Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Virology

Expansion of the metazoan virosphere: progress, pitfalls, and prospects

Darren J Obbard

Metagenomic sequencing has led to a recent and rapid expansion of the animal virome. It has uncovered a multitude of new virus lineages from under-sampled host groups, including many that break up long branches in the virus tree, and many that display unexpected genome sizes and structures. Although there are challenges to inferring the existence of a virus from a 'virus-like sequence', in the absence of an isolate the analysis of nucleic acid (including small RNAs) and sequence data can provide considerable confidence. As a consequence, this period of molecular natural history is helping to reshape our views of deep virus evolution.

Address
Institute of Evolutionary Biology, and Centre for Immunity, Infection and Evolution, The University of Edinburgh, Charlotte Auerbach Road, Edinburgh EH9 3FL, United Kingdom

Corresponding author: Obbard, Darren J (darren.obbard@ed.ac.uk)

Current Opinion in Virology 2018, **31**:17–23

This review comes from a themed issue on **Viral evolution**

Edited by **Gustavo Palacios** and **Jens H Kuhn**

https://doi.org/10.1016/j.coviro.2018.08.008

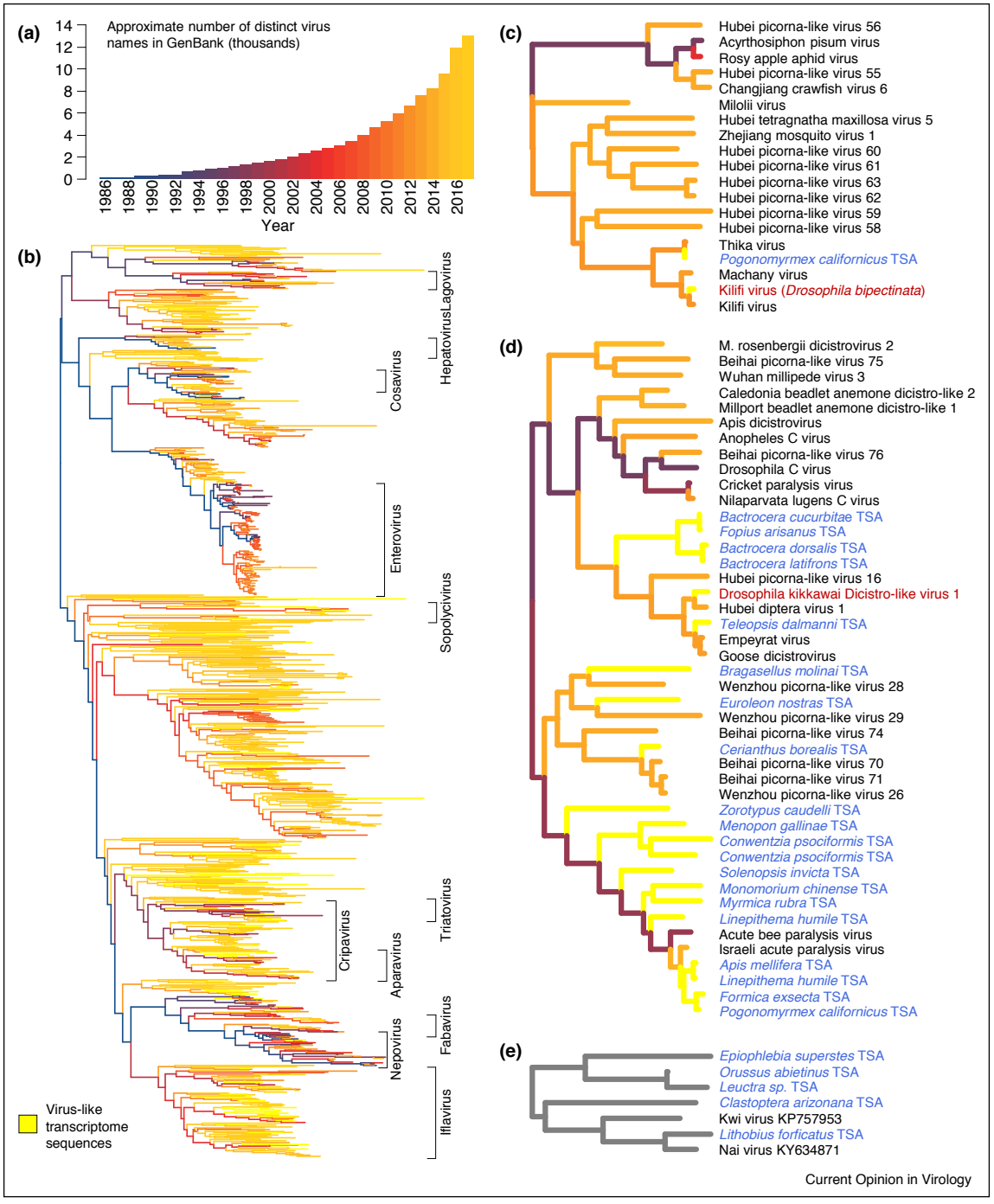
Explosive metagenomic growth

It is 120 years since the word 'virus' was first applied specifically to a viral pathogen [1], but the number of known viruses is growing faster than ever (Figure 1a; [2••]). Much of this growth is through metagenomic discovery: the undirected large-scale sequencing of nucleic acids sampled from potential hosts or their environment [2••,3,4••]. Pioneered by studies of bacteriophage in the marine environment [5], recent years have witnessed an explosion in metagenomic sampling of the metazoan virosphere. This boom has focussed first on viruses likely to infect us and our livestock, particularly the virome of mammalian faeces [e.g. 6], on putative disease reservoirs such as bats [e.g. 7,8], and on arbovirus vectors [e.g. 9]. Subsequently, the focus has expanded to include neglected animal lineages, identifying hundreds of new RNA viruses in arthropods and other invertebrates [10,11••,12,13], and recently in divergent and under-sampled chordates [14•,15].

Compared to the isolation of new virus cultures, metagenomic discovery seems (relatively) cheap, easy, and (virtually) guaranteed — sequences often appear 'for free' when sequencing genomes and transcriptomes (Figure 1b–e) [10,16–18]. Nevertheless, there are clearly limitations to metagenomic discovery — especially for important applied questions such as 'Where is the pandemic coming from?' [2••]. With an isolate in hand we would have more than just a 'virus-like sequence': we could unambiguously confirm the host, be confident we hadn't been misled by a computational artefact, and study viral replication, host range and immunity [19–21]. However, our catalogue of the virosphere is in its infancy, and there are still great gains to be made from simple 'molecular natural history'. Fewer than 5 thousand viruses have received formal taxonomic recognition [22] and only around 15 thousand have even been named informally (Figure 1a). This is less comprehensive than the 17th century view of plant diversity, even in absolute terms [ca. 18 thousand species, 23], but few biologists today would claim the naturalists of subsequent centuries wasted their effort when making herbarium collections. And a modern evolutionary virologist can probably learn more from a virus genome than a 17th century botanist could from a dried specimen.

Metagenomic discovery has already had a huge impact on our knowledge of virus diversity. It has 'filled in' shallower parts of the tree, finding close relatives of iconic human pathogens, such as new influenzas in toads and eels [14•]. It has also discovered new deep branches, such as clades of insect-infecting Partitiviruses [10,11••] and Luteo/Sobemo-like viruses [10,24], and whole new families, such as the Chuviruses [25]. This in turn has led to renewed interest in inferring deep viral phylogenies [11••,26•], and has prompted proposals for large-scale updates of higher-level virus taxonomy [27•]. More importantly, metagenomics now contributes to our thinking on virus evolution. It has provided a better perspective on host-association and host-switching [14•,28,29], found familiar virus lineages with unexpected genome sizes and structures [11••,25,30], and uncovered an unexpectedly dynamic history of 'modular' protein swapping [11••,26•]. Finally, merely having a PCR product from a metagenomic sample can provide an experimental route to the functional biology of an uncultured virus [31].

Figure 1



Panel a: The number of distinct names for viruses (excluding phage) in the GenBank nucleotide database, by year (colours provide a scale for panels b–d). Counts were obtained by finding the record creation date and GenBank 'species' (collapsing strain identifiers) for each of 2.6 million virus sequences. Exclusion of unrecognised species names and the merging of divergent strains are likely to make this an underestimate. **Panel b**: Midpoint-rooted maximum likelihood phylogeny of picorna-like viruses and caliciviruses, inferred from approximately 250 amino acids of the polymerase. Branches are coloured by the year in which the lineage was first recorded in GenBank (colours provided by panel a). Approximately 8000 picorna-like polymerase sequences from the NCBI non-redundant protein (nr) and transcriptome shotgun assembly (tsa_nt) databases were identified by blastp and tblastn. These were collapsed into 1140 clusters at a threshold of 96% identity, with one representative of each cluster used to infer the tree. Around 10% of the represented picorna-like lineages are known only as unannotated virus-like sequences from transcriptomes (pale yellow; viruses from transcriptome datasets are treated as unpublished and given a more recent date). Note that the short conserved-sequence length leads to poor resolution and fails to recover some named genera, and that similarity criteria for inclusion means that some picornavirus groups were excluded. **Panels c,d**: To illustrate with ease with which new virus-like sequences can be found in public