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

- Softwares used in the study:
  - Spades (v3.14)
  - Metabat2 (v2.15)
  - Bowtie2 (v2.3.4.1)
  - Checkm (v1.1.2)
  - Genemark (v3.36)
  - BLAST+ (v2.8.1)
  - Hmmer (v3.2.1)
  - TopGO (v2.40.0)
  - Diamond (v0.9.31.132)
  - Interproscan (v5.39-77)
  - Emsapper-1.03
  - Bornnap (v0.9)
  - ViralrecaII (v2.0)
  - Integrative Genome Viewer (2.8.6)
  - Samtools (v1.3.1)
  - MAFFT (v7.407)
  - ClipKIT (v1.3.0)
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

Supplementary data files can be accessed here: https://doi.org/10.6084/m9.figshare.10464917. Large genome fragments and annotations were deposited under the study PRJEB47746 with the following accessions: ERS10539964, ERS10539963, ERS10539962, ERS10539961, ERS10539960, ERS10539959, ERS10539958, ERS10539957. Source data are provided with this paper. In addition, previously published public data used for analysis includes: Genbank NR (from June 2020), GVMAGs [https://figshare.com/s/14788156283d65466731] and https://genome.jgi.doe.gov/portal/GVMAGs/GVMAGs.home.html], VOG orthogroups [https://vogdb.org/], Refseq protein database [from March 2020], EgpNOG [v5], GEVEs [https://zenodo.org/record/3975964#.XeFJoH17mfZ], JGI IMG/M [database 2021, https://img.jgi.doe.gov/].

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

The study provides computational analyses of large DNA viral genomes in permafrost metagenomics data and comparison to other environments. No replicates were used in this study.

Research sample

Among the 12 permafrost samples generated by a previous study [https://doi.org/10.1093/femsml/uqac003], eleven were kept for analysis. One sample (sample B) was excluded since it was older than the other ones and contained less data with shorter reads. For comparison, other terrestrial samples were also collected from the JGI IMG/M databases. All available samples with contigs over 10 kb were analyzed.

Sampling strategy

As described in [https://doi.org/10.1093/femsml/uqac003] no statistical method was used for the sampling procedure. We used samples corresponding to 3 different forages and two to three samples by forage. We were able to reveal a large variety of Nucleocytoviricota sequences even with this small dataset.

Data collection

Data used in this study were retrieved from a previous work [https://doi.org/10.1093/femsml/uqac003]. These samples were collected by Alexander Morawitz. Furthermore Dr. Jens Strauss and Dr. Guido Grosse performed the forage, storage and transportation of the samples.

Timing and spatial scale

As described in the previous study [https://doi.org/10.1093/femsml/uqac003] for which samples were collected, the sampling took place in July 2017 and March 15, as part of a programmed expedition considering climatic and administrative constraints. The surface cryosol samples were collected on July 2017 at these locations (GPS coordinates):
- Sample C: 54.549722°N, 160.581944°E
- Sample D: 55.098611°N, 160.349444°E
- Sample E: 55.115°N, 159.963333°E
The permafrost samples were collected on March 2015 at these locations [GPS coordinates]:
- Samples G and M: 61.76086°N, 130.47466°E
- Samples P and K: 61.75967°N, 130.47438°E
- Samples L, N and R: 61.76490°N, 130.46503°E

Data exclusions

One sample was excluded from the original permafrost dataset because it was sequenced using a different technology with shorter reads.

Reproducibility

The sequencing data from the previous study [https://doi.org/10.1093/femsml/uqac003] was not replicated to optimize the quantity and diversity of sequenced DNA. The present study does not contain experimental results but computational procedures to reveal
Nucleocytoviricota sequences were applied on different datasets. Bioinformatics procedures are described in the methods section and can be replicated.

Randomization
We did not employ any randomization in this study since we did not have different treatment groups.

Blinding
Blinding was not relevant for this study as all datasets were treated the same way.

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 |