Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity

Rayan Chikhi[1], Brice Raffestin[1], Anton Korobeynikov[3], Robert Edgar[4], and Artem Babaian[5, 6]

[1]Institut Pasteur, Université Paris Cité, Paris, France
[3]Independent, Vancouver, Canada
[4]Independent
[5]Department of Molecular Genetics, University of Toronto, Toronto, Canada
[6]The Donnelly Centre for Cellular + Biomolecular Research, University of Toronto, Toronto, Canada

July 30, 2024

Abstract

The NCBI Sequence Read Archive (SRA) is the largest public repository of DNA sequencing data, containing the most comprehensive snapshot of Earth's genetic diversity to date. As its size exceeds 50.0 petabases across >27 million sequencing datasets, the entirety of these data cannot be searched for genetic sequences of interest in a reasonable time. To drastically increase the accessibility of this data we perform genome assembly over each SRA dataset using massively parallel cloud computing. The resulting Logan assemblage is the largest dataset of assembled sequencing data to date, and we believe will enable a new-era of accessible petabase-scale computational biology inquiry. We provide free and unrestricted access to the Logan assemblage and disseminate these datasets to foster early adoption. To illustrate the usefulness of Logan we align a diverse set of sequence queries across all of the SRA, completing queries in as little as 11 hours.

1 Introduction

The INSDC Sequence Read Archive (SRA [12]) is the largest repository of DNA and RNA sequences, as of December, 2023 the SRA contains 50.0 petabases (Pbp, 5.00×10^{16}) of raw sequencing data, or approximately 20 petabytes in the binary-compressed format. The SRA is 4 orders of magnitude larger than GenBank (2.6 terabases, Tbp) [25] which is the database upon which the BLAST web server is based [24]. Streaming the entire SRA through a 10 Gbits Internet connection at full speed would require 6 months of download time. For nearly all institutions the SRA is too large to be either hosted locally or streamed quickly, therefore its utility is effectively limited to accessing small subsets of experiments. SRA-scale analyses of the raw data are inaccessible outside of a handful of data centers worldwide where the entire dataset has been mirrored.

A wealth of genomic diversity is present in the SRA, ranging from public human sequencing projects and cell lines, to eukaryotic samples from all across the Tree of Life, environmental samples, and viral amplicons. Several studies have performed one-off investigations over a subset of the

SRA focusing on particular organisms. Serratus [7] mined all SRA RNA-sequencing experiments prior to 2020 and uncovered 9x more RNA viral species than previously known. NCBI STAT [13] provides an approximate taxonomic overview of each SRA accession. BIGSI [2] surveyed antibiotic resistance in bacterial and viral accessions. Other studies mined SRA metagenomes to e.g. uncover environmental diversity [19], human gut dark matter [18], and soil diversity [15].

Methods for searching subsets of the SRA efficiently are beginning to emerge. In Serratus [7] an SRA-wide RNA search was performed using off-the-shelf read alignment tools [3]. Each search took several days over tens of thousands of CPU cores using a complex cloud computing architecture, which makes the system difficult to deploy broadly. A similar approach was developed by NCBI and run on-premises to screen a fraction of the SRA (300,000 metagenome samples) for the presence of *Candida auris*, a multidrug-resistant fungus [16]. In MetaGraph [11] a public interface is available for sequence search over subsets of the SRA: all fungi, microbes, metazoa, gut metagenomes, using k-mer indexing and therefore restricted to high-homology matches. NCBI Pebblescout [27] offers a similar service to MetaGraph over all metagenomes and human RNA-seq experiments prior to 2021. The NCBI SRA BLAST service performs BLAST alignment on a single SRA accession at a time [23].

Other groups have realized the value of providing assemblies for SRA accessions. ENA hosts >50,000 assemblies of metagenomes (less than 10% of all metagenomic runs), 88% of which were generated by MGnify [22]. IMG/M hosts >30,000 metagenome assemblies and >100,000 bacterial isolates [4]. AllTheBacteria is a recent collection of 1,932,812 assembled bacterial isolates [8]. NCBI WGS is a superset of GenBank containing 25 terabases of assembled sequences of over 2 million biological samples [25]. All these collections are dwarfed by the 27.3 million accessions in the SRA.

As the SRA is too big to be wholly explored in its current form, a radical transformation of its content is needed. Here we perform for the first time a SRA-wide genome assembly across the entire SRA, using massive cloud resources. We processed raw reads into assembled unitigs and contigs thereby increasing sequence lengths and greatly decreasing redundancy and data volume: the 50 petabases of SRA raw data become 384 terabytes of compressed contigs, a 130x reduction in size. We present the v1 release of Logan as a cloud-hosted, publicly available dataset of assemblies for nearly all SRA accessions. This unlocks SRA-wide sequence search at affordable costs and in turn, planetary-scale biological investigations.

2 Results

2.1 Performance-optimized and cloud-accelerated genome assembly of all SRA accessions

We designed a cloud architecture to perform SRA-wide genome assembly efficiently and in parallel. Figure 1 (top of middle panel) describes the workflow. Briefly, a Docker container processes each SRA accession independently: 1) raw reads of an accession are downloaded from a cloud mirror of the SRA, 2) a conservative assembly (unitigs) of the accession is made from the reads, 3) a consensus assembly (contigs) is made from the unitigs, and 4) both the unitigs and contigs are compressed and uploaded to a public repository.

We executed this system at SRA-scale using cloud resources. Containers are executed in parallel over tens of thousands of cloud computers through a container orchestration system, and a set of dashboard were deployed to monitor the execution. Figure A3 shows some key statistics of the

execution. We optimized the system to use as many CPU cores in parallel as possible, as opposed to running it at smaller scale over a longer period of time, to take advantage of lower computation costs and higher availability of cloud instances during night time.

Using this system we performed genome assembly over the entire SRA and report results for each accession in two forms: unitigs and contigs. Contigs are the usual product of genome assembly tools, they represent consensus sequences of the genomic material present in the reads of an accession. Unitigs on the other hand are shorter sequences constituent of contigs, as well as additional discarded variants (e.g. SNPs, small indels) and more generally any sequence seen more than twice in the reads. Formally, unitigs are the maximal non-branching paths of the de Bruijn graph, that we constructed for a single k value of 31. The rationale for this dual set of results is to preserve nearly all information from the reads of the original accession in unitigs, and perform more aggressive contiguity and assembly size optimizations in contigs.

In total 27.3 million accessions were assembled into unitigs, representing 96.00% of the SRA in size as of December 2023. Some accessions resulted in too many unitigs to fit the assembly graph in memory, hence were not further assembled into contigs at this time. In total 26.8 million accessions were assembled into contigs, representing 88% of the SRA in size. The total cloud computation time was around 30 million CPU hours.

Assembly statistics for the unitigs and contigs are reported in Figure 2. In a nutshell, contigs provide draft-level genome assemblies, and unitigs are overall shorter but still provide a reduction of redundancy over raw reads. Contigs for Whole-Genome Sequencing/Amplification (WGS/WGA) accessions are generally longer than those of RNA-Seq accessions or other sequencing types, as expected by the longer sequenced molecules. Note that non-circular contigs shorter than 150 bp that are isolated nodes in the assembly graph were discarded by the assembler as they were more likely to be artifacts than actual biological material.

Dataset	**Type**	**Reads**	**# accessions**	**Index/Seqs**	**Compress.**
Sourmash Branchwater	Index	1.4 Pbp	682,688	7 TB	191x
MetaGraph-SRA	Index	3.3 Pbp	1,891,328	7 TB	473x
Pebblescout-SRA	Index	3.7 Pbp	4,141,058	171 TB	21x
NCBI SRA (Dec 2023)	Raw data	50.0 Pbp	27,764,169	19 PB	3x
Logan, Unitigs (v1)	Assembly	48.3 Pbp	27,311,279	2 PB	22x
Logan, Contigs (v1)	Assembly	44.1 Pbp	26,785,902	385 TB	114x

Table 1: **Size of existing indexed data vs Logan.** The MetaGraph-SRA and Pebblescout-SRA rows refer to all SRA accessions indexed by MetaGraph and Pebblescout respectively. The NCBI SRA row refers to all public accessions from the SRA as of December 10th 2023. The Reads column refers to the number of bases in SRA reads for the considered dataset. The Index/Seqs column indicates the sum of all sub-indices sizes (for Branchwater, MetaGraph and Pebblescout) or the size of compressed sequences for all accessions (for SRA and Logan). The "Compress." column gives the compression ratio between the size of reads (Reads column) as if each base was stored using 8 bits, and the Index/Seqs column.

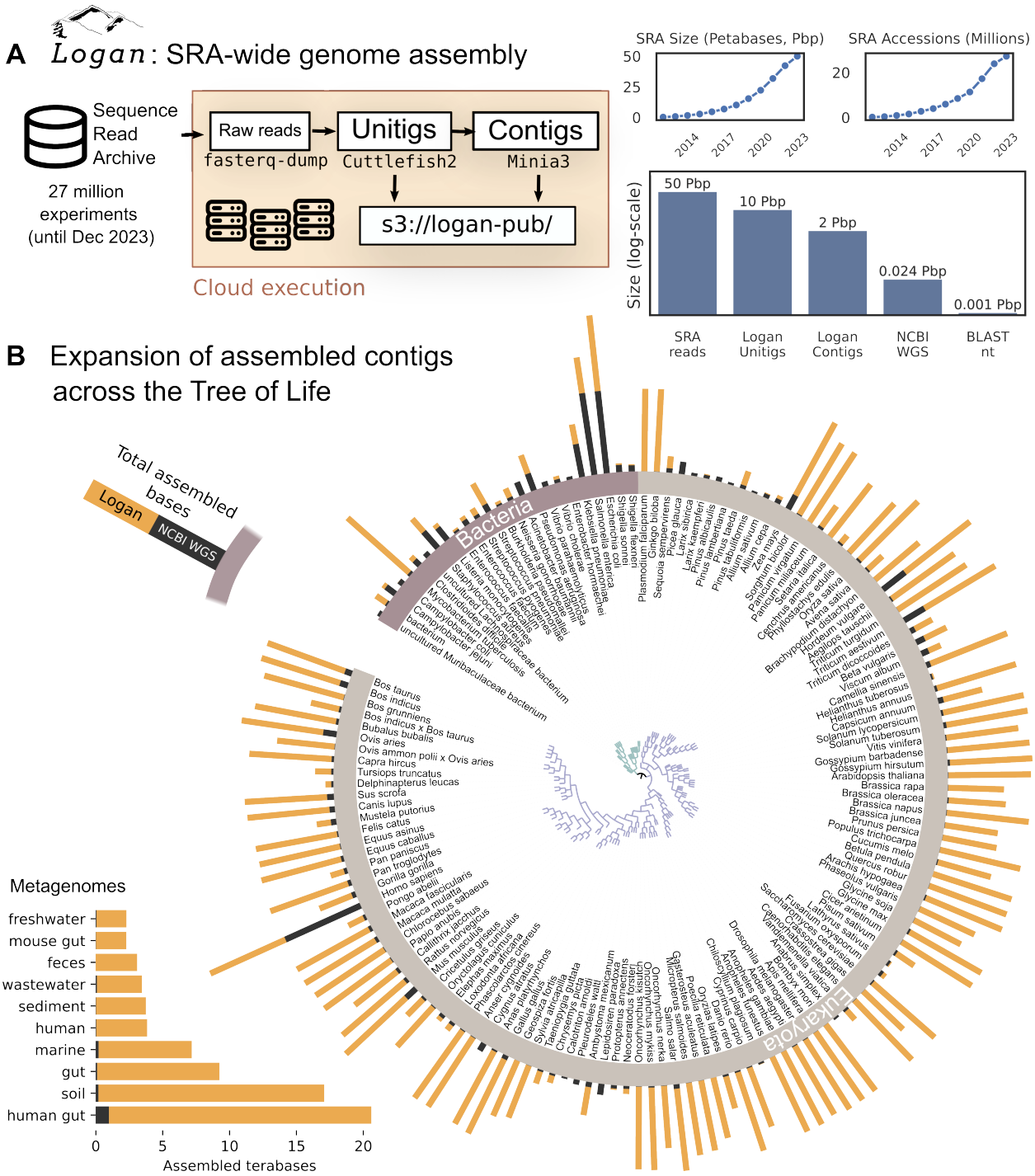


Figure 1: **Assembling all accessions of the SRA using a cloud architecture into unitigs and contigs.** Panel A: Left diagram describes the cloud computation workflow of Logan, starting from SRA reads, then computing unitigs and contigs assemblies, and finally uploading data to our public repository. Right graphs describe the Sequence Read Archive (SRA) and its near-exponential growth in terms of number of accessions and cumulative size of raw data. Bottom right graph reports the size of the SRA compared to Logan assembled unitigs and contigs in sum of bases, and WGS and BLAST databases. Panel B: Tree of Life sampled with the 116 most abundant taxon from NCBI WGS as well as 116 most abundant taxon in Logan assemblies, according to NCBI taxonomy. Black bars represent the total number of assembled bases in NCBI WGS, and orange bars the additional number of bases in Logan contigs. All bars are clamped at 1 terabases for visibility of less abundant taxons. Assembled bases for a subset of metagenome types are represented separately as a barplot (bottom left).

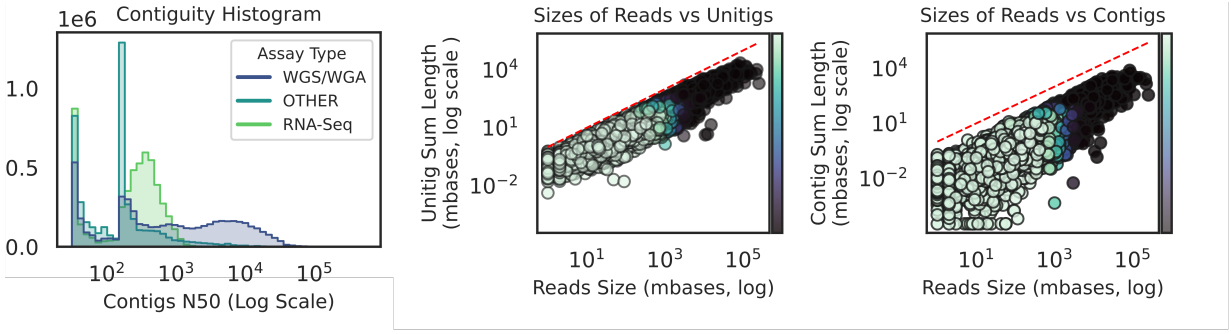


Figure 2: **Statistics of SRA assemblies in Logan.** Left panels shows the frequency distribution of contiguity (contig N50) of contigs, for Whole-Genome Sequencing and Whole-Genome Amplification (WGS/WGA), RNA-Sequencing (RNA-Seq) and other (non-ampliconic) sequencing experiments. Middle (resp. right) panel shows the size of raw reads versus the size of unitigs (resp. contigs) for a random subsample of SRA accessions.

2.2 Order-of-magnitude more data processed than in other SRA-derived resources

The Logan dataset represents the most extensive set of assembled sequences to date. It is two orders of magnitude larger than NCBI WGS, the set of all assembled sequences in GenBank [25]. Table 1 compares Logan to three of the most comprehensive previous efforts at indexing the SRA: Sourmash Branchwater, MetaGraph and Pebblescout [9, 11, 27]. Compared to those, Logan contains an order of magnitude more SRA data, both in number of accessions and total size, albeit it is not searchable. Figure 1B represents a subsample of the Tree of Life with the 116 most abundant taxa of Logan and of NCBI WGS, annotated by the volume of data in Logan and WGS. Note that the most abundant organism from the Archea domain (*C. Thermoplasmatota*) appears at position 1550 in the list of abundant taxa hence Archaea are not represented in the tree. Figure A1 shows a histogram of the number of accessions for the most dominant taxonomic group per accession, reflecting that coronavirus, human, and mouse datasets are the most abundant in the SRA, and all the next 7 most abundant taxa are bacterial.

2.3 SRA-wide sequence alignment over diverse targets

We have performed sequence alignment over the entire set of SRA contigs. A search database was created from all known RNA viral palmprints cores (palmcores) [1], 16S and ITS sequences from the BOLD database [21], and viral sequences of papillomavirus and anellovirus as well as from the recently discovered Obelisks elements [28]. By streaming all Logan contigs and aligning them using DIAMOND2 to this search database, we collected a terabyte-sized collection of hits to all SRA accessions. In Figure 3 we report the number of hits per sequencing data type.

As expected, Obelisks are mostly found in (human) metagenomes. Logan greatly expands the set of accessions where Obelisks are detected: from 34 datasets using Pebblescout [28] and 4,505 datasets using Serratus search over SRA reads (confident Oblin-1 hits in [28]), to 62,282 accessions in Logan contigs (Oblin-1, E-value $< 10^{-10}$). Palmcores reflect the presence of RNA viruses, and are found in nearly all RNA-sequencing experiments. Further exploration of these hits will consist

of collecting Logan contigs to retrieve longer barcode sequences or viral genomes, and examine any putatively new host, expanding our genomic and biogeographic knowledge of these organisms.

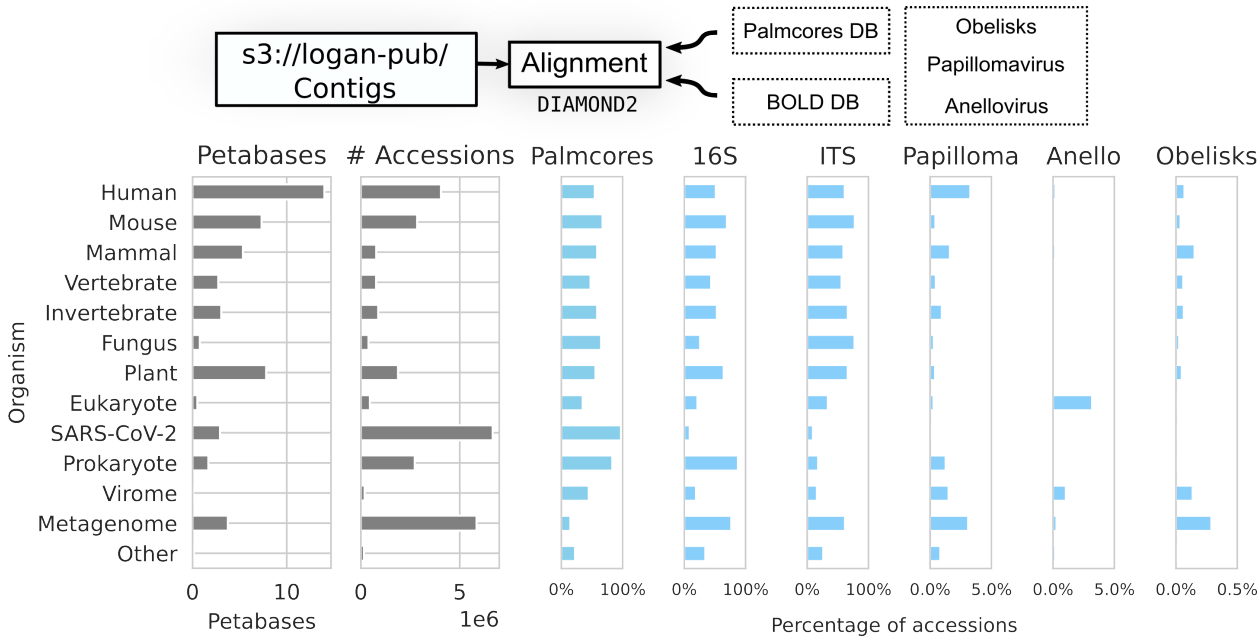


Figure 3: **Survey of viral diversity, 16S and ITS markers, Papillomavirus and Anellovirus sequences and Obelisks viroid-like elements across the entire SRA.** Top panel shows a cloud alignment workflow that aligns Logan contigs to an aggregated database of reference sequences of interest. Bottom first two plots show the (raw reads) size and number of SRA accessions split by organism indicated by SRA metadata. Next plots show the percentage of SRA accessions having at least one DIAMOND hit (E-value $< 10^{-10}$) for the SRA-wide search against palmcores (viral barcodes), 16S and ITS marker genes, Papillomavirus and Anellovirus sequences and Obelisks elements.

3 Methods

3.1 Input data

We selected all public samples from the Sequence Read Archive on December 10th, 2023, with read length above 31 bp. Accessions with shorter reads than 31 bp would yield no useable k-mers in the downstream assembly step. The list of accessions was obtained from the NCBI SRA metadata table, using AWS Athena with SQL query filter: `WHERE consent = 'public' and avgspotlen >= 31`. This resulted in 27,764,169 accessions totalling 50,304,659,857 bases in reads. In the rest of this manuscript we refer to this dataset as 'the SRA', although the current-day SRA has since been updated with new samples.

3.2 Assembly tools

Unitigs (non-branching paths of the de Bruijn graph, here k=31) were constructed using a modified version of Cuttlefish2 (commit `9401ef5` of forked repository `github.com/rchikhi/cuttlefish`) which records approximate mean k-mer abundance per unitig. We modified the k-mer counting method KMC3 [14] integrated inside of Cuttlefish2 to stream SRA files directly through a piped call to `fasterq-dump` with parameters `--seq-defline '>' --fasta-unsorted --stdout`, avoiding a prior decompression step to disk, and discarding on the fly FASTQ headers and quality values.

In Cuttlefish2, abundances were recorded per k-mer then averaged over all k-mers of a unitig. Per k-mer abundances are approximate due to two heuristics: 1) Cuttlefish2 records abundances for k+1-mers and not k-mers, hence the abundance of a k-mer was obtained by summing all the abundances of the k+1-mers it appears in, then divided by two. 2) To save memory during Cuttlefish2, small abundances of k-mers were accurate within 5% error, and large abundances were capped at 50,000. To remove some of the likely sequencing errors, k-mers seen only once in an accession were discarded from unitigs.

Unitigs were given as input to Minia3 [5] (commit `71484e8` of original repository), which performs de Bruijn graph simplifications inspired by the SPAdes assembler [20], and outputs assembled contigs. Contigs shorter than 150 bp which are not connected to any other contig are discarded. For both unitigs and contigs, the FASTA headers contain BCALM2-style link information that enable to reconstruct the assembly graph in GFA format.

These tools were selected for their memory and running time frugality, and conservativeness in results quality. A comparison with two other state-of-the-art assemblers (Penguin [10], rnaviralSPAdes [17]) is provided in A2, showing that substantial cloud computing costs were saved in our pipeline.

Unitigs and contigs were compressed using a new block variant of the Zstandard algorithm [6] (`https://github.com/asl/f2sz`), compatible with the `zstd` command line tool. The compressor creates FASTA-aligned blocks allowing for faster random access to any subset of contigs.

3.3 Cloud infrastructure

For SRA-scale assembly we have a set up a cloud infrastructure on Amazon Web Services (AWS), to perform assembly of each SRA accession in an embarrassingly parallel fashion. In a nutshell, the infrastructure is based on a Docker container executing a set of Python scripts responsible for calling child programs for unitig and contig assembly and validation. The AWS Batch execution system handles scheduling of containers across a pool of cloud computers (AWS EC2 instances). Each container is executed independently for each SRA accession. Each instance is equipped with temporary network storage (EBS). The number of CPUs, RAM, and storage for each job were set according to size of input reads measured in megabases. The Batch instances pool was set to the `c6g` and `c7g` families (AWS Graviton-based instances), with sizes `4xlarge` and above to target larger instances and thus limit the number of instance creation API calls.

Executions were monitored primarily through live dashboards on AWS CloudWatch and Batch services, as well as a DynamoDB database recording runtime and assembly statistics for each accession. Global statistics such as number of processed accessions and total size of raw data assembled were recorded in real-time by sending messages from each container to a global partitioned database. Summary metrics were aggregated, enabling to monitor computation speed during execution and potentially stop all jobs should metrics fall behind projected estimates.

To limit the number of simultaneous queries to NCBI servers, two mechanisms were implemented: (1) raw `.sra` files were directly downloaded from the AWS Registry of Open Data cloud mirror of the NCBI SRA to a cloud instance in the same data center (`us-east-1`), and (2) special `.sra` files containing alignments to RefSeq were handled by downloading references from a S3 mirror of RefSeq. Aligned `.sra` file containing references from the NCBI WGS database were discarded in the later runs, but some were processed in the earlier runs before we identified that they incurred (rate-limited) queries to NCBI servers.

Results (unitigs, contigs) have been deposited in a public repository. Detailed instructions to download the data are provided here: `https://github.com/IndexThePlanet/Logan`. Logan essentially recapitulates the information contained in the entire SRA, at the same geographical location as one of its cloud mirrors, in 114x smaller space.

While the Logan cloud infrastructure is solely intended for internal usage due to its high execution costs, its source code is publicly available at `https://gitlab.pasteur.fr/rchikhi_pasteur/erc-unitigs-prod/`. At the moment the Logan dataset is not planned to be updated.

3.4 Assembly quality assessment

Standard assembly metrics were computed using seqkit [26]: number of unitigs (resp. contigs), N50, total length, longest sequence. In addition, FASTA file sizes before and after Zstandard compression are also recorded. All these statistics are stored on a AWS DynamoDB database then exported to a S3 bucket in the Parquet format, enabling users to link this database with other NCBI SRA databases such as STAT [13] or SRA metadata.

In a unitigs file of an accessions all 31-mers are distinct, by construction. In contigs, the same rule applies in principle. We noticed *a posteriori* a bug in some of the circular connected components assembled by Minia3 within each accession: sequences marked as circular (i.e. where the suffix of the sequence (k-1)-overlaps its prefix) were "looped" 2 to 3 times, e.g. given a circular sequence s the contig sequence is reported incorrectly as $s's$ or $s's's$ where s' is s truncated of its last (k-1) characters. This bug leads to repeated 31-mers within contigs of an accession. Only the contigs are affected; the unitigs are unaffected. We note that the looped circular sequences are still biologically correct hence we did not attempt to correct this artifact in the assemblies, but we provide a tool to do so (`https://gitlab.pasteur.fr/rchikhi_pasteur/logan-circles`).

3.5 SRA-wide sensitive sequence alignment

DIAMOND2 was run with parameters `-b 0.4 --masking 0 -s 1 --sensitive` to balance speed and sensitivity, given the large index size. A custom cloud pipeline was set up to perform the search, which took approximatively 11 hours on 60k vCPUs.

4 Data Availability

Logan is publicly available on AWS Registry of Open Data at `https://registry.opendata.aws/pasteur-logan/`. There are no egress charges and anonymous access is permitted. A data access tutorial is provided at `https://github.com/IndexThePlanet/Logan`.

5 Acknowledgments

We thank Greg Autric, Maxime Hugues, Dorian Schaal and Adrien Lainé from AWS for support, Thomas Menard and Stéphane Fournier from Pasteur DSI for IT support, Peter Schmiedeskamp, Chris Stoner and Erin Chu from AWS Registry of Open Data for data hosting, Ryan Connor and Yuriy Skripchenko from NCBI for assistance with SRA operations, and Matthieu Falce for custom AWS tooling. Administrative support was provided by Melanie Ridel, Loïc Orellou, and Florence Percie du Sert. We acknowledge the help of the HPC Core Facility of the Institut Pasteur for this work.

R.C was supported by ANR grants ANR-22-CE45-0007, ANR-19-CE45-0008, PIA/ANR16-CONV-0005, ANR-19-P3IA-0001, ANR-21-CE46-0012-03, and Horizon Europe grants No. 872539, 956229, 101047160 and 101088572 (ERC IndexThePlanet).

References

[1] Artem Babaian and Robert Edgar. Ribovirus classification by a polymerase barcode sequence. *PeerJ*, 10:e14055, 2022.

[2] Phelim Bradley, Henk C Den Bakker, Eduardo PC Rocha, Gil McVean, and Zamin Iqbal. Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology*, 37(2):152–159, 2019.

[3] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015.

[4] I-Min A Chen, Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, Stephan J Ritter, Cody Webb, Dongying Wu, et al. The IMG/M data management and analysis system v. 7: content updates and new features. *Nucleic Acids Research*, 51(D1):D723–D732, 2023.

[5] Rayan Chikhi and Guillaume Rizk. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for Molecular Biology*, 8:1–9, 2013.

[6] Yann Collet. Rfc 8878: Zstandard compression and the'application/zstd'media type, 2021.

[7] Robert C Edgar, Brie Taylor, Victor Lin, Tomer Altman, Pierre Barbera, Dmitry Meleshko, Dan Lohr, Gherman Novakovsky, Benjamin Buchfink, Basem Al-Shayeb, et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature*, 602(7895):142–147, 2022.

[8] Martin Hunt, Leandro Lima, Wei Shen, John Lees, and Zamin Iqbal. AllTheBacteria-all bacterial genomes assembled, available and searchable. *bioRxiv*, pages 2024–03, 2024.

[9] Luiz Irber, N Tessa Pierce-Ward, and C Titus Brown. Sourmash branchwater enables lightweight petabyte-scale sequence search. *bioRxiv*, pages 2022–11, 2022.

[10] Annika Jochheim, Florian E Jochheim, Alexandra Kolodyazhnaya, Etienne Morice, Martin Steinegger, and Johannes Soeding. Strain-resolved de-novo metagenomic assembly of viral genomes and microbial 16S rRNAs. *bioRxiv*, pages 2024–03, 2024.

[11] Mikhail Karasikov, Harun Mustafa, Daniel Danciu, Christopher Barber, Marc Zimmermann, Gunnar Rätsch, and André Kahles. Metagraph: Indexing and analysing nucleotide archives at petabase-scale. *BioRxiv*, pages 2020–10, 2020.

[12] Kenneth Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, and Christopher O'Sullivan. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Research*, 50(D1):D387–D390, 2022.

[13] Kenneth S Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, and Christopher O'Sullivan. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. *Genome Biology*, 22:1–15, 2021.

[14] Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.

[15] Bin Ma, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, Xiaofei Lv, Ronghui Pan, Jiabao Zhang, et al. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nature Communications*, 14(1):7318, 2023.

[16] Jorge E Mario-Vasquez, Ujwal R Bagal, Elijah Lowe, Aleksandr Morgulis, John Phan, D Joseph Sexton, Sergey Shiryev, Rytis Slatkevičius, Rory Welsh, Anastasia P Litvintseva, et al. Finding candida auris in public metagenomic repositories. *Plos One*, 19(1):e0291406, 2024.

[17] Dmitry Meleshko, Iman Hajirasouliha, and Anton Korobeynikov. coronaSPAdes: from biosynthetic gene clusters to RNA viral assemblies. *Bioinformatics*, 38(1):1–8, 2022.

[18] Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S Pollard, and Nikos C Kyrpides. New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753):505–510, 2019.

[19] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542, 2017.

[20] Andrey Prjibelski, Dmitry Antipov, Dmitry Meleshko, Alla Lapidus, and Anton Korobeynikov. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, 70(1), June 2020.

[21] Sujeevan Ratnasingham, Catherine Wei, Dean Chan, Jireh Agda, Josh Agda, Liliana Ballesteros-Mejia, Hamza Ait Boutou, Zak Mohammad El Bastami, Eddie Ma, Ramya Manjunath, et al. Bold v4: A centralized bioinformatics platform for dna-based biodiversity data. In *DNA Barcoding: Methods and Protocols*, pages 403–441. Springer, 2024.

[22] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759, 2023.

[23] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 40(D1):D13–D25, 2012.

[24] Eric W Sayers, Jeff Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Donald C Comeau, Ryan Connor, Michael DiCuccio, Catherine M Farrell, Michael Feldgarden, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 52(D1):D33, 2024.

[25] Eric W Sayers, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, and Ilene Karsch-Mizrachi. GenBank 2024 update. *Nucleic Acids Research*, 52(D1):D134–D137, 2024.

[26] Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS One*, 11(10):e0163962, 2016.

[27] Sergey A Shiryev and Richa Agarwala. Indexing and searching petabase-scale nucleotide resources. *Nature Methods*, pages 1–9, 2024.

[28] Ivan N Zheludev, Robert C Edgar, Maria Jose Lopez-Galiano, Marcos De la Peña, Artem Babaian, Ami S Bhatt, and Andrew Z Fire. Viroid-like colonists of human microbiomes. *BioRxiv*, 2024.

A Appendix

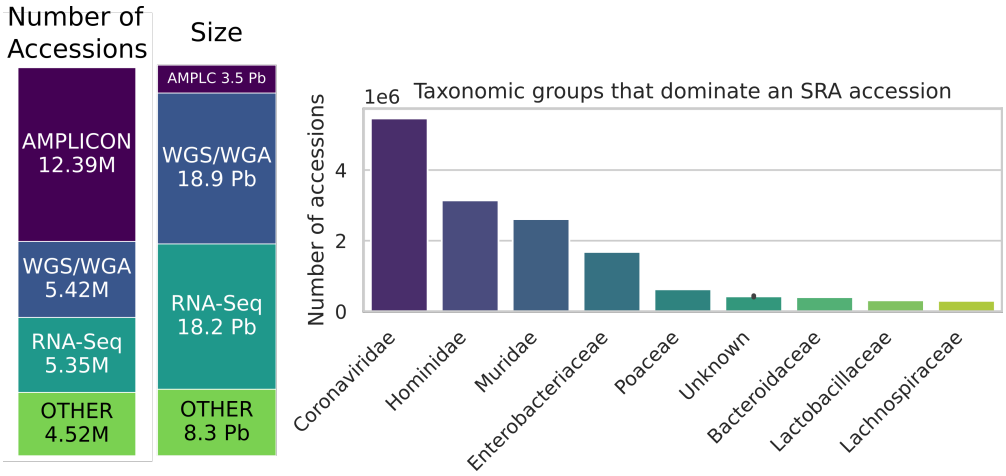


Figure A1: **Multiple datatypes and diversity in SRA samples.** Left panel is breakdown of SRA number of accessions and sizes per sequencing data type, as of December 2023. Right panel is an histogram of taxonomic groups that dominate each SRA accession, as estimated by STAT [13].

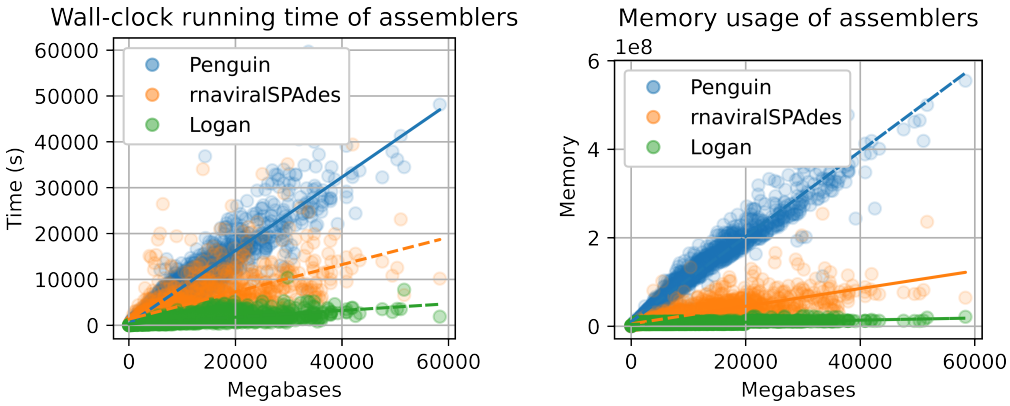
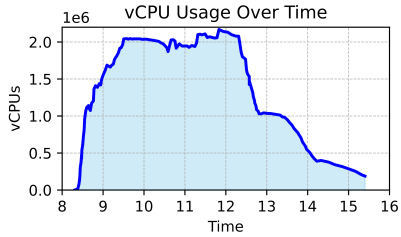


Figure A2: **Time and memory usage of assembly methods.** We evaluated Penguin, rnaviralSPAdes, and the Logan pipeline (Cuttlefish2 and Minia3) on a pilot set of 1,403 RNA-sequencing accessions. Note that the tools were run with per-accession customized number of threads. For Penguin and rnaviralSPAdes: the maximum memory usage was predicted using a linear regression, and the number of threads were set to 1/5th of the predicted memory usage (in GB). For Logan, accessions smaller than 20 Gbp were run with 4 threads and 8 GB maximum memory, otherwise with 8 threads and 16 GB maximum memory.

Infrastructure statistics for computation over the entire SRA

Global statistics

Input SRA Accessions	27 million
Input SRA size	50 petabases
Total CPU Hours	~30 million
Number of Runs	6
Total Runtime	30 hours



Run 6 statistics

Input data	19.6 petabases
Runtime*	7 hours
Peak Number of Instances	73,100
Peak Number of vCPUs	2.18 million
Peak Total EBS storage	52 petabytes

Figure A3: **Statistics of Logan computation.** Left panel shows global statistics for all Logan computation, aggregated over 6 runs. Middle panel shows the number of CPUs across instances during the 6th (and last) run. Right panel shows selected statistics for the 6th run. Runtime is reported from the beginning of the run until $< 10\%$ of the peak number of instances remain.