nature portfolio

Corresponding author(s): Artem Babaian

Last updated by author(s): Nov 22, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Serratus (v0.3.0) is available at https://github.com/ababaian/serratus. Electronic notebooks for experiments are available at https://github.com/ababaian/serratus.
Data analysis	Archival copies of SerraTax, SerraPlace, Batch Assembly workflow, and DARTH are available at https://github.com/serratus-bio/. coronaSPADES (2020-07-15) is available at https://cab.spbu.ru/software/coronaspades. palmDB sequence database (2021-03-14) is available at https://github.com/rcedgar/palmdb and Palmscan (v1.0.0) https://github.com/rcedgar/palmscan. Software used in analysis: bcftools (v1.7), Bowtie2 (v2.4.1), BWA (v0.7.17), CD-HIT (v4.8.1), CheckV (v0.6.0), coronaSPAdes (v3.15.3), D3.js (5.16.0), DARTH (maul), DIAMOND (v2.0.8), Dustmasker (ncbi-blast:2.10.0), EPA-ng (v0.3.7), Gappa (v0.6.1), getorf (EMBOSS:6.6.0.0), Grafana (8.2.5), HMMER3 (v3.3), Infernal (v1.1.4), IQTREE (v.1.6.6), MAFFT (v.7.407), MEGAHIT (v1.2.9), ModelTest-NG (v0.1.3), MUSCLE (v3.8), nidhoggr (v0.1), Palmscan (v1.0.0), ParGenes (v1.1.2), PostgreSQL (10.14), Prodigal (v2.6.3), Prometheus (2.5.0), RAxML-NG (v0.9.0), React (16.13.1), rnaviralSPAdes (v3.15.3), SAMtools (v1.7), seqkit (v0.12.0), seqtk (github@7c04ce7), Serratus (v0.3.0), snakemake (v6.6.0), TrimAL (v1.14), UBLAST (usearch v11.0.667), UCHIME2 (usearch v8.0.1623), USEARCH (v11.0.667), VADR (1.1), Virsorter2 (v2.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All Serratus data, raw and processed, is released into the public domain immediately in accordance with the Bermuda Principles and freely available at https://serratus.io/access. Assembled genomes for this study are available on GenBank under project PRJEB44047.

SRA datasets analyzed in detail: ERR2756788, ERR866585, SRR12063536, SRR12300397, SRR2136906, SRR5001850, SRR5864109, SRR6201737, SRR6943136, SRR7170939, SRR7286070, SRR7910143, SRR8242383, SRR8739608, SRR8840728, SRR8924823, SRR8954566; and the sequencing libraries in BioProjects PRJEB9357 and PRJEB34360.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Metagenomic re-analysis of 5.7 million public sequencing datasets to uncover viral sequence diversity. Study design is consistent with a discovery oriented project.
Research sample	Data was obtained from the Sequence Read Archive (SRA); RNA-seq, metagenomic, and metatranscriptomic runs were gathered with queries as defined in Extended Table 1a. In brief, queries for Human (n = 837 694), Mouse (1 058 559), Bat (14 103), Vertebrate (114 078), Invertebrate (184 729), Eukaryotes (184 729), Prokaryotes (2 672 802), Metagenome (566 826), and Virome (52 072); Mammalian DNA (14 103) were searched for the RNA viral hallmark gene, RNA dependent RNA polymerase.
Sampling strategy	No sample-size calculations were performed. We opted to search all available/relevant data exhaustively. For the RNA virus search we limited our search to datasets derived from RNA as the starting material, or metagenomes, with the exception of the Mammalian DNA sequencing control set.
Data collection	Data was collected by running Serratus command-line and documented via a Jupyter electronic notebook by A. Babaian. Notebooks are available at https://github.com/ababaian/serratus/tree/master/notebook
Timing and spatial scale	The underlying sequencing datasets were generated and shared by the global biology community rangng from 2007-2021. This data spans all continents. Data collection from the SRA was performed between 2020-05-30 and 2020-07-11 for the nucleotide search 2021-01-11 and 2021-01-21 for the RdRP search.
Data exclusions	Data was limited to sequencing runs on the ILLUMINA platform to allow for uniform data processing. Whole Genome Sequencing data was not searched except for Chordata (bat) samples or for the Mammalian WGS control experiment. Exclusion criteria was established.
Reproducibility	The SRA Run Info tables to reproduce our search are available at https://serratus.io/access. Note: due to ongoing data migrations, some sequencing runs in the SRA may be temporarily unavailable and need to be re-attempted after a few days. Data in the SRA is archived and freely available for reproduction.
Randomization	No randomization was performed in this study and no controlling for covariants is not relevant to this study design.
Blinding	Blinding does not apply to this study as it is discovery-oriented.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging