

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

Data analysis

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](#)

Databases our study used include (1) the TARA Oceans metagenomes and metatranscriptomes (<https://www.ebi.ac.uk/ena/browser/view/PRJEB402>), (2) publicly available Nucleocytoviricota MAGs (<https://www.nature.com/articles/s41586-020-1957-x> and <https://www.nature.com/articles/s41467-020-15507-2>), (3) and Virus-Host DB (<https://www.genome.jp/virushostdb/>), (4) RefSeq (<https://ftp.ncbi.nlm.nih.gov/refseq/>), (5) UniRef90 (<https://ftp.ebi.ac.uk/pub/databases/uniprot/uniref/uniref90/>), (6) NCVOG (<https://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG/>) and (7) NCBI nr database (<https://ftp.ncbi.nih.gov/blast/db/>). Data our study generated has been made publicly available at <https://doi.org/10.6084/m9.figshare.20284713>. This link provides access to (1) the RNAPolB genes reconstructed from the Tara Oceans assemblies (along with references), (2) individual FASTA files for the 1,593 non-redundant marine Nucleocytoviricota and mirusvirus MAGs (including the 697 manually curated MAGs from our survey) and 224 reference Nucleocytoviricota genomes contained in the GOEV database, (3) the GOEV anv'io CONTIGS database, (4) genes and proteins found in the GOEV database, (5) manually curated hallmark genes, (6) predicted 3D structures of the Duplodnaviria virion module (includes proteins and their alignments), (7) phylogenies and associated anv'io PROFILE databases with metadata, (8) HMMs for hallmark genes, (9) a FASTA file for the near-complete contiguous genome (SAMEA2619782_METAG_scaffold_2), (10) and all the supplemental tables.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<input type="text" value="Not applicable"/>
Population characteristics	<input type="text" value="Not applicable"/>
Recruitment	<input type="text" value="Not applicable"/>
Ethics oversight	<input type="text" value="Not applicable"/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	<input mirusviricota"="" type="text" value="The study is based on metagenomic data generated by the Tara Oceans consortium over the years. We characterized and manually curated environmental genomes for giant viruses as well as a previously unknown clade dubbed "/> .
Research sample	<input type="text" value="Sunlit oceans (plankton)"/>
Sampling strategy	<input type="text" value="The study did not involve any sampling, and we used data generated by the Tara Oceans consortium"/>
Data collection	<input type="text" value="We used all Tara Oceans metagenomes. Those were generated in our institute s part of previous publications, so we had direct access to the data. The data is also publicly available to others."/>
Timing and spatial scale	<input type="text" value="The study did not involve any sampling or other data collection."/>
Data exclusions	<input type="text" value="No data was excluded."/>
Reproducibility	<input type="text" value="All data is available, and the tool anv'io is available for all to reproduce our findings."/>
Randomization	<input type="text" value="We worked on all the metagenomic legacy of Tara Oceans, so their is no randomization."/>
Blinding	<input type="text" value="We worked on all the metagenomic legacy of Tara Oceans, so their is no blinding."/>

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 | Involvement in the study |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

Methods

- | n/a | Involvement in the study |
|-------------------------------------|---|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |