bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Petabase-scale sequence alignment catalyses viral discovery

Robert C. Edgar[1], Jeff Taylor[1], Victor Lin[1], Tomer Altman[2], Pierre Barbera[3], Dmitry Meleshko[4,5], Dan Lohr[1], Gherman Novakovsky[6], Benjamin Buchfink[7], Basem Al-Shayeb[8], Jillian F. Banfield[9], Marcos de la Peña[10], Anton Korobeynikov[4,11], Rayan Chikhi[12], and Artem Babaian[1,‡]

All authors contributed equally to this work
[1]Unaffiliated
[2]Altman Analytics LLC, San Francisco, California, USA
[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]Computational Biology Group, Max Planck Institute for Developmental Biology, Tübingen, Germany
[8]Department of Plant and Microbial Biology, University of California, Berkeley, USA
[9]Department of Earth and Planetary Science, University of California, Berkeley, USA
[10]Instituto de Biología Molecular y Celular de Plantas, Universidad Politécnica de Valencia-CSIC, Valencia, Spain
[11]Department of Statistical Modelling, St. Petersburg State University, St. Petersburg, Russia
[12]Institut Pasteur, CNRS, Paris, France
[‡]Corresponding Author: ababaian@bccrc.ca

March 14, 2021

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Public databases contain a planetary collection of nucleic acid sequences, but their systematic exploration has been inhibited by a lack of efficient methods for searching this corpus, now exceeding multiple petabases and growing exponentially [1, 2]. We developed a cloud computing infrastructure, `Serratus`, to enable ultra-high throughput sequence alignment at the petabase scale. We searched 5.7 million biologically diverse samples (10.2 petabases) for the hallmark gene RNA dependent RNA polymerase, identifying well over 10^5 novel RNA viruses and thereby expanding the number of known species by roughly an order of magnitude. We characterised novel viruses related to coronaviruses and to hepatitis δ virus, respectively and explored their environmental reservoirs. To catalyse a new era of viral discovery, we established a free and comprehensive database of these data and tools. Expanding the known sequence diversity of viruses can reveal the evolutionary origins of emerging pathogens and improve pathogen surveillance for the anticipation and mitigation of future pandemics.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Viral zoonotic disease has had a major impact on human health over the past century despite dramatic advances in medical science, notably the 1918 Spanish influenza, AIDS, SARS, Ebola, and COVID-19. There are an estimated 3×10^5 mammalian virus species [3] from which infectious diseases in humans may arise [4], of which only a fraction is currently known. Global survey and monitoring of virus diversity is required for improved prediction and prevention of future epidemics; this effort is a focus of international consortia and hundreds of research laboratories worldwide [5, 6].

Virus discovery can be aided through re-analysis of the petabases of high-throughput sequencing data available in public databases such as the Sequence Read Archive (SRA) [1, 7–12]. This data spans millions of ecologically diverse biological samples, many of which capture viral transcripts incidental to the goals of the original studies [13]. To catalyse global virus discovery, we developed the `Serratus` cloud computing infrastructure for ultra-high throughput sequence alignment.

We screened 5.7 million libraries comprising 10.2 petabases (10.2×10^{15} bases) of sequencing data, and report 883,502 RNA-dependent RNA polymerase (RdRP)-containing contigs which cluster at 90% amino-acid identity into 140,208 species-like operational taxonomic units (sOTUs), of which 132,260 were novel, i.e. >10% diverged from a previously known RdRP. We assembled 11,120 runs identified in this screen, recovering coronavirus and corona-like virus contigs including sequences from nine novel sOTUs. To demonstrate the broader utility of our approach, we also report 53 novel delta- and delta-like viruses related to the human pathogen hepatitis δ virus, and 252 representatives of a recently characterised family of huge bacteriophages.

Virus discovery is a fundamental step in preparing for the next pandemic. We lay the foundations for years of future research by enabling direct access to hundreds of thousands of virus sequences, captured from the collective efforts of over a decade of high-throughput sequencing studies. A free, interactive repository of `Serratus` data is available at `https://serratus.io`.

Results

Accessing the planetary virome with petabase-scale alignment

`Serratus` is a free, open-source cloud-computing infrastructure optimised to enable petabase-scale sequence alignment against a set of query sequences. 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). We used a widely available commercial computing service to deploy up to 22,250 virtual CPUs simultaneously (see Methods), leveraging SRA data mirrored onto cloud platforms as part of the NIH STRIDES initiative [14].

Our search space spans data deposited over 13 years from every continent and ocean, and all kingdoms of life (Figure 1). We applied `Serratus` in two of many possible configurations. First, to identify libraries containing known or closely related viruses we searched 3,837,755 (ca. May 2020) public RNA-seq, meta-genome, meta-transcriptome, and meta-virome datasets (termed sequencing runs [1]) against a nucleotide pangenome of all known coronavirus sequences plus all vertebrate RefSeq viruses (Extended Figure 3). We then aligned 5,686,715 runs (ca. January 2021) against known viral RdRP amino acid sequences, completing this search within 11 days (Figure 1a and Methods).

Previous approaches for identifying sequences across the entire SRA rely on pre-computed indexes [15–17] requiring exact substring or hash-based matches which limit sensitivity to diverged sequences (Extended Figure 1f). Pre-assembled reads (e.g. NCBI Transcriptome Shotgun Assembly database [2]) enable efficient alignment-based searches, [8], but are currently available only for a small fraction of the SRA. `Serratus` aligns a query of up to hundreds of Mb against unassembled libraries, achieving much greater sensitivity to diverged viruses compared to substring (k-mer) indexes while using far less computational resources than assembly (Figure 1g, and Methods).

A sketch of viral RNA dependent RNA polymerase

Viral RdRP is a hallmark gene of RNA viruses which lack a DNA stage of replication [18]. We identified RdRP by a well-conserved amino acid sub-sequence we call the "palmprint". Palmprints are delineated by three essential motifs which together form the catalytic core in the RdRP structure (Figure 2 and [19]). We constructed species-like operational taxonomic units (sOTUs) by clustering palmprints at a threshold of 90% amino-acid identity, chosen to approximate taxonomic species [19].

3,376,880 (59.38%) sequencing runs contained $\geq$1 reads mapping to the RdRP query (E-value $\leq$1e-4). We assembled aligned reads from each library (and their mate-pairs when available), yielding 4,261,616 contigs. 881,167

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

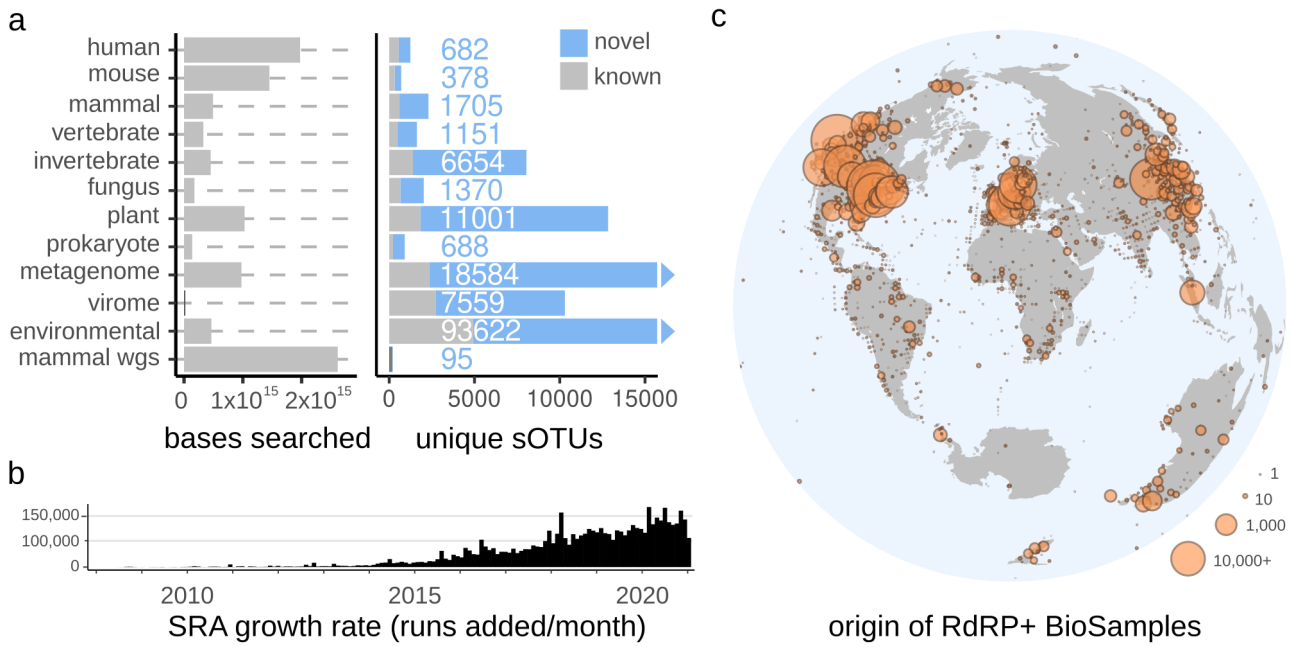


Figure 1: **Searching the planetary virome**
a Total bases searched from the 5,686,715 SRA sequencing runs analysed in the viral RdRP search grouped by sample taxonomy, where available (see Extended Figures 1 and 3, and Extended Table 1). 8,871/15,016 (59%) of known RdRP species-like operational taxonomic units (sOTUs) were observed in the SRA, and 131,957 unique and novel RdRP sOTUs were identified (see Extended Figure 2). sOTUs identified in multiple taxonomic groups are counted in each group separately, numbers shown indicate the number of novel sOTUs in each group. **b** Release dates of the runs included in the analysis reflecting the growth rate of available data. **c** Sample locations for 635,656 RdRP-containing contigs (27.8% of samples lacked geographic metadata). The high density of RdRP seen in North America, Western Europe and Eastern Asia reflects the substantial acquisition bias for samples originating from these regions. Interactive map is available at `https://serratus.io/geo`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

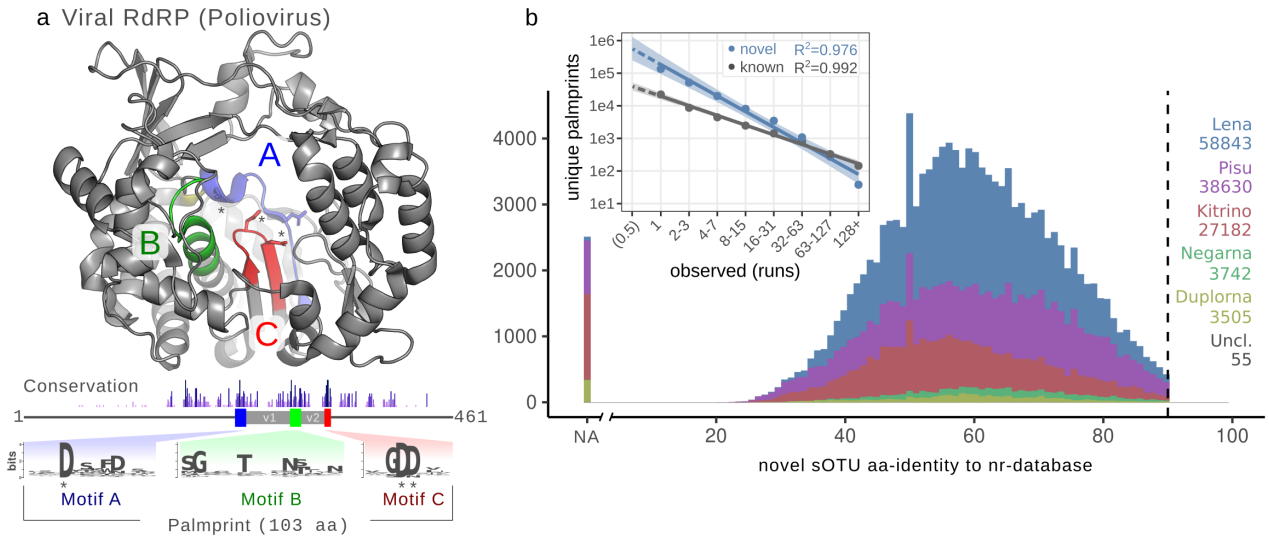


Figure 2: **RNA dependent RNA Polymerase in the Sequence Read Archive**
a The RdRP palmprint is the protein sequence spanning three well-conserved sequence motifs (A, B, and C), including intervening variable regions, exemplified within full-length poliovirus RdRP structure with essential aspartic acid residues(*) (pdb: 1RA6 [20]). Conservation was calculated from RdRP alignment in [21], trimmed to the poliovirus sequence; motif sequence logos are shown below. **b** Per-phylum histogram of amino acid identity of novel species-like operational taxonomic units (sOTUs) aligned to the NCBI non-redundant protein database. **inlay** Preston plot of palmprint abundances indicates that singleton palmprints (i.e., observed in exactly one run) occur within 95% confidence intervals of the value predicted by extrapolation from high-abundance palmprints (linear regression applied to log-transformed data), and this distribution is consistent through time (Extended Figure 2).

(20.7%) contained a high-confidence palmprint identified by `Palmscan` (false discovery rate = 0.001, further discussed in a companion manuscript [19]), representing 260,808 unique palmprints. Applying `Palmscan` to reference sequence databases[2, 21, 22] we obtained 45,824 unique palmprints, which clustered into 15,016 known sOTUs. If a newly acquired palmprint aligned to a known palmprint at ≥90% identity, it was assigned membership to that reference sOTU, otherwise it was designated novel. We clustered novel palmprints at 90% identity, obtaining 131,957 novel sOTUs, representing an increase of known RNA viruses by a factor ~9.8. Clustering novel palmprints at genus-like 75% and family-like 40% thresholds yielded 78,485 and 3,599 novel OTUs, representing increases of 8.0x and 1.9x, respectively (Figure 2b).

We extracted host, geospatial, and temporal metadata for each biological sample when available (Figure 1c), noting that the majority (88%) of novel RdRP sOTUs were observed from metagenomic or environmental runs, where accurate host inference is challenging. Mapping observations of virus marker genes across time and space suggests ecological niches for these viruses, while improved characterisation of sequence diversity can improve PCR primer design for *in situ* virus identification.

We estimate that ~1% of sOTUs are endogenous virus elements (EVEs), i.e. viral RdRPs which have serendipitously reverse-transcribed into a host germline. We did not attempt to systematically distinguish EVEs from virus RdRPs, noting that EVEs with intact catalytic motifs are likely to be recent insertions which can serve as a representative sequence for related exogenous viruses.

Most (60.5%) recovered palmprints were found in exactly one run (singletons), and are observed within the expected frequency range predicted by extrapolating from more abundant sequences (Figure 2b). The abundance distribution of distinct palmprints is consistent with log-log-linear for each year from 2015 to 2020 (Extended Figure 2e), and over time, singletons are confirmed by subsequent runs at an approximately constant rate (Extended Figure 2g). The majority of novel viruses will be singletons until the diversity represented by the search query and the fraction of the planetary virome sampled in the SRA both approach saturation. Extrapolating one year forward, when the SRA is expected to double in size, we project 430,000 (95% CI [330K, 561K]) additional unique palmprints will be identified by running `Serratus` with its current query (Figure 2b).

The total number of virus species is estimated to be 10^8 to 10^{12} [23], thus our expansion captured at most 0.1% of the global virome. However, if exponential data growth continues, we are at the cusp of identifying a significant fraction of Earth's total genetic diversity with tools such as `Serratus`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Expanding the scope of known Coronaviridae

The SARS-CoV-2 pandemic has significantly impacted human society. We further exemplify the potential of `Serratus` for virus discovery with *Coronaviridae* (CoV), including a recently proposed sub-family [24] which contains a CoV-like virus, Microhyla alphaletovirus 1 (MLeV), in the frog *Microhyla fissipes*, and Pacific salmon nidovirus (PsNV) described in the endangered *Oncorhynchus tshawytscha* [25].

First, we identified 52,772 runs containing $\geq$ 10 CoV-aligned reads or $\geq$ 2 CoV k-mers (32-mer, [17]). These runs were *de novo* assembled with a new version of synteny-informed `SPAdes` called `coronaSPAdes` (discussed in a companion manuscript [26]). This yielded 11,120 identifiable CoV contigs which we annotated for a comprehensive assemblage of *Coronaviridae* in the SRA (see Methods for discussion). With this training data we defined a scoring function to predict subsequent success of assembly (Extended Figure 3c).

CoV and neighbouring palmprints comprise 70 sOTUs, 44 of which are described in public databases. 17 CoV sOTUs contained partial RdRP (inclusive of full palmprint) from an amplicon-based virus discovery study not yet publicly deposited [27]. The remaining 9 sOTUs are novel viruses, with protein domains consistent with a CoV or CoV-like genome organisation (Extended Figure 4).

We operationally designate MLeV, PsNV and the nine novel viruses broadly as group *E*, noting that all were found in samples from non-mammalian aquatic vertebrates (Figure 3). Notably, *Ambystoma mexicanum* (axolotl) nidovirus (AmexNV) was assembled in 18 runs, 11 of which yielded common ~19 kb contigs. Easing the criteria of requiring an RdRP match in a contig, 28/44 (63.6%) of the runs from the associated studies were AmexNV positive [28–30]. Consistent assembly breakpoints in AmexNV, PsNV and similar viruses suggests that the viral genomes of this clade of CoV-like viruses are organised in at least two segments, one containing ORF1ab with RdRP, and a shorter segment containing a lamin-associated domain protein, spike and N' accessory genes (Figure 3). An assembly gap with common breakpoints is present in the published PsNV genome [25]. Together these seven monophyletic species likely represent a distinct clade of segmented CoV-like nidoviruses, although molecular validation of this hypothesis is warranted.

In addition to identifying genetic diversity within CoV, we cross-referenced CoV+ library meta-data to identify possible zoonoses and vectors of transmission. Discordant libraries, one in which a CoV is identified and the viral expected host [31] does not match the sequencing library source taxa, were rare, accounting for only 0.92% of cases (Extended Table 1e).

In data from a 2010 virome sequencing study of children with febrile illness [32], we identified sequencing runs from two children, one febrile (SRR057949) and one afebrile (SRR057961) with reads mapping to the β-CoV, 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* [33], but clinical infection may be rare and missed by targeted diagnostic assays. Infectious agents are the leading cause of pyrexia of unknown origin (PUO) in children and immunocompromised adults [34]. This highlights the need for rapid and unbiased metagenomic sequence diagnostics, technically akin to `Serratus`. This would help resolve the etiology of a sub-set of PUO and serve as an early-warning surveillance system for zoonoses, enabling more efficacious public-health intervention.

An important limitation 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 five libraries in which a porcine, avian or bat coronavirus was found in plant samples. The parsimonious explanation is that CoV was present in faeces/fertiliser originating from a mammalian or avian host applied to these plants. However, this exemplifies a merit of unbiased search in identifying transmission vectors and monitoring the geo-temporal boundaries of a virus.

Rapid expansion into the viral unknowns

The global mortality from viral hepatitis exceeds that of HIV/AIDS, tuberculosis or malaria [35]. Hepatitis δ virus (HDV) has a small circular RNA genome (~1.7 knt) which folds into a rod-like shape and encodes three genes: a delta antigen protein, and two self-cleaving ribozymes (drbz) [36].

Prior to 2018, HDV was the sole known member of its genus; 13 drbz-containing members have since been characterised [37–42], and recently a second class of ribozyme (known as hammerhead or hhrbz) characteristic of plant viroids was identified in delta-like viruses we refer to as epsilon viruses [43]. By sequence search for the delta antigen protein and ribozymes, we identified 14 deltaviruses, 39 epsilonviruses and 311 enigmatic sequences with deltavirus-like synteny we term zeta viruses (Figure 4, Extended Figure 5). The evolutionary histories of these mammalian deltaviruses are explored further in a companion paper [41].

The zeta virus circular genomes are highly compressed, ranging from 324-789 nt and predicted to fold into rod-like structures. They contain a hhrbz in each orientation and encode two ORFs, one sense and one anti-sense. Both

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

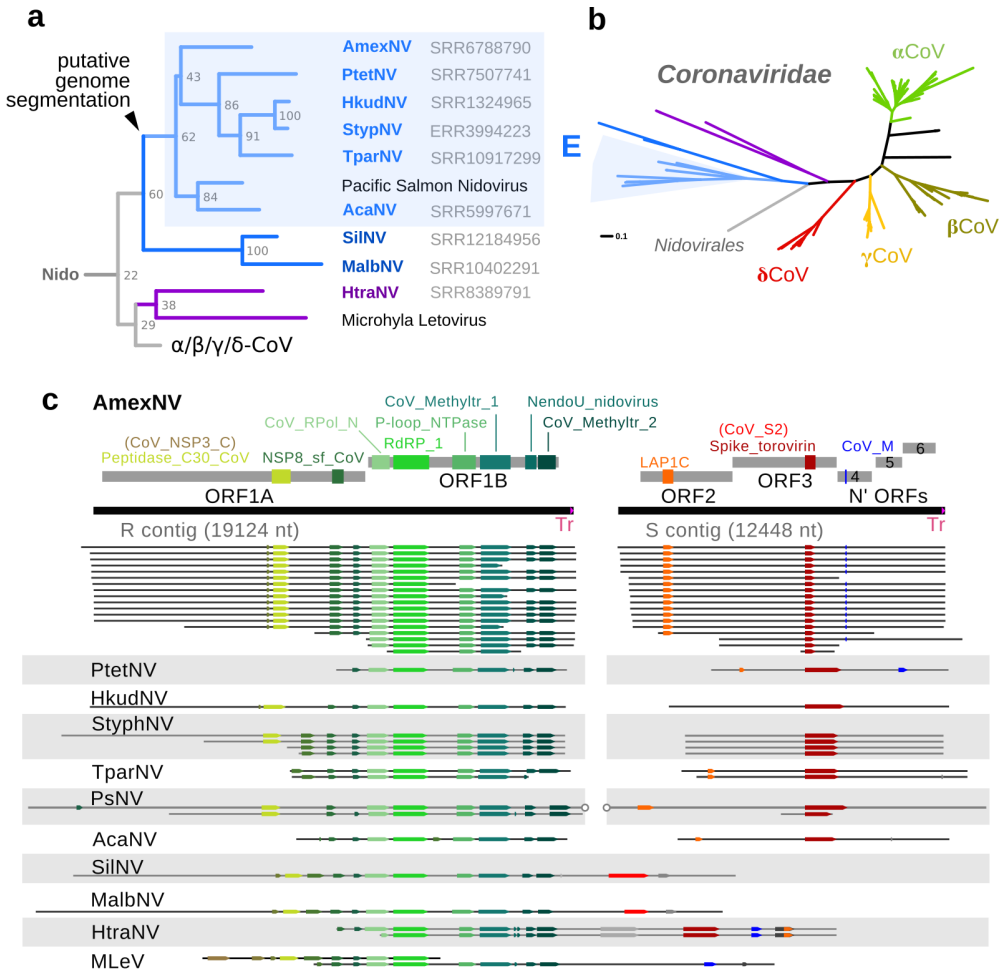


Figure 3: **Expanding Coronaviridae**

a Phylogram for group *E* sequences. Six viruses were similar to PsNV in *Ambystoma mexicanum* (axolotl; AmexNV), *Puntigrus tetrazona* (tiger barb; PtetNV), *Hippocampus kuda* (seahorse; HkudNV), *Syngnathus typhle* (broad-nosed pipefish; StypNV), *Takifugu pardalis* (fugu fish; TparNV), and the *Acanthemblemaria sp.* (blenny; AcaNV). More distant members identified were in *Hypomesus transpacificus* (the endangered delta smelt; HtraNV), *Silurus sp.* (catfish) SilNV, and *Monopterus albus* (asian swamp eel) MalbNV. **b** Unrooted phylogram for *Coronaviridae* annotated with genera (Greek letters) and group *E* CoV-like nidoviruses (see also Extended Figure 4). Maximum likelihood tree generated by clustering the RdRp amino acid sequences at 97% identity to show ~sub-species variability. **c** Genome structure of AmexNV and the contigs recovered from group *E* CoV-like viruses annotated with hidden-Markov model matches. AmexNV contigs contain an identical 129 nt trailing sequence (Tr). All the putatively segmented CoV-like are monophyletic with PsNV. A gap in the PsNV reference sequence [25] is shown with circles, overlapping the common contig ends seen in these viruses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

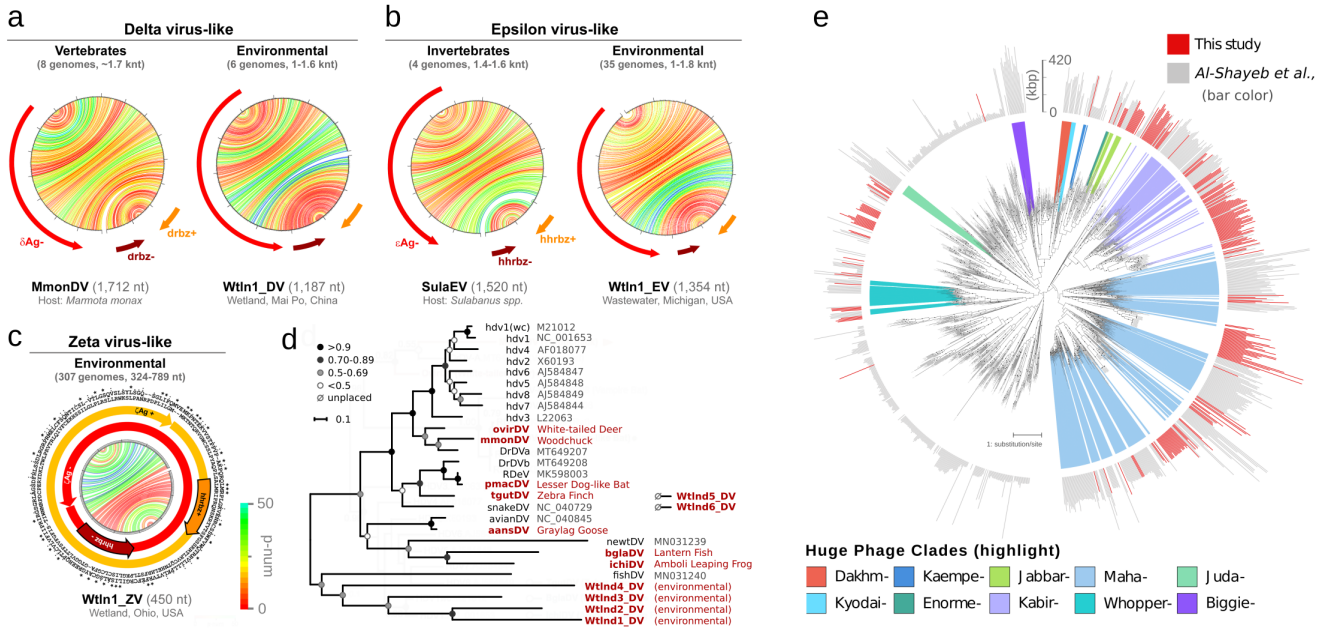


Figure 4: **Expanding deltaviruses and hugephages**

a Genome structure for the *Marmota monax* Delta virus (MmonDV) and a DV-like genome detected in an environmental dataset each containing a negative-sense delta-antigen (δAg) ORF; two delta ribozymes (dvrbz); and characteristic rod-like folding, where each line shows the predicted base-pairing within the RNA genome, coloured by base-pairing confidence score (p-num) [44]. **b** Similar genome structure for the *Sulabanus* spp. Epsilon virus-like (SulaEV) and an EV-like genome from an environmental dataset each containing a negative-sense epsilon-antigen (ϵAg) ORF; two hammerhead ribozymes (hhrbz); and rod-like folding. **c** Example of the compact genome structure of a Zeta virus-like from an environmental dataset containing two predicted zeta-antigen (ζAg+/-, protein alignment is shown in the outer circles) ORFs without stop codons; two hhrbz overlapping with the ORFs; and rod-like folding. Further novel genomes are shown in Extended Figures 5 and 6. **d** Maximum-likelihood phylogenetic tree of DVs derived from a delta-antigen protein alignment with bootstrap values. Two divergent environmental DV could not yet be placed. **e** Tree showing huge phage clade expansion. Black dots indicate branches with bootstrap values >90. Outer ring indicates genome or genome fragment length: gray are sequences from [45] 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 Extended Figure 7.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ORFs generally lack stop codons and encompass the entire genome, potentially producing an endless tandem-repeat of antigen. The atypical coiled-coil domain of the HDV antigen [46] is conserved in the antigens of new delta and epsilon viruses, whereas epsilon and zeta genomes show analogous hhrbzs (Extended Figure 6), supporting that these sequences may share common ancestry. These abundant elements may help to solve a long-standing question about the origins of circular RNA subviral agents in higher eukaryotes (Extended Figure 6), historically regarded as molecular fossils of a prebiotic RNA world [47]

To evaluate the feasibility of applying `Serratus` in the context of microbiome research, we sought to locate bacteriophages related to recently reported huge phages [45], searching for terminase amino acid sequences. Targeted assembly of 287 high-scoring runs returned 252 terminase-containing contigs $\geq$140 kbp. Phylogenetics of these sequences resolved new groups of phages with large genomes (Figure 4e). While most phages were from a single animal genus, we identified closely related phages crossing animal orders, including related phages in a human from Bangladesh (ERR866585) and groups of cats (PRJEB9357) and dogs (PRJEB34360) from England, sampled 5 years apart. Similarly, we recovered two 554 kbp Lak megaphage genomes (among the largest animal microbiome phages reported to date) that are extremely closely related to sequences previously reported from pigs, baboons and humans (Extrended Figure 7), [48]. These two genomes were circularised and manually curated to completion. The large carrying capacity of such phages and broad distribution underlines their potential for extensive lateral gene transfer amongst animal microbiomes and modification of host bacterial function. The newly-recovered sequences substantially expand and augment the inventory of phages with genomes whose length range overlaps with those of bacteria.

Discussion

Since the completion of the the human genome, growth of DNA sequencing databases has outpaced Moore's Law [49]. `Serratus` provides rapid and focused access to genomic sequences captured over more than a decade by the global research community which would otherwise be inaccessible in practice. This work and further extensions of petabase scale genomics [16, 17, 50] are shaping a new era in computational biology, enabling expansive gene discovery, pathogen surveillance, and pangemomic evolutionary analyses.

Optimal translation of such massive datasets into meaningful biomedical advances requires free and open collaboration amongst scientists [51, 52]. The current pandemic underscores the need for prompt, unrestricted and transparent data sharing. With these goals in mind, we deposited 7.3 terabytes of virus alignments and assemblies into an open-access database which can be explored via a graphical web interface at `https://serratus.io` or programatically through the `tantalus R` package and its `PostgreSQL` interface.

The "metagenomics revolution" of virus discovery is accelerating [22, 53]. Innovative fields such as high-throughput viromics [54] can leverage vast collections of virus sequences to inform policies that predict and mitigate emerging pandemics [55, 56]. Combining ecoinformatics with virus, host, and geotemporal metadata offers a proof of concept for a global pathogen surveillance network, arising as a byproduct of centralised and open data sharing.

Human population growth and encroachment on animal habitats is bringing more species into proximity, leading to increased rate of zoonosis [4] and accelerating the Anthropocene mass extinction [57, 58]. `Serratus` enhances our capability to chronicle the full genetic diversity of our planet's diminishing biosphere. The need to invest in collection and curation of biologically diverse samples with emphasis on geographically under-represented regions has never been more pressing. If not for the conservation of endangered species, then to better conserve our own.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. 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).
2. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. eng. *Nucleic Acids Research* **46,** D8–D13. ISSN: 1362-4962 (Jan. 2018).
3. 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).
4. Jones, K. E. *et al.* Global trends in emerging infectious diseases. eng. *Nature* **451,** 990–993. ISSN: 1476-4687 (Feb. 2008).
5. 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).
6. 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).
7. Levi, K., Rynge, M., Abeysinghe, E. & Edwards, R. A. *Searching the Sequence Read Archive using Jetstream and Wrangler* in *Proceedings of the Practice and Experience on Advanced Research Computing* (Association for Computing Machinery, New York, NY, USA, July 2018), 1–7. ISBN: 978-1-4503-6446-1. https://doi.org/10.1145/3219104.3229278 (2021).
8. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. en. *Nature* **556,** 197–202. ISSN: 1476-4687. https://www.nature.com/articles/s41586-018-0012-7 (2020) (Apr. 2018).
9. Wahba, L. *et al.* An Extensive Meta-Metagenomic Search Identifies SARS-CoV-2-Homologous Sequences in Pangolin Lung Viromes. en. *mSphere* **5.** ISSN: 2379-5042. https://msphere.asm.org/content/5/3/e00160-20 (2020) (June 2020).
10. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. eng. *Nucleic Acids Research* **48,** D570–D578. ISSN: 1362-4962 (Jan. 2020).
11. Chen, I.-M. A. *et al.* The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Research* **49,** D751–D763. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkaa939 (2021) (Jan. 2021).
12. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. en. *Cell* **184,** 1098–1109.e9. ISSN: 0092-8674. https://www.sciencedirect.com/science/article/pii/S0092867421000726 (2021) (Feb. 2021).
13. 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).
14. *STRIDES Initiative — Data Science at NIH* https://datascience.nih.gov/strides (2021).
15. Debat, H. J. Expanding the size limit of RNA viruses: Evidence of a novel divergent nidovirus in California sea hare, with a ~35.9 kb virus genome. en. *bioRxiv,* 307678. https://www.biorxiv.org/content/10.1101/307678v1 (2020) (Apr. 2018).
16. Karasikov, M. *et al.* MetaGraph: Indexing and Analysing Nucleotide Archives at Petabase-scale. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.10.01.322164. https://www.biorxiv.org/content/10.1101/2020.10.01.322164v2 (2021) (Nov. 2020).
17. Katz, K. S. *et al.* A fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next generation sequence submissions. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory. https://www.biorxiv.org/content/early/2021/02/17/2021.02.16.431451 (2021).
18. Koonin, E. V. & Dolja, V. V. Virus world as an evolutionary network of viruses and capsidless selfish elements. eng. *Microbiology and molecular biology reviews: MMBR* **78,** 278–303. ISSN: 1098-5557 (June 2014).
19. Babaian, A. & Edgar, R. C. Ribovirus classification by a polymerase barcode sequence. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2021.03.02.433648. https://www.biorxiv.org/content/10.1101/2021.03.02.433648v1 (2021) (Mar. 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

20. Thompson, A. A. & Peersen, O. B. Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. *The EMBO Journal* **23,** 3462–3471. ISSN: 0261-4189. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC516629/ (2021) (Sept. 2004).

21. Wolf, Y. I. *et al.* Origins and Evolution of the Global RNA Virome. en. *mBio* **9.** Publisher: American Society for Microbiology Section: Research Article. ISSN: 2150-7511. https://mbio.asm.org/content/9/6/e02329-18 (2021) (Dec. 2018).

22. Wolf, Y. I. *et al.* Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. en. *Nature Microbiology.* ISSN: 2058-5276. http://www.nature.com/articles/s41564-020-0755-4 (2020) (July 2020).

23. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. en. *Microbiology and Molecular Biology Reviews* **84.** Publisher: American Society for Microbiology Section: Review. ISSN: 1092-2172, 1098-5557. https://mmbr.asm.org/content/84/2/e00061-19 (2021) (May 2020).

24. 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).

25. 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).

26. Meleshko, D., Hajirasouliha, I. & Korobeynikov, A. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.07.28.224584. https://www.biorxiv.org/content/10.1101/2020.07.28.224584v2 (2021) (Feb. 2021).

27. Tao, Y. *et al.* Broad-Range Virus Detection and Discovery Using Microfluidic PCR Coupled with High-throughput Sequencing. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2020.06.10.145052. https://www.biorxiv.org/content/10.1101/2020.06.10.145052v1 (2021) (June 2020).

28. 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).

29. 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).

30. 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).

31. Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research* **49,** D723–D733. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkaa983 (2021) (Jan. 2021).

32. 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).

33. 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).

34. 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).

35. 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).

36. Taylor, J. M. Infection by Hepatitis Delta Virus. en. *Viruses* **12,** 648. https://www.mdpi.com/1999-4915/12/6/648 (2020) (June 2020).

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

38. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

39. 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).

40. 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).

41. Bergner, L. M. *et al.* Diversification of mammalian deltaviruses by host shifting. en. *Proceedings of the National Academy of Sciences* **118.** Publisher: National Academy of Sciences Section: Biological Sciences. ISSN: 0027-8424, 1091-6490. https://www.pnas.org/content/118/3/e2019907118 (2021) (Jan. 2021).

42. Iwamoto, M. *et al.* Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution. *Virus Evolution* **7.** ISSN: 2057-1577. https://doi.org/10.1093/ve/veab003 (2021) (Jan. 2021).

43. De la Peña, M., Ceprian, R., Casey, J. L. & Cervera, A. Hepatitis delta virus-like circular RNAs from diverse metazoans encode conserved hammerhead ribozymes. en. *Virus Evolution* **7,** veab016. https://academic.oup.com/ve/article/7/1/veab016/6143048 (2021) (Feb. 2021).

44. 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).

45. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. eng. *Nature* **578,** 425–431. ISSN: 1476-4687 (2020).

46. Zuccola, H. J., Rozzelle, J. E., Lemon, S. M., Erickson, B. W. & Hogle, J. M. Structural basis of the oligomerization of hepatitis delta antigen. en. *Structure* **6,** 821–830. https://www.sciencedirect.com/science/article/pii/S0969212698000847 (1998) (July 1998).

47. Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R. & Elena, S. F. Viroids: survivors from the RNA world? en. *Annual Review of Microbiology* **68,** 395–414. https://www.annualreviews.org/doi/10.1146/annurev-micro-091313-103416 (2014) (June 2014).

48. 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).

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

50. Bradley, P., Den Bakker, H. C., Rocha, E. P. C., McVean, G. & Iqbal, Z. Ultra-fast search of all deposited bacterial and viral genomic data. *Nature biotechnology* **37,** 152–159. ISSN: 1087-0156. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6420049/ (2021) (Feb. 2019).

51. 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).

52. Baker, D. *et al.* No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics. en. *PLOS Pathogens* **16.** Publisher: Public Library of Science, e1008643. ISSN: 1553-7374. https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1008643 (2021) (Aug. 2020).

53. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2,** 63–77. ISSN: 1879-6257. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102772/ (2021) (Feb. 2012).

54. 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).

55. 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).

56. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

57. 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).

58. 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).

59. 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).

60. 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).

61. Schatz, M. C. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* **25,** 1363–1369. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btp236 (2021) (June 2009).

62. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btq461 (2010).

63. 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).

64. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. en. *PLOS ONE* **11.** Publisher: Public Library of Science, e0163962. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163962 (2021) (Oct. 2016).

65. 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 (2015).

66. 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).

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

68. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. metaviralSPAdes: assembly of viruses from metagenomic data. en. *Bioinformatics.* https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa490/5837667 (2020).

69. 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).

70. 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).

71. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34,** 867–868 (2018).

72. Edgar, R. C. UCHIME2: improved chimera prediction for amplicon sequencing. *BioRxiv.* Publisher: Cold Spring Harbor Laboratory, 074252 (2016).

73. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9.** Publisher: Springer, 1–13 (2021).

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

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

76. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. eng. *Trends in genetics: TIG* **16,** 276–277. ISSN: 0168-9525 (June 2000).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

78. Schäffer, A. A. *et al.* VADR: validation and annotation of virus sequence submissions to GenBank. en. *bioRxiv,* 852657. https://www.biorxiv.org/content/10.1101/852657v2 (2020) (Mar. 2020).

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

80. 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).

81. 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).

82. 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).

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

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

85. 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).

86. 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).

87. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods (San Diego, Calif.)* **102,** 3–11. https://www.ncbi.nlm.nih.gov/pubmed/27012178 (2016).

88. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28,** 2223–2230 (2012).

89. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49,** D480–D489. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkaa1100 (2021) (Jan. 2021).

90. Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D. A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* **14,** 112. ISSN: 1471-2105. https://doi.org/10.1186/1471-2105-14-112 (2020) (Mar. 2013).

91. Li, W., Fu, L., Niu, B., Wu, S. & Wooley, J. Ultrafast clustering algorithms for metagenomic sequence analysis. en. *Briefings in Bioinformatics* **13,** 656–668. ISSN: 1467-5463, 1477-4054. https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs035 (2017) (Nov. 2012).

92. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. en. *Bioinformatics* **34.** Publisher: Oxford Academic, 2490–2492. ISSN: 1367-4803. https://academic.oup.com/bioinformatics/article/34/14/2490/4916099 (2020) (July 2018).

93. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973. ISSN: 1367-4803. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2712344/ (2020) (Aug. 2009).

94. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. eng. *Molecular Biology and Evolution* **32,** 268–274. ISSN: 1537-1719 (Jan. 2015).

95. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology.* https://doi.org/10.1038/s41587-020-00774-7 (Dec. 2020).

96. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. eng. *Database: The Journal of Biological Databases and Curation* **2020.** ISSN: 1758-0463 (Jan. 2020).

97. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkh340 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

98. 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. *Bioinformatics.* ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btz305 (2019).

99. Barbera, P. *et al.* EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic Biology* **68** (ed Posada, D.) 365–369. https://doi.org/10.1093/sysbio/syy054 (Mar. 2019).

100. Czech, L., Barbera, P. & Stamatakis, A. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* **36,** 3263–3265. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btaa070 (2021) (May 2020).

101. Morel, B., Kozlov, A. M. & Stamatakis, A. ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* **35** (ed Schwartz, R.) 1771–1773. https://doi.org/10.1093/bioinformatics/bty839 (Oct. 2018).

102. Darriba, D. *et al.* ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution* **37** (ed Crandall, K.) 291–294. https://doi.org/10.1093/molbev/msz189 (Aug. 2019).

103. Felsenstein, J. Confidence limits on ohylogenies: 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).

104. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556,** 452–456. https://doi.org/10.1038/s41586-018-0043-0 (Apr. 2018).

105. Crisci, M. A. *et al.* Wide distribution of alternatively coded Lak megaphages in animal microbiomes. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 2021.01.08.425732. https://www.biorxiv.org/content/10.1101/2021.01.08.425732v1 (2021) (Jan. 2021).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Materials and Methods

1.1 Serratus alignment architecture

`Serratus` (`https://github.com/ababaian/serratus`) is an open-source cloud-infrastructure designed for ultra-high throughput sequence alignment against a query sequence or pangenome (Extended Figure 1). `Serratus` compute costs are dependent on search parameters (expanded discussion available: `https://github.com/ababaian/serratus/wiki/pangenome_design`). The nucleotide vertebrate viral pangenome search (`bowtie2`, database size: 79.8 Mb) reached processing rates of 1.29 million SRA runs in 24-hours at a cost of $0.0062 US dollars per dataset (Extended Figure 1). The translated-nucleotide RdRP search (`DIAMOND`, database size: 7.1 Mb) reached processing rates exceeding 0.5 million SRA runs in 12-hours at a cost of $0.0042 per dataset. All 5,686,715 runs analysed in the RdRP search were completed within 11 days for a total cost of $23,980 or $2,350 per petabase. For a detailed breakdown of `Serratus` project costs and recommendations for managing cloud-computing costs, see Serratus wiki: `https://github.com/ababaian/serratus/wiki/budget`.

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 Amazon Web Services (AWS) Simple Storage Service (S3) mirror of the Sequence Read Archive (SRA), decompresses to FASTQ with `fastq-dump`, and splits the data into chunks of 1 million reads or read-pairs (`fq-blocks`) into a temporary S3 cache bucket. To mitigate excessive disk usage caused by a few large datasets, a total 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`) [59] for nucleotide search, or `DIAMOND` (v2.0.6 - development version, `--mmap-target-index --target-indexed --masking 0 --mid-sensitive -s 1 -c1 -p1 -k1 -b 0.75`) [60] 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.

AWS Elastic Compute Cloud (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 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 or a `Hadoop` distributed filesystem [61].

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 system problems in realtime and resume data processing. For example, shutting down the `align` modules to hotfix a genome indexing problem without having to re-run the `dl` modules, or if an alignment instance is killed by a Spot termination, only that block needs to be reprocessed instead of the entire sequencing run.

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 queries 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 worst-case response rate in the hundreds of milliseconds, keeping cluster throughput high. Each task typically lasts several minutes depending on the pangenome.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. Up to 300,000 SRA jobs can be processed in the work queue per batch process.

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 self shuts-down. Finally a "job cleaner" component checks the active jobs against currently running instances. If an instance has disappeared 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 obtained by 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,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

1.1.6 Sensitivity to novel viruses as a function of identity

We aimed to assess the sensitivity of our pipeline as a function of sequence identity by asking what fraction of novel viruses is detected at increasingly low identities compared to the reference sequences used for the search. Several variables other than identity affect sensitivity, including read length, whether reads are mate-paired, sequencing error rate, coverage bias, and presence of other similar viruses which may cause some variants to be unreported in the contigs. Coverage bias can render a virus with high average read depth undetectable, in particular if the query is RdRP-only and the RdRP gene has low coverage or is absent from the reads. Successful detection might be defined in different ways, depending on the goals of the search; e.g. a single local alignment of a reference to a read (maximising sensitivity, but not always useful in practice); a micro-assembled palmprint; a full assembly contig that contains a complete palmprint or otherwise classifiable fragment of a maker gene; or an assembly of a complete genome. We assessed alignment sensitivity of `bowtie2 --very-sensitive-local` and Serratus-optimised `DIAMOND` as a function of identity by simulating typical examples in representative scenario: unpaired reads of length 100 with a base call error rate of 1%. We manually selected test-reference pairs of RefSeq complete Ribovirus genomes at RdRP aa identities 100%, 95% ... 20%, generating simulated length-100 reads at uniformly-distributed random locations in the test genome with a mean coverage of 1000x. For `bowtie2`, the complete reference genome was used as a reference; for `DIAMOND` the reference was the translated amino acid sequence of the RdRP gene (400aa), which was identified by aligning to the Wolf18 dataset. These choices model the coronavirus pan-genome used as a `bowtie2` query and the rdrp1 protein reference used as a `DIAMOND` query, respectively. Sensitivity was assessed as the fraction of reads aligned to the reference. With `bowtie2`, the number of unmapped reads reflects a combination of lack of alignment sensitivity and divergence in gene content as some regions of the genome may lack homology to the reference. With `DIAMOND`, the number of unmapped reads reflects a combination of lack of alignment sensitivity and the fraction of the genome which is not RdRP, which varies by genome length 1g. They show that the fraction of aligned reads by `bowtie2` drops to around 2% to 4% at 90% RdRP aa identity, and maps no reads for most of the lower identity test-reference pairs. `DIAMOND` maps around 5% to 10% of reads down to 50% RdRP aa identity, then less than 1% at lower identities; around 30% to 35% is the lower limit of practical detection.

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` [62] and masked by `Dustmasker` [63] 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). `SeqKit` (v.0.15) was used

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for working with fasta files [64].

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 viral RNA-dependent RNA polymerase search panproteome

For the translated-nucleotide search of viral RNA-dependent RNA polymerase (RdRP; hereinafter viral RdRP is implied) we combined sequences from several sources. 1) The 'wolf18' collection is a curated snapshot (ca. 2018) of RdRP from GenBank ([21] accessed: `ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/rnavir18/RNAvirome.S2.afa` 2) The 'wolf20' collection is RdRPs from assembled from marine metagenomes ([22] accessed: `ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/yangshan/gb_rdrp.afa`) 3) All viral GenBank protein sequences were aligned with `diamond --ultra-sensitive` against the combined 'wolf18' and 'wolf20' sequences (E-value <1e-6). These produced local alignments which contained truncated RdRP, so each RdRP-containing GenBank sequence was then re-aligned to the 'wolf18' and 'wolf20' collection to "trim" them to 'wolf' RdRP boundaries. 4) The above algorithm was also applied to all viral GenBank nucleotide records to capture additional RdRP not annotated as such by GenBank . A region of HCV capsid protein shares similarity to HCV RdRP, sequences annotated as HCV-capsid were therefore removed. Eight novel coronavirus RdRP sequences identified in a pilot experiment were added manually. The combined RdRP sequences from the above collections were clustered (uclust) at 90% amino acid identity and the resulting representative sequences (centroids, N = 14 653) used as the rdrp1 search query.

In addition, we added Deltavirus antigen proteins from NC_001653, M21012, X60193, L22063, AF018077, AJ584848, AJ584847, AJ584844, AJ584849, MT649207, MT649208, MT649206, NC_040845, NC_040729, MN031240, MN031239, MK962760, MK962759, and eight additional homologs we identified in a pilot experiment.

1.2.3 SRA search space and queries

To run `Serratus`, a target list of SRA run accessions is required. We defined eleven (not-mutually exclusive) queries as our search space which were named human, mouse, mammal, vertebrate, invertebrate, eukaryotes, prokaryotes/others, bat (including genomic sequences), virome, metagenome and mammalian genome (Extended Table 1 c). Our search was restricted to Illumina sequencing technologies and to RNA-seq, meta-genomic, and meta-transcriptome library types for these organisms (except for mammalian genome query which was genome or exome). Prior to each `Serratus` deployment, target lists were depleted of accessions already analysed. Reprocessing of a failed accession was attempted at least twice. In total, we aligned 3,837,755/4,059,695 (94.5%) of the runs in our nucleotide-pangenome search (ca. May 2020) and 5,686,715/5,780,800 (98.37%) of the runs in our translated-nucleotide RdRP search (ca. January 2021).

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, the .bam nucleotide alignment files from 3.84 million SRA runs are stored in `s3://lovelywater/bam/X.bam`; the protein .summary files are in `s3://lovelywater/psummary/X.psummary`, where `X` is a SRA run accession. These FAIR and structured releases enable downstream and third-party programmatic access to the data.

Summary files for every searched SRA dataset are parsed into a publicly accessible AWS Relational Database (RDS) instance which can be queried remotely via any `PostgreSQL` client. 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, one can query for 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:

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

```
SELECT genbank_id, run_id, tax_id, n_reads
FROM nsequence
JOIN srarun ON (nsequence.run_id = srarun.run)
WHERE n_reads >= 20
```

For users unfamiliar with SQL we developed `Tantalus` (`https://github.com/serratus-bio/tantalus`, an `R` programming-language package which directly interfaces the `Serratus` PostgreSQL database to retrieve summary information as data-frames. `Tantalus` also offers functions to explore and visualise the data.

Finally, the `Serratus` data can be explored via a graphical web interface by accession, virus, or viral family at `https://serratus.io/explorer`. Under the hood, we developed a REST API to query the database from the website. The website uses React+D3.js to serve graphical reports 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`.

1.3.1 Geocoding BioSamples

To generate the map in figure 1c, we parsed and extracted geographic information from all 16 million BioSample XML submissions. Geographic information is either in the form of coordinates (latitude/longitude) or free-form text (e.g. "France", "Great Lakes"). For each BioSample, coordinate extraction was attempted using regular expressions. If that failed, text extraction was attempted using a manually curated list of keywords that capture BioSample attribute names likely to contain geographic information. If that failed, then we were unable to extract geographic information for that BioSample. Geocoding the text to coordinates was done using Amazon Location Service on a reduced set of distinct filtered text values (52,028 distinct values from 2,760,241 BioSamples with potential geographic text). BioSamples with geocoded coordinates were combined with BioSamples with submitted coordinate information to form a set of 5,325,523 geospatial BioSamples. This is then cross-referenced with our subset of SRA accessions with an RdRP match to generate the figure.

All intermediate and resulting data from this step is stored on the SQL database described in 1.3. Development work is public at `https://github.com/serratus-bio/biosample-sql`.

1.4 Viral alignment, assembly and annotation

Upon identificationn of CoV reads in a run from alignment, we assembled 52,772 runs containing at ≥ 10 reads which aligned to our CoV pan-genome or ≥ 2 reads with CoV-positive k-mers. ([17]). 11,120 of the resulting assemblies contained identifiable CoV contigs, of which only 4,179 (37.58%) contained full-length CoV RdRP (Extended Table 1). The discrepancy between alignment-positive, assembly-positive and RdRP-positive libraries arises due random sampling of viral reads and assembly fragmentation. In this respect, alignment or k-mer based methods are more sensitive than assembly in detecting for the presence of low-abundance viruses (genome coverage <1) with high identity to a reference sequence. Scoring libraries for genome-coverage and depth is a good predictor of ultimate assembly success (Extended Figure 3) thus, it can be used to efficiently prioritise computationally expensive assembly in the future.

1.4.1 `DIAMOND` optimisation and output

To optimise `DIAMOND` for small (< 10 Mb) databases such as the RdRP search database, we built a probabilistic hash set which stores 8-bit hash values for the database seeds, using SIMD instructions for fast probing. This index is loaded as a memory mapped file to be shared among processes and allows us to filter the query reads for seeds contained in the database, thus omitting the full construction of the query seed table. We also eliminated the overhead of building seed distribution histograms that is normally required to allocate memory and construct the query table in a single pass over the data using a deque-like data structure. In addition, query reads were not masked for simple repeats, as the search database is already masked. These features are available starting from `DIAMOND` v2.0.8 with the command line flags `--target-indexed --masking 0`. In a benchmark of 4 sets of 1 million read from a bat metagenome (ERR2756788), optimisation reduced `DIAMOND` runtime against RdRP-search from 197.96s (s.d=0.18s), to 21.29s (s.d=0.23s) per million reads, a speed-up of a factor of 9.3. This effectively reduced the computational cost of translated-nucleotide search for `Serratus` from $0.03, to $0.0042 per library.

`DIAMOND` output files (we label `.pro`) were specified with the command `-f 6 qseqid qstart qend qlen qstrand sseqid sstart send slen pident evalue cigar qseq_translated full_qseq full_qseq_mate`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1.4.2 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 [65]. Technical bias is confounded by biological complexity such as intra-sample sequence variation due to transcript isoforms 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 [26]. In brief, `rnaviralSPAdes` and the more specialized variant, `coronaSPAdes`, combines algorithms and methods from several previous approaches based on `metaSPAdes` [66], `rnaSPAdes` [67] and `metaviralSPAdes` [68] 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 [67] and subspecies-bases variation [68].

To deal with possible misassemblies and high-covered sequencing artefacts, 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 [69], `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 [70], 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, the HMMs from this set 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.

1.4.3 Micro-assembly of RdRP-aligned reads

Reads aligned by `DIAMOND` in the translated-nucleotide RdRP search are stored in the `.pro` alignment file. All sets of mapped reads (3,379,127 runs) were extracted, and each non-empty set was assembled with `rnaviralSPAdes` [26] using default parameters. This process is referred to as *micro-assembly* since a collection of `DIAMOND` hits is orders of magnitude smaller than the original SRA accession (40±534 KB compressed size, ranging from a single read up to 53 MB). Then `bowtie2` [59] (default parameters) was used to align the `DIAMOND` read hits of an accession back to the micro-assembled contigs of that accession. `Palmscan --hicon` was run on micro-assembled contigs, resulting in high-confidence palmprints for 337,344 contigs. Finally `mosdepth` [71] was used to calculate a coverage pileup for each palmprint hit region within micro-assembled contigs.

1.4.4 Classification of assembled RdRP sequences

Our methods for RdRP classification are described and validated in a companion paper [19]. Briefly, we defined a barcode sequence, the polymerase palmprint (PP), as a ~100 aa segment of the RdRP palm sub-domain delineated by well-conserved catalytic motifs. We implemented an algorithm, `Palmscan`, to identify palmprint sequences and discriminate RdRPs from reverse transcriptases. The combined set of RdRP palmprints from public databases and our assemblies were classified by clustering into operational taxonomic units (OTUs) at 90%, 75% and 40% identity, giving species-like, genus-like and family-like clusters (sOTUs, gOTUs and fOTUs), respectively. Tentative taxonomy of novel OTUs was assigned by aligning to palmprints of named viruses and taking a consensus of the top hits above the identity threshold for each rank.

1.4.5 Quality control of assembled RdRP sequences

Our goal was to identity novel viral RdRP sequences and novel sOTUs in SRA libraries. From this perspective, we considered the following to be erroneous to varying degrees: sequences which are (a) not polymerases, (b) not viral, (c) with differences due to experimental artefacts, or (d) with sufficient differences to cause a spurious inference of a novel sOTU. We categorised potential sources of such errors and implemented quality control procedures to identify and mitigate them, as follows.

Point errors are single-letter substitution and indel errors which may be caused by PCR or sequencing *per se*. Random point errors are not reproduced in multiple non-PCR duplicate reads and are unlikely to assemble because such errors almost always induce identifiable structures in the assembly graph (tips and bubbles) which are pruned during graph simplification. In rare cases, a contig may contain a read with random point errors. Such contigs

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

will have low coverage ~1, and we therefore recorded coverage as a QC metric and assessed whether low-coverage assemblies were anomalous compared to high-coverage assemblies by measures such as the frequencies with which they are reproduced in multiple libraries compared to exactly one library, finding no noticeable difference when coverage is low.

Chimeras of polymerases from different species could arise from PCR amplification or assembly. We used the UCHIME2 algorithm [72] to screen assembled palmprint sequences, finding no high-scoring putative chimeras. Mosaic sequences formed by joining a polymerase to unrelated sequence would either have an intact palmprint, in which case the mosaic would be irrelevant to our analysis, or would be rejected by `Palmscan` due to the lack of delimiting motifs.

Reverse transcriptases (RTs) are homologous to RdRP. Retroviral insertions into host genomes induce ubiquitous sequence similarity between host genomes and viral RdRP. `Palmscan` was designed to discriminate RdRP from sequences of RT origin. Testing on a large decoy set of non-RdRP sequences with recognisable sequence similarity showed that the `Palmscan` false discovery rate for RdRP identification is 0.001. We estimated the probability of false positives matches in unrelated sequence by generating sufficient random nucleotide and amino acid sequences to show that the expected number of false positive palmprint identifications is zero in a dataset of comparable size to our assemblies. We also regard the low observed frequency of palmprints in DNA WGS data (in 2.6 Pbp or 25.8% of reads, accounted for 100 known palmprints and 95 novel palmprints or 0.13% of the total identified) as a *de facto* confirmation of the low probability false positives in unrelated sequence.

Endogenous viral elements (EVEs, i.e. insertions of viral sequence into host genomes which are potentially degraded and non-functional) cannot be distinguished from viral genomes on the basis of the palmprint sequence alone. To assess the frequency of EVEs in our data, we re-assembled 890 randomly-chosen libraries yielding one or more palmprints using all reads, extracted the 23 530 resulting contigs with a positive palmprint hit by `Palmscan`, and classified them using Virsorter2 [73]. Of these contigs, 11,914 were classified as viral, confirming the `Palmscan` identification; 49 as Viridiplantae (green plants); 46 as Metazoa; 25 as Fungi and the remainder were unclassified. Thus, $120/12034 = 1\%$ of the classified contigs were predicted as non-viral, suggesting that the frequency of EVEs in the reported palmprints is ~1%.

1.4.6 Annotation of CoV assemblies

Accurate annotation of CoV genomes is challenging due to ribosomal frameshifts and polyproteins which are cleaved into maturation proteins [74], 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` [75][1] which leverages both reference-based and *ab initio* annotation approaches.

In brief, `DARTH` consists of the following phases: standardise 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 standardisation, `DARTH` generates the six-frame translation of the contigs using the `transeq` [76] and uses `HMMER3` [77] to search the translations for Pfam domain models specific to CoV [70]. DARTH compares the Pfam accessions from the 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` [78], which provides a set of genome models for all CoV RefSeq genomes [79]. `VADR` provides annotations of gene coordinates, polyprotein cleavage sites, and functional annotation of all proteins. `DARTH` supplements the `VADR` annotation by using `Infernal` [80] to scan the contigs against the SARS-CoV-2 Rfam release [81] 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` [82], 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` [83] representing self-alignment of the trimmed reads to the canonicalized assembly using `bowtie2` [59], 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` [84] or `IGV` [85]. As the final step `DARTH` generates a single Generic Feature Format (GFF) 3.0 file [86] containing combined set of annotation information described above, ready for

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

use in a genome browser, or for submitting the annotation and sequence to a genome repository.

1.4.7 Phage assembly

Each metagenomic dataset was individually *de novo*-assembled using `MEGAHIT` [87], and filtered to remove contigs smaller than 1 kbp in size. ORFs were then predicted on all contigs using `Prodigal` v2.6.3 [88] with the following parameters: `-m -p meta`. Predicted ORFs were initially annotated using `USEARCH` [62] to search all predicted ORFs against `UniProt` [89], `UniRef90` and `KEGG` [90]. Sequencing coverage of each contig was calculated by mapping raw reads back to assemblies using `Bowtie2` [59]. Terminase sequences from Al-Shayeb *et al.* [45] were clustered at 90% amino acid identity to reduce redundancy using `CD-HIT` [91], and HMM models were built with `hmmbuild` (from the `HMMER3` suite [77]) from the resulting set. Terminases in the assemblies from `Serratus` were identified using `hmmsearch`, retaining representatives from contigs greater than 140 kbp in size. Some examples of prophage and large phages that did not co-cluster with the sequences from Al-Shayeb *et al.*, were also recovered because they were also present in a sample that contained the expected large phages. The terminases were aligned using `MAFFT` v.7.407 [92] and filtered by TrimAL [93] to remove columns comprised of more than 50% gaps, or 90% gaps, or using the automatic `gappyout` setting to retain the most conserved residues. Maximum likelihood trees were built from the resulting alignments using `IQTREE` v.1.6.6 [94].

1.4.8 Deploying the assembly and annotation workflow

The `Serratus` search for known or closely related viruses identified 37,131 libraries (14,304 by nucleotide and 23,898 by amino acid) as potentially positive for CoV (score $\geq$20 and $\geq$10 reads). To supplement this search we also employed a recently developed index of the SRA called STAT [17] 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 [95] 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 [95], 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.

The average cost of assembly was between \$0.30-\$0.40 per library, varying depending on library-type (RNA-seq versus metagenomic). This places an estimate of 46-95 fold higher cost for assembly alone compared to a cost of \$0.0042 or \$0.0065 for an alignment based search.

1.5 Taxonomic and phylogenetic analyses

1.5.1 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

[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`

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

genomes; report best-estimate predictions balancing over-classification and under-classification (too many and too few ranks, respectively); and assign an NCBI Taxonomy Database [96] 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 annotated as CoV were downloaded from `UniProt` [89], 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` [62] to expand the repertoire of domain sequences. The reference sequences were clustered using `UCLUST` [62] at 97% sequence identity to reduce redundancy.

For a given query genome, open reading frames (ORFs) are extracted using the `EMBOSS getorf` software [76]. 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` [97] 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.

1.5.2 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 ([62], `UCLUST` algorithm, `v11.0.667`), resulting in 546 centroid sequences. Subsequently we performed multiple sequence alignment on the clustered sequences using `MUSCLE` ([97], `v3.8.31`). We then performed maximum likelihood tree inference using `RAxML-NG` ([98], `PROTGTR+FO+G4`, `v0.9.0`), resulting in our reference tree.

To apply `SerraPlace` to a given genome, we first use HMMER ([77], `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` ([99], `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` ([100], `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.

1.5.3 Phylogenetic inference

We performed phylogenetic inferences using a custom snakemake pipeline (available at `https://github.com/lczech/nidhoggr`), using `ParGenes` [101], `v1.1.2`. `ParGenes` is a treesearch orchestrator, built on top of `ModelTest-NG` [102] and `RAxML-NG`, enabling higher levels of parallelisation for a given tree search.

To infer the maximum likelihood phylogenetic trees , 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 for any given treesearch. The pipeline also automatically produces versions of the best maximum likelihood tree annotated with Felsenstein's Bootstrap ([103]) support values, and Transfer Bootstrap Expectation ([104]) values.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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, S. Jackman, I. Birol, K. Wellman, O. Fornes, C. Xu, M. Huss, K. Ha, M. Krzywinski, E. Nawrocki, R. McLaughlin, C. Morgan-Lang, C. Blumberg, and the J. Brister lab. A. Rodrigues, S. McMillan, V. Wu, C. Kennett, K. Chao, and N. Pereyaslavsky for AWS support. We would also like to thank the J. Joy lab, G. Mordecai, J. Taylor, S. Roux, N. Kyrpides, T. Reddy, 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, Inception and PRAIRIE grants (PIA/ANR-16-CONV-0005, ANR-18-CE45-0020, ANR-19-P3IA-0001), MdlP by Ministerio de Economía y Competitividad of Spain and FEDER grant (BFU2017-87370-P), 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 support and computing resources were kindly provided by the University of British Columbia Community Health and Wellbeing Cloud Innovation Centre, powered by AWS.

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 virus pangenomes and RdRP query. 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. BB optimised `DIAMOND` algorithm for RdRP search. 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, MdlP, AK, and RC analysed the coronavirus, RdRP and deltavirus data. BAS and JB designed the phage pan-proteome, assembled phage genomes, and conducted phage phylogenetic analyses. All authors contributed to data interpretation and writing the manuscript.

Correspondence

Correspondence should be addressed to AB.

Reporting Summary

Does not apply.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

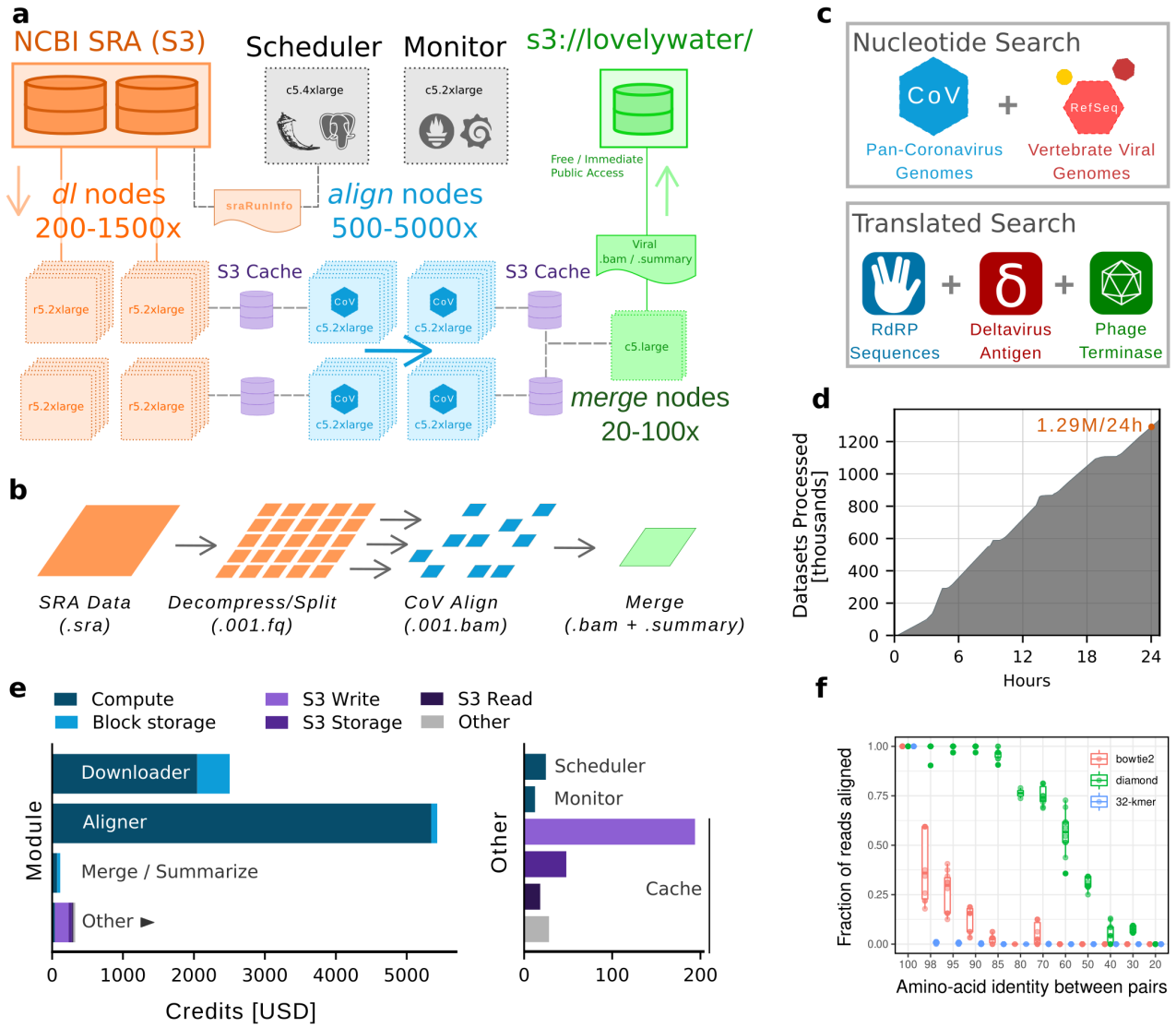
Extended Table 1: **SRA run queries and CoV assembly table** Queries and accessions from this study. **a** SRA queries to retrieve collections of datasets. **b** Run accessions, assembly statistics and select meta-data for the 11,120 runs for which *Coronaviridae*, or *Coronaviridae*-like sequences were assembled. **c** Assignment of assembled runs to operational taxonomic units (OTUs) based on 97% identity of the RNA dependent RNA polymerase (RdRp) domain. **d** Assignment of GenBank records to RdRp OTUs. **e** Assignment of expected viral host for GenBank records based on Sequence Read Archive and JGI GOLD metadata [2, 31]. **f** Taxonomic source for RdRp containing assemblies.}

Extended Tables

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

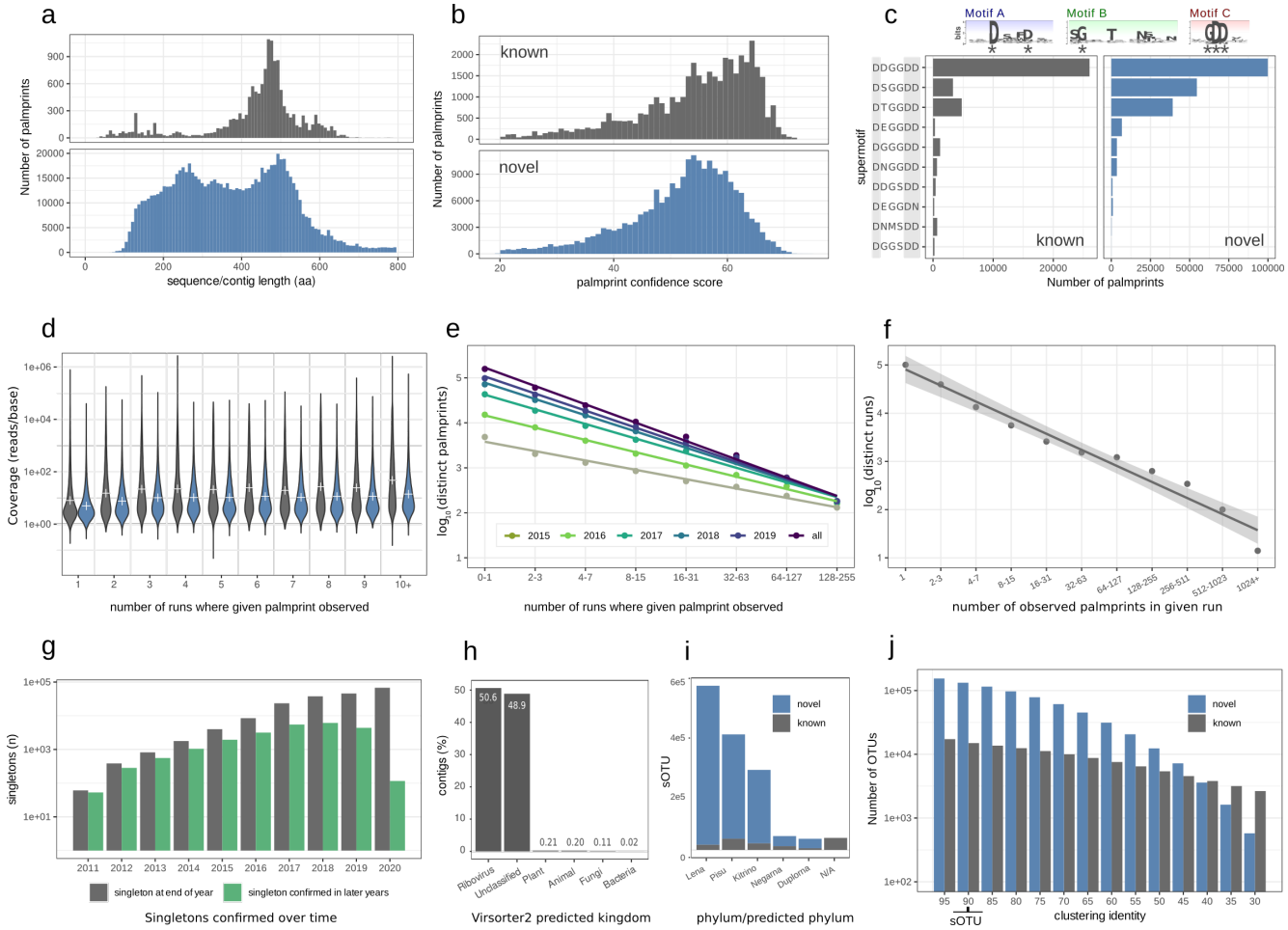
2 Extended Figures

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



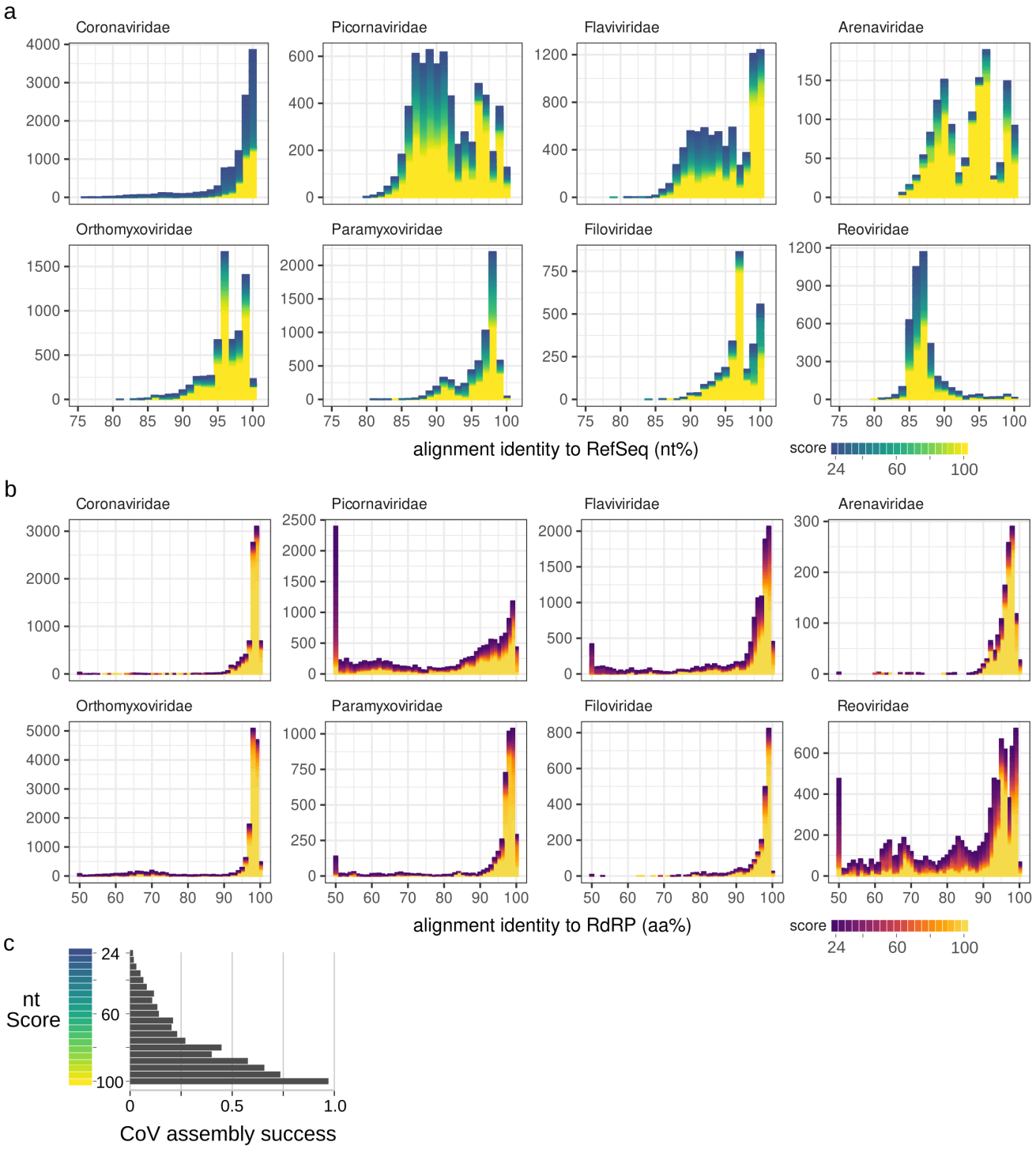
Extended Figure 1: **Overview of the `Serratus` infrastructure.**
a Schematic and data workflow (**b**) as described in the methods for sequence alignment. **c** The `align` module accepts either a nucleotide or protein sequence query. **d** A nucleotide alignment completion rate for *Serratus* shows stable and linear performance to complete 1.29 million SRA accessions in a 24-hour period and the **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 for nucleotide search or \$0.0042 for translated-nucleotide search. **f** Biological cross-validation to measure alignment sensitivity for `bowtie2` (nucleotide search), `diamond` (translated nucleotide search) or `32 kmer` for exact search. Briefly, two RdRP sequences sharing the nominal amino acid identities form a "pair". 100 bp reads were simulated from the coding sequence of one pair and mapped onto the second pair, with the fraction of reads mapped reported. A value of 50% indicates that half the simulated reads at the given RdRP percent identity are mappable and thus detectable (see Methods).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Extended Figure 2: **Analysis of palmprint contigs recovered by `Serratus`**
a Length distribution of amino acid sequences in the rdrp1 query (upper histogram) and micro-assembled contigs (lower histogram, length=nucleotides/3). **b** Distribution of `Palmscan` confidence scores. **c** Observations of the 10 most frequent "super-motifs" (six well-conserved residues marked with asterisk) reported by `Palmscan`. **d** Distribution of coverage vs. abundance (number of runs where a given palmprint is observed), showing that palmprints have similar underlying coverage distributions at all abundances. **e** Preston plot of distinct palmprints vs. abundance exhibiting similar, approximately log-log-linear relationships to totals at end-of-year 2015 to 2019 and final totals at approx. end of 2020 (all). **f** Preston plot of number of distinct palmprints observed in a given run vs. number of runs. **g** Numbers of singletons and second observations (confirmations) at the end of each year showing that the growth in singletons is matched by a comparable growth in confirmations. **h** Kingdom predicted by `Virsorter2`for RdRP+ contigs (by `Palmscan`) obtained by full assembly of 880 randomly-chosen RdRP+ runs. **i** Number of palmprints in each phylum assigned by taxonomy (known) or predicted (novel). **j** Number of OTUs as a function of clustering identity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

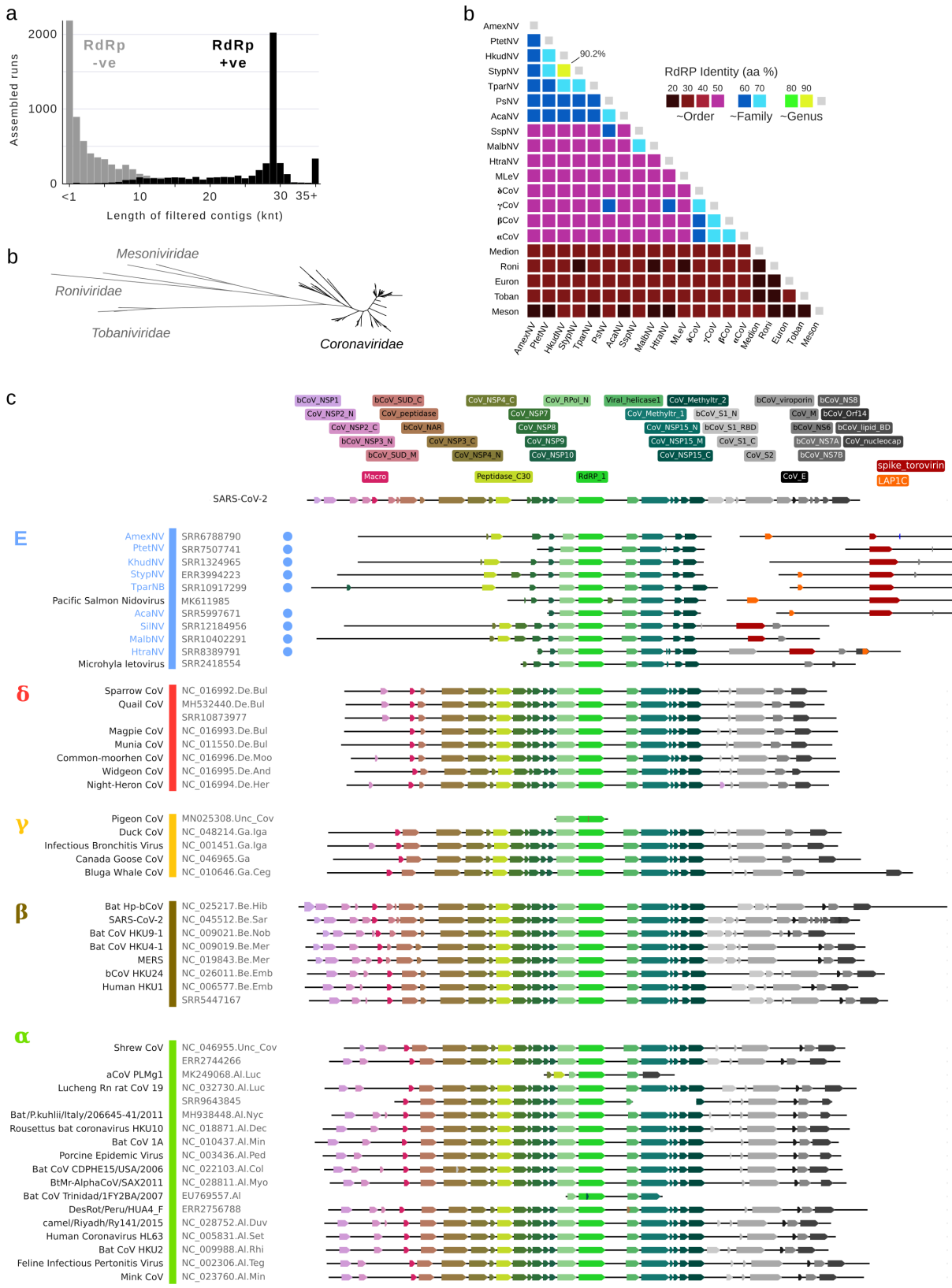


Extended Figure 3: **Distribution of select RNA virus families**

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Extended Figure 3: **Distribution of select RNA virus families (cont'd)** Histogram of datasets matching select RNA viral family by (**a**) nucleotide search against RefSeq pangenomes or (**b**) translated-nucleotide search against RdRP query, binned by the average nucleotide or amino acid identity, respectively. Score (gradient colouring) function approximates pangenome/gene coverage (see methods) used for manual inspection and to prioritise assembly. Coarsely, nucleotide identity between 80-100% approximate amino acid identity of 90-100%. Interactive and queryable versions of these plots for extended virus families are available at `https://serratus.io/explorer`. **c** Relationship between the pangenome score function and the subsequent assembly success (defined by the presence of an RdRP+ contig) measured from 52,772 libraries with reads aligning to Coronaviridae.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

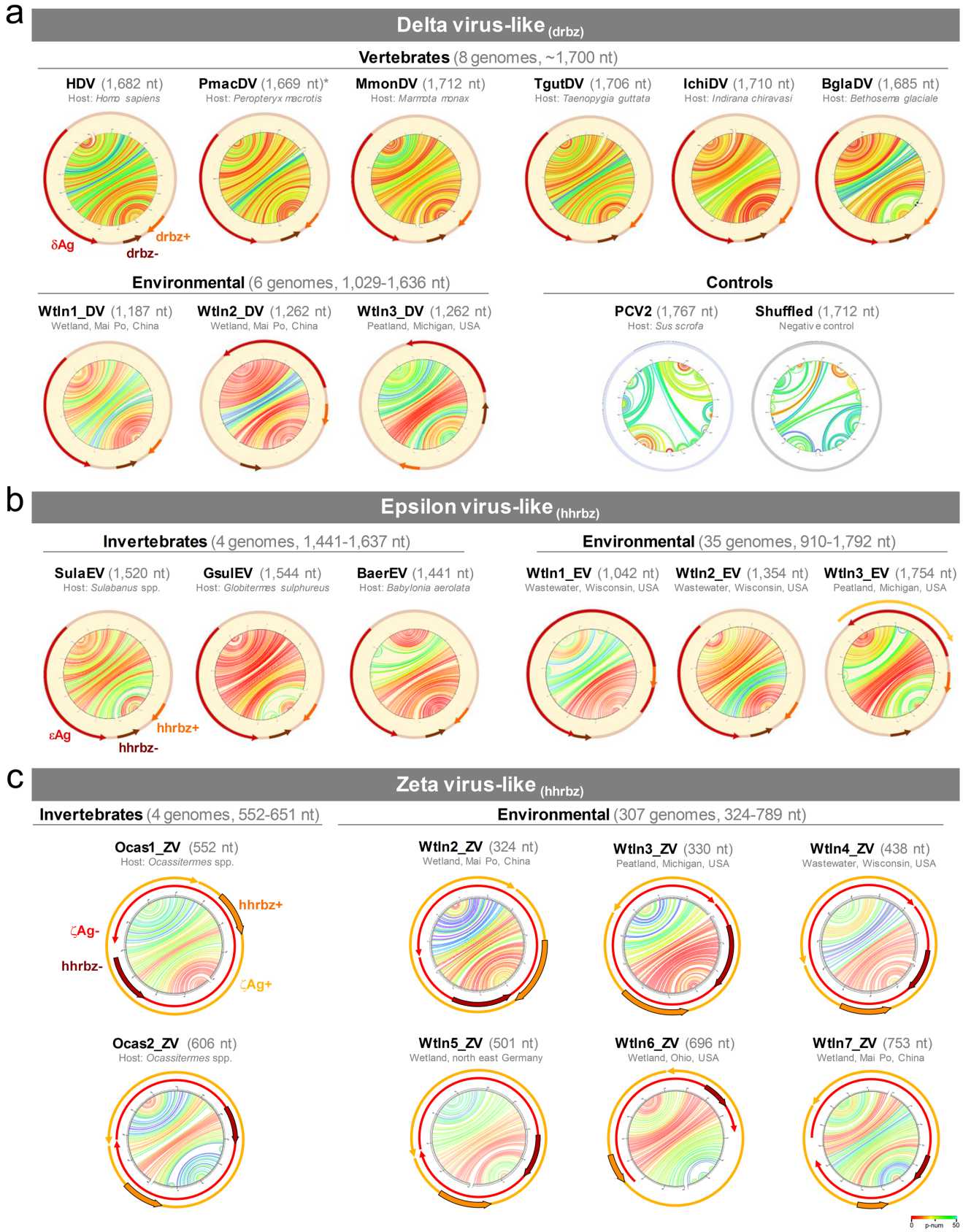


Extended Figure 4: **Genome organisation of *Coronaviridae* and neighbours**

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Extended Figure 4: **Genome organisation of *Coronaviridae* and neighbours (cont'd) a** Length distribution for 11,120 assembled contigs classified as CoV-positive, showing a peak around the typical CoV genome length, 4,179 (37.58%) of contigs also contained a match for RdRP. **b** Phylogram shown in Figure 3 showing the *Mesoniviridae*, *Tobaniviridae*, and *Roniviridae* outgroups. **c** Triangular matrix showing median RdRp sequence identities between selected *Nidovirales* and group *E* sequences. **d** Hidden Markov Model (HMM) protein domain matches from the RdRp in exemplar sequences (contigs or GenBank sequences), grouped by genus. Novel sOTUs identified in this analysis indicated by a coloured circle.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

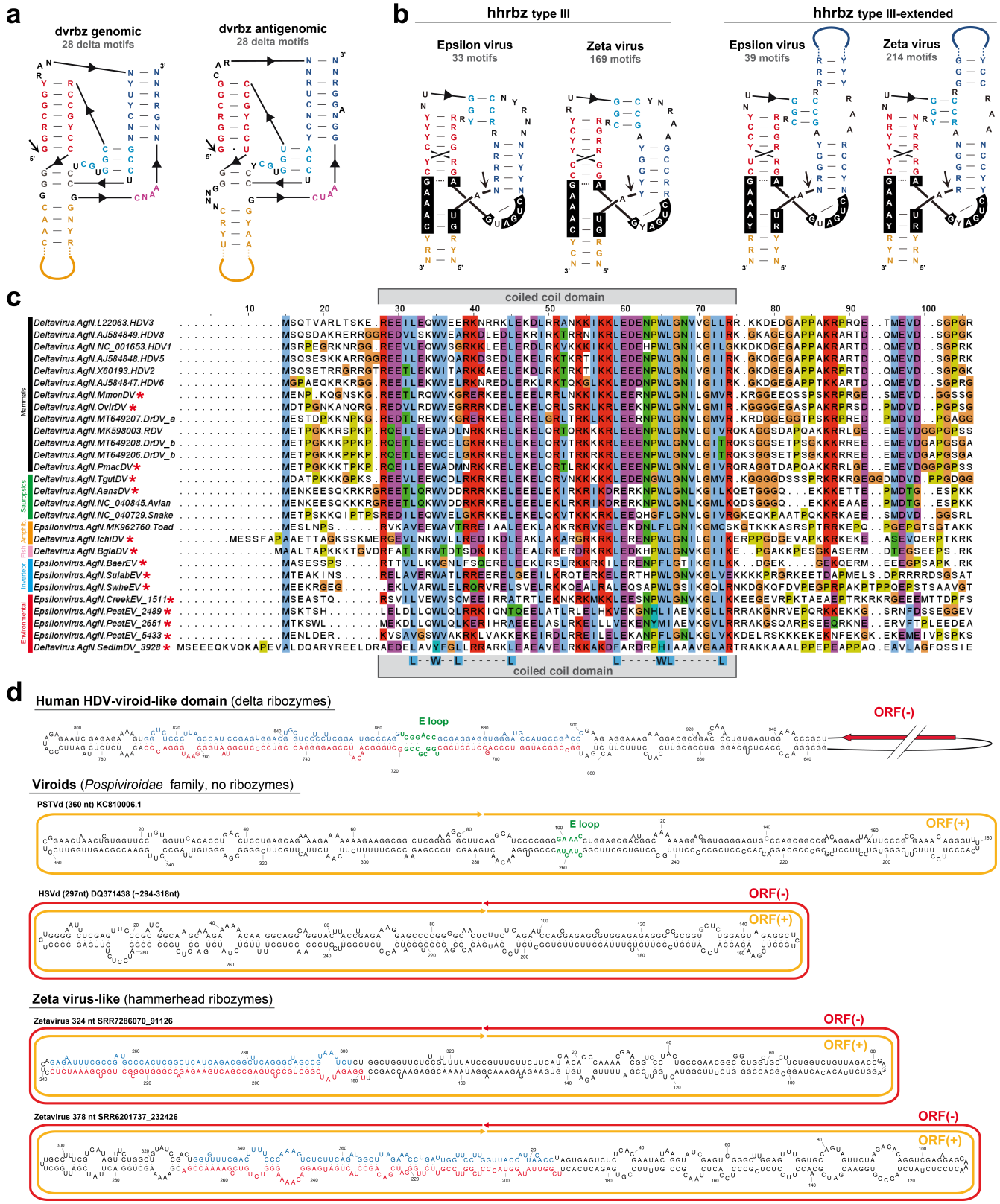


Extended Figure 5: **Newly characterised Deltavirus and Deltavirus-like genomes**

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Extended Figure 5: **Newly characterised Deltavirus and Deltavirus-like genomes (cont'd)** Structure and organisation of selected examples from the 14 delta virus-, 39 epsilon virus- and 311 zeta virus-like genomes identified in our study. **a** Similar to human deltavirus (HDV), delta virus-like genomes from vertebrates (PmacDV SRR7910143; MmonDV SRR2136906; TgutDV SRR5001850; IchiDV SRR8954566 and BglaDV SRR8242383) and environmental datasets (SRR7286070 and SRR6943136) share similar predicted stable rod-like folding, a predicted ORF coding for the delta antigen (δAg) and a delta ribozyme (dvrbz) on each polarity. Folding of the circular DNA virus Porcine Circovirus 2 (PCV2) and a shuffled MmonDV sequence are shown as negative controls. **b** Epsilon virus-like genomes detected in invertebrates (SulaEV SRR8739608; GsulEV SRR7170939 and BaerEV SRR12300397) and environmental datasets (SRR8840728 and SRR6943136) show similar structure and organization to deltaviruses, with one or two predicted ORFs (epsilon antigen or ϵAg) and two hammerhead ribozymes (hhrbz) in equivalent genomic regions. **c** Zeta virus-like genomes detected in invertebrate (Ocassitermes sp. ZVs SRR8924823) and environmental datasets (SRR7286070, SRR6943136, SRR8840728, SRR6201737, SRR5864109 and SRR12063536) are smaller than delta and epsilon agents. Up to 90% of the zeta genomes have sizes multiple of 3 and predicted ORFs without stop codons, capable to encode endless tandem-repeated zeta antigens in both polarities (ζAg+ and ζAg– shown as yellow and red arrows, respectively). Both genomic zeta polarities keep hhrbzs (shown as arrows overlapping the ORFs) similar to the epsilon ribozymes (Extended Fig 6). Larger zeta virus-like genomes (>651 nt) were less abundant (7% of all zeta genomes) and frequently show stop codons, or their sizes are not multiple of 3.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

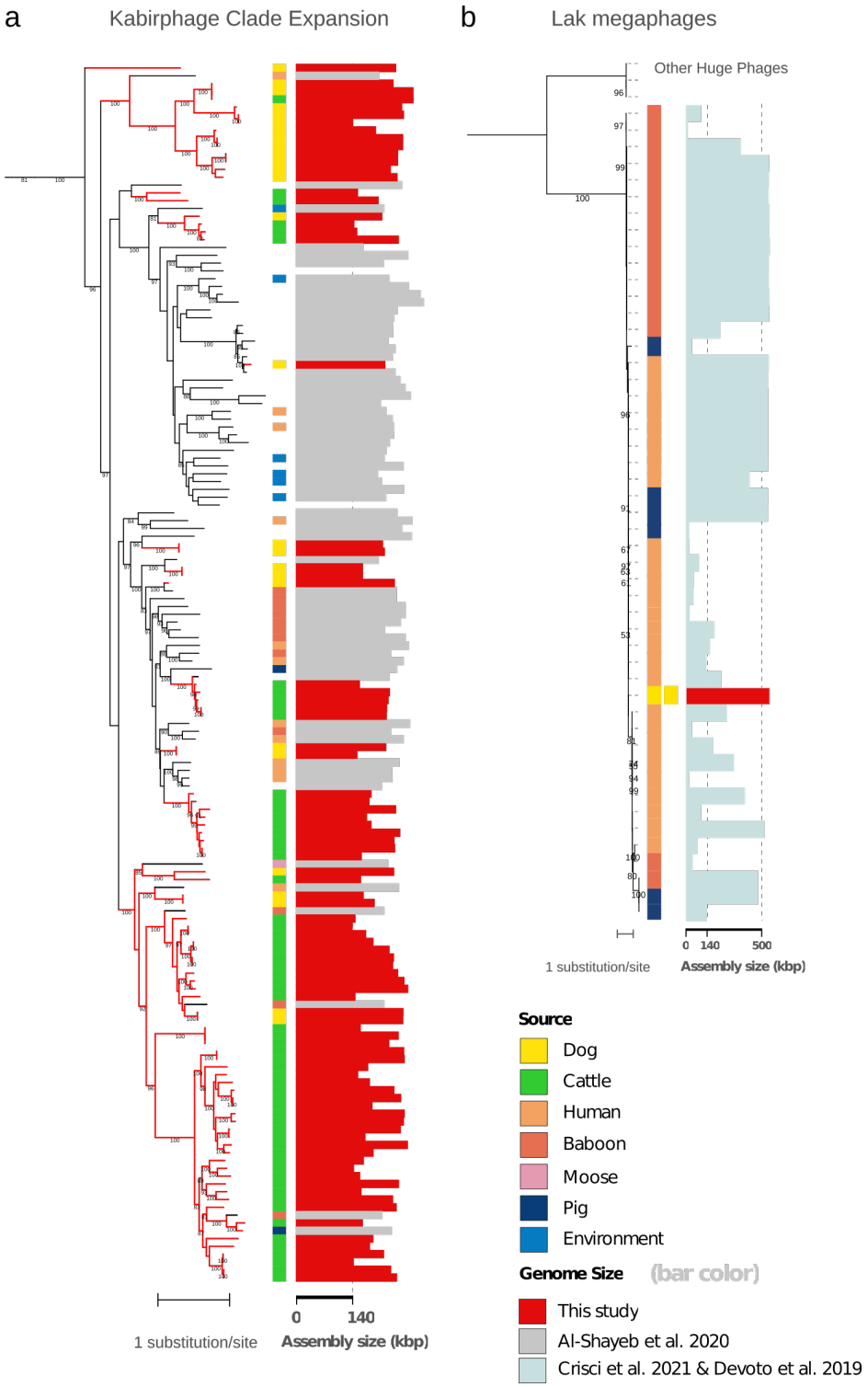


Extended Figure 6: **Evolutionary history of deltavirus-like agents**

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Extended Figure 6: **Evolutionary history of deltavirus-like agents (cont'd) a** Consensus structures (weighted nucleotide conservation threshold of 90%) of deltavirus ribozymes, including the 14 genomes described in this work. **b** Consensus structures of the two hammerhead ribozyme families (type III and extended-type III [43]) detected in epsilon and zeta agents. Most positions of epsilon and zeta motifs are sequence conserved for each ribozyme family. **c** MSA of the predicted antigen (N-term domain) from delta and epsilon agents (genomes detected in this study are indicated with a red asterisk). The antiparalel coiled-coil of the HDV is delimited with a grey box, and conserved residues involved in hydrophobic interactions are shown at the bottom [46], supporting a highly divergent connection between delta and epsilon genomes. **d** Human HDV deltavirus is known to contain a viroid-like domain related to the *Pospiviroidae* family of plant viroids. Both families of agents conserve a tertiary structure reminiscent of the E-loop 5S rRNA (nucleotides in green) and are replicated by the RNA Pol II of the host [47]. *Pospiviroids*, despite of lacking hhrbzs, share with zeta genomes a small rod structure, and in some cases, the presence of predicted endless tandem-repeat ORFs, most notably in both polarities of numerous variants of the Hop Stunt Viroid (HSVd). Whereas viroids have been historically regarded as non protein coding RNAs, our reported observations warrant further investigation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.07.241729; this version posted March 14, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Extended Figure 7: **Huge phage and Lak phage detail** Expanded view of maximum likelihood terminase large subunit protein phylogenetic trees for (**a**) 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 previously reported genomes from [45]. **b** Publicly available Lak phage genomes [105] with sequences of two newly reconstructed complete Lak megaphage genomes. These are the first reported Lak megaphages from dogs (assembled from fecal sample metagenome reads from Allaway et al. 2020). The genomes have identical terminase sequences (at the nucleotide level) although the dogs were in different housing areas and were sampled at different times (D Allaway, personal communication).