Petabase-scale sequence alignment catalyses viral discovery

Robert C. Edgar[1], Jeff Taylor[1], Tomer Altman[2], Pierre Barbera[3], Dmitry Meleshko[4,5], Victor Lin[1], Dan Lohr[1], Gherman Novakovsky[6], Basem Al-Shayeb[7], Jillian F. Banfield[8], Anton Korobeynikov[4,9], Rayan Chikhi[10], and Artem Babaian[3,*]

[0]All authors contributed equally to this work
[1]Unaffiliated
[2]Altman Analytics LLC
[3]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
[4]Center for Algorithmic Biotechnology, St. Petersburg State University, St. Petersburg, Russia
[5]Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, USA
[6]Department of Medical Genetics, University of British Columbia. Vancouver, BC, Canada
[7]Department of Plant and Microbial Biology, University of California, Berkeley
[8]Department of Earth and Planetary Science, University of California, Berkeley
[9]Department of Statistical Modelling, St. Petersburg State University, St. Petersburg, Russia
[10]Institut Pasteur, CNRS, Paris, France
[*]Corresponding Author: ababaian@bccrc.ca

August 07 2020

Abstract

Public sequence data represents a major opportunity for viral discovery, but its exploration has been inhibited by a lack of efficient methods for searching this corpus, which is currently at the petabase scale and growing exponentially. To address the ongoing pandemic caused by Severe Acute Respiratory Syndrome Coronavirus 2 and expand the known sequence diversity of viruses, we aligned pangenomes for coronaviruses (CoV) and other viral families to 5.6 petabases of public sequencing data from 3.8 million biologically diverse samples. To implement this strategy, we developed a cloud computing architecture, `Serratus`, tailored for ultra-high throughput sequence alignment at the petabase scale. From this search, we identified and assembled thousands of CoV and CoV-like genomes and genome fragments ranging from known strains to putatively novel genera. We generalise this strategy to other viral families, identifying several novel deltaviruses and huge bacteriophages. To catalyse a new era of viral discovery we made millions of viral alignments and family identifications freely available to the research community. Expanding the known diversity and zoonotic reservoirs of CoV and other emerging pathogens can accelerate vaccine and therapeutic developments for the current pandemic, and help us anticipate and mitigate future ones.

Introduction

Viral zoonotic disease has had a major impact on human health over the past century despite dramatic advances in medical science, notably by the Spanish Flu, AIDS, SARS, Ebola and COVID-19 pandemics. There are an estimated 320,000 mammalian viruses [1] from which emerging infectious diseases in humans may arise [2]. Uncovering this viral biodiversity is a prerequisite for predicting and preventing future epidemics and is therefore the focus of consortia such as USAID PREDICT [3] and the Global Virome Project [4] as well as hundreds of government and academic research projects worldwide.

These efforts can be aided through re-analysis of petabases of high-throughput sequencing data available in public databases such as the Sequence Read Archive (SRA) [5]. This data spans millions of ecologically diverse biological samples, many of which capture viral transcripts that may be incidental to the goals of the original studies [6]. To expand the known repertoire of viruses and catalyse global virus discovery, in particular for *Coronaviridae* (CoV) family, we developed the `Serratus` cloud computing architecture for ultra-high throughput sequence alignment.

From a screen of 3.8 million libraries comprising 5.6 petabases of sequencing reads, we report 11,120 assemblies, including sequences from 13 previously uncharacterised or unavailable CoV or CoV-like operational taxonomic units (OTUs), defined by clustering amino sequences of the RNA dependent RNA polymerase (RdRp) gene at 97% identity. To demonstrate the broader utility of our approach, we also report six novel deltaviruses related to the human pathogen Hepatitis δ Virus (HDV), and expand the described members of the recently characterised family of huge bacteriophages (phages).

Viral discovery is a first step in preparing for the next pandemic. Sequencing reads for thousands of uncharacterised viruses already exist and require careful curation. To accelerate this process, we established a freely available and explorable resource of all vertebrate viral alignment data generated by `Serratus` at `https://serratus.io`. This work lays the foundation for years of future research by enabling the exploration of viruses which have been captured by more than a decade of high-throughput sequencing studies.

Results

Petabase-scale alignment enables coronavirus discovery

`Serratus` is a freely available, open-source cloud-computing platform designed to enable petabase-scale sequence alignment against a set of references. Using `Serratus`, we aligned in excess of one million short-read sequencing datasets per day for under 1 US cent per dataset (Extended Figure 1). This was achieved by leveraging commercially available computing infrastructure to employ up to 22,250 virtual CPUs simultaneously (see Methods).

We aligned 3,837,755 public RNA-seq, meta-genome, meta-virome and meta-transcriptome datasets (termed a sequencing run [5]) against a collection of viral family pangenomes comprising all GenBank CoV records clustered at 99% identity plus all non-retroviral RefSeq records for vertebrate viruses (see Methods and Extended Table 1). To uncover more divergent viruses, we re-analysed 370,014 runs in a translated nucleotide search against a query comprising panproteomes for CoV and other families. We performed *de novo* assembly on 52,772 runs potentially containing CoV sequencing reads by combining 37,131 SRA accessions identified by the `Serratus` search with 18,584 identified by an ongoing cataloguing initiative of the SRA called STAT [5]. 11,120 of the resulting assemblies contained putative CoV contigs, of which 4,179 aligned to CoV RdRp (Extended Table 2). Of these, we identified 13 OTUs from a total of 129, i.e. not represented by *Coronaviridae* in GenBank (Figure 1a and Extended Figure 2). The protein domains of these OTU are consistent with a CoV or CoV-like genome organisation (Extended Figure 3).

Three of the novel CoV OTUs fell within the *Alphacoronavirus* (αCoV) genus. The first (exemplar run: ERR2756788) was from two *Desmodus rotundus* bat metagenomes yielding 29.1 and 25.4 kb CoV contigs respectively in the *Nyctacovirus* subgenus. These CoV were noted by the data-collectors, [7], but the sequences were not public and thus novel to our analysis. The second OTU (SRR9643845) was from a *Pipistrellus pipistrellus* bat metagenome collected in 2016 in China. Finally, from five libraries (ERR2744266) generated for a study on the metagenomic effects of the burying beetle *Nicrophorus vespilloides* on a mouse carcass, we assembled a *Luchacovirus* related to the rodent Lucheng Rn rat coronavirus (83% genome nucleotide identity to NC_032730.1).

From a rodent virome study which identified several novel CoV [8], a sample from an unknown species contained a βCoV Embecovirus (SRR5447167), with the closest matching genome matching an unclassified βCoV from Vietnam (77.52% to MH687971). Finally, the δCoV OTU (SRR5447167) appears to be from a currently unpublished avian virome study in China.

We designated the eight remaining OTUs as group *E*, noting that all were found in samples from non-mammal aquatic vertebrates falling outside of δCoV in the tree (Extended Figure 2). A sister taxon to *Coronaviridae*

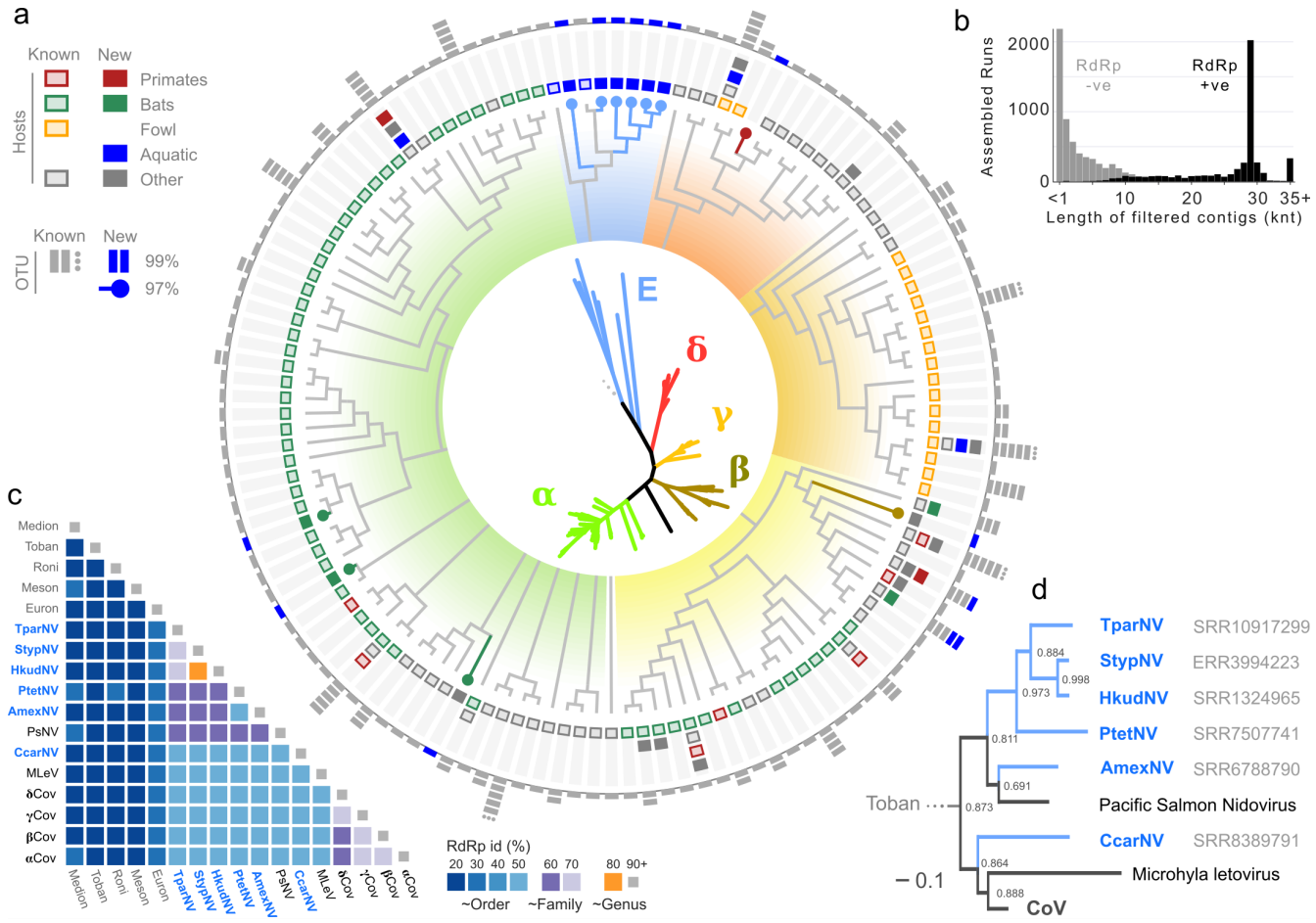


Figure 1: **Expanded characterisation of CoV and related OTUs**
a Radial cladogram derived from maximum likelihood tree of CoV and related OTUs. Inset is a phylogram of the same tree annotated with CoV genera (Greek letters) and group *E* CoV-like nidoviruses. OTUs were generated by clustering the RdRp gene at 97% identity. Diversity within each such 97% OTU was characterised by counting the number of 99% identity OTUs it contained. An OTU (97% or 99%) was considered to be known if it contained a GenBank sequence, otherwise to be a novel OTU discovered by `Serratus`. Hosts were considered novel if the source organism annotated by the SRA belonged to a species not annotated as a host in any GenBank record, noting that the annotated source may differ from the viral host (e.g., faecal contamination in a plant sample). Hosts are classified as Primates, Fowl (*Galliformes*), Bats (*Chiroptera*, Aquatic (*Amphibia* and *Osteichthyes*), or Other. **b** Length distribution for assemblies of SRA datasets classified as likely CoV-positive, showing a peak around the typical CoV genome length ~30knt. **c** Triangular matrix showing median RdRp sequence identities between selected nidovirales and group *E* viruses. **d** Phylogram of group *E* CoV-like nidoviruses.

was recently proposed [9] following the characterisation of a corona-like virus, Microhyla alphaletovirus 1 (MLeV), in the frog *Microhyla fissipes*, and soon after a related Pacific salmon nidovirus (PsNV) was described in the endangered *Oncorhynchus tshawytscha* [10]. Two of our OTUs were in these host species and the described viruses proved to be near-perfect matches. We expand this recently characterised group with six additional members, five similar to PsNV in; *Takifugu pardalis* (fugu fish; TparNV), *Syngnathus typhle* (broad-nosed pipefish; StypNV), *Hippocampus kuda* (seahorse; HkudNV) [11], *Puntigrus tetrazona* (tiger barb; PtetNV), *Ambystoma mexicanum* (axolotl; AmexNV), and a more distant member in *Caretta caretta* (loggerhead sea turtle; CcarNV).

Notably, the *Ambystoma mexicanum* (Axolotl) nidovirus (AmexNV) was assembled in 18 runs, 11 of which yielded ~19 kb contigs. Easing the criteria of requiring an RdRp match, 28/44 (63.6%) of the runs from the associated studies were AmexNV positive. Gene structure of the AmexNV and related contigs suggests that there is genomic segmentation within this clade (Extended Figure 3), with a homologous assembly gap is present in the published PsNV genome [10]. These contigs were obtained from experimental animals from two different research groups [12–14], the common factor is the animal stock centre used by these studies which is therefore likely to be the source of the virus. Axolotl are critically endangered in the wild; determining the distribution and pathophysiology of AmexNV in these animals can assist with conservation efforts.

Infectious agents are the leading cause of pyrexia of unknown origin (PUO) in children and immunocompromised adults [15]. In addition to identifying genetic diversity within CoV, we cross-referenced CoV+ library meta-data to identify possible zoonoses and infer vectors of transmission. Discordant libraries, one in which a CoV is identified and the viral expected host does not match the sequencing library source taxa, were rare, accounting for only 0.92% of cases (Extended Table 2e).

In a 2010 virome sequencing study [16] of children with febrile illness, we identified sequencing runs from two children, one febrile (id:9007) and one afebrile (id:9090) with reads mapping to the (βCoV), Murine Hepatitis Virus (MHV). We assembled a complete 31.3 kb MHV genome from each replicate taken from the febrile child and a partial genome from the afebrile child. MHV can infect human cells *in vitro* [17], but may be rare in humans, highlighting how rapid and unbiased meta-genomic sequence analysis can not only resolve the etiology of a sub-set of PUO, centralisation of these data (stripped of human-identifying reads) also serves as a public-health surveillance system for zoonosis.

An important consideration for these analyses is that the nucleic acid reads do not prove viral infection has occurred in the nominal host species. For example, we identified four libraries in which a porcine or avian coronavirus was found in plant samples. A more likely explanation than cross-kingdom CoV transmission is that CoV was present in faeces/fertiliser originating from a mammalian or avian host.

Coronaviridae is a well-characterised family (Figure 2 and Extended Figure 4), yet our re-analysis of the SRA yielded eleven novel or under-reported OTUs. There are at least 4,497 more high-confidence (score $\geq$80) and diverged ($\leq$90% identity) virus-containing datasets. In particular *Picornaviridae* and *Reoviridae* are enriched and numerous within this category (Figure 2). `Serratus` exploration of under-characterised viruses can potentially fill these gaps in our knowledge.

Expanding the boundaries of known viral biodiversity

The global mortality from viral hepatitis exceeds that of HIV/AIDS, tuberculosis or malaria, due to acute and chronic liver cirrhosis and subsequent hepatocellular carcinoma [18]. Hepatitis delta virus (HDV) is a small (~1.7 knt) RNA satellite virus infecting hepatocytes. Alone, HDV is unable to produce infectious viral particles, as it requires the envelope protein from its helper virus, Hepatitis B (HBV) [19]. HDV infection aggravates liver cirrhosis caused by HBV and worsens clinical outcomes [20].

Prior to 2018, HDV was the sole known member of its genus; ten members have since been characterised [21–25]. We identified an additional six deltaviruses (Figure 3a) and assembled complete circular genomes for five (Extended Figure 6). The evolutionary histories of these deltaviruses are explored further in a companion manuscript [24]. One of these novel deltaviruses, MmonDV, was identified in *Marmota monax* (Eastern woodchuck), a model organism used over the last three decades for the study of viral-induced hepatitis and hepatocellular carcinoma following Woodchuck Hepatitis Virus (WHV) infection, a hepadnavirus similar to HBV [26].

From a 2015 study of 24 woodchucks born in captivity and experimentally infected with WHV [30], liver biopsy RNA-seq from four (16.7%) animals contained >5 MmonDV-mapping reads in at least one time-point of the 26 week study (Figure 3c). Woodchuck Hepatitis Virus can support replication of human HDV, it is in fact a model for HDV pathogenesis [31, 32], so it is probable that WHV is also the helper virus for MmonDV. Inter-animal variation of WHV-induced liver cirrhosis can be substantial [30] and cryptic MmonDV infection may have be underlying some of this variability from the past three decades of research using this model system, which warrants further investigation.

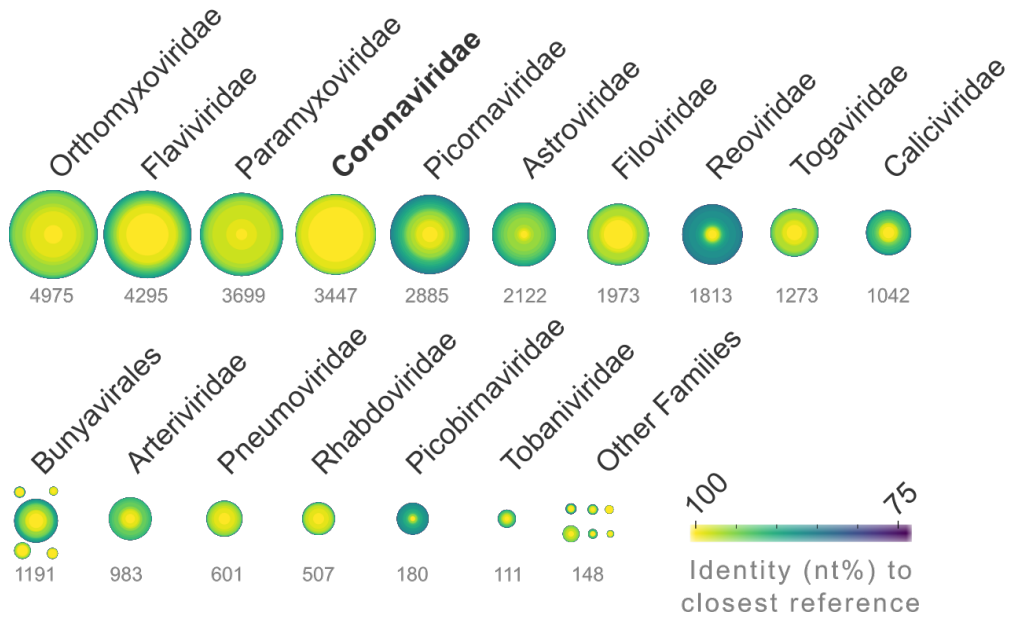


Figure 2: **Estimated assemblable RNA virome across the SRA** Per-family distribution of the 31,074 high-confidence matches (score ≥80, indicating the presence of a virus with a high probability of being assembled) in the 3.84 million SRA datasets searched by `Serratus`. Colour from yellow=100% to purple=75% indicates percentage nucleotide identity, with area proportional to count. *Picornaviridae* and *Reoviridae* are enriched for high-confidence, low-identity viruses relative to other RNA viruses. See Extended Figure 4 for a detailed breakdown of RNA-viruses and Extended Figure 5 for DNA-viruses.

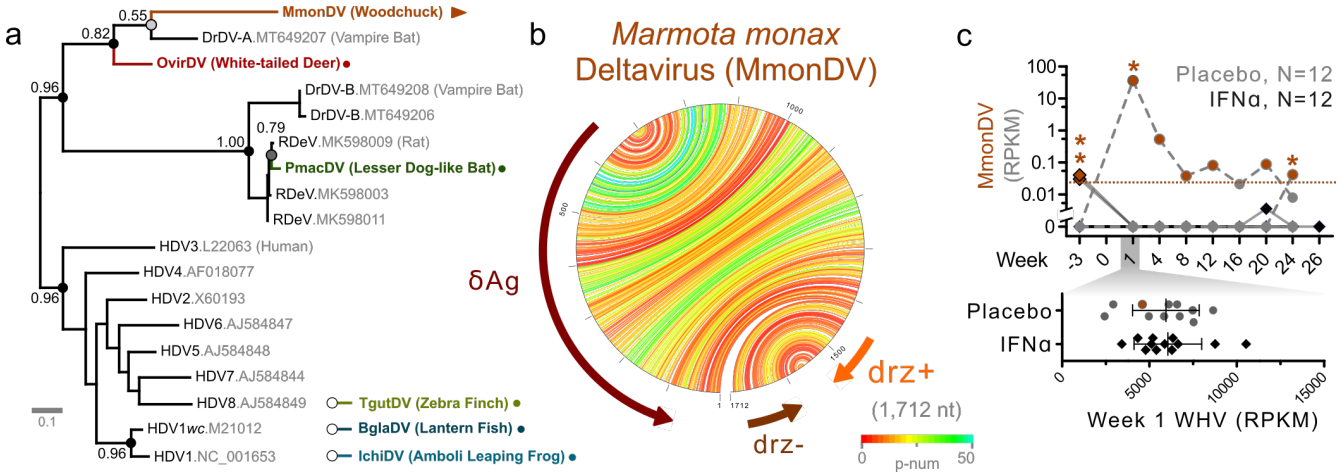


Figure 3: **Novel deltaviruses a** Maximum-likelihood phylogenetic tree of mammalian deltaviruses (DV) derived from a delta-antigen protein alignment; bootstrap values are shown at selected nodes. Labels are MmonDV: *Marmota monax* Deltavirus; DrDV: *Desmodus Rotundus* Deltavirus-A and B; RDeV: Spiny rat Deltavirus; OvirDV: *Odocoileus virginianus* Deltavirus [27]; PmacDV: *Peropteryx macrotis* Deltavirus; HDV: Hepatitis Deltavirus clades; and three unplaced non-mammalian DV identified in this study; TgutDV: *Taeniopygia guttata* Deltavirus [28]; BglaDV: *Benthosema glaciale* Deltavirus; IchiDV: *Indirana chiravasi* Deltavirus. See Extended Figure 7 for supporting ribozyme phylogeny. **b** Genome structure for MmonDV and other novel DV (Extended Figure 6) containing; a negative-sense delta-antigen (δAg) open reading frame; two ribozymes; and characteristic rod-like folding, where each connecting line shows the predicted base-pairing within the single stranded RNA genome, coloured by confidence (p-num)[29]. The . **c** MmonDV was identified in reads from a study using RNA-seq to determine the effect of interferon-α treatment on animals infected with Woodchuck Hepatitis Virus (WHV)[30]. MmonDV reads were present in samples from four separate animals, all co-infected with WHV, with viral expression measured in reads per kilobase million (RPKM).

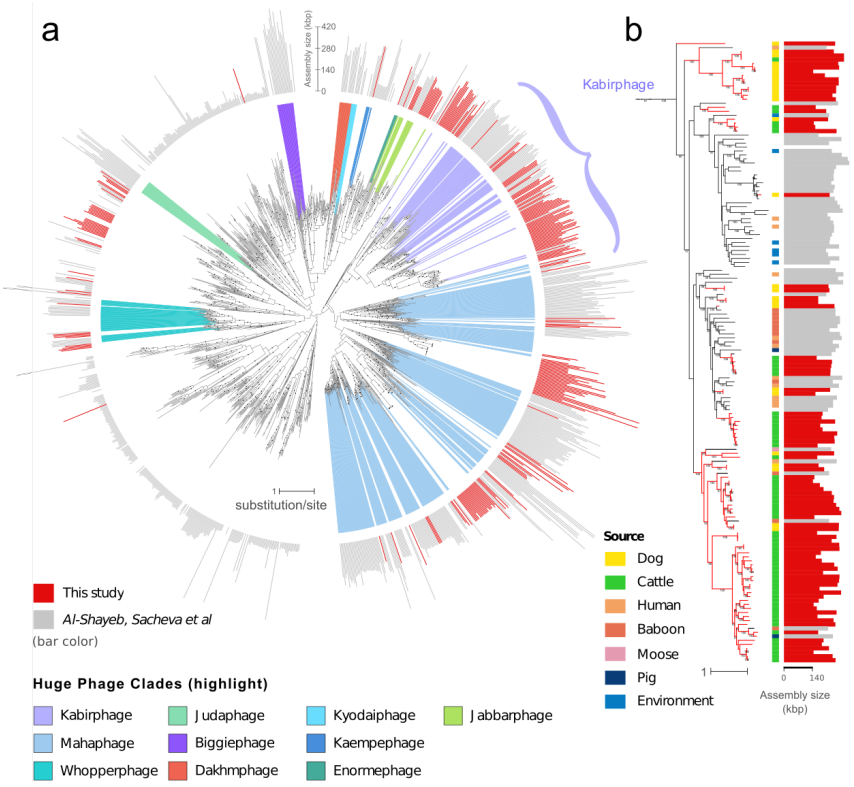


Figure 4: **Phylogenetic analysis using the terminase protein of newly recovered and previously reported phages. a** Tree showing phage clade expansion through the addition of new sequences. Black dots indicate branches with bootstrap values >90. Outer ring indicates genome or genome fragment length: gray are sequences from [33] and reference sequences, shadings indicate previously defined clades of phages with very large genomes (200 - 735 kbp). The Kabirphages (light purple) are shown in expanded view in **b** The expansion of the Kabirphage clade by newly recovered sequences from different animal types (colored dots). Red branches are public data recovered by `Serratus`, black branches indicate the genomes with terminase sequences that were a included in the phage panproteome.

To explore the utility of broad-scale read archive searches for microbiome research, we sought to locate phages whose genomes encode proteins related to the terminases and major capsid proteins from recently reported huge phages [33]. To focus on phages whose genomes are substantially larger than normal (the average size is 52 kbp [33]), we prioritised assembled sequences of $\geq$140 kbp (Figure 4a). Assembly of 287 high-scoring runs returned 252 terminase-containing long contigs, primarily from cats, dogs, cattle and whales. The phylogenetic analysis of these sequences resolves new groups of phages with large genomes, some of which are comprised only of sequences only from one animal genus. However, in a few cases we identified closely related phages in different animal orders, including one case where related phages were found in a human from Bangladesh (ERR866585) and groups of cats (PRJEB9357) and dogs (PRJEB34360) from England, sampled 5 years apart. This result parallels the finding of ~545 kbp Lak phage genomes in pigs, baboons and humans [34]. These newly recovered sequences substantially expand the previously defined clades and reveal members of these clades in new habitats (Figure 4b). Overall, these findings amplify that phages with large genomes are prevalent in human and animal microbiomes.

Discussion

Since the completion of the initial draft of the human genome, the cost of DNA sequencing has outpaced Moore's Law with a corresponding increase in the sizes of sequence databases [35]. `Serratus` offers researchers access to over a decade of data collected by the global research community in a rapid and a cost-effective manner. While our first priority was viral discovery in the context of an ongoing global health crisis, we believe that `Serratus` and further extensions of petabase scale metagenomics will shape a new era in computational biology, and enable radically new approaches to gene discovery, pathogen surveillance, pangenomic evolutionary analysis amongst other applications.

Rapid translation of large datasets, such as those generated by `Serratus`, into meaningful biomedical advances requires concerted collaboration between specialists [36] and underscores a greater need for prompt, free and unrestricted data sharing in the community, not only of raw data (reads) but also of analyses such as assemblies and annotations. To facilitate such progress, we established a data warehouse of the 5.7 terabytes of viral alignments containing known, and yet to be characterized, viral species, each requiring domain expertise for curation. These data can be explored via a graphical web interface at `https://serratus.io` or programatically through the `R` package `tantalus` (`https://github.com/serratus-bio/tantalus`) which interfaces to a `PostgreSQL`-server hosting high-level data summaries.

Computational biology is outpacing the rate at which classical isolation- or culture-based validation can be performed. Reverse genetics and synthetic nucleic acids offer a path to biological validation when virions are unavailable, such as those predicted from sequence alone [37, 38]. Innovative fields such as high-throughput functional viromics [39] leverage these broad and rapidly growing collections of viral sequences, and can inform evidence-based policies responding to emerging pandemics [40, 41].

Human population growth and encroachment on animal habitats is bringing more species into proximity, leading to increased zoonosis [2] and accelerating the Anthropocene mass extinction [42, 43]. While the availability of computation and data analysis is increasing, the opportunity to capture the rich genetic diversity of endangered species and their associated microorganism biodiversity is not. The need to invest in field studies for the collection and curation of rare and biologically diverse samples has never been as pressing as it is today. If not for the conservation of endangered species, then to conserve our own.

References

1. Anthony, S. J. *et al.* A Strategy To Estimate Unknown Viral Diversity in Mammals. en. *mBio* **4.** ISSN: 2150-7511. https://mbio.asm.org/content/4/5/e00598-13 (2020) (Nov. 2013).
2. Jones, K. E. *et al.* Global trends in emerging infectious diseases. eng. *Nature* **451,** 990–993. ISSN: 1476-4687 (Feb. 2008).
3. Johnson, C. K. *et al.* Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proceedings of the Royal Society B: Biological Sciences* **287,** 20192736. https://royalsocietypublishing.org/doi/10.1098/rspb.2019.2736 (2020) (Apr. 2020).
4. Carroll, D. *et al.* The Global Virome Project. en. *Science* **359,** 872–874. ISSN: 0036-8075, 1095-9203. https://science.sciencemag.org/content/359/6378/872 (2020) (Feb. 2018).
5. Leinonen, R., Sugawara, H. & Shumway, M. The Sequence Read Archive. *Nucleic Acids Research* **39,** D19–D21. ISSN: 0305-1048. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013647/ (2020) (Jan. 2011).
6. Moore, R. A. *et al.* The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue. eng. *PloS One* **6,** e19838. ISSN: 1932-6203 (2011).
7. Bergner, L. M. *et al.* Demographic and environmental drivers of metagenomic viral diversity in vampire bats. en. *Molecular Ecology* **29,** 26–39. ISSN: 1365-294X. https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15250 (2020) (2020).
8. Wu, Z. *et al.* Comparative analysis of rodent and small mammal viromes to better understand the wildlife origin of emerging infectious diseases. *Microbiome* **6,** 178. ISSN: 2049-2618. https://doi.org/10.1186/s40168-018-0554-9 (2020) (Oct. 2018).
9. Bukhari, K. *et al.* Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the Coronavirinae, the proposed genus Alphaletovirus. *Virology* **524,** 160–171. ISSN: 0042-6822. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7112036/ (2020) (Nov. 2018).
10. Mordecai, G. J. *et al.* Endangered wild salmon infected by newly discovered viruses. *eLife* **8.** ISSN: 2050-084X. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6721791/ (2020).
11. Romiguier, J. *et al.* Comparative population genomics in animals uncovers the determinants of genetic diversity. en. *Nature* **515,** 261–263. ISSN: 0028-0836, 1476-4687. http://www.nature.com/articles/nature13685 (2020) (Nov. 2014).
12. Tsai, S. L., Baselga-Garriga, C. & Melton, D. A. Blastemal progenitors modulate immune signaling during early limb regeneration. eng. *Development (Cambridge, England)* **146.** ISSN: 1477-9129 (2019).
13. Tsai, S. L., Baselga-Garriga, C. & Melton, D. A. Midkine is a dual regulator of wound epidermis development and inflammation during the initiation of limb regeneration. *eLife* **9** (eds Newmark, P. A., Akhmanova, A. & Currie, J.) e50765. ISSN: 2050-084X. https://doi.org/10.7554/eLife.50765 (2020) (Jan. 2020).
14. Sabin, K. Z., Jiang, P., Gearhart, M. D., Stewart, R. & Echeverri, K. AP-1 cFos/JunB /miR-200a regulate the pro-regenerative glial cell response during axolotl spinal cord regeneration. en. *Communications Biology* **2,** 1–13. ISSN: 2399-3642. https://www.nature.com/articles/s42003-019-0335-4 (2020) (Mar. 2019).
15. Fernandez, C. & Beeching, N. J. Pyrexia of unknown origin. *Clinical Medicine* **18,** 170–174. ISSN: 1470-2118. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6303444/ (2020) (Apr. 2018).
16. Wylie, K. M., Mihindukulasuriya, K. A., Sodergren, E., Weinstock, G. M. & Storch, G. A. Sequence Analysis of the Human Virome in Febrile and Afebrile Children. *PLoS ONE* **7.** ISSN: 1932-6203. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3374612/ (2020) (June 2012).
17. Koetters, P. J., Hassanieh, L., Stohlman, S. A., Gallagher, T. & Lai, M. M. Mouse hepatitis virus strain JHM infects a human hepatocellular carcinoma cell line. eng. *Virology* **264,** 398–409. ISSN: 0042-6822 (Nov. 1999).
18. Stanaway, J. D. *et al.* The global burden of viral hepatitis from 1990 to 2013: findings from the Global Burden of Disease Study 2013. English. *The Lancet* **388,** 1081–1088. ISSN: 0140-6736, 1474-547X. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)30579-7/abstract (2020) (Sept. 2016).
19. Taylor, J. M. Infection by Hepatitis Delta Virus. en. *Viruses* **12,** 648. https://www.mdpi.com/1999-4915/12/6/648 (2020) (June 2020).

20. Palom, A. *et al.* Long-term clinical outcomes in patients with chronic hepatitis delta: the role of persistent viraemia. eng. *Alimentary Pharmacology & Therapeutics* **51,** 158–166. ISSN: 1365-2036 (2020).

21. Szirovicza, L. *et al.* Snake Deltavirus Utilizes Envelope Proteins of Different Viruses To Generate Infectious Particles. eng. *mBio* **11.** ISSN: 2150-7511 (2020).

22. Wille, M. *et al.* A Divergent Hepatitis D-Like Agent in Birds. *Viruses* **10.** ISSN: 1999-4915. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6315422/ (2020) (Dec. 2018).

23. Chang, W.-S. *et al.* Novel hepatitis D-like agents in vertebrates and invertebrates. en. *Virus Evolution* **5.** https://academic.oup.com/ve/article/5/2/vez021/5532287 (2020) (July 2019).

24. Bergner, L. M. *et al.* Satellite virus diversification through host shifting revealed by novel deltaviruses in vampire bats. en. *bioRxiv,* 2020.06.17.156745. https://www.biorxiv.org/content/10.1101/2020.06.17.156745v2 (2020) (June 2020).

25. Paraskevopoulou, S. *et al.* Mammalian deltavirus without hepadnavirus coinfection in the neotropical rodent Proechimys semispinosus. en. *Proceedings of the National Academy of Sciences.* ISSN: 0027-8424, 1091-6490. https://www.pnas.org/content/early/2020/07/09/2006750117 (2020) (July 2020).

26. Popper, H., Roth, L., Purcell, R. H., Tennant, B. C. & Gerin, J. L. Hepatocarcinogenicity of the woodchuck hepatitis virus. en. *Proceedings of the National Academy of Sciences* **84,** 866–870. ISSN: 0027-8424, 1091-6490. https://www.pnas.org/content/84/3/866 (2020) (Feb. 1987).

27. Seabury, C. M. *et al.* Genome-wide polymorphism and comparative analyses in the white-tailed deer (Odocoileus virginianus): a model for conservation genomics. eng. *PloS One* **6,** e15811. ISSN: 1932-6203 (Jan. 2011).

28. Newhouse, D. J., Hofmeister, E. K. & Balakrishnan, C. N. Transcriptional response to West Nile virus infection in the zebra finch (Taeniopygia guttata). eng. *Royal Society Open Science* **4,** 170296. ISSN: 2054-5703 (June 2017).

29. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. en. *Nucleic Acids Research* **31,** 3406–3415. ISSN: 0305-1048. https://academic.oup.com/nar/article/31/13/3406/2904217 (2020) (July 2003).

30. Fletcher, S. P. *et al.* Intrahepatic Transcriptional Signature Associated with Response to Interferon-α Treatment in the Woodchuck Model of Chronic Hepatitis B. en. *PLOS Pathogens* **11,** e1005103. ISSN: 1553-7374. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1005103 (2020) (Sept. 2015).

31. Negro, F. *et al.* Hepatitis delta virus (HDV) and woodchuck hepatitis virus (WHV) nucleic acids in tissues of HDV-infected chronic WHV carrier woodchucks. eng. *Journal of Virology* **63,** 1612–1618. ISSN: 0022-538X (Apr. 1989).

32. Aldabe, R., Suárez-Amarán, L., Usai, C. & González-Aseguinolaza, G. Animal Models of Chronic Hepatitis Delta Virus Infection Host–Virus Immunologic Interactions. *Pathogens* **4,** 46–65. ISSN: 2076-0817. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384072/ (2020) (Feb. 2015).

33. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. en. *Nature* **578,** 425–431. ISSN: 1476-4687. https://www.nature.com/articles/s41586-020-2007-4 (2020) (Feb. 2020).

34. Devoto, A. E. *et al.* Megaphages infect Prevotella and variants are widespread in gut microbiomes. en. *Nature Microbiology* **4,** 693–700. ISSN: 2058-5276. https://www.nature.com/articles/s41564-018-0338-9 (2020) (Apr. 2019).

35. *The Cost of Sequencing a Human Genome* en. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (2020).

36. Milham, M. P. *et al.* Assessment of the impact of shared brain imaging data on the scientific literature. en. *Nature Communications* **9,** 1–7. ISSN: 2041-1723. https://www.nature.com/articles/s41467-018-04976-1 (2020) (July 2018).

37. Lee, Y. N. & Bieniasz, P. D. Reconstitution of an infectious human endogenous retrovirus. eng. *PLoS pathogens* **3,** e10. ISSN: 1553-7374 (Jan. 2007).

38. Becker, M. M. *et al.* Synthetic recombinant bat SARS-like coronavirus is infectious in cultured cells and in mice. eng. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 19944–19949. ISSN: 1091-6490 (Dec. 2008).

39. Letko, M., Seifert, S. N., Olival, K. J., Plowright, R. K. & Munster, V. J. Bat-borne virus diversity, spillover and emergence. en. *Nature Reviews Microbiology* **18,** 461–471. ISSN: 1740-1534. https://www.nature.com/articles/s41579-020-0394-z (2020) (Aug. 2020).

40. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. en. *Nature Microbiology* **5,** 562–569. ISSN: 2058-5276. https://www.nature.com/articles/s41564-020-0688-y (2020) (Apr. 2020).

41. Damas, J. *et al.* Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. *bioRxiv.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7263403/ (2020) (Apr. 2020).

42. Díaz, S. *et al.* Pervasive human-driven decline of life on Earth points to the need for transformative change. eng. *Science (New York, N.Y.)* **366.** ISSN: 1095-9203 (2019).

43. Chase, J. M., Blowes, S. A., Knight, T. M., Gerstner, K. & May, F. Ecosystem decay exacerbates biodiversity loss with habitat loss. en. *Nature,* 1–6. ISSN: 1476-4687. https://www.nature.com/articles/s41586-020-2531-2 (2020) (July 2020).

44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature Methods* **9,** 357–359. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.1923 (2020) (Apr. 2012).

45. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. en. *Nature Methods* **12,** 59–60. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.3176 (2020) (Jan. 2015).

46. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461. ISSN: 1367-4803. eprint: https://academic.oup.com/bioinformatics/article-pdf/26/19/2460/16896486/btq461.pdf. https://doi.org/10.1093/bioinformatics/btq461 (Aug. 2010).

47. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. eng. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **13,** 1028–1040. ISSN: 1066-5277 (June 2006).

48. Hunt, M. *et al.* IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* **31,** 2374–2376. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btv120 (Feb. 2015).

49. Sawicki, S. G. & Sawicki, D. L. in *Corona- and Related Viruses: Current Concepts in Molecular Biology and Pathogenesis* (eds Talbot, P. J. & Levy, G. A.) 499–506 (Springer US, Boston, MA, 1995). ISBN: 978-1-4615-1899-0. https://doi.org/10.1007/978-1-4615-1899-0_79.

50. Meleshko, D. & Korobeynikov, A. coronaSPAdes: from biosynthetic gene clusters to coronaviral assemblies. *bioRxiv.* https://www.biorxiv.org/content/early/2020/07/28/2020.07.28.224584 (2020).

51. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27,** 824–834. http://genome.cshlp.org/content/27/5/824.abstract (2017).

52. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience* **8.** giz100. ISSN: 2047-217X. https://doi.org/10.1093/gigascience/giz100 (Sept. 2019).

53. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics.* btaa490. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btaa490 (May 2020).

54. Meleshko, D. *et al.* BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. *Genome Research* **29,** 1352–1362. http://genome.cshlp.org/content/29/8/1352.abstract (2019).

55. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Research* **47,** D427–D432. ISSN: 0305-1048. https://doi.org/10.1093/nar/gky995 (Oct. 2018).

56. Thiel, V. *et al.* Mechanisms and enzymes involved in SARS coronavirus genome expression. *Journal of General Virology* **84,** 2305–2315. https://doi.org/10.1099/vir.0.19424-0 (Sept. 2003).

57. Altman, T. *DARTH Coronavirus Annotation Pipeline* https://bitbucket.org/tomeraltman/darth/src/master/ (2020).

58. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16,** 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2 (June 2000).

59. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7** (ed Pearson, W. R.) e1002195. https://doi.org/10.1371/journal.pcbi.1002195 (Oct. 2011).

60. Team, T. P. *Pfam SARS-CoV-2 special update (part 2)* en. Library Catalog: xfam.wordpress.com. Apr. 2020. https://xfam.wordpress.com/2020/04/06/pfam-sars-cov-2-special-update-part-2/ (2020).

61. Schäffer, A. A. *et al.* VADR: validation and annotation of virus sequence submissions to GenBank. *BMC Bioinformatics* **21,** 211. ISSN: 1471-2105. https://doi.org/10.1186/s12859-020-3537-3 (2020) (May 2020).

62. Nawrocki, E. *Coronavirus annotation using VADR* en. Library Catalog: github.com. https://github.com/nawrockie/vadr/wiki/Coronavirus-annotation#build (2020).

63. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29,** 2933–2935. ISSN: 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810854/ (2020) (Nov. 2013).

64. Team, T. R. *Rfam Coronavirus Special Release* en. Library Catalog: xfam.wordpress.com. Apr. 2020. https://xfam.wordpress.com/2020/04/27/rfam-coronavirus-release/ (2020).

65. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. en. *Nucleic Acids Research* **38.** Publisher: Oxford Academic, e191–e191. ISSN: 0305-1048. https://academic.oup.com/nar/article/38/20/e191/1317565 (2020) (Nov. 2010).

66. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. eng. *Bioinformatics (Oxford, England)* **25,** 2078–2079. ISSN: 1367-4811 (Aug. 2009).

67. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. eng. *Genome Biology* **17,** 66. ISSN: 1474-760X (Apr. 2016).

68. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. en. *Cancer Research* **77.** Publisher: American Association for Cancer Research Section: Focus on Computer Resources, e31–e34. ISSN: 0008-5472, 1538-7445. https://cancerres.aacrjournals.org/content/77/21/e31 (2020) (Nov. 2017).

69. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology* **6,** R44. ISSN: 1474-760X. https://doi.org/10.1186/gb-2005-6-5-r44 (2020) (Apr. 2005).

70. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Research* **40,** D136–D143. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/40/D1/D136/9480848/gkr1178.pdf. https://doi.org/10.1093/nar/gkr1178 (Dec. 2011).

71. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47,** D506–D515. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/47/D1/D506/27437297/gky1049.pdf. https://doi.org/10.1093/nar/gky1049 (Nov. 2018).

72. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797. ISSN: 0305-1048. eprint: https://academic.oup.com/nar/article-pdf/32/5/1792/7055030/gkh340.pdf. https://doi.org/10.1093/nar/gkh340 (Mar. 2004).

73. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. en. *Bioinformatics* **35,** 4453–4455. ISSN: 1367-4803. https://academic.oup.com/bioinformatics/article/35/21/4453/5487384 (2020) (Nov. 2019).

74. Barbera, P. *et al.* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. en. *Systematic Biology* **68,** 365–369. ISSN: 1063-5157. https://academic.oup.com/sysbio/article/68/2/365/5079844 (2020) (Mar. 2019).

75. Czech, L., Barbera, P. & Stamatakis, A. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* (ed Schwartz, R.) ISSN: 1367-4803. https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa070/5722201 (Feb. 2020).

76. Morel, B., Kozlov, A. M. & Stamatakis, A. ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. en. *Bioinformatics* **35,** 1771–1773. ISSN: 1367-4803. https://academic.oup.com/bioinformatics/article/35/10/1771/5132696 (2020) (May 2019).

77. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. en. *Molecular Biology and Evolution* **37,** 291–294. ISSN: 0737-4038. https://academic.oup.com/mbe/article/37/1/291/5552155 (2020) (Jan. 2020).

78. Felsenstein, J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. en. *Evolution* **39,** 783–791. ISSN: 00143820. http://doi.wiley.com/10.1111/j.1558-5646.1985.tb00420.x (2020) (July 1985).

79. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. en. *Nature* **556,** 452–456. ISSN: 1476-4687. https://www.nature.com/articles/s41586-018-0043-0 (2020) (Apr. 2018).

1 Materials and Methods

1.1 Serratus alignment architecture

`Serratus` (`https://github.com/ababaian/serratus`) is an open-source tool designed for ultra-high throughput sequence alignment against a query sequence or pangenome (Extended Figure 1).

1.1.1 Computing cluster architecture

The processing of each sequencing library is split into three modules `dl` (download), `align`, and `merge`. The `dl` module acquires compressed data (.sra format) via `prefetch`, from the AWS S3 mirror of the SRA, decompresses to FASTQ, and splits the data into `fq-blocks` of 1 million reads or read-pairs into a temporary S3 cache bucket. To mitigate excessive disk usage caused by a few large datasets, a limit of 100 million reads per dataset was imposed. The `align` module reads individual fq-blocks and aligns to an indexed database of user-provided query sequences using either `bowtie2` (v2.3.4.1, `--very-sensitive-local`) [44] for nucleotide search, or `diamond` (v0.9.31, `--unal 0 -k 1 -b 0.35`) [45] for translated-protein search. Finally, the `merge` module concatenates the aligned blocks into a single output file (`.bam` for nucleotide, or `.pro` for protein) and generates alignment statistics with a Python script (see `Summarizer` below).

1.1.2 Computing resource allocation

Each component is launched from a separate AWS autoscaling group with its own launch template, allowing the user to tailor instance requirements per task. This enabled us to minimise the use of costly block storage during compute-bound tasks such as alignment. We used the following Spot instance types; `dl`: 250GB SSD block storage, 8vCPUs, 32GB RAM (r5.xlarge) ~1300 instances; `align`: 10GB SSD block storage, 8vCPUs, 8GB RAM (c5.xlarge) ~4,300 instances; `merge`: 150GB SSD block storage, 4vCPUs, 4GB RAM (c5.large) ~60 instances. Users should note that it may be necessary to submit a service ticket to access more than the default 20 EC2 instance limit.

EC2 instances have higher network bandwidth (up to 1.25 GB/s) than block storage bandwidth (250 MB/s). To exploit this, we used S3 buckets as a data buffering and streaming system and to transfer data between instances following methods developed in a previous cloud architecture (`https://github.com/FredHutch/sra-pipeline`). This, combined with splitting of FASTQ files into individual blocks, effectively eliminated file input/output (i/o) as a bottleneck, since the available i/o is multiplied per running instance (conceptually analogous to a RAID0 configuration).

Using S3 as a buffer also allowed us to decouple the input and output of each module S3 storage is cheap enough that in the event of unexpected issues (e.g., exceeding EC2 quotas) we could resolve problems and resume processing. For example, shutting done the `align` modules to hotfix a genome indexing problem without having to re-run the `dl` modules.

1.1.3 Work queue and scheduling

The `Serratus` scheduler node controls the number of desired instances to be created for each component of the workflow, based on the available work queue. We implemented a pull-based work queue. Upon boot-up each instance launches a number of worker threads equal to the number of CPU available. Each worker independently manages itself via a boot script, and query the scheduler for available tasks. Upon completion of the task, the worker updates the scheduler of the result: success, or fail, and queries for a new task. Under ideal conditions, this allows for a response time in the hundreds of milliseconds, worst case, keeping cluster throughput high. Each task typically lasts several minutes.

The scheduler itself was implemented using Postgres (for persistence and concurrency) and Flask (to pool connections and translate REST queries into SQL). The Flask layer allowed us to scale the cluster past the number of simultaneous sessions manageable by a single Postgres instance. The work queue can also be managed manually by the user, to perform operations such as re-attempt downloading of an SRA accession upon a failure or to pause an operation while debugging.

The system is designed to be fully self-scaling. An "autoscaling controller" was implemented which scales-in or scales-out the desired number of instances per task every five minutes based on the work queue. As a backstop, when all workers on an instance fail to receive work instructions from the scheduler, the instance is shut-down. Finally a "job cleaner" component checks the active jobs against currently running instances. If an instance has disappear due to SPOT termination or manual shutdown, it resets the job allowing it to be processed up by the next available instance.

To monitor cluster performance in real-time, we used `Prometheus` and `node_exporter` to retrieve CPU, disk, memory, and networking statistics from each instance, `postgres_exporter` to expose performance information about the work queue, and Python exporter to export information from the Flask server. This allowed us to identify and diagnose performance problems within minutes to avoid costly overruns.

1.1.4 Generating viral summary reports

We define a viral pangenome as the entire collection of reference sequences belonging to a taxonomic viral family, which may contain both full-length genomes and sequence fragments such as those based on RdRp amplicon sequencing.

We developed a `Summarizer` module written in Python to provide a compact, human- and machine-readable synopsis of the alignments generated for each SRA dataset. The method was implemented in `serratus_summarizer.py` for nucleotide alignment and `serratus_psummarizer.py` for amino acid alignments. Reports generated by the `Summarizer` are text files with three sections described in detail online (`https://github.com/ababaian/serratus/wiki/.summary-Reports`). In brief, each contains a header section with alignment meta-data and one-line summaries for each virus family pangenome, reference sequence and gene respectively, with gene summaries provided for protein alignments only.

For each summary line we include descriptive statistics gathered from the alignment data such as the number of aligned reads, estimated read depth, mean alignment identity, and coverage, i.e. the distribution of reads across each reference sequence or pangenome. Coverage is measured by dividing a reference sequence into 25 equal bins and depicted as an ASCII text string of 25 symbols, one per bin; for example `oaooomoUU:oWWUUWOWamWAAUW`. Each symbol represents $\lfloor log_2(n+1) \rfloor$ where n is the number of reads aligned to a bin in this order: `_.:uwaomUWAOM^`. Thus, '_' indicates no reads, '`.`' exactly one read, '`:`' two reads, '`u`' 3-4 reads, '`w`' 5-7 reads and so on; '`^`' represents $> 2^{13} = 8,192$ reads in the bin. For a pangenome, alignments to its reference sequences are projected onto a corresponding set of 25 bins. For a complete genome, the projected pangenome bin number $1, 2, \ldots, 25$ is the same as the reference sequence bin number. For a fragment, a bin is projected onto the pangenome bin implied by the alignment of the fragment to a complete genome. For example, if the start of a fragment aligns half way into a complete genome, bin 1 of the fragment is projected to bin $\lfloor \frac{25}{2} \rfloor = 12$ of the pangenome. The introduction of pangenome bins was motivated by the observation that `bowtie2` selects an alignment at random when there are two or more top-scoring alignments, which tends to distribute coverage over several reference sequences when a single viral genome is present in the reads. Coverage of a single reference genome may therefore be fragmented, and binning to a pangenome better assesses coverage over a putative viral genome in the reads while retaining pangenome sequence diversity for detection.

1.1.5 Identification of viral families within a sequencing dataset

The `Summarizer` implements a binary classifier predicting the presence or absence of each virus family in the query. For a given family F, the classifier reports a score in the range [0,100] with the goal of assigning a high score to a dataset if it contains F and a low score if it does not. Setting a threshold on the score divides datasets into disjoint subsets representing predicted positive and negative detections of family F. The choice of threshold implies a trade-off between false positives and false negatives. Sorting by decreasing score ranks datasets in decreasing order of confidence that F is present in the reads.

Naively, a natural measure of the presence of a virus family is the number of alignments to its reference sequences. However, alignments may be induced by non-homologous sequence similarity, for example low-complexity sequence. The score for a family was therefore designed to reflect the overall coverage of a pangenome because coverage across all or most of a pangenome is more likely to reflect true homology, i.e. the presence of a related virus. Ideally, coverage would be measured individually for each base in the reference sequence, but this could add undesirable overhead in compute time and memory for a process which is executed in the Linux alignment pipe (FASTQ decompression → aligner → `Summarizer` → alignment file compression). Coverage was therefore measured by binning as described above, which can be implemented with minimal overhead.

A virus that is present in the reads with coverage too low to enable an assembly may have less practical value than an assembled genome. Also, genomes with lower identity to previously known sequences will tend to contain more novel biological information than genomes with high identity and will tend to have fewer alignments highly diverged segments. With these considerations in mind, the classifier was designed to give higher scores when coverage is high, read depth is high, and/or identity is low. This was accomplished as follows. Let H be the number of bins with at least 8 alignments to F, and L be the number of bins with from 1 to 7 alignments. Let S be the mean alignment percentage identity, and define the identity weight $w = (\frac{S}{100})^{-3}$, which is designed to give higher

weight to lower identities, noting that w is close to one when identity is close to 100% and increases rapidly at lower identities. The classification score for family F is calculated as $Z_F = max(w(4H + L)), 100)$. By construction, Z_F has a maximum of 100 when coverage is consistently high across a pangenome, and is also high when identity is low and coverage is moderate, which may reflect high read depth but many false negative alignments due to low identity. Thus, Z_F is greater than zero when there is at least one alignment to F and assigns higher scores to SRA datasets which are more likely to support successful assembly of a virus belonging to F.

1.2 Defining viral pangenomes and the SRA search space

1.2.1 Nucleotide search pangenomes

To create a collection of viral pangenomes, a comprehensive set of complete and partial genomes representing the genetic diversity of each viral family, we used two approaches.

For *Coronaviridae*, we combined all RefSeq (n = 64) and GenBank (n = 37,451) records matching the NCBI Nucleotid server query `"txid11118[Organism:exp]"` (date accessed: June 1st 2020). Sequences <200 nt were excluded as well as sequences identified to contain non-CoV contaminants during preliminary testing (such as plasmid DNA or ribosomal RNA fragments). Remaining sequences were clustered at 99% identity with `UCLUST` [46] and masked by `Dustmasker` [47] with `--window 30` and `--window 64`. The final query contained 10,101 CoV sequences (accessions in Extended Table 1a, masked coordinates in Extended Table 1b).

For all other vertebrate viral family pangenomes, RefSeq sequences (n = 2,849) were downloaded from the NCBI Nucleotide server with the query `"Viruses[Organism] AND srcdb_refseq[PROP] NOT wgs[PROP] NOT cellular organisms[ORGN] NOT AC_000001:AC_999999[PACC] AND ("vhost human"[Filter] AND "vhost vertebrates"[Filter])"` (date accessed: May 17th 2020). Retroviruses (n = 80) were excluded as preliminary testing yielded excessive numbers of alignments to transcribed endogenous retroviruses. Each sequence was annotated with its taxonomic family according to its RefSeq record; those for which no family was assigned by RefSeq (n = 81) were designated as "unknown".

The collection of these pangenomes was termed `cov3m`, and was the sequence reference used for this study.

1.2.2 Amino acid search panproteome

The protein search query was composed of the following sequences: (i) CoV proteins (method described under `SerraTax`); (ii) all CDS annotated from RefSeqs in the nucleotide query; (iii) deltavirus antigen proteins from accessions NC_001653, M21012, X60193, L22063, AF018077, AJ584848, AJ584847, AJ584844, AJ584849, MT649207, MT649208, MT649206, NC_040845, NC_040729, MN031240, MN031239, MK962760, and MK962759 (iv) large terminase (TerL), viral DNA-packaging protein and major capsid protein (MCP) for the huge phage clades reported in [Al-Shayeb & Sachdeva et al 2020 TODO citation].

1.2.3 SRA search space

To run `Serratus`, a target list of SRA run accessions is required. For this work, we designed target lists broadly classified as human, mouse, mammal, vertebrate, invertebrate, bat (including genome sequencing libraries), virome and metagenome (Extended Table 1c). Each list contained accessions of RNA-seq, meta-genomic, and meta-transcriptome runs for these organisms; some run accessions appeared in more than one list. Prior to each `Serratus` run, the lists were depleted for accessions already analyzed. Re-processing of a failed dataset was attempted at least twice. In total we were able to generate alignments to the query pangenomes for 3,837,755/4,059,695 (94.5%) of the targeted SRA accessions.

1.3 User interfaces for the Serratus databases

We implemented an on-going, multi-tiered release policy for code and data generated by this study, as follows. All code, electronic notebooks and raw data is immediately available at `https://github.com/ababaian/serratus` and on the `s3://serratus-public/` bucket, respectively. Upon completion of a project milestone, a structured data-release is issued containing raw data into our viral data warehouse `s3://lovelywater/`. For example, at the time of writing the .bam alignment files from 3.84 million SRA runs are stored in `s3://lovelywater/bam/X.bam`; .summary files are `s3://lovelywater/summary/X.summary`, where `X` is a SRA run accession. These structured releases enable downstream and third-party programmatic access to the data.

Summary files for every searched SRA dataset are parsed into a `PostgreSQL` relational database which can be queried remotely via an AWS Relational Database (RDS) server. This enables users and programs to perform

complex operations such as retrieving summaries and meta-data for all SRA runs matching a given reference sequence with above a given classifier score threshold. For example, all records containing at least 20 aligned reads to Hepatitis Delta Virus (NC_001653.2) and the associated host taxonomy for the corresponding SRA datasets.

For users unfamiliar with SQL queries we developed `Tantalus` (`https://github.com/serratus-bio/tantalus`, an `R` programming-language package which directly interfaces the `Serratus` RDS server to retrieve summary information as data-frames. `Tantalus` also offers functions to explore and visualize the data. Finally, the `Serratus` data can be explored via a graphical web interface by accession, virus, or viral family at `https:/serratus.io`. The website uses `javascript` to access the RDS server and create a graphical report with an overview of viral families found in each SRA accession matching a user query.

All four data access interfaces are under ongoing development, receiving community feedback via their respective GitHub issue trackers to facilitate the translation of this data collection into an effective viral discovery resource. Documentation for data access methods is available at `https://serratus.io/access`

1.4 Viral assembly and annotation

1.4.1 `coronaSPAdes`

RNA viral genome assembly faces several distinct challenges stemming from technical and biological bias in sequencing data. During library preparation, reverse transcription introduces 5′ end coverage bias, and GC-content skew and secondary structures lead to unequal PCR amplification [48]. Technical bias is confounded by biological complexity such as intra-sample sequence variation due to transcript isoforms, as found in CoV [49] and/or to presence of multiple strains.

To address the assembly challenges specific to RNA viruses, we developed `coronaSPAdes`, described in detail in a companion manuscript [50]. In brief, `rnaviralSPAdes` and the more specialized variant, `coronaSPAdes`, combines algorithms and methods from several previous approaches based on `metaSPAdes` [51], `rnaSPAdes` [52] and `metaviralSPAdes` [53] with a `HMMPathExtension` step. `coronaSPAdes` constructs an assembly graph from a RNA-sequencing dataset (transcriptome, meta-transcriptome, and meta-virome are supported), removing expected sequencing artifacts such as low-complexity (poly-A / poly-T) tips, edges, single-strand chimeric loops or double-strand hairpins [52] and subspecies-bases variation [53].

To deal with possible misassemblies and high-covered sequencing artifacts, a secondary `HMMPathExtension` step is performed to leverage orthogonal information about the expected viral genome. Protein domains are identified on all assembly graphs using a set of viral hidden Markov models (HMMs), and similar to biosyntheticSPAdes [54], `HMMPathExtension` attempts to find paths on the assembly graph which pass through significant HMM matches in order.

`coronaSPAdes` is bundled with the Pfam SARS-CoV-2 set of HMMs [55], although these may be substituted by the user. This latter feature of `coronaSPAdes` was utilized for HDV assembly, where the HMM model of HDAg, the Hepatitis Delta Antigen, was used instead of Pfam SARS-CoV-2 set. Note that despite the name, these HMMs are quite general, modeling domains found in all coronavirus genera in addition to RdRp, which is found in many RNA virus families. Hits from these HMMs cover most bases in most known coronaviruse genomes, enabling the recovery of strain mixtures and splice variants.

1.4.2 Annotation of CoV assemblies

Accurate annotation of CoV genomes is challenging due to ribosomal frameshifts and polyproteins which are cleaved into maturation proteins [56], and thus previously-annotated viral genomes offer a guide to accurate gene-calls and protein functional predictions. However, while many of the viral genomes we were likely to recover would be similar to previously-annotated genomes in Refseq or GenBank, we anticipated that many of the genomes would be taxonomically distant from any available reference. To address these constraints, we developed an annotation pipeline called `DARTH` [57][1] which leverages both reference-based and *ab initio* annotation approaches.

In brief, `DARTH` consists of the following phases: canonicalize the ordering and orientation of assembly contigs using conserved domain alignments, perform reference-based annotation of the contigs, annotate RNA secondary structure, *ab intio* gene-calling, generate files for aiding assembly and annotation diagnostics, and generate a master annotation file. It is important to put the contigs in the "expected" orientation and ordering to facilitate comparative analysis of synteny and as a requirement for genome deposition. To perform this canonicalization, `DARTH` generates the six-frame translation of the contigs using the `transeq` [58] and uses HMMER3 [59] to search the translations for Pfam domain models specific to CoV [60]. DARTH compares the Pfam accessions from the

[1] `https://bitbucket.org/tomeraltman/darth/`

HMMER alignment to the NCBI SARS-CoV-2 reference genome (NCBI Nucleotide accession NC_045512.2) to determine the correct ordering and orientation, and produces an updated assembly FASTA file. `DARTH` performs reference-based annotation using `VADR` [61], which provides a set of genome models for all CoV RefSeq genomes [62]. `VADR` provides annotations of gene coordinates, polyprotein cleavage sites, and functional annotation of all proteins. `DARTH` supplements the `VADR` annotation by using `Infernal` [63] to scan the contigs against the SARS-CoV-2 Rfam release [64] which provides updated models of CoV 5′ and 3′ untranslated regions (UTRs) along with stem-loop structures associated with programmed ribosomal frame-shifts. While `VADR` provides reference-based gene-calling, `DARTH` also provides *ab initio* gene-calling by using `FragGeneScan` [65], a frameshift-aware gene caller. `DARTH` also generates auxiliary files which are useful for assembly quality and annotation diagnostics, such as indexed BAM files created with `SAMtools` [66] representing self-alignment of the trimmed reads to the canonicalized assembly using `bowtie2` [44], and variant-calls using `bcftools` from `SAMtools`. `DARTH` generates these files so that the can be easily loaded into a genome browser such as `JBrowse` [67] or `IGV` [68]. As the final step `DARTH` generates a single Generic Feature Format (GFF) 3.0 file [69] containing combined set of annotation information described above, ready for use in a genome browser, or for submitting the annotation and sequence to a genome repository.

1.4.3 Deploying the assembly and annotation workflow

The `Serratus` searches described above identified 37,131 libraries (14,304 by nucleotide and 23,898 by amino acid) as potentially positive for CoV (score ≥20 and ≥10 reads). To supplement this search we also employed a recently developed index of the SRA called STAT [5] with which identified an additional 18,584 SRA datasets not in the defined SRA search space. The STAT BigQuery was `WHERE tax_id=11118 AND total_count >1"` accessed on June 24th 2020.

We used AWS Batch to launch thousands of assemblies of NCBI accessions simultaneously. The workflow consists of four standard parts: a job queue, a job definition, a compute environment, and finally, the jobs themselves. A CloudFormation template[2] was created for building all parts of the cloud infrastructure from the command line.

The job definition specifies a Docker image, and asks for 8 virtual CPUs (vCPUs, corresponding to threads) and 60 GB of memory per job, corresponding to a reasonable allocation for `coronaSPAdes`. The compute environment is the most involved component. We set it to run jobs on cost-effective Spot instances (`optimal` setting) with an additional cost-optimization strategy (`SPOT_CAPACITY_OPTIMIZED` setting), and allowing up to 40,000 vCPUs total. In addition, the compute environment specifies a launch template which, on each instance, i) automatically mounts an exclusive 1 TB EBS volume, allowing sufficient disk space for several concurrent assemblies, and ii) downloads the 5.4 GB CheckV database, to avoid bloating the Docker image.

The peak AWS usage of our Batch infrastructure was ~28,000 vCPUs, performing ~3,500 assemblies simultaneously. A total of 46,861 accessions out of 55,715 were assembled in a single day. They were then analysed by two methods to detect putative CoV contigs. The first method is CheckV, followed selecting contigs associated to known CoV genomes. The second method is a custom script[3] that parses `coronaSPAdes` BGC candidates and keeps contigs containing CoV domain(s). For each accession, we kept the set of contigs obtained by the first method (CheckV) if it is non-empty, and otherwise we kept the set of contigs from the second method (BGC). A majority (76%) of the assemblies were discarded for one of the following reasons: i) no CoV contigs were found by either filtering method, ii) reads were too short to be assembled, iii) Batch job or SRA download failed, or iv) `coronaSPAdes` ran out of memory. A total of 11,120 assemblies were considered for further analysis.

1.5 Per-base quality assessment of assembly contigs

With RNA-seq metagenomic reads, the number of reads per base may be highly variable at different locations in a viral genome. Regions of high coverage may be adjacent to regions with low coverage or no reads, causing breaks between contigs. Thus, a given base in a contig may have only one or very few reads as evidence, and as a consequence the reliability of base calls may be low in some regions of the assembly which could degrade inference of biological variations between genomes. The assemblers used in this work do not provide a per-base quality score, and to address this issue we used two complementary approaches: (1) reporting contig average coverage as a proxy for quality, and (2) self-aligning reads to the assembly sequence and calling variants to enable facile visual inspection of per-base coverage levels and significant variants in genome browsers (see Section 1.4.2).

[2] `https://gitlab.pasteur.fr/rchikhi_pasteur/serratus-batch-assembly/-/blob/master/template/template.yaml`

[3] `https://gitlab.pasteur.fr/rchikhi_pasteur/serratus-batch-assembly/-/blob/master/stats/bgc_parse_and_extract.py`

1.6 Taxonomy prediction for coronavirus genomes

We developed a module, `SerraTax`, to predict taxonomy for CoV genomes and assemblies (`https://github.com/ababaian/serratus/tree/master/containers/serratax`). `SerraTax` was designed with the following requirements in mind: provide taxonomy predictions for fragmented and partial assemblies in addition to complete genomes; report best-estimate predictions balancing over-classification and under-classification (too many and too few ranks, respectively); and assign an NCBI Taxonomy Database [70] identifier (`TaxID`).

Assigning a best-fit `TaxID` was not supported by any previously published taxonomy prediction software to the best of our knowledge; this requires assignment to intermediate ranks such as sub-genus and ranks below species (commonly called strains, but these ranks are not named in the Taxonomy database), and to unclassified taxa, e.g. `TaxID 2724161`, `unclassified Buldecovirus`, in cases where the genome is predicted to fall inside a named clade but outside all named taxa within that clade.

`SerraTax` uses a reference database containing domain sequences with `TaxID`s. This database was constructed as follows. Records annoated as CoV were downloaded from `UniProt` [71], and chain sequences were extracted. Each chain name, e.g. `Helicase`, was considered to be a separate domain. Chains were aligned to all complete coronavirus genomes in `GenBank` using `UBLAST` [46] to expand the repertoire of domain sequences. The reference sequences were clustered using `UCLUST` [46] at 97% sequence identity to reduce redundancy.

For a given query genome, open reading frames (ORFs) are extracted using the `EMBOSS getorf` software [58]. ORFs are aligned to the domain references and the top 16 reference sequences for each domain are combined with the best-matching query ORF. For each domain, a multiple alignment of the top 16 matches plus query ORF is constructed on the fly by `MUSCLE` [72] and a neighbour-joining tree is inferred from the alignment, also using `MUSCLE`. Finally, a consensus prediction is derived from the placement of the ORF in the domain trees. Thus, the presence of a single domain in the assembly suffices to enable a prediction; if more domains are present they are combined into a consensus.

1.7 Taxonomic assignment by phylogenetic placement

To generate an alternate taxonomic annotation of an assembled genome, we created a pipeline based on phylogenetic placement, `SerraPlace`.

To perform phylogenetic placement, a reference phylogenetic tree is required. To this end, we collected 823 reference amino acid RdRp sequences, spanning all *Coronaviridae*. To this set we added an outgroup RdRp sequence from the Torovirus family (NC_007447). We clustered the sequences to 99% identity using USEARCH ([46], `UCLUST` algorithm, `v11.0.667`), resulting in 546 centroid sequences. Subsequently we performed multiple sequence alignment on the clustered sequences using `MUSCLE` ([72], `v3.8.31`). We then performed maximum likelihood tree inference using `RAxML-NG` ([73], `PROTGTR+FO+G4`, `v0.9.0`), resulting in our reference tree.

To apply `SerraPlace` to a given genome, we first use HMMER ([59], `v3.3`) to generate a reference HMM, based on the reference alignment. We then split each contig into ORFs using `esl-translate`, and use `hmmsearch` (p-value cutoff 0.01) to identify those query ORFs that align with sufficient quality to the previously generated reference HMM. All ORFs that pass this test are considered valid input sequences for phylogenetic placement. Subsequently, we use `EPA-ng` ([74], `v0.3.7`) to place each sequence on the RdRp reference tree. This produces a set of likely placement locations on the tree, with an associated likelihood weight. We then use `Gappa` ([75], `v0.6.1`) to assign taxonomic information to each query, using the taxonomic information for the reference sequences. `Gappa` assigns taxonomy by first labelling the interior nodes of the reference tree by a consensus of the taxonomic labels of all descendant leaves of that node. If 66% of leaves share the same taxonomic label up to some level, then the internal node is assigned that label. Then, the likelihood weight associated with each sequence is assigned to the labels of internal nodes of the reference tree, according to where the query was placed.

From this result, we select that taxonomic label that accumulated the highest total likelihood weight as the taxonomic label of a sequence. Note that multiple ORFs of the same genome may result in a taxonomic label, in which case, we select the longest sequence as the source of the taxonomic assignment of the genome.

1.8 Phylogenetic inference

We performed phylogenetic inferences using a custom snakemake pipeline (available at `https://github.com/lczech/nidhoggr`), using `ParGenes` ([76], `v1.1.2`). `ParGenes` is a treesearch orchestrator, build on top of `ModelTest-NG` [77] and `RAxMLNG`, enabling higher levels of parallelisation for a given tree search.

To infer the maximum likelihood phylogenetic tree displayed in Extended Figure 2, we performed a tree search comprising 100 distinct starting trees (50 random, 50 parsimony), as well as 1000 bootstrap searches. We used

`ModelTestNG` to automatically select the best evolutionary model, which in this case was `LG+IU+G4m`. The pipeline also automatically produces versions of the best maximum likelihood tree annotated with Felsenstein's Bootstrap ([78]) support values, and Transfer Bootstrap Expectation ([79]) values, the latter of which was used in Extended Figure 2.

Data availability

Archival copies of all code generated for this study is available at `https://github.com/serratus-bio`. Electronic notebooks for experiments are available at `https://github.com/ababaian/serratus`. Access to all data generated in this study can be accessed at `https://serratus.io/access`. Assembled genomes contigs for this study are available at `https://serratus.io/access` pending deposition into public repositories.

Acknowledgments

The `Serratus` project is an initiative of the hackseqRNA genomics hackathon (`https://www.hackseq.com`). We would like to thank the many contributors for code snippets and bioinformatic discussion; E. Erhan, J. Chu, I. Birol, K. Wellman, C. Xu, M. Huss, K. Ha, E. Nawrocki, R. McLaughlin, C. Morgan-Lang, C. Blumberg, and the J. Brister lab. A. Rodrigues, S. McMillan, V. Wu, C. Kennet, K. Chao, and N. Pereyaslavsky for AWS support. We would also like to thank the J. Joy lab, G. Mordecai, J. Taylor, S. Roux, L. Bergner, R. Orton, and D. Streicker for virology discussions. We are grateful to the entire team managing the NCBI SRA. TA is grateful for the Advanced Research Computing resource at the University of British Columbia. PB was financially supported by the Klaus Tschira Foundation, RC by ANR Transipedia and Inception grants (PIA/ANR-16-CONV-0005, ANR-18-CE45-0020), AK and DM were supported by the Russian Science Foundation (grant 19-14-00172) and computation was carried out in part by Resource Centre "Computer Centre of SPbU". AK and DM are grateful to Saint Petersburg State University for the overall support of this work (project id: 51555639). Project support and computing resources were kindly provided by the University of British Columbia Community Health and Wellbeing Cloud Innovation Centre, powered by AWS. And special thanks to our patient and understanding partners.

Author Contributions

AB conceived and led the study. AB and JT designed and implemented the `Serratus` architecture. AB and RCE constructed the viral pangenomes and panproteomes. RCE developed the `SerraTax` and `Summarizer` modules. PB developed the `SerraPlace` tree placement and taxonomy prediction code and calculated maximum likelihood trees. TA developed the `DARTH` annotation pipeline and submitted the annotated genomes to ENA. DM and AK developed the `coronaSPAdes` assembler. RC implemented the assembly pipeline, and deployed the assembly and annotation pipeline. AB, VL, and DL designed and developed https://serratus.io and the SQL server. AB and GN developed the `Tantalus R` package. AB, RCE, TA, PB, DM, AK, and RC analysed the coronavirus and deltavirus data. BAS and JB designed the phage panproteome, assembled phage genomes, and conducted phylogenetic analyses. All authors contributed to data interpretation and writing the manuscript.

Correspondence

Correspondence should be addressed to AB.

Reporting Summary

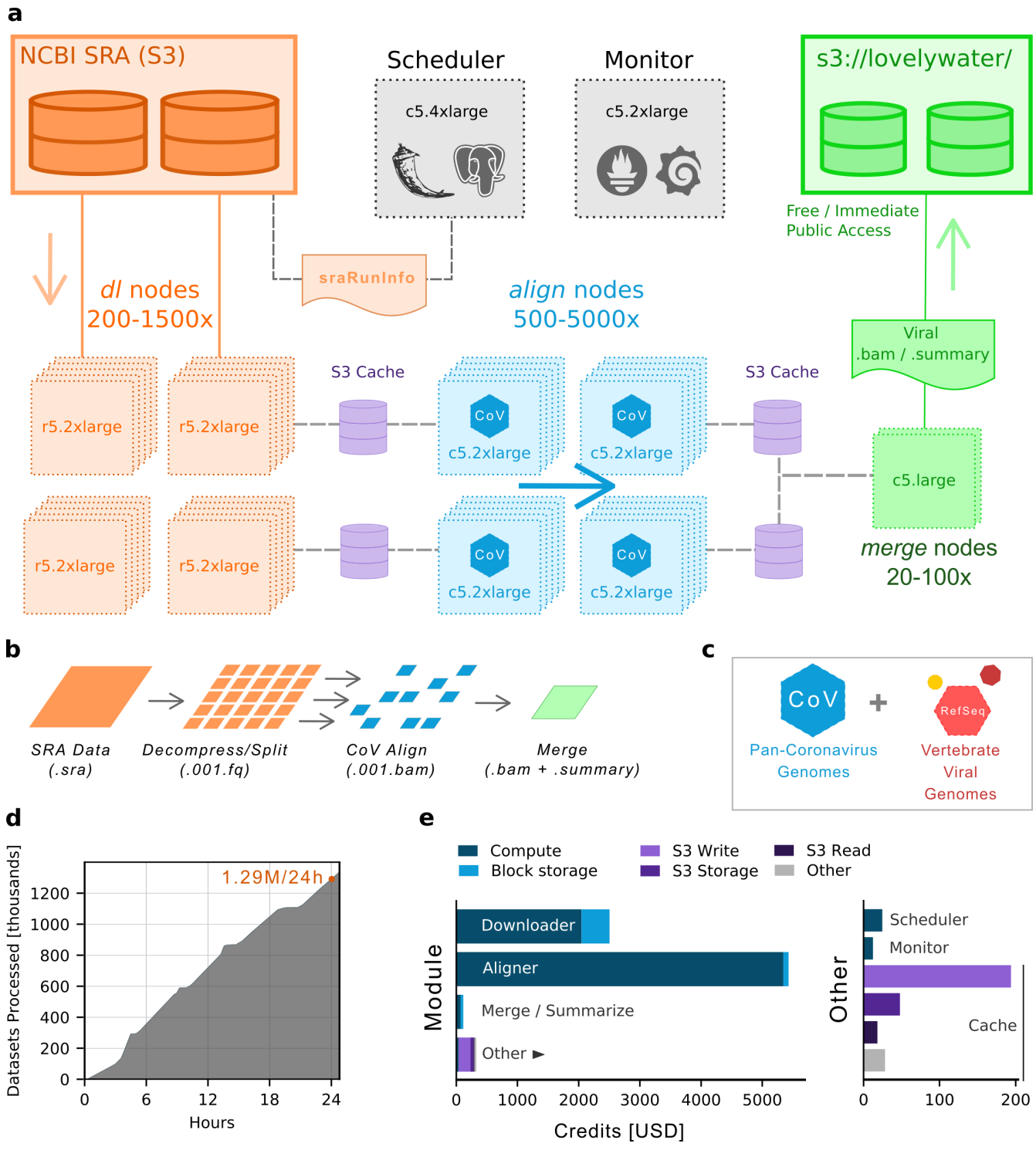
Does not apply.

Extended Table 1: **SRA run queries and search nucleotide accessions.** Queries and accessions from this study. **a** SRA queries to retrieve collections of datasets. **b** Nucleotide accessions compiled into the `cov3ma` reference query and c the sequence masked applied to those sequences.

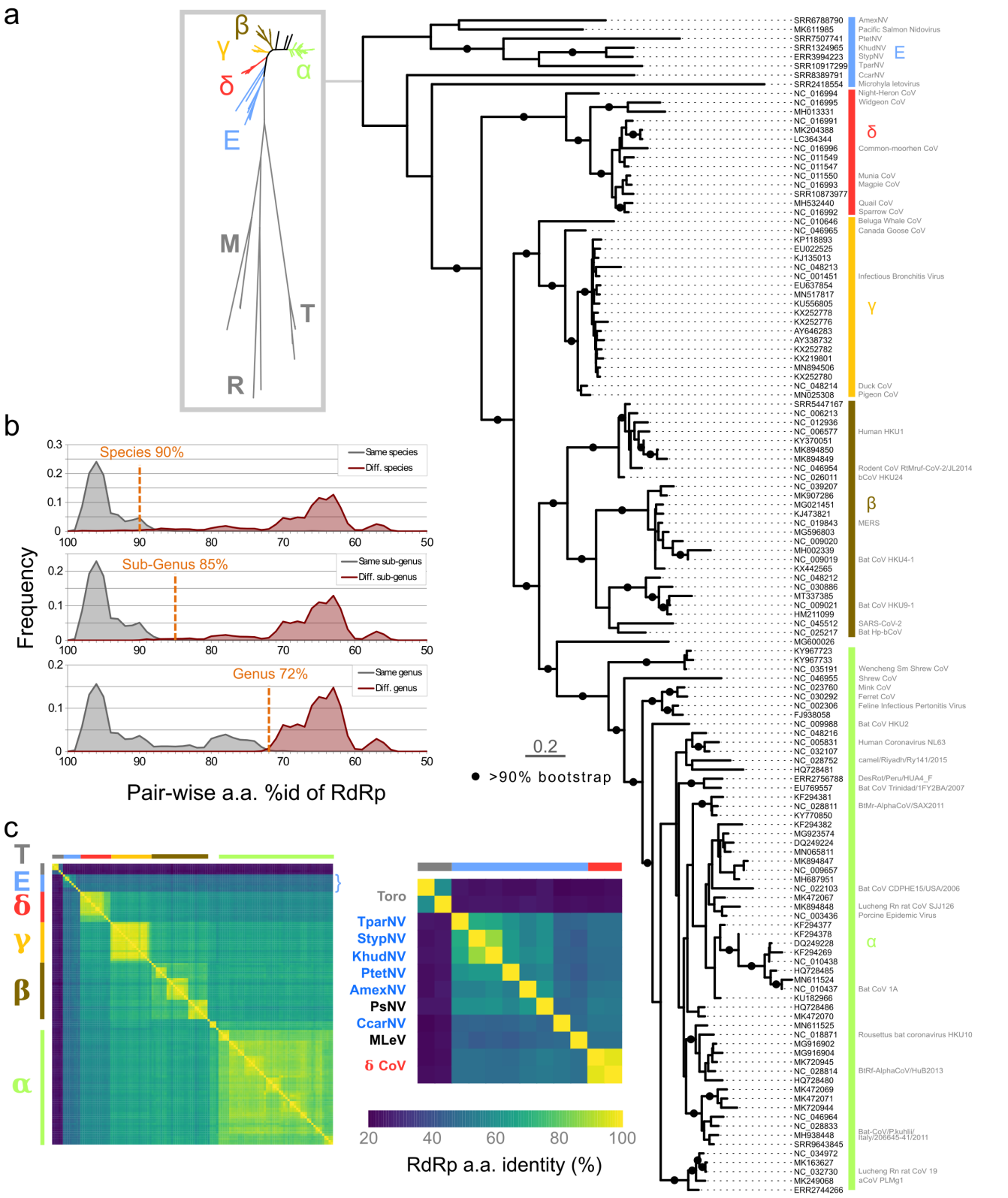
Extended Table 2: **Assembled *Coronaviridae* in the SRA.** **a** Run accessions, assembly statistics and select meta-data for the 11,120 runs for which *Coronaviridae*, or *Coronaviridae*-like sequences were assembled. **b** Assignment of assembled runs to operational taxonomic units (OTUs) based on 97% identity of the RNA dependent RNA polymerase (RdRp) domain. **c** Assignment of GenBank records to RdRp OTUs. **d** Assignment of expected viral host for GenBank records. **e** Taxonomic source for RdRp containing assemblies. **f** Supporting data for Figure 1.

Extended Tables

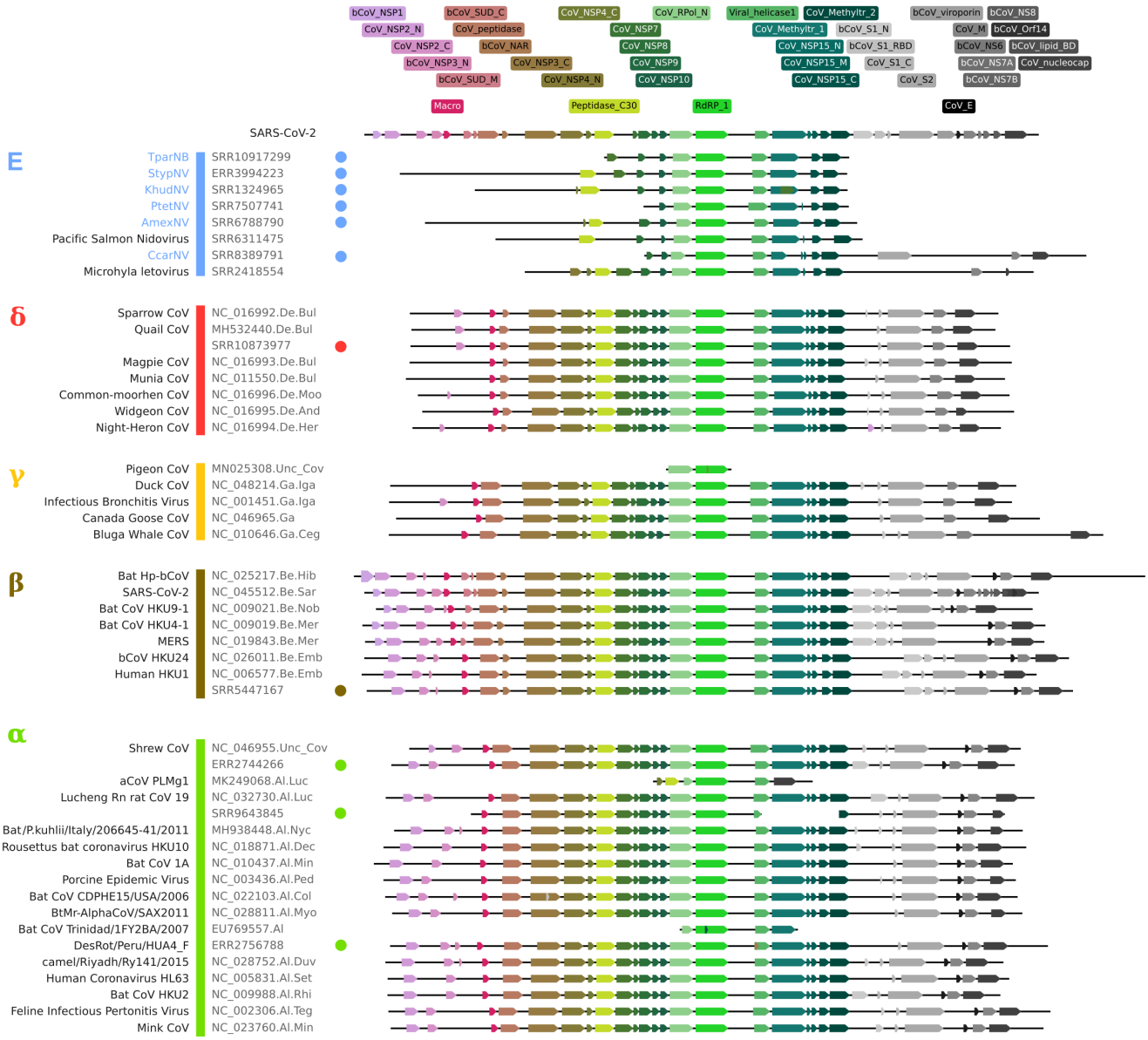
Extended Figures



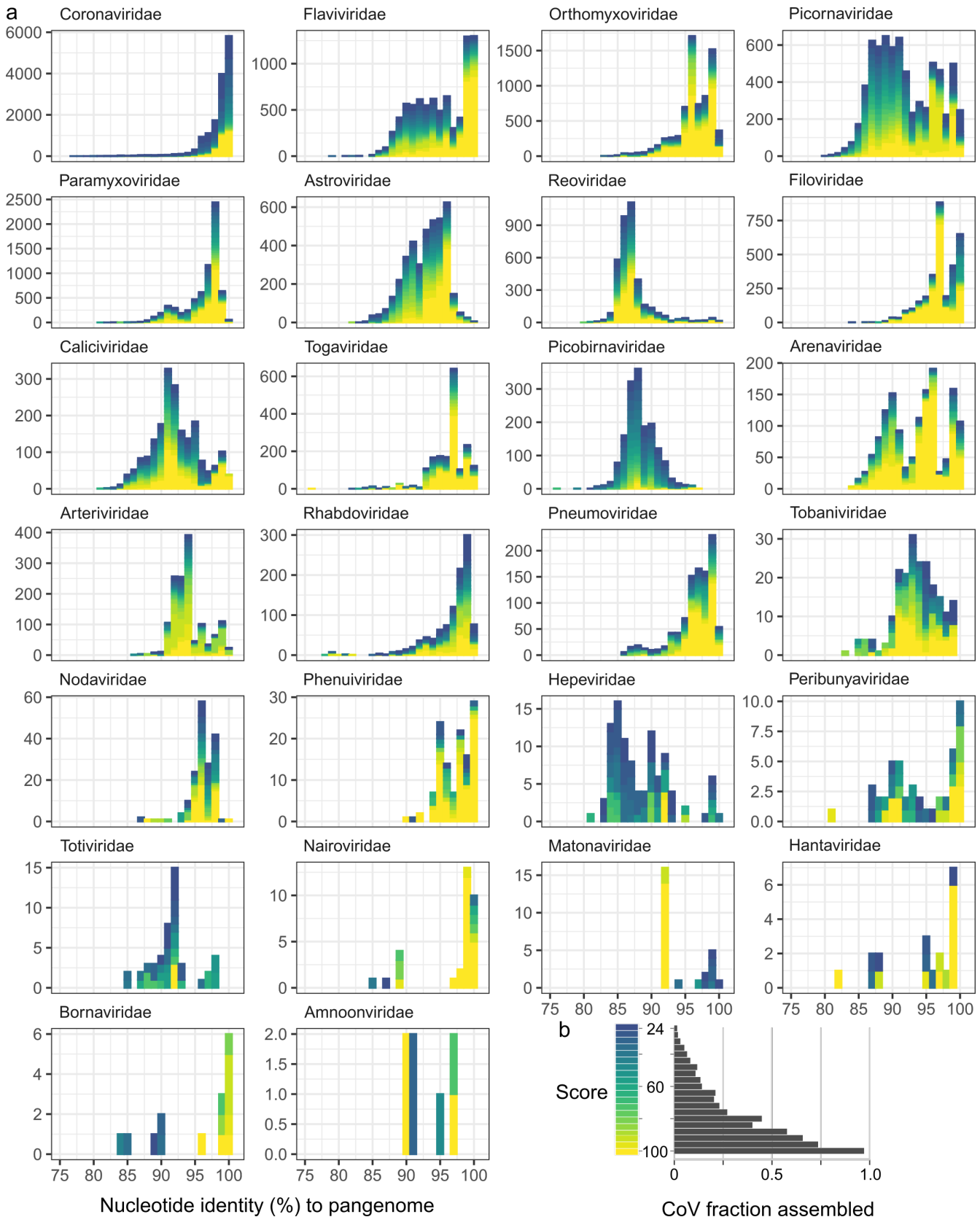
Extended Figure 1: **Overview of the *Serratus* architecture. a** Schematic and data workflow (**b**) as described in the methods for aligning to the viral pangenome (**c**). **d** A nucleotide alignment completion rate for *Serratus* shows stable and linear performance to complete 1.29 million SRA accessions in a 24-hour period. **e** Cost breakdown for this run. Compute costs between modules are an approximate comparison of CPU requirements of each step. The total average cost per completed SRA accession was $0.0062 US dollars or $0.1892 US dollars per terabase processed.



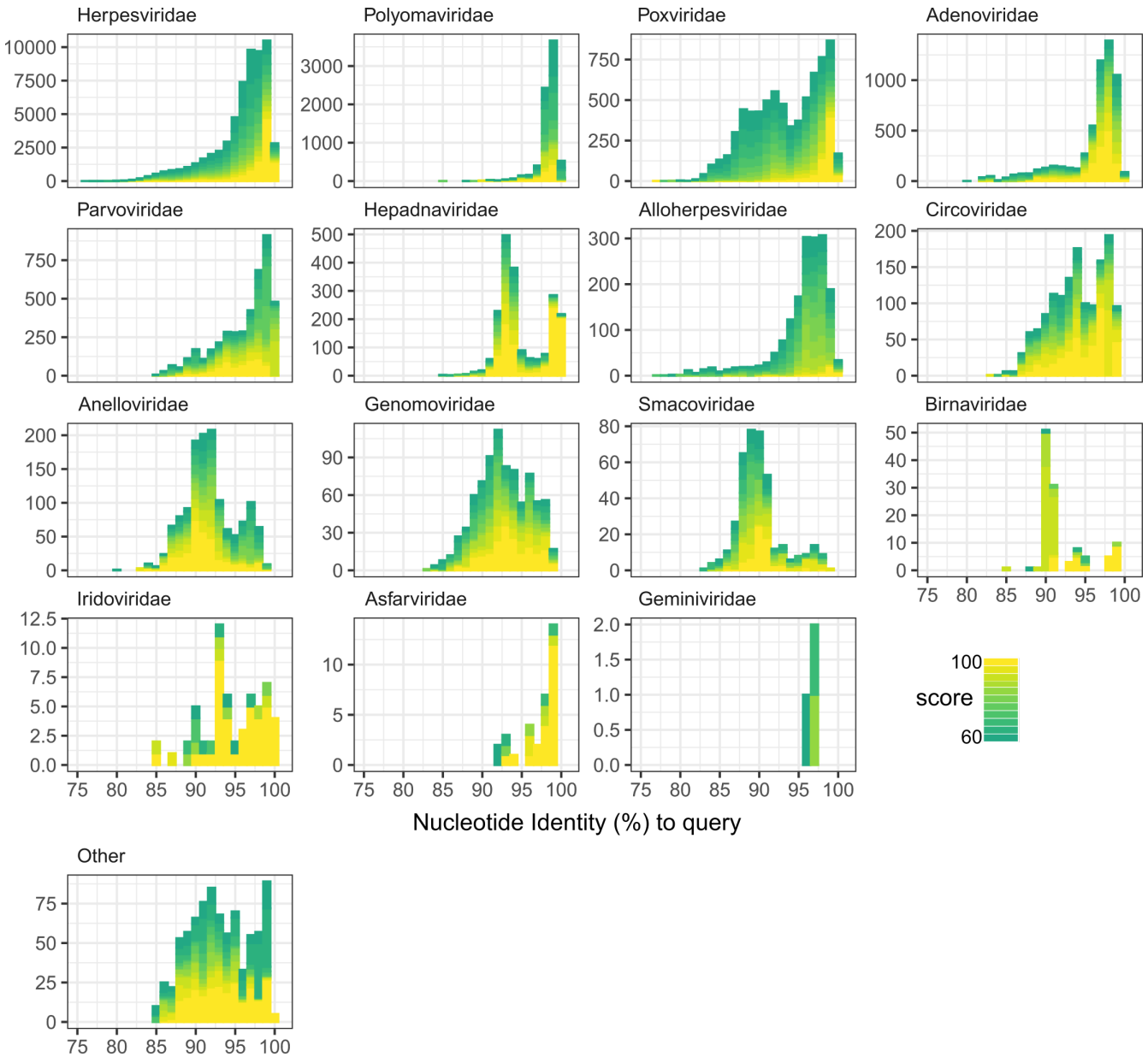
Extended Figure 2: **Phylogenetic relationships of *Coronaviridae* and related OTUs. a** Maximum likelihood (ML) tree of aligned OTU exemplar sequences from newly assembled genomes (SRR and ERR prefix) and GenBank. Dots indicate edges with transfer bootstrap > 0.9. See Section 1.8 for a detailed description of tree construction with inlay of unrooted tree including outgroups *Mesoniviridae* (M), *Tobaniviridae* (T), and *Roniviridae* (R). **b** Distribution of pair-wise sequence identities for RdRp sequences within and between distinct taxa at species, subgenus and genus rank, respectively. **c** Distribution of pair-wise RdRp identities for *Coronaviridae* genera and Group E (left) with expanded view of intra-group E identities (right).



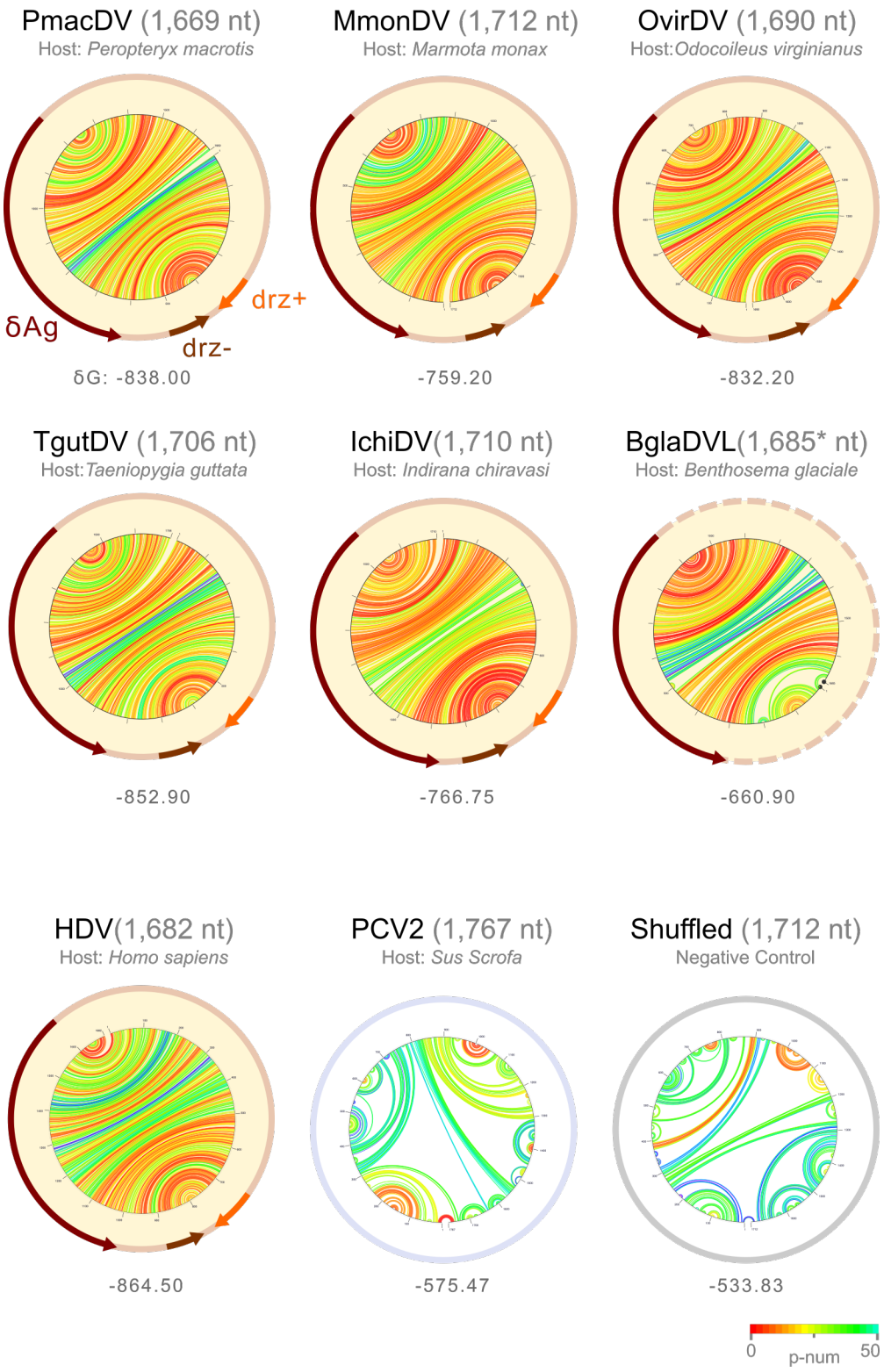
Extended Figure 3: **Genome Organisation *for Coronaviridae* and neighbours.** Hidden Markov Model (HMM) protein domain matches from the RdRp containing contigs or reference sequences for 47 exemplar operational taxonomic units (OTUs) grouped by genus (Extended Figure 2. OTUs identified in this study are indicated with a coloured circle.



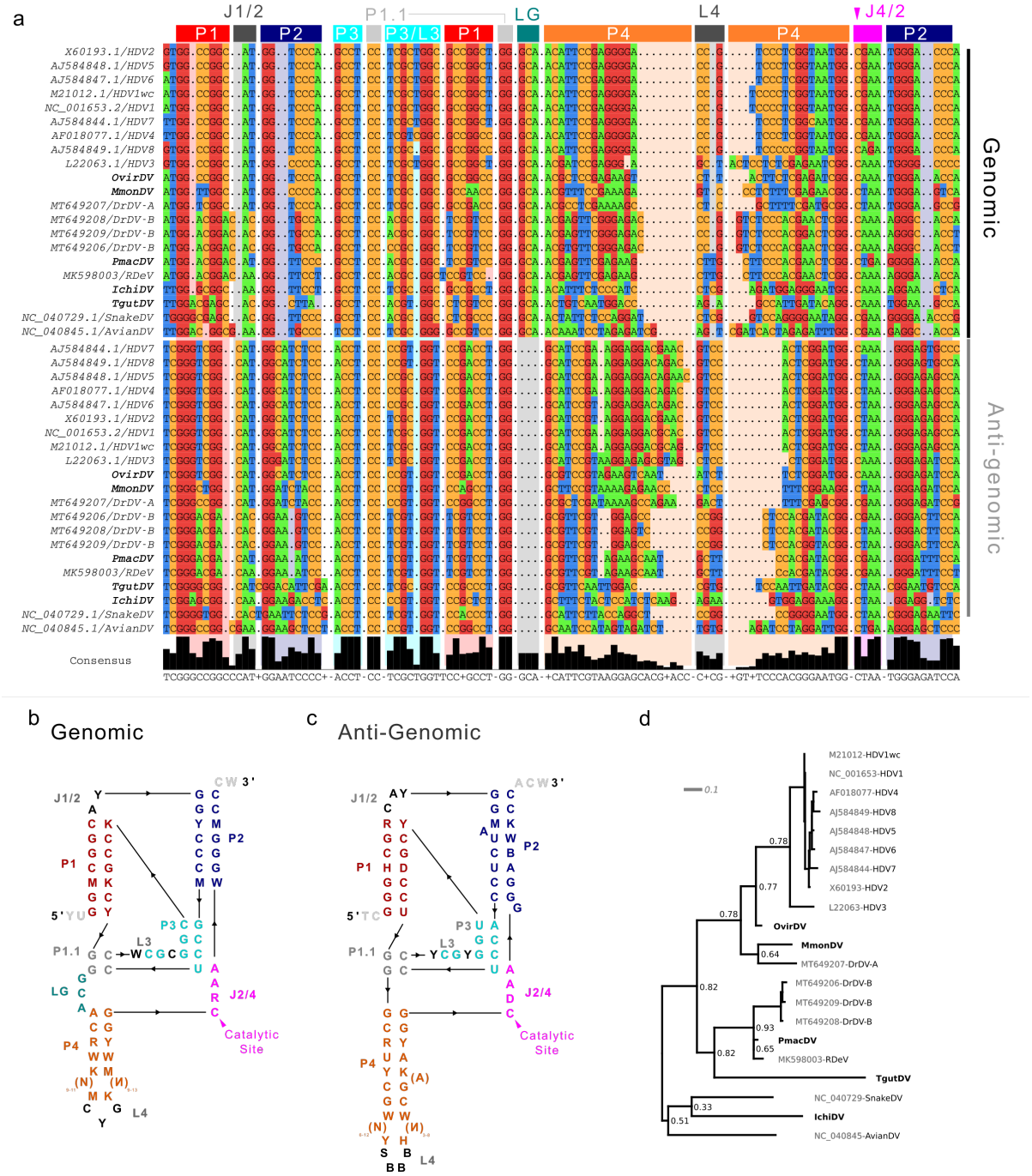
Extended Figure 4: **Distribution of RNA viral families in the SRA** The total number of datasets matching each RNA viral pangenome, binned by the average nucleotide identity taken from the alignment CIGAR string. Score (gradient coloring) function approximates pangenome coverage (see methods) for use to prioritise assembly and manual inspection. An interactive and queryable version of this plot is available at `https://serratus.io/family`.



Extended Figure 5: **Distribution of DNA and Other viral families in the SRA** The total number of datasets matching each DNA or other viral pangenome, binned by the average nucleotide identity and colored by score (see methods). An interactive and queryable version of this plot is available at `https://serratus.io/family`.



Extended Figure 6: **Newly characterised Deltavirus genomes** Genome structure and organisation of the five Deltaviruses (PmacDV SRR7910143; MmonDV SRR2136906; OvirDV SRR4256033; TgutDV SRR5001850; and IchiDV SRR8954566) and one Deltavirus-like (BglaDVL SRR8242383; for which we could not identify a ribozyme sequence) sequence identified in our study. Each circular RNA virus shows characteristic rod-like genome folding and low free-energy (δG), similar to a Hepatitis Delta Virus positive control, and two ribozymes and in contrast to negative controls of the circular DNA virus Porcine Circovirus 2 (PCV2) or shuffled MmonDV sequence.



Extended Figure 7: **Deltavirus ribozymes evolutionary history a** Multiple sequence alignment of the genomic and anti-genomic deltavirus ribozymes based on MUSCLE [72] and refined manually based on secondary structure. The shortening of the J1/2 loop and presence of the LG loop is specific to and conserved within the genomic ribozyme. Consensus secondary structure of the **b** genomic and **c** anti-genomic ribozymes. **d** Maximum-likelihood tree based on concatenated ribozyme sequences supports the topology of the δAg amino-acid tree (Figure 3)