Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 May 2018 doi:10.20944/preprints201805.0278.v1

Expansion of the Metazoan Virosphere: Progress, Pitfalls, and Prospects

Darren J Obbard[1*]

[1]Institute of Evolutionary Biology, and Centre for Immunity, Infection and Evolution
The University of Edinburgh
Charlotte Auerbach Road
Edinburgh, EH9 3FL
*darren.obbard@ed.ac.uk

Abstract

Metagenomic sequencing has led to a recent and rapid expansion in the animal virome. It has uncovered a multitude of new virus lineages from under-sampled host lineages, including many that break up long branches among previously known clades, and many with genomes that display unexpected sizes and structures. Although there are challenges to inferring the existence of a virus from a virus-like sequence, the analysis of nucleic acid (including small RNAs) and sequence data can give us considerable confidence in the absence of an isolate. As a consequence, this period of 'molecular natural history' is helping to reshape our views of deep virus evolution. Nevertheless, there is a limit to what metagenomic discovery alone can tell us, and some open questions will require experimental isolates.

Key words: virus; metagenomic; RNAi; host-range; picornavirales; drosophila

Explosive metagenomic growth

It is 120 years since the word 'virus' was first applied specifically to a viral pathogen (Bos 1999), but the number of viruses is growing faster than ever (figure 1A; (Greninger 2018)). Much of this growth is through metagenomic discovery: the undirected large-scale sequencing of nucleic acids sampled from potential hosts or their environment (Rose et al. 2016; Simmonds et al. 2017; Greninger 2018). Pioneered by studies of bacteriophage in the marine environment (Breitbart et al. 2002), recent years have witnessed an explosion in metagenomic sampling of the metazoan virosphere. This boom focussed first on viruses likely to infect us and our livestock, particularly the virome of mammalian faeces (e.g. Williams et al. 2018), putative disease reservoirs such as bats (e.g. Berto et al. 2018; Zheng et al. 2018), and arbovirus vectors (e.g. Tokarz et al. 2018). The focus has subsequently expanded to include neglected animal lineages, identifying hundreds of new RNA viruses in arthropods and other invertebrates (Webster et al. 2015; Shi et al. 2016a; Roberts et al. 2018; Waldron et al. 2018), and recently in divergent and under-sampled chordates (Geoghegan et al. 2018; Shi et al. 2018).

Compared to isolating new virus cultures, metagenomic discovery seems cheap, easy, and (virtually) guaranteed—sequences often appear 'for free' when sequencing genomes and transcriptomes (Figure 1B-E) (Longdon et al. 2015; Webster et al. 2015; François et al. 2016; Kapun et al. 2018). There are clearly limitations to metagenomic discovery, especially for important applied questions such as "Where is the pandemic coming from?" (Greninger 2018). With an isolate in hand we would have more than just a 'virus-like sequence': we could study replication, host range, and immunity, and be confident we haven't been misled by a computational artefact (Ladner et al. 2014; Murphy 2016; van Regenmortel 2016). However, our catalogue of the virosphere is in its infancy, and there are still great gains to be made from 'molecular natural history'. Fewer than 5 thousand viruses have received formal taxonomic recognition (King et al. 2018) 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, Ray 2014), but few would claim the naturalists of subsequent centuries wasted their effort in 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. It has 'filled in' shallower parts of the tree, finding

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

close relatives of iconic human pathogens, such as new influenzas in toads and eels (Shi et al. 2018). It has also discovered new deep branches, such as clades of insect-infecting Partitiviruses (Webster et al. 2015; Shi et al. 2016a) and Luteo/Sobemo-like viruses (Tokarz et al. 2014; Shi et al. 2016a), and whole new families, such as the Chuviruses (Li et al. 2015). This in turn has led to renewed interest in inferring deep viral phylogenies (Koonin et al. 2015; Shi et al. 2016a), and has prompted proposals for large-scale updates of higher-level virus taxonomy (Aiewsakun and Simmonds 2018). More importantly, metagenomics now contributes to our thinking on virus evolution. It has provided a better perspective on host-association and switching (Geoghegan et al. 2017; Dolja and Koonin 2018; Shi et al. 2018), found familiar virus lineages with unexpected genome sizes and structures (Li et al. 2015; Shi et al. 2016a; Shi et al. 2016b), and uncovered an unexpectedly dynamic history of 'modular' protein swapping (Koonin et al. 2015; Shi et al. 2016a). Finally, merely having a PCR product from a metagenomic sample can provide an experimental route to the functional biology of an uncultured virus (van Mierlo et al. 2014).

Potential pitfalls

The recent viral bonanza partly reflects advances in nucleic acid sequencing, a field that has left Moore's Law—that computational power doubles every 2 years—far behind (Wetterstrand 2018). But sequencing is just one of the challenges to exploring the virosphere. The lack of a viable 'meta-barcoding' sequence means that virus discovery often takes a full metagenomic approach, sequencing total (or virus-enriched) nucleic acid, and subsequently assigning sequences through inferred homology (e.g. Rose et al. 2016; Paez-Espino et al. 2017; Nooij et al. 2018). This is challenging, because high divergence means that only the most conserved sequences are recognisable (e.g. RNA virus polymerases), and even then, only at the protein level. Sensitive surveys therefore benefit from assembled contigs rather than raw reads (so that divergent genes are linked to recognisable ones) and protein rather than nucleic-acid similarity searches (because divergence is high). This can be done using off-the-shelf assemblers and search algorithms such as SPADes (Bankevich et al. 2012) or Trinity (Grabherr et al. 2011), and Diamond (Buchfink et al. 2014), but there is also a growing ecosystem of virus-specific metagenomic packages and pipelines available (Nooij et al. 2018).

As with any field in rapid development, best practice is uncertain and fluid, and there are pitfalls for the unwary (Rose et al. 2016). For example, although virus (especially RNA virus) assembly is facilitated by their small and generally unrepetitive genomes, the high complexity of metagenomic pools tends to promote artefactual and chimeric contigs (Simmonds et al. 2017; Tithi et al. 2018). These can unite viral sequences with non-viral ones, especially high-copy-number host sequences such as those from mitochondria and ribosomes. Such 'wide' chimeras are partly mitigated by the use of paired-end and strand-specific reads, ensuring effective adaptor removal, and (when possible) removing host reads before assembly. Chimeric mis-assemblies among divergent viruses or viral segments are also possible, especially when they share near identical stretches of sequence, such as structural RNA motifs or terminal repeats. These are harder to diagnose, and may ultimately require PCR verification, but can often be flagged by comparison with close relatives (if available), unexpected local variation in read-depth, and comparison across metagenomic samples.

These challenges aside, discovering a 'virus-like sequence' remains easier than confirming its status as an infectious agent of the targeted host. First, even if a sequence is 'virus-derived' in an evolutionary sense, its immediate origin may have been an Endogenous Viral Element (EVE) (Katzourakis and Gifford 2010). If expressed and/or 'domesticated' by the recipient genome, these may be represented at high levels and retain open reading frames (Katzourakis and Gifford 2010; Palatini et al. 2017). Conversely, host sequences—especially transposable elements (TEs)—are often incorporated into large DNA viruses and can move freely between hosts and viruses (Gilbert et al. 2016). Second, the host can be misassigned if samples contain multiple hosts, either naturally or through contamination. Technical nucleic acid contamination can be minimised by good laboratory practice (though see Naccache et al. 2013), but high-throughput sequencing technologies are prone to cross-contamination at the point of sequencing. For example, in the absence of dual indexing, 'barcode-switching' in some Illumina platforms can misattribute reads among libraries at a rate of up to 0.3-1% (Kircher et al. 2012). Multi-host samples are usually explicitly recognised as such, for example viruses from 'holobionts' such as anemones (Brüwer and Voolstra 2018). However, the multi-host nature of other samples is often downplayed. For example, faecal samples are often dominated by viruses infecting the host's diet and/or gut microbiota (Zhang et al. 2006; Li et al. 2010), but virus-like sequences are sometimes reported (at least in the headline) as if they were viruses of the faecal donor itself. And, if nucleic acids or virions are prepared from whole host individuals, viruses in faecal matter and viral infections of parasites (notably nematodes, platyhelminthes, and microscopic arthropods) and pathogens (fungi, trypanosomatids, apicomplexans, amoebae, and many others), will also be represented among the sequences. For example, the only dimarhabdovirus recorded from a plant sample derives from RNA contaminated with thrips (Longdon et al. 2015). The potential for viral infections of eukaryotic parasites means that even 'clean' dissected tissue may be cryptically multi-host.

Going beyond 'virus-like sequences'

Such pitfalls make some authors (justifiably) hesitant to proclaim a new virus from metagenomic sequencing alone, and many choose to report 'virus-like sequences'—providing an implicit *caveat emptor*. But in the absence of an isolate, sequence data and nucleic acid analysis can support the existence of a free/replicating virus. First, the nature and quantity of the nucleic acid provides useful clues. Endogenous DNA copies can be identified by a comparison of PCR and RT-PCR (or DNA and RNA sequencing) (Webster et al. 2015; Shi et al. 2016a; Medd et al. 2018; Waldron et al. 2018). Functional DNA viruses must express their proteins, so the absence of viral mRNAs argues against active replication. Active replication also affects strand-bias in RNA viruses, so that strand-specific PCR (Plaskon et al. 2009) or RNAseq (Medd et al. 2018) can identify the negative-sense replication intermediates of positive-sense single-stranded (+ss) RNA viruses, and quantitative analyses can detect the presence of coding products from -ssRNA and dsRNA viruses (Medd et al. 2018). And, for both DNA and RNA viruses, contaminating sequences are likely to be at relatively low titre while the copy-number of inherited EVEs will match the host genome. This means that high copy-number itself provides an argument in favour of viral status (Shi et al. 2016a).

Second, contigs that encode complete viral genomes with intact open reading frames are more consistent with functional viruses than with EVEs. Although whole viruses can be (retro-)copied into a host's genome, there is rarely selective pressure to maintain the virus genome intact. Even expressed and functional (i.e. 'domesticated') EVEs generally only provide the host with one or two beneficial sequences (Katzourakis and Gifford 2010; Palatini et al. 2017). Complete or near-complete virus genomes can also rule out the misattribution of host TEs as viral sequences, as TEs from viral genomes are unlikely to be detected in the absence of other virus genes. Third, the distribution of virus-like sequences across metagenomic pools and host individuals (e.g. surveyed by PCR) can help to confirm a genuine viral origin, and narrow down the true host (Webster et al. 2015; Waldron et al. 2018). Presence/absence patterns can help to weed out EVEs, as—unless it is very recent in origin—an EVE insertion is likely to be present in all host genomes, but virus prevalence is likely be below 100% and variable among populations and over time (Webster et al. 2015; Waldron et al. 2018). Across host individuals, the co-occurrence of virus-like and other sequences can be used to correctly infer hosts, as viruses that infect a contaminating microparasite will co-occur with it. Patterns of co-occurrence can also help to identify missing parts of the viral genome, such as fragments of incompletely assembled genomes and components of segmented viruses that are not recognisable using sequence conservation (Webster et al. 2015).

Finally, perhaps the ultimate evidence of infection is recognition by the host antiviral immune system (Aguiar et al. 2015; Webster et al. 2015). In nematodes and arthropods, antiviral RNA-interference (RNAi) processes viral genomes into distinctive small RNAs (viRNAs)(Félix et al. 2011; Lewis et al. 2018). These can be sequenced from the metagenomic discovery RNA pool, and the reads mapped to RNAseq assemblies (their small size and patchy distribution mean that viRNA assemblies are fragmentary) (Aguiar et al. 2015; Webster et al. 2015). Because Dicer-mediated viRNA biogenesis targets dsRNA such as replication intermediates, viRNAs can demonstrate both an antiviral response and replication. Importantly, viRNAs usually have a tight and characteristic length distribution (e.g. 20nt in Lepidoptera, 21nt in *Drosophila*, 22nt in C. elegans) (Félix et al. 2011; Lewis et al. 2018) and a 3' 2-O-methyl group, making them distinguishable from degradation products. Their size distribution and base composition also distinguish them from TE- and EVE-derived piwi-associated RNAs (Palatini et al. 2017; Lewis et al. 2018; Waldron et al. 2018).

How many animal viruses are there, and what are they doing?

Our expanded view of the animal virosphere has already started to answer old questions and provoke new ones, but these two stand out. What is needed to answer them? Given any definition of 'different virus' (van Regenmortel 2016; Simmonds et al. 2017), whether based on an operational taxonomic unit or a functional biological definition, virus lineages are countable. Sampling of nine virus families to near-saturation from one bat species in Bangladesh identified 55 different viruses, and implied an estimate of 320 thousand viruses infecting mammals (Anthony et al. 2013)—under the assumption of no geographic variation, complete host specificity, and the absence of other virus families. A more confident estimate could be made from unbiased metagenomic samples of the joint distribution of prevalence across host and virus lineages, across their geographic range.

Such metagenomic surveys may soon be possible for a few carefully-considered host groups, but they would still miss the viral 'dark matter': those virus sequences that we cannot see because they have no detectable homology with known viruses (Krishnamurthy and Wang 2017). It remains unknown what proportion of 'dark' sequences represent poorly-conserved regions of otherwise recognisable viruses (François et al. 2018), versus completely new virus lineages. One approach is to search for deeper homologies, such as those provided by protein structure (Yutin et al. 2018). Another is to consider unclassifiable sequences that are processed by the antiviral RNAi pathway, such as the viRNA-based 'candidate viruses' reported from *Drosophila* (Figure 1E) (Webster et al. 2015). Subsequently discovered relatives have now identified around half of these contigs as fragments or segments of known lineages. This leaves open the possibility that some of the 'dark matter' sequences do represent genuinely new viruses (Figure 1E), but suggests that most do not. The confirmation of genuinely novel virus lineages may represent a case in which viral isolates are unequivocally necessary.

What are these viruses doing to their hosts? It is almost axiomatic that viruses are parasites, but micro-organisms are often mutualist or commensal, and although viruses necessarily use host resources, their impact on host fitness may be negligible and/or outweighed by provision of some unknown benefit (Roossinck and Bazán 2017). At first glance it might seem that elucidating the fitness consequences of infection must also require isolates for experimentation. However, experimental studies are rarely useful for inferring real-world fitness. First, most studies measure traits such as survival or reproduction in place of fitness, and so misinterpret life-history tradeoffs—such as mistaking a host response to mitigate cost (e.g. terminal investment) with a virus-derived benefit (increased early-life reproduction). Second, they tend to be under-powered: an absence of detectable harm does not imply an absence of cost, only a small one. The ultimate arbiter of costliness must be natural selection: if the presence of the virus can select for host resistance, then the virus imposes a net fitness cost. A resistance mutation is expected to spread if its fitness benefit substantially exceeds the impact of genetic drift (i.e. $N_e s >> 1$ where N_e is effective population size and s is the selective benefit). Very conservatively, a fitness cost to infection of 0.1% in *Drosophila* (or *Arabidopsis*) would select strongly for host resistance, but this cost is probably an order of magnitude too small to measure experimentally in a multicellular organism (Gallet et al. 2012). This means that, far from requiring more isolates, our best approach to understand fitness costs could be to add a metagenomic screening component to fitness studies of animals in the wild (e.g. Knowles et al. 2012).

Bibliography

Aguiar, E., et al. (2015). "Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host." Nucleic Acids Research **43**(13): 6191-6206.

Aiewsakun, P. and P. Simmonds (2018). "The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification." Microbiome **6**(1): 38.

Anthony, S. J., et al. (2013). "A Strategy To Estimate Unknown Viral Diversity in Mammals." mBio **4**(5).

Bankevich, A., et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." Journal of computational biology **19**(5): 455-477.

Berto, A., et al. (2018). "Detection of potentially novel paramyxovirus and coronavirus viral RNA in bats and rats in the Mekong Delta region of southern Viet Nam." Zoonoses and Public Health **65**(1): 30-42.

Bos, L. (1999). "Beijerinck's work on tobacco mosaic virus: historical context and legacy." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **354**(1383): 675-685.

Breitbart, M., et al. (2002). "Genomic analysis of uncultured marine viral communities." Proceedings of the National Academy of Sciences **99**(22): 14250-14255.

Brüwer, J. D. and C. R. Voolstra (2018). "First insight into the viral community of the cnidarian model metaorganism Aiptasia using RNA-Seq data." PeerJ **6**: e4449.

Buchfink, B., et al. (2014). "Fast and sensitive protein alignment using DIAMOND." Nature Methods **12**: 59.

Dolja, V. V. and E. V. Koonin (2018). "Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer." Virus research **244**: 36-52.

Félix, M.-A., et al. (2011). "Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses." PLoS biology **9**(1): e1000586.

François, S., et al. (2018). "Increase in taxonomic assignment efficiency of viral reads in metagenomic studies." Virus research **244**: 230-234.

François, S., et al. (2016). "Discovery of parvovirus-related sequences in an unexpected broad range of animals." Scientific Reports **6**: 30880.

Gallet, R., et al. (2012). "Measuring selection coefficients below 10− 3: method, questions, and prospects." Genetics **190**(1): 175-186.

Geoghegan, J. L., et al. (2017). "Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families." PLOS Pathogens **13**(2): e1006215.

Geoghegan, J. L., et al. (2018). "Virological Sampling of Inaccessible Wildlife with Drones." Preprints 2018, 2018050184.

Gilbert, C., et al. (2016). "Continuous Influx of Genetic Material from Host to Virus Populations." PLOS Genetics **12**(2): e1005838.

Grabherr, M. G., et al. (2011). "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data." Nature biotechnology **29**(7): 644.

Greninger, A. L. (2018). "A decade of RNA virus metagenomics is (not) enough." Virus Research **244**: 218-229.

Kapun, M., et al. (2018). "Genomic analysis of European Drosophila melanogaster populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection." bioRxiv.

Katzourakis, A. and R. J. Gifford (2010). "Endogenous Viral Elements in Animal Genomes." PLOS Genetics **6**(11): e1001191.

King, A. M. Q., et al. (2018). "Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018)." Archives of Virology.

Kircher, M., et al. (2012). "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform." Nucleic Acids Research **40**(1): e3-e3.

Knowles, S. C., et al. (2012). "Epidemiology and fitness effects of wood mouse herpesvirus in a natural host population." Journal of General Virology **93**(11): 2447-2456.

Koonin, E. V., et al. (2015). "Origins and evolution of viruses of eukaryotes: The ultimate modularity." Virology **479-480**: 2-25.

Krishnamurthy, S. R. and D. Wang (2017). "Origins and challenges of viral dark matter." Virus research **239**: 136-142.

Ladner, J. T., et al. (2014). "Standards for Sequencing Viral Genomes in the Era of High-Throughput Sequencing." mBio **5**(3).

Lewis, S. H., et al. (2018). "Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements." Nature ecology & evolution **2**(1): 174.

Li, C. X., et al. (2015). "Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses." Elife **4**.

Li, L. L., et al. (2010). "Bat Guano Virome: Predominance of Dietary Viruses from Insects and Plants plus Novel Mammalian Viruses." Journal of Virology **84**(14): 6955-6965.

Longdon, B., et al. (2015). "The evolution, diversity, and host associations of rhabdoviruses." Virus Evolution **1**(1): 12.

Ma, S., et al. (2018). "Comparative transcriptomics across 14 Drosophila species reveals signatures of longevity." Aging cell: e12740.

Medd, N. C., et al. (2018). "The virome of Drosophila suzukii, an invasive pest of soft fruit." Virus Evolution **4**(1): vey009-vey009.

Murphy, F. A. (2016). Chapter Five - Historical Perspective: What Constitutes Discovery (of a New Virus)? Advances in Virus Research. M. Kielian, K. Maramorosch and T. C. Mettenleiter, Academic Press. **95:** 197-220.

Naccache, S. N., et al. (2013). "The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns." Journal of Virology **87**(22): 11966-11977.

Nooij, S., et al. (2018). "Overview of Virus Metagenomic Classification Methods and Their Biological Applications." Frontiers in Microbiology **9**(749).

Paez-Espino, D., et al. (2017). "Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data." Nature Protocols **12**: 1673.

Palatini, U., et al. (2017). "Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus." BMC genomics **18**(1): 512.

Plaskon, N. E., et al. (2009). "Accurate Strand-Specific Quantification of Viral RNA." PLOS ONE **4**(10): e7468.

Ray, J. (2014). Methodus Plantarum Nova, The Ray Society.

Roberts, J. M. K., et al. (2018). "Metagenomic analysis of Varroa-free Australian honey bees (Apis mellifera) shows a diverse Picornavirales virome." Journal of General Virology.

Roossinck, M. J. and E. R. Bazán (2017). "Symbiosis: Viruses as Intimate Partners." Annual Review of Virology **4**(1): 123-139.

Rose, R., et al. (2016). "Challenges in the analysis of viral metagenomes." Virus Evolution **2**(2): vew022-vew022.

Shi, M., et al. (2018). "The evolutionary history of vertebrate RNA viruses." Nature **556**(7700): 197-+.

Shi, M., et al. (2016a). "Redefining the invertebrate RNA virosphere." Nature **540**(7634): 539-+.

Shi, M., et al. (2016b). "Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses." Journal of Virology **90**(2): 659-669.

Simmonds, P., et al. (2017). "Virus taxonomy in the age of metagenomics." Nature Reviews Microbiology **15**: 161.

Tithi, S. S., et al. (2018). "FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data." PeerJ **6**: e4227.

Tokarz, R., et al. (2018). "Identification of Novel Viruses in Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis Ticks." Msphere **3**(2).

Tokarz, R., et al. (2014). "Virome Analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis Ticks Reveals Novel Highly Divergent Vertebrate and Invertebrate Viruses." Journal of Virology **88**(19): 11480-11492.

van Mierlo, J. T., et al. (2014). "Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi." Plos Pathogens **10**(7): 13.

van Regenmortel, M. H. (2016). "Classes, taxa and categories in hierarchical virus classification: a review of current debates on definitions and names of virus species." Bionomina **10**(1): 1-21.

Waldron, F. M., et al. (2018). "Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes." bioRxiv.

Webster, C. L., et al. (2015). "The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster." Plos Biology **13**(7): 33.

Wetterstrand, K. (2018). "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) " Retrieved 14th May 2018, from https://www.genome.gov/sequencingcostsdata.

Williams, S. H., et al. (2018). "Viral Diversity of House Mice in New York City." Mbio **9**(2): 17.

Yutin, N., et al. (2018). "Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis." Virology Journal **15**(1): 67.

Zhang, T., et al. (2006). "RNA viral community in human feces: Prevalence of plant pathogenic viruses." Plos Biology **4**(1): 108-118.

Zheng, X. Y., et al. (2018). "Viral metagenomics of six bat species in close contact with humans in southern China." Archives of Virology **163**(1): 73-88.

Acknowledgements

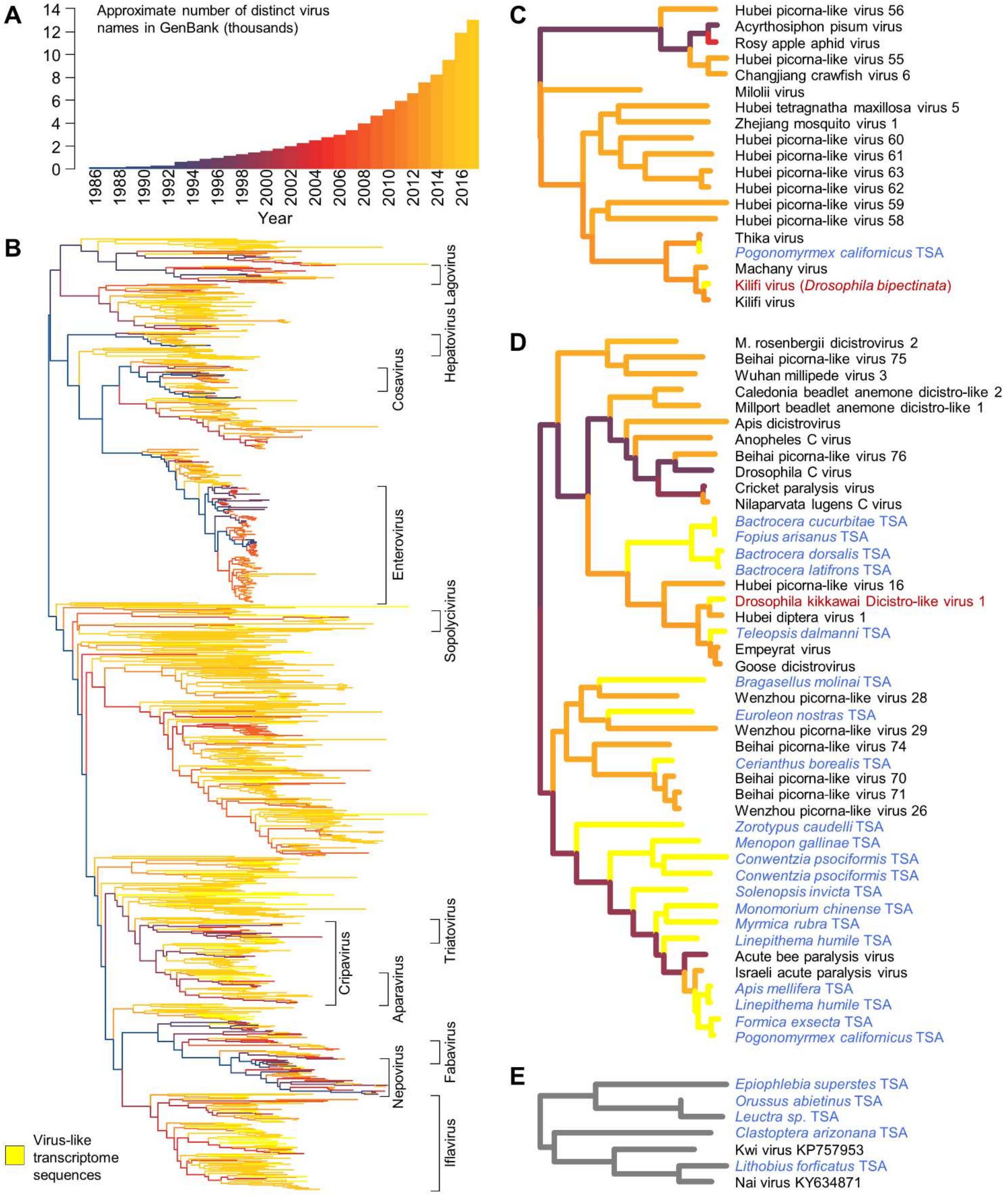
I thank Alistair Greaves for preparing the phylogeny of Picornavirales in Figure 1B and Ben Longdon for comments on the first draft of the manuscript. I apologise to the many authors whose work could not be cited due to restrictions on space and timeframe.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Figure Legend:

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 GenBank 'species' (collapsing strain identifiers) and record creation-date 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 (scale 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). Note that the short conserved-sequence length leads to poor resolution and fails to recover some named genera. **Panels C and D:** To illustrate with ease with which new virus-like sequences can be found in public datasets, I obtained the most recently deposited *Drosophila* RNAseq dataset (PRJNA414017 (Ma et al. 2018)), performed a de-novo assembly using Trinity (Grabherr et al. 2011), and identified virus-like sequences using Diamond (Buchfink et al. 2014). I found complete genomes for two picorna-like viruses (red labels; MH320557 and MH320558): a divergent sequence of Kilifi virus from *D. bipectinata* (previously known from *D. melanogaster*) and a novel Dicistro-like virus from *D. kikkawai*, related to Hubei diptera virus 1 (Shi et al. 2016a). Maximum-likelihood phylogenies for these two sequences were inferred from around 700 amino acids of the polymerase, mid-point rooted, and coloured as in panel B. These trees illustrate the dominance of recent discoveries, including the many virus-like sequences in transcriptome assemblies (blue taxon labels). They also illustrate the potential confusion introduced by naming faecal-sample viruses after the faecal donor (all close relatives of Goose Dicistrovirus infect invertebrates). **Panel E:** Phylogeny of two putative 'dark matter' viruses from *Drosophila,* including related transcriptome sequences. These putative viruses each comprise four 1.5Kb segments encoding a single long open reading frame, but lack detectable homology with any known virus lineage and were inferred to be viral on the basis of viRNA profiles (Webster et al. 2015). Data associated with this figure are available via FigShare https://dx.doi.org/10.6084/m9.figshare.6272066.


A
Approximate number of distinct virus names in GenBank (thousands)
Year
B
Lagovirus
Hepatovirus
Cosavirus
Enterovirus
Sopolycivirus
Triatovirus
Cripavirus
Aparavirus
Fabavirus
Nepovirus
Iflavirus
🟨 Virus-like transcriptome sequences
C
Hubei picorna-like virus 56
Acyrthosiphon pisum virus
Rosy apple aphid virus
Hubei picorna-like virus 55
Changjiang crawfish virus 6
Milolii virus
Hubei tetragnatha maxillosa virus 5
Zhejiang mosquito virus 1
Hubei picorna-like virus 60
Hubei picorna-like virus 61
Hubei picorna-like virus 63
Hubei picorna-like virus 62
Hubei picorna-like virus 59
Hubei picorna-like virus 58
Thika virus
Pogonomyrmex californicus TSA
Machany virus
Kilifi virus (Drosophila bipectinata)
Kilifi virus
D
M. rosenbergii dicistrovirus 2
Beihai picorna-like virus 75
Wuhan millipede virus 3
Caledonia beadlet anemone dicistro-like 2
Millport beadlet anemone dicistro-like 1
Apis dicistrovirus
Anopheles C virus
Beihai picorna-like virus 76
Drosophila C virus
Cricket paralysis virus
Nilaparvata lugens C virus
Bactrocera cucurbitae TSA
Fopius arisanus TSA
Bactrocera dorsalis TSA
Bactrocera latifrons TSA
Hubei picorna-like virus 16
Drosophila kikkawai Dicistro-like virus 1
Hubei diptera virus 1
Teleopsis dalmanni TSA
Empeyrat virus
Goose dicistrovirus
Bragasellus molinai TSA
Wenzhou picorna-like virus 28
Euroleon nostras TSA
Wenzhou picorna-like virus 29
Beihai picorna-like virus 74
Cerianthus borealis TSA
Beihai picorna-like virus 70
Beihai picorna-like virus 71
Wenzhou picorna-like virus 26
Zorotypus caudelli TSA
Menopon gallinae TSA
Conwentzia psociformis TSA
Conwentzia psociformis TSA
Solenopsis invicta TSA
Monomorium chinense TSA
Myrmica rubra TSA
Linepithema humile TSA
Acute bee paralysis virus
Israeli acute paralysis virus
Apis mellifera TSA
Linepithema humile TSA
Formica exsecta TSA
Pogonomyrmex californicus TSA
E
Epiophlebia superstes TSA
Orussus abietinus TSA
Leuctra sp. TSA
Clastoptera arizonana TSA
Kwi virus KP757953
Lithobius forficatus TSA
Nai virus KY634871