth (4, 5). Their genomes vary by length (2 kilobases to 1.2 megabases), gene content (encoding one to ~1,000 proteins), genomic organisation (1–12 segments, both linear and circular), and even the nature of their genomic material (DNA/RNA, single/double stranded, positive sense/negative sense/ambisense) (6). However, with the exception of large DNA viruses (e.g. mimiviruses, poxviruses) (6), viral genomes share several characteristics that require a different approach for sequencing compared with prokaryotic and eukaryotic genomes. One of the most obvious differences is that viral genomes are generally much shorter (two to seven orders of magnitude) than those of bacterial and eukaryotic genomes. This difference in length means that the resolution of repetitive regions, which is one of the dominant factors in defining different categories of prokaryotic and eukaryotic genome sequencing standards (1), is largely irrelevant for viral genome sequence quality.

However, despite their short length, sufficient sequencing coverage is hard to obtain because of the difficulty in separating the viral target genome from host genomic material. As virus replication is intimately tied to the host's cellular machinery, separation of the two genomes is nearly impossible. A variety of techniques has been designed to enrich viral genomic material (e.g. filtration, ultracentrifugation, viral nucleic acid amplification, host nucleic acid depletion) (7), but these procedures are imperfect and cannot always be routinely applied (e.g. within containment suites, in resource-limited settings, or in the absence of sufficient amounts of sample). Obtaining the necessary quantity of genetic material needed for high-throughput library preparation is difficult due to the short lengths of viral genomes. Thus, some form of amplification is necessary, which can lead to sequencing bias (i.e. over-representation of particular portions of the viral genome or of some constituents of the viral population). In general, the genomic segment termini and regions assuming strong secondary structures are the most difficult to resolve because of significantly lower-than-average coverage.

Moreover, viral genomes evolve much more rapidly than most prokaryote and eukaryote genomes. Low-fidelity polymerases and high replication speeds result in mutation rates ranging from 10^{-4} to 10^{-8} substitutions per nucleotide per cell infection (8). At the higher end of this mutation spectrum, each copy of a 10 kb viral genome will differ from its parent by an average of one nucleotide. Thus, any virus sample is truly a genetically diverse collection of viral genomes that is likely to change rapidly in response to a variety of selective pressures within a particular cell type, tissue or organism. Therefore, a single consensus genome sequence is an incomplete characterisation of a viral infection or stock. To obtain a full picture of the infection, population-level characterisation is necessary.

With these caveats and needs in mind, the authors have defined a set of viral genome finishing standards for the era of HTS that will enable genome quality to be easily compared across different research groups, sequencing platforms and assembly techniques. Given the short length of viral genomes, a single viral genome standard would seem sufficient; however, given the difficulties described above, completely finishing every viral genome is often not time-efficient or cost-effective. Depending on the intended use of the sequence, genome finishing may be unnecessary. The categories chosen are designed to encompass the levels of completeness most likely to be encountered in viral sequencing projects, and are defined using technology-agnostic criteria. Each viral taxon comes with its own challenges and complications (e.g. taxon-specific genomic secondary structures, repetitions or guanidine-cytosine [GC] content). Therefore, the authors provide only loose guidance on the depth of sequence coverage likely to be required to obtain different levels of finishing. In reality, a similar amount of data will generate genome assemblies with different levels of finishing for various viruses.

To alleviate reliance on particular aspects of the different sequencing technologies, the authors made two assumptions that should be valid in the majority of viral sequencing projects but which may slightly limit their applicability under certain circumstances. The first assumption is that researchers have a basic understanding of the genomic structure of the virus being sequenced, including the expected length of the genome, the number of genomic segments, and the number and distribution of major open reading frames. Although new viruses continue to be discovered, the need to establish a novel viral family, or even a novel order or genus, remains relatively uncommon (9). Combined with the fact that the basic viral genome structure is highly conserved at higher taxonomic levels, it is reasonable to assume that this type of information will be available in most sequencing projects. In the absence of such information, the defined standards can still be applied following further analyses to determine genome structure.

The second assumption is that the genetic material of the virus under study can be accurately separated from the genomes of the host and/or other microbes, either physically or bioinformatically. Depending on the technology used, it is critical that the potential for cross-contamination of samples during the sample indexing/barcoding process and sequencing procedure is addressed by appropriate internal controls and procedural methods (e.g. tag-based

identifiers, quality restrictions on index reads, support for multiple barcodes) (10). Separation of virus and host or other genomes is especially important when the target virus has a multi-segmented genome and/or occurs in diverse populations, such as in environmental samples or microbiomes. It is difficult to move beyond this second assumption without transitioning to a system based on community-level genomic profiles.

Although the initial rationale for developing these standards was for use with human viral pathogens, the categories are not taxon-specific. Therefore, these categories can be applied to all viruses, including zoonotic or epizootic viruses under the One Health concept (11). Given the economic importance of viruses, application of similar levels of precision, curation and technology is vital when categorising sequences of viruses that affect plant and animal health.

Categories for whole-genome sequencing of viruses

Figure 1 and Table I illustrate the proposed virus sequencing standard categories.

Standard draft

Standard draft (SD) is the most basic of the five proposed categories. This category is used primarily to distinguish genome assemblies from the targeted amplification of partial viral sequences, the latter of which have been commonly used in molecular epidemiology studies. SD genomes should contain at least one contig for each viral genomic segment and cover most of a virus's genome. To avoid the inclusion of short pieces of genomes as 'drafts', the authors suggest that at least 50% of the genome should

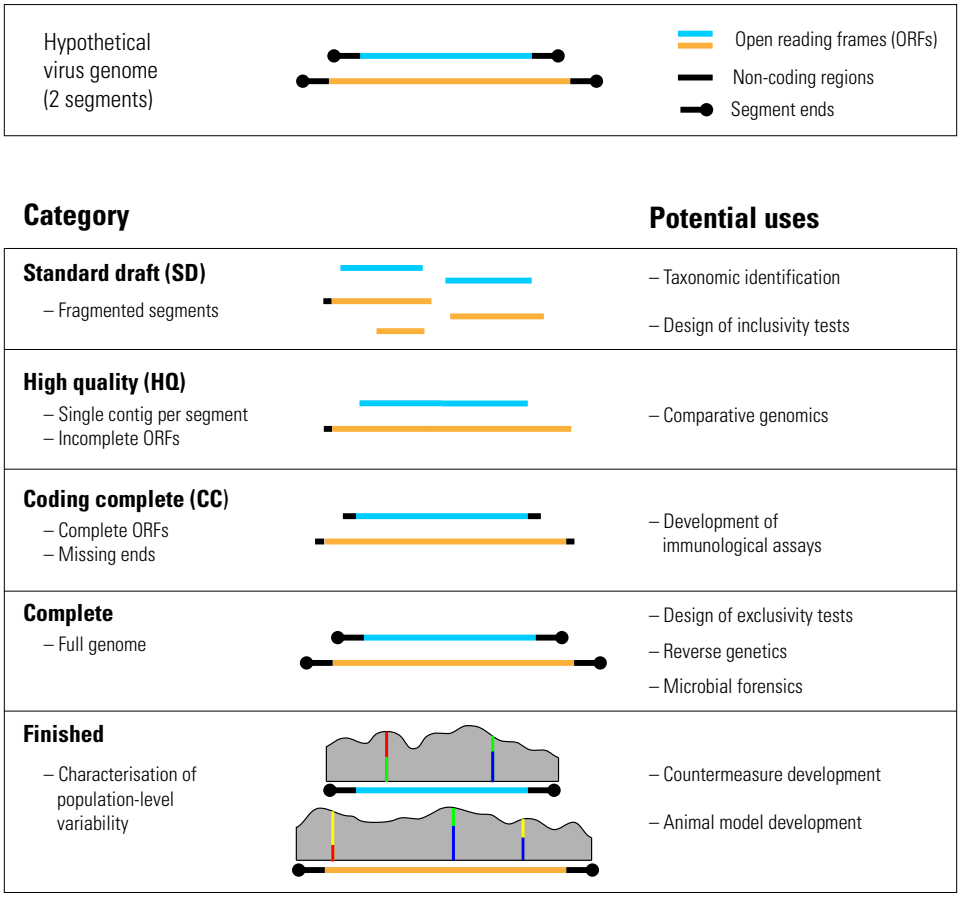

Fig. 1
Graphical representation of viral genome standards
Bullet points on the left represent primary distinctions between categories and bullet points on the right indicate potential downstream applications of genomes in each category
Source: Ladner *et al.* (3)

Table I
Viral genome standards overview
Genome percentages are estimates of the expected levels of coverage

Feature	Standard draft*	High quality*	Coding complete*	Complete	Finished
Contigs	>1 for some segments	1 per segment	1 per segment	1 per segment	1 per segment
Open reading frames	Incomplete	Incomplete	Complete	Complete	Complete
Estimated % genome covered	>50%	~80–90%	~90–99%	100%	100%
Population-level characterisation	Optional	Optional	Optional	Optional	Required
Contaminant analysis	Optional	Optional	Optional	Optional	Optional

*All bases included in an incomplete genome should minimally include ≥5 reads supporting the consensus base call with individual base qualities ≥20 on the Phred-scale

be described. SD category genomes are often generated from whole genome shotgun datasets from clinical and environmental samples (12), which are likely to contain low viral titres, and for genomes with regions that are difficult to sequence due to secondary structure (e.g. internal ribosome entry sites) (13).

High quality

Sequenced viral genomes are classified as high quality (HQ) when each genome segment is represented by a single contig, i.e. when the resulting sequences do not contain gaps. However, individual open reading frames may be incompletely sequenced at their termini. HQ genomes may be established via Sanger sequencing of an SD genome or with a medium amount of HTS coverage, often around 15 to 30× read depth.

Coding complete

Coding-complete (CC) genomes are now frequently demanded by the International Committee on Taxonomy of Viruses for virus classification. CC genomes are HQ genomes with all open reading frames fully resolved. Achieving the CC category may require targeted sequencing of open reading frame termini with conserved polymerase chain reaction (PCR) primers, but may also be possible with high levels of HTS coverage (often >100× read depth).

Complete

Complete genomes are CC genomes for which all non-coding sequences have been determined, in addition to gap-less coding sequences. Techniques such as rapid amplification of complementary DNA ends (RACE) typically need to be performed to determine the terminal sequences.

Finished

Finished genomes provide a complete overview of a viral population in a single stock preparation. The genome is complete, but genomic diversity has also been determined at the population level (typically requires 400–1,000× coverage).

Other applications of high-throughput sequencing technologies for viral genome characterisation

High-throughput sequencing technologies also provide detailed characterisation of viral samples in ways that have largely been unavailable until now. These in-depth characterisations can be performed on samples with genomes finished to any of the consensus categories defined above (although, typically, when in-depth characterisation is required, it will also be important to obtain a consensus genome sequence from one of the top three categories, i.e. coding complete, complete or finished).

Population-level characterisation

One of the more exciting applications of HTS is to describe the genetic diversity present within a viral population, especially when the population will subsequently be subjected to different types of selective pressures. Such characterisation is crucial for understanding viral evolution (see below) and studying virulence and pathogenesis (14, 15). Moreover, these types of studies are integral to the understanding of the development of viral resistance to antiviral drugs (16) or of viral escape (17). Very high levels of sequencing coverage are required to describe population-level genetic diversity (18, 19); the exact level depends on the background error profiles of the sequencing technology used and the desired level of sensitivity in the detection of minority viral populations. For instance, approximately 400× and 1,000× coverage should be achieved to identify

viral minor variants present at a frequency of 1% and 0.5%, respectively, with 99.999% confidence using pyrosequencing data (18). To accomplish such coverage, viral genomes need to be amplified or enriched in a targeted manner. Typically, population-level analysis is performed by characterising unphased single nucleotide polymorphisms; however, single-molecule technologies with read lengths of >20 kb are beginning to enable complete genome haplotype phasing (20).

Contamination analysis

Newly isolated viruses are usually grown in bulk stock tissue cultures for later experimental use. Unfortunately, stocks can easily become contaminated with other organisms either through the source material or through experimental oversight. A careful contamination analysis of viral stocks is warranted because contaminants may influence cellular responses to infection, replication and disease in experimentally infected animals during development of medical countermeasures. Luckily, HTS enables the genomic sequencing of both the target organism and (in combination with random amplification) that of any other organism in the same sample, thus enabling identification of contaminants. Contamination analysis can be performed with a low depth of coverage to identify major, high-abundance contaminants (e.g. mycoplasma, a common tissue-culture contaminant) or with a high depth of coverage to ensure detection of low-abundance contaminants.

Recommended downstream application standards

Novel virus description

At the time of writing, fewer than 4,000 viruses have been described and classified (21). This number is far lower than the actual number of viruses on earth, particularly when one considers that almost any multicellular organism is or can be infected by a specific virus. The discrepancy between the number of known and yet to be discovered viruses is best exemplified by recent metagenomic studies demonstrating viral diversity in sea-water samples (4, 5, 22, 23, 24). In addition, genetic drift, recombination and reassortment will lead to the evolution of novel viruses (9, 25, 26, 27, 28). This evolution, together with the fact that most currently known viruses are not yet genomically characterised, means that novel virus description will remain an essential task of genomic sequencing for many more years to come. In particular, this sequencing should include viruses not known to be economically, agriculturally or medically 'important' because these agents may turn out to be the emergent pathogens of tomorrow (29, 30, 31). Comprehensive and thoroughly annotated virus genome reference databases are required for the continued characterisation of viral diversity. The importance of an unknown sequence will typically not be recognised in the absence of near neighbours with significantly conserved sequence similarity (32). Indeed, given the overall paucity of viral reference sequences in American, Asian and European databases, virus identification based on sequence similarity will fail for 10–90% of sequences from metagenomic datasets (33). The sheer number of novel viruses still to be discovered and the need to quickly fill the virological sequence space in reference databases pose a dilemma. More complete finishing of genomic sequences requires more time, better technology, more experienced bioinformaticists and more experimentation; however, faster and cheaper approaches could lead to lower quality or partially incomplete sequences. The authors recommend using CC genomes to populate databases because most comparative analyses are based on the nucleotide or deduced amino acid sequences of open reading frames and very rarely depend on the sequence of genomic termini. Complete genomes are desirable for reference viruses of completely new taxa and if downstream experimentation, such as reverse genetics, is planned.

Molecular epidemiology

Viral molecular epidemiology is the process of determining viral transmission patterns by sequencing and comparing viral genomes, ideally in real time, in thousands of clinical samples. Early studies attempted to track virus spread and evolution using targeted short pieces of viral genomes. However, these studies often failed because, in general, few genomic substitutions occur between transmission events, resulting in the appearance of genetic homogeneity when only a small portion of the virus genome is analysed. The Broad Institute and the J. Craig Venter Institute have pioneered the database deposition of large numbers of CC-type viral sequences that typically suffice for epidemiological enquiries. Similar recent efforts have greatly helped to reconstruct transmission chains and to understand intercountry- and within-country spread of Ebola virus during the recent 2013–2016 Ebola virus disease outbreak across Western Africa (12, 34, 35, 36, 37, 38).

Medical countermeasure development

High-throughput sequencing is highly informative for medical countermeasure development, as a spreading pathogen can be identified in real time based on genomic signature alignment with reference sequences (genomics-based diagnostics). Such genomics-based diagnostics are often successful with SD or HQ-level genomes and may be complemented by standard, targeted PCR. CC genomes can inform the development of immunological/serological diagnostics (e.g. enzyme-linked immunosorbent assays,

immunofluorescence assays). Obtaining complete genomes may aid the design of exclusivity tests, the establishment of reverse genetics capabilities or the design of robust forensics protocols to determine the origin of, for instance, a deliberately spread virus. HTS can also be used to predict whether certain existing candidate vaccines or therapeutics may be efficacious against a novel pathogen, or whether candidate countermeasure escape mutants are or could be evolving (17). Finished genomes with concomitant contamination analysis of viral challenge stocks used for animal experimentation and viruses from post-exposure samples will greatly help to address these concerns.

Viral genome repositories and data curation

Researchers should be aware that routinely submitting raw sequencing reads to databases such as GenBank is of paramount importance to the bioinformatics community to independently verify, and thereby improve, the quality of an assembly. The authors suggest that the community consider implementing Wiki-like crowd-sourcing strategies for genome assembly similar to those already adopted for specific genomes of high interest (39, 40). Likewise, Wiki-like genomic discussion forums such as http://virological.org should be further publicised and expanded. Such Wiki-like international discussion of viruses may also lead to the automatic, non-authority-driven evolution of virus-isolate naming standards as a first step of database entry annotation. For instance, non-centralised, community-wide discussions have helped establish the current GenBank filovirus isolate naming standard that is now under adoption for other viruses (41, 42).

Although these standards were developed primarily with human pathogens and emerging infectious diseases in mind, these same standards are equally applicable to all animal and plant viruses.

Acknowledgements

This work was partly funded by Defense Threat Reduction Agency Project No. 1881290 and by Battelle Memorial Institute's prime contract with the US National Institute of Allergy and Infectious Diseases under Contract No. HHSN272200700016I. A subcontractor who performed this work is JHK, an employee of Tunnell Government Services. The authors thank Laura Bollinger for editing this manuscript.

Catégories de référence pour la finition du séquençage à haut débit des génomes viraux

J.T. Ladner, J.H. Kuhn & G. Palacios

Résumé

Le séquençage des génomes viraux est devenu la pierre angulaire de pratiquement toutes les facettes de la virologie. En particulier, le séquençage à haut débit de nouvelle génération est désormais une partie intégrante des enquêtes d'épidémiologie moléculaire relatives aux foyers de maladies virales, par exemple les récentes épidémies du syndrome respiratoire du Moyen-Orient, la maladie due au virus Ebola ou l'infection par le virus Zika. Nombre d'institutions ont acquis les compétences techniques et les infrastructures nécessaires pour réaliser ce type d'enquêtes, comme en témoigne l'accumulation de milliers de séquences virales nouvelles obtenues en un laps de temps de plus en plus court. Les auteurs ont récemment élaboré une nomenclature constituée de cinq catégories de référence décrivant la qualité des séquences d'un génome viral obtenues par séquençage à haut débit. Ces cinq catégories (ébauche de référence, séquence de haute qualité, séquence codante complète, séquence complète et séquence finie) couvrent toutes les étapes de la finition du génome

viral et s'appliquent quelle que soit la plateforme technologique ou la technique d'assemblage utilisée pour déterminer la séquence.

Mots-clés
Assemblage de génomes – Élaboration de contre-mesures médicales – Épidémiologie moléculaire – Génomique des populations – Organisation des données – Séquençage à haut débit – Virus.

Categorías de referencia relativas al acabado de la secuenciación de alto rendimiento de genomas víricos

J.T. Ladner, J.H. Kuhn & G. Palacios

Resumen
La secuenciación del genoma vírico se ha erigido a día de hoy en la piedra angular de casi todos los aspectos de la virología. La secuenciación de alto rendimiento de próxima generación, en particular, es ahora un componente integral de las investigaciones de epidemiología molecular sobre brotes de enfermedades víricas como los registrados últimamente de síndrome respiratorio de Oriente Medio, enfermedad por el virus del Ebola o infección por el virus Zika. Numerosas instituciones se han dotado de las competencias técnicas y la infraestructura necesaria para llevar a cabo tales investigaciones, como deja patente la acumulación de miles de nuevas secuencias víricas en periodos de tiempo cada vez más cortos. En fechas recientes los autores han propuesto una nomenclatura compuesta de cinco categorías de referencia que sirven para describir la calidad de las secuencias de genoma vírico determinadas por secuenciación de alto rendimiento. Estas cinco categorías (borrador normal, gran calidad, codificación completa, completa y acabada) cubren toda la gradación de acabados en la secuenciación de genoma vírico y pueden ser aplicadas a las secuencias obtenidas por cualquier dispositivo técnico o cualquier técnica de ensamblaje.

Palabras clave
Elaboración de contramedidas médicas – Ensamblaje del genoma – Epidemiología molecular – Genómica de poblaciones – Mantenimiento y organización de datos – Secuenciación de alto rendimiento – Virus.

References

1. Chain P.S.G., Grafham D.V., Fulton R.S., FitzGerald M.G., Hostetler J., Muzny D., Ali J., Birren B., Bruce D.C., Buhay C., Cole J.R., Ding Y., Dugan S., Field D., Garrity G.M., Gibbs R., Graves T., Han C.S., Harrison S.H., Highlander S., Hugenholtz P., Khouri H.M., Kodira C.D., Kolker E., Kyrpides N.C., Lang D., Lapidus A., Malfatti S.A., Markowitz V., Metha T., Nelson K.E., Parkhill J., Pitluck S., Qin X., Read T.D., Schmutz J., Sozhamannan S., Sterk P., Strausberg R.L., Sutton G., Thomson N.R., Tiedje J.M., Weinstock G., Wollam A., Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium & Detter J.C. (2009). – Genome project standards in a new era of sequencing. *Science*, **326** (5950), 236–237. doi:10.1126/science.1180614.

2. Brister J.R., Bao Y., Kuiken C., Lefkowitz E.J., Mercier P.L., Leplae R., Madupu R., Scheuermann R.H., Schobel S., Seto D., Shrivastava S., Sterk P., Zeng Q., Klimke W. & Tatusova T. (2010). – Towards viral genome annotation standards. Report from the 2010 NCBI annotation workshop. *Viruses*, **2** (10), 2258–2268. doi:10.3390/v2102258.

3. Ladner J.T., Beitzel B., Chain P.S.G., Davenport M.G., Donaldson E., Frieman M., Kugelman J., Kuhn J.H., O'Rear J., Sabeti P.C., Wentworth D.E., Wiley M.R., Yu G.Y., Threat Characterization Consortium, Sozhamannan S., Bradburne C. & Palacios G. (2014). – Standards for sequencing viral genomes in the era of high-throughput sequencing. *mBio*, **5** (3), e01360-14. doi:10.1128/mBio.01360-14.

4. Suttle C. (2005). – The viriosphere: the greatest biological diversity on earth and driver of global processes. *Environ. Microbiol.*, **7** (4), 481–482. doi:10.1111/j.1462-2920.2005.803_11.x.

5. Culley A.I. (2006). – Metagenomic analysis of coastal RNA virus communities. *Science*, **312** (5781), 1795–1798. doi:10.1126/science.1127404.

6. Pérez-Ruiz M., Navarro-Marí J.M., Sánchez-Seco M.P., Gegúndez M.I., Palacios G., Savji N., Lipkin W.I., Fedele G. & de Ory-Manchón F. (2012). – Lymphocytic choriomeningitis virus–associated meningitis, southern Spain. *Emerg. Infect. Dis.*, **18** (5), 855–858. doi:10.3201/eid1805.111646.

7. Hall R.J., Wang J., Todd A.K., Bissielo A.B., Yen S., Strydom H., Moore N.E., Ren X., Huang Q.S., Carter P.E. & Peacey M. (2014). – Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods*, **195**, 194–204. doi:10.1016/j.jviromet.2013.08.035.

8. Sanjuan R., Nebot M.R., Chirico N., Mansky L.M. & Belshaw R. (2010). – Viral mutation rates. *J. Virol.*, **84** (19), 9733–9748. doi:10.1128/JVI.00694-10.

9. Woolhouse M., Scott F., Hudson Z., Howey R. & Chase-Topping M. (2012). – Human viruses: discovery and emergence. *Philos. Trans. Roy. Soc. Lond., B, Biol. Sci.*, **367** (1604), 2864–2871. doi:10.1098/rstb.2011.0354.

10. Kircher M., Sawyer S. & Meyer M. (2012). – Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, **40** (1), e3–e3. doi:10.1093/nar/gkr771.

11. Webster J.P., Gower C.M., Knowles S.C.L., Molyneux D.H. & Fenton A. (2016). – One Health: an ecological and evolutionary framework for tackling neglected zoonotic diseases. *Evol. Appl.*, **9** (2), 313–333. doi:10.1111/eva.12341.

12. Ladner J.T., Wiley M.R., Mate S., Dudas G., Prieto K., Lovett S., Nagle E.R., Beitzel B., Gilbert M.L., Fakoli L., Diclaro J.W., Schoepp R.J., Fair J., Kuhn J.H., Hensley L.E., Park D.J., Sabeti P.C., Rambaut A., Sanchez-Lockhart M., Bolay F.K., Kugelman J.R. & Palacios G. (2015). – Evolution and spread of Ebola virus in Liberia, 2014–2015. *Cell Host Microbe*, **18** (6), 659–669. doi:10.1016/j.chom.2015.11.008.

13. Sánchez A.B. & de la Torre J.C. (2006). – Rescue of the prototypic arenavirus LCMV entirely from plasmid. *Virology*, **350** (2), 370–380. doi:10.1016/j.virol.2006.01.012.

14. Domingo E., Martín V., Perales C., Grande-Pérez A., García-Arriaza J. & Arias A. (2006). – Viruses as quasispecies: biological implications. *In* Quasispecies: concept and implications for virology (E. Domingo, ed.). Springer-Verlag, Berlin/Heidelberg, 51–82. Available at: http://link.springer.com/10.1007/3-540-26397-7_3 (accessed on 23 February 2016).

15. Vignuzzi M., Stone J.K., Arnold J.J., Cameron C.E. & Andino R. (2006). – Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, **439** (7074), 344–348. doi:10.1038/nature04388.

16. Warren T., Jordan R., Lo M., Ray A., Mackman R., Soloveva V., Siegel D., Perron M., Bannister R., Hui H., McMullan L., Chen S., Fearns R., Swaminathan S., Mayers D., Spiropoulou C., Lee W., Nichol S.T., Cihlar T. & Bavari S. (2016). – Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature*. E-pub.: 2 March. doi:10.1038/nature17180.

17. Kugelman J.R., Kugelman-Tonos J., Ladner J.T., Pettit J., Keeton C.M., Nagle E.R., Garcia K.Y., Froude J.W., Kuehne A.I., Kuhn J.H., Bavari S., Zeitlin L., Dye J.M., Olinger G.G., Sanchez-Lockhart M. & Palacios G.F. (2015). – Emergence of Ebola virus escape variants in infected nonhuman primates treated with the MB-003 antibody cocktail. *Cell Reports*, **12** (12), 2111–2120. doi:10.1016/j.celrep.2015.08.038.

18. Wang C., Mitsuya Y., Gharizadeh B., Ronaghi M. & Shafer R.W. (2007). – Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.*, **17** (8), 1195–1201. doi:10.1101/gr.6468307.

19. Macalalad A.R., Zody M.C., Charlebois P., Lennon N.J., Newman R.M., Malboeuf C.M., Ryan E.M., Boutwell C.L., Power K.A., Brackney D.E., Pesko K.N., Levin J.Z., Ebel G.D., Allen T.M., Birren B.W. & Henn M.R. (2012). – Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput. Biol.*, **8** (3), e1002417. doi:10.1371/journal.pcbi.1002417.

20. Roberts R.J., Carneiro M.O. & Schatz M.C. (2013). – The advantages of SMRT sequencing. *Genome Biol.*, **14** (6), 405. doi:10.1186/gb-2013-14-6-405.

21. King A.M.Q., Adams M.J., Carstens E.B. & Lefkowitz E.J. (eds) (2012). – Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses. Academic Press, London, Waltham, Massachusetts, 1272 pp.

22. Suttle C.A. (2013). – Viruses: unlocking the greatest biodiversity on earth. *Genome*, **56** (10), 542–544. doi:10.1139/gen-2013-0152.

23. Labonté J.M., Hallam S.J. & Suttle C.A. (2015). – Previously unknown evolutionary groups dominate the ssDNA gokushoviruses in oxic and anoxic waters of a coastal marine environment. *Front. Microbiol.*, **6**, 315. doi:10.3389/fmicb.2015.00315.

24. Chow C.E.T., Winget D.M., White R.A., Hallam S.J. & Suttle C.A. (2015). – Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front. Microbiol.*, **6**, 265. doi:10.3389/fmicb.2015.00265.

25. Anthony S.J., Epstein J.H., Murray K.A., Navarrete-Macias I., Zambrana-Torrelio C.M., Solovyov A., Ojeda-Flores R., Arrigo N.C., Islam A., Ali Khan S., Hosseini P., Bogich T.L., Olival K.J., Sanchez-Leon M.D., Karesh W.B., Goldstein T., Luby S.P., Morse S.S., Mazet J.A.K., Daszak P. & Lipkin W.I. (2013). – A strategy to estimate unknown viral diversity in mammals. *mBio*, **4** (5), e00598-13. doi:10.1128/mBio.00598-13.

26. Lipkin W.I. (2013). – The changing face of pathogen discovery and surveillance. *Nat. Rev. Microbiol.*, **11** (2), 133–141. doi:10.1038/nrmicro2949.

27. Nelson M.I. & Holmes E.C. (2007). – The evolution of epidemic influenza. *Nat. Rev. Genet.*, **8** (3), 196–205. doi:10.1038/nrg2053.

28. Palacios G., Tesh R., Travassos da Rosa A., Savji N., Sze W., Jain K., Serge R., Guzman H., Guevara C., Nunes M.R.T., Nunes-Neto J.P., Kochel T., Hutchison S., Vasconcelos P.F.C. & Lipkin W.I. (2011). – Characterization of the Candiru antigenic complex (*Bunyaviridae*: *Phlebovirus*), a highly diverse and reassorting group of viruses affecting humans in tropical America. *J. Virol.*, **85** (8), 3811–3820. doi:10.1128/JVI.02275-10.

29. Guan Y. (2003). – Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*, **302** (5643), 276–278. doi:10.1126/science.1087139.

30. Chan J.F.W., Lau S.K.P. & Woo P.C.Y. (2013). – The emerging novel Middle East respiratory syndrome coronavirus: the 'knowns' and 'unknowns'. *J. Formosan Med. Assoc.*, **112** (7), 372–381. doi:10.1016/j.jfma.2013.05.010.

31. Wang C., Wang J., Su W., Gao S., Luo J., Zhang M., Xie L., Liu S., Liu X., Chen Y., Jia Y., Zhang H., Ding H. & He H. (2014). – Relationship between domestic and wild birds in live poultry market and a novel human H7N9 virus in China. *J. Infect. Dis.*, **209** (1), 34–37. doi:10.1093/infdis/jit478.

32. Fancello L., Raoult D. & Desnues C. (2012). – Computational tools for viral metagenomics and their application in clinical research. *Virology*, **434** (2), 162–174. doi:10.1016/j.virol.2012.09.025.

33. Huson D.H., Auch A.F., Qi J. & Schuster S.C. (2007). – MEGAN analysis of metagenomic data. *Genome Res.*, **17** (3), 377–386. doi:10.1101/gr.5969107.

34. Quick J., Loman N.J., Duraffour S., Simpson J.T., Severi E., Cowley L., Bore J.A., Koundouno R., Dudas G., Mikhail A., Ouédraogo N., Afrough B., Bah A., Baum J.H.J., Becker-Ziaja B., Boettcher J.P., Cabeza-Cabrerizo M., Camino-Sánchez Á., Carter L.L., Doerrbecker J., Enkirch T., Dorival I.G., Hetzelt N., Hinzmann J., Holm T., Kafetzopoulou L.E., Koropogui M., Kosgey A., Kuisma E., Logue C.H., Mazzarelli A., Meisel S., Mertens M., Michel J., Ngabo D., Nitzsche K., Pallasch E., Patrono L.V., Portmann J., Repits J.G., Rickett N.Y., Sachse A., Singethan K., Vitoriano I., Yemanaberhan R.L., Zekeng E.G., Racine T., Bello A., Sall A.A., Faye O., Faye O., Magassouba N., Williams C.V., Amburgey V., Winona L., Davis E., Gerlach J., Washington F., Monteil V., Jourdain M., Bererd M., Camara A., Somlare H., Camara A., Gerard M., Bado G., Baillet B., Delaune D., Nebie K.Y., Diarra A., Savane Y., Pallawo R.B., Gutierrez G.J., Milhano N., Roger I., Williams C.J., Yattara F., Lewandowski K., Taylor J., Rachwal P., Turner D.J., Pollakis G., Hiscox J.A., Matthews D.A., Shea M.K.O., Johnston A.M., Wilson D., Hutley E., Smit E., Di Caro A., Wölfel R., Stoecker K., Fleischmann E., Gabriel M., Weller S.A., Koivogui L., Diallo B., Keïta S. *et al.* (2016). – Real-time, portable genome sequencing for Ebola surveillance. *Nature*, **530** (7589), 228–232. doi:10.1038/nature16996.

35. Kugelman J.R., Wiley M.R., Mate S., Ladner J.T., Beitzel B., Fakoli L., Taweh F., Prieto K., Diclaro J.W., Minogue T., Schoepp R.J., Schaecher K.E., Pettitt J., Bateman S., Fair J., Kuhn J.H., Hensley L., Park D.J., Sabeti P.C., Sanchez-Lockhart M., Bolay F.K. & Palacios G., on behalf of US Army Medical Research Institute of Infectious Diseases, National Institutes of Health & Integrated Research Facility–Frederick Ebola Response Team 2014–2015 (2015). – Monitoring of Ebola virus Makona evolution through establishment of advanced genomic capability in Liberia. *Emerg. Infect. Dis.*, **21** (7), 1135–1143. doi:10.3201/eid2107.150522.

36. Park D.J., Dudas G., Wohl S., Goba A., Whitmer S.L.M., Andersen K.G., Sealfon R.S., Ladner J.T., Kugelman J.R., Matranga C.B., Winnicki S.M., Qu J., Gire S.K., Gladden-Young A., Jalloh S., Nosamiefan D., Yozwiak N.L., Moses L.M., Jiang P.P., Lin A.E., Schaffner S.F., Bird B., Towner J., Mamoh M., Gbakie M., Kanneh L., Kargbo D., Massally J.L.B., Kamara F.K., Konuwa E., Sellu J., Jalloh A.A., Mustapha I., Foday M., Yillah M., Erickson B.R., Sealy T., Blau D., Paddock C., Brault A., Amman B., Basile J., Bearden S., Belser J., Bergeron E., Campbell S., Chakrabarti A., Dodd K., Flint M., Gibbons A., Goodman C., Klena J., McMullan L., Morgan L., Russell B., Salzer J., Sanchez A., Wang D., Jungreis I., Tomkins-Tinch C., Kislyuk A., Lin M.F., Chapman S., MacInnis B., Matthews A., Bochicchio J., Hensley L.E., Kuhn J.H., Nusbaum C., Schieffelin J.S., Birren B.W., Forget M., Nichol S.T., Palacios G.F., Ndiaye D., Happi C., Gevao S.M., Vandi M.A., Kargbo B., Holmes E.C., Bedford T., Gnirke A., Ströher U., Rambaut A., Garry R.F. & Sabeti P.C. (2015). – Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell*, **161** (7), 1516–1526. doi:10.1016/j.cell.2015.06.007.

37. Simon-Loriere E., Faye O., Faye O., Koivogui L., Magassouba N., Keita S., Thiberge J.M., Diancourt L., Bouchier C., Vandenbogaert M., Caro V., Fall G., Buchmann J.P., Matranga C.B., Sabeti P.C., Manuguerra J.C., Holmes E.C. & Sall A.A. (2015). – Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature*, **524** (7563), 102–104. doi:10.1038/nature14612.

38. Carroll M.W., Matthews D.A., Hiscox J.A., Elmore M.J., Pollakis G., Rambaut A., Hewson R., García-Dorival I., Bore J.A., Koundouno R., Abdellati S., Afrough B., Aiyepada J., Akhilomen P., Asogun D., Atkinson B., Badusche M., Bah A., Bate S., Baumann J., Becker D., Becker-Ziaja B., Bocquin A., Borremans B., Bosworth A., Boettcher J.P., Cannas A., Carletti F., Castilletti C., Clark S., Colavita F., Diederich S., Donatus A., Duraffour S., Ehichioya D., Ellerbrok H., Fernandez-Garcia M.D., Fizet A., Fleischmann E., Gryseels S., Hermelink A., Hinzmann J., Hopf-Guevara U., Ighodalo Y., Jameson L., Kelterbaum A., Kis Z., Kloth S., Kohl C., Korva M., Kraus A., Kuisma E., Kurth A., Liedigk B., Logue C.H., Lüdtke A., Maes P., McCowen J., Mély S., Mertens M., Meschi S., Meyer B., Michel J., Molkenthin P., Muñoz-Fontela C., Muth D., Newman E.N.C., Ngabo D., Oestereich L., Okosun J., Olokor T., Omiunu R., Omomoh E., Pallasch E., Pályi B., Portmann J., Pottage T., Pratt C., Priesnitz S., Quartu S., Rappe J., Repits J., Richter M., Rudolf M., Sachse A., Schmidt K.M., Schudt G., Strecker T., Thom R., Thomas S., Tobin E., Tolley H., Trautner J., Vermoesen T., Vitoriano I., Wagner M., Wolff S., Yue C. *et al*. (2015). – Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature*, **524** (7563), 97–101. doi:10.1038/nature14594.

39. Lee E., Helt G.A., Reese J.T., Muñoz-Torres M.C., Childers C.P., Buels R.M., Stein L., Holmes I.H., Elsik C.G. & Lewis S.E. (2013). – Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14** (8), R93. doi:10.1186/gb-2013-14-8-r93.

40. Winsor G.L., Lam D.K.W., Fleming L., Lo R., Whiteside M.D., Yu N.Y., Hancock R.E.W. & Brinkman F.S.L. (2011). – Pseudomonas Genome Database: improved comparative analysis and population genomics capability for *Pseudomonas* genomes. *Nucleic Acids Res.*, **39** (Suppl. 1), D596–D600. doi:10.1093/nar/gkq869.

41. Kuhn J.H., Bao Y., Bavari S., Becker S., Bradfute S., Brister J.R., Bukreyev A.A., Chandran K., Davey R.A., Dolnik O., Dye J.M., Enterlein S., Hensley L.E., Honko A.N., Jahrling P.B., Johnson K.M., Kobinger G., Leroy E.M., Lever M.S., Mühlberger E., Netesov S.V., Olinger G.G., Palacios G., Patterson J.L., Paweska J.T., Pitt L., Radoshitzky S.R., Saphire E.O., Smither S.J., Swanepoel R., Towner J.S., van der Groen G., Volchkov V.E., Wahl-Jensen V., Warren T.K., Weidmann M. & Nichol S.T. (2013). – Virus nomenclature below the species level: a standardized nomenclature for natural variants of viruses assigned to the family *Filoviridae*. *Arch. Virol.*, **158** (1), 301–311. doi:10.1007/s00705-012-1454-0.

42. Kuhn J., Andersen K., Bào Y., Bavari S., Becker S., Bennett R., Bergman N., Blinkova O., Bradfute S., Brister J., Bukreyev A., Chandran K., Chepurnov A., Davey R., Dietzgen R., Doggett N., Dolnik O., Dye J., Enterlein S., Fenimore P., Formenty P., Freiberg A., Garry R., Garza N., Gire S., Gonzalez J.P., Griffiths A., Happi C., Hensley L., Herbert A., Hevey M., Hoenen T., Honko A., Ignatyev G., Jahrling P., Johnson J., Johnson K., Kindrachuk J., Klenk H.D., Kobinger G., Kochel T., Lackemeyer M., Leroy E., Lever M., Mühlberger E., Netesov S., Olinger G., Omilabu S., Palacios G., Panchal R., Park D., Patterson J., Paweska J., Peters C., Pettitt J., Pitt L., Radoshitzky S., Ryabchikova E., Ollmann Saphire E., Sabeti P., Sealfon R., Smither S., Sullivan N., Swanepoel R., Takada A., Towner J., van der Groen G., Volchkov V., Volchkova V., Wahl-Jensen V., Warren T., Warfield K., Weidmann M., Nichol S., Fackner D. & Shestopalov A. (2014). – Filovirus RefSeq entries: evaluation and selection of filovirus type variants, type sequences, and names. *Viruses*, **6** (9), 3663–3682. doi:10.3390/v6093663.