Detecting contamination in viromes using ViromeQC

To the editor

Eukaryotic viruses and bacteriophages have important roles in microbiomes, but characterization of viruses in metagenomics data is difficult. Viral-like particle (VLP) purification enables enrichment for viruses from microbiome samples before sequencing but contamination can result in misleading conclusions. We present a software tool named ViromeQC for analyzing virome data. Here, we demonstrate the utility of ViromeQC by applying it to 2,050 human, animal and environmental samples from 35 metagenomic virome sequencing studies that used one of the available VLP enrichment techniques. The resulting analysis reveals these viromes are rife with bacterial, archaeal and fungal contamination. Most samples show only modest virus enrichment, and such enrichment is very variable between viromes in the same study. To address these issues, we present a validated contamination quality-control pipeline to enable more robust virome metagenomic analyses.

Viruses affect the ecology and composition of microbial communities[1,2]. Bacteriophages (viruses of bacteria and archaea) are extremely abundant and diverse, and affect microbiomes in several ways including transduction, which is an important mechanism of lateral gene transfer[3]. Metagenomics can be used to characterize phage populations, but phage are so diverse, and evolve so rapidly, that they are poorly represented in sequence databases. Also, there are no universal viral genetic markers and the overall biomass of viruses, compared with other microorganisms in a sample, is low. For these reasons, phage sequences are difficult to identify in metagenomes although specific methods that are partly based on sequence characteristics of known phages have been reported[4,5].

VLP purification can be used to enrich microbiome samples for viral nucleic acids[6], thereby improving virus detection. VLP protocols have various goals, ranging from untargeted analyses of highly purified phage populations to targeted identification of rare sequences of viral pathogens in diagnostic samples. These methods typically include filtration through small pore size filters that retain bacteria, cesium chloride gradient purification, treatment with chloroform to disrupt membranes, and exposure to nucleases to reduce free DNA and RNA concentration. If the aim is to use metagenomics to detect known viral pathogens, a low-purity sample may suffice because identification will be by alignment of sequence reads to viral databases. However, if the aim is to detect unknown viruses or report all viruses in a sample, a high purity sample is required. When coupled with untargeted shotgun sequencing[7], VLP enrichment has underpinned numerous studies in human[8,9], environmental [10,11], and built-environment settings[12], but there is no single VLP enrichment protocol that is optimal for all sample types.

Regardless of the VLP protocol, non-viral genetic material remains after enrichment[13]. These unwanted nucleic acids are contaminants, and their presence particularly confounds the *de novo* discovery of phages in untargeted virome sequencing. If the VLP virome is pure, it is possible to assemble reads into possibly fragmented viral genomes without using computational prediction approaches, which are unavoidably affected by low-confidence calls and false negatives[4,5]. The fraction of viruses in the VLP sample is associated with improved *de-novo* recovery of new viruses, but methods for evaluating VLP purity in samples have not been systematically explored. Studies have assessed contamination of VLP-preparations by PCR-amplification of prokaryotic 16S rRNA gene sequences before virome sequencing[11,14–19]. Others have mapped the NGS virome sequencing output against the 16S rRNA gene, or a different marker[9,20–24].

However, these studies haven't provided a validated pipeline to quantify viral enrichment in viromes or unenriched samples. Although efforts towards VLP-protocol optimization have been reported[24], the largest meta-analysis of post-sequencing non-viral quantification to date considered just 67 viromes[13]. As the use of VLP enrichment for virome sequencing is increasing, we set out to evaluate non-viral contamination in >2,000 virome samples.

To assess the enrichment rates of publicly available viromes, we applied our method (**Supplementary Methods**) on a collection of 2,050 VLP samples (**Supplementary Table 1**). As controls, we included 2,189 metagenomes that were not enriched for viruses from the *curatedMetagenomicData*[25] and the National Center for Biotechnology Information Shortread Archive (NCBI-SRA)[26] repositories, as well as 108 publicly accessible synthetic metagenomes[27,28] and one mock community (**Supplementary Table 2**). After uniform preprocessing to remove low-quality reads (**Supplementary** **Methods**), we computed the percentage of raw reads in each sample that align to the small subunit ribosomal RNA gene (SSU rRNA), which has never been found in a virus genome. This provided a proxy for non-viral microbial sequence abundance[13]. We estimated the abundance of the bacterial and archaeal 16S and micro-eukaryotic 18S ribosomal genes in all of the viromes and metagenomes. Unenriched metagenomes provided a baseline estimation of the environment-specific rRNA gene abundance, from which we calculated the relative enrichment of viromes with respect to the metagenomes. Environmental and human/animal unenriched metagenomes had a median rRNA gene abundance of 0.08% (n=320, interquartile-range=0.07%) and 0.25% (n=1,551, interquartile-range=0.1%) (**Fig. 1**).

Prokaryotic and micro-eukaryotic contamination of viromes estimated by the quantification of the SSU-rRNA revealed a wide range of enrichment efficiencies, with a large fraction of samples (n=567, 28.7%) having no virus enrichment at all, and >50% (n=990) having less than threefold enrichment. A substantially smaller fraction of samples (n=339, 17.15%) showed high enrichment (>100-fold). Differences in enrichment rates were not clearly associated with any one VLP-purification method, although the heterogeneity of protocols makes it difficult to provide statistical support to this observation. According to taxonomic annotations of the rRNA gene sequences retrieved in viromes, the largest source of contamination was bacterial DNA (1,466 samples), with 88 samples having higher abundances of eukaryotic associated SSU rRNAs (**Supplementary Table 3**). The rRNA gene abundance variability was higher in viromes than in metagenomes (Mann–Whitney U test p-value = $7.5e^{-8}$, **Supplementary Figure 1**), revealing not only that many viromes are poorly enriched for viruses, but also that the level of bacterial and archaeal contamination is unpredictable.

The intra-dataset enrichment efficiencies were extremely variable, spanning more than two orders of magnitude in 48.7% of the studies, which shows that even the same virome-enrichment protocol applied to samples from the same study can still have vastly different levels of contamination. For example, the 91 stool samples from the dataset of Ly *et al.*[18] had a 16S rRNA gene abundance standard deviation equal to 4.6 times the average (**Figure 1**; ref. 38). This suggests that quality-benchmarking viromes after sequencing is crucial to evaluate the presence of contaminants, and that intra-dataset variability should be carefully considered in downstream analyses of untargeted viromes.

Four VLP datasets were highly enriched in rRNA genes with a median abundance > 10% and peaks of 90% reads aligning to either the 16S/18S or 23S/28S rRNA gene subunits (datasets 36, 47, 50 and 51, see **Supplementary Table 1**). Conversely, the median rRNA gene abundance observed in unenriched real and synthetic metagenomes never exceeded 1% (**Supplementary**

Table 2). The experimental design of these four studies pointed at the likely cause of contamination because they involved DNA and RNA co-extraction, with DNA and retro-transcribed cDNA sequenced together. We hypothesize that higher rRNA abundance was observed due to incompletely depleted structural rRNA in the samples. In a further 25 RNA viromes, we also found higher rRNA abundances than would be expected (4.18% median abundance when considering both rRNA subunits, maximum of 67.5%, **Supplementary Table 4**). As it was not possible to evaluate the VLP enrichment efficiency by estimating rRNA abundances for samples with atypically high levels of rRNA, we excluded datasets with more than 10% median abundance of rRNA genes from the downstream analysis because viromes with such high levels of rRNA genes are unlikely to be useful in downstream genome assembly and analysis. In total, 307 samples were removed, all of which were from studies that sequenced DNA and RNA together. Although protocols of this type cannot be evaluated with our approach, they may be useful for some tasks such as sequence-based detection of known pathogens.

To improve virus enrichment estimates we next calculated the abundance of the large ribosomal subunit rRNA gene (LSU-rRNA), encoding for prokaryotic 23S and eukaryotic 28S rRNAs (**Fig. 2a**) and of 31 single-copy universal markers from bacterial and archaeal ribosomal proteins[29] (**Supplementary Figure 2**). Because some ribosomal proteins are occasionally found in viral genomes[30], it is plausible that this might result in assigning viral genomes as contaminants. However, extensive mapping of these universal ribosomal markers against viral repositories provided evidence that the rare inclusion of a marker gene in a viral genome is unlikely to affect the results (**Supplementary Note 1, Supplementary Fig. 3, Supplementary Table 5**), especially when considering all 31 single-copy universal markers. Although a few samples (11.8%) still harbored high levels of rRNA genes (i.e., >5% abundance, **Supplementary Fig. 4b, Supplementary Fig. 5**), the abundance quantification of rRNA genes (SSU and LSU) and genes encoding single-copy proteins were in agreement for most viromes. In 75.3% of the viromes, rRNA genes and single-copy marker abundances were either both below (67.1%) or above (8%) the reference unenriched-metagenomes medians (**Supplementary Fig. 4**). The abundance of the individual markers was highly correlated (**Fig. 2b**), as were the abundances of SSU rRNA and single-copy markers (Spearman's coefficient = 0.72 when considering the abundance of all 31 markers together). A weaker correlation was observed between LSU rRNA and single-copy markers (**Fig. 2b,** Spearman's coefficient = 0.47). Although rRNA and single-copy marker abundances were generally in agreement, we propose that a multi-marker approach is required to accurately estimate viral enrichment. For example, one of the datasets we examined[9] had substantial amounts of LSU rRNA genes, but was found to be highly virus-enriched if only SSU rRNA were quantified.

Finally, we defined a comprehensive enrichment score that includes rRNA large and small subunit abundances and single-copy markers. This score expresses virus enrichment in a sample compared with the medians observed in unenriched metagenomes, and was computed by taking the minimum across the three single enrichment scores for both viromes and metagenomes (see **Supplementary Methods**). Fewer than 50% of viromes that we analyzed had an overall enrichment greater than threefold, fewer than 15% reached 30-fold enrichment, and only 10% of the viromes were more than 50-fold enriched. Most of the viromes failed to meet even a low level of enrichment (two- to threefold; **Fig. 2c**). Most studies had mixed enrichment levels across samples (average of 55.41 s.d. 76.5 samples per dataset), with samples within the same dataset spanning between one- and 100-fold virus-enrichment, confirming what we observed previously on enrichments based on the SSU-rRNA gene only (**Fig. 2d**),. This further underscores how samples that underwent the same VLP-technique might have widely different levels of non-viral contamination.

To highlight the importance of quality control in untargeted virome metagenomics, we investigated the extent to which the viral enrichment score is connected with success in computational identification of viral genomes from virome samples subjected to metagenomic assembly. We assembled 1,445 untargeted virome samples and classified each of the resulting 2.09×10^7 contigs as viral or not-viral using VirSorter[4] (**Supplementary Methods**). The proportion of viral and potentially-viral contigs increased from an average of 7.9% to an average of 31% for samples with viral enrichment-scores of 1–2-fold and 5–9-fold, respectively. However, the proportion of predicted viral contigs did not substantially increase at higher enrichment values (**Supplementary Fig. 6**). Indeed, in most samples enriched by a factor of 100-fold or more, for which there are, at best, just traces of ribosomal genes from prokaryotes and eukaryotes, fewer than 25% of the assembled nucleotides could be classified as “potentially viral” (i.e., VirSorter *category 1, 2 or 3*), and fewer than 4% was ‘surely viral’ (i.e,. *category 1).* At such high enrichment rates, assembled contigs could all be considered viral, which means there is a substantial false negative rate. This is likely due to viral genomes not displaying enough similarity with known reference viruses, and to the limitation of contig-based viral detection tools when analyzing contigs with relatively short length[4]. Conversely, 55 of the 475 lowly enriched samples (i.e. less than threefold) had more than 20% of the assembled nucleotides labelled as potentially viral, which is inconsistent with the high abundance of prokaryotic organisms with much longer genomes and could suggest the presence of false positives. Caution is needed when interpreting the results of viral mining software and incorporating virome-enrichment into untargeted virome analyses should improve downstream analyses.

Our analysis should serve to raise awareness of the potential for prokaryotic and eukaryotic contamination in viromes. Unfortunately, post-sequencing evaluation of non-viral contaminants in viromes before contig-based virus classification is rarely performed. Our read-based estimates of non-viral contamination could be used to guide the selection of tools and thresholds for downstream viral contig detection. We caution that if metagenomic assembly is carried out on poorly enriched samples, it increases the number of contigs that are wrongfully assigned as viral by computational predictions.

We urge researchers to apply quality control to viromes before genome analysis. This is particularly important when datasets are retrieved from public sources, and when metagenomic assembly is used to characterize unknown viruses in samples. The computational pipeline we introduce to analyze the enrichment of viromes differs from previous methods that focused on only 16S rRNA genes to address microbial contamination. ViromeQC integrates the abundances of 16S/18S rRNA genes, 23S/28S rRNA genes, and a panel of 31 universal bacterial genes. ViromeQC software is freely available at http://segatalab.cibio.unitn.it/tools/viromeqc.

Moreno Zolfo[1], Federica Pinto[1], Francesco Asnicar[1], Paolo Manghi[1], Adrian Tett[1], Frederic D. Bushman[2] &Nicola Segata[1,]*

[1]Department CIBIO, University of Trento, Trento, Italy

[2]Department of Microbiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

** Corresponding author N.S. (e-mail: nicola.segata@unitn.it)*

1. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The 'Known Unknown' of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).

2. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801 (2007).

3. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1**, 147 (2010).

4. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

5. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).

6. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).

7. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

8. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).

9. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).

10. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).

11. Watkins, S. C. *et al.* Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Mar. Freshwater Res.* **67**, 1700–1708 (2016).

12. Rosario, K., Fierer, N., Miller, S., Luongo, J. & Breitbart, M. Diversity of DNA and RNA Viruses in Indoor Air As Assessed via Metagenomic Sequencing. *Environmental Science and Technology* **52**, 1014–1027 (2018).

13. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).

14. Minot, S. *et al.* The human gut virome : Inter-individual variation and dynamic response to diet The human gut virome : Inter-individual variation and dynamic response to diet.

Genome Res. 1616–1625 (2011).

15. Emerson, J. B. *et al.* Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl. Environ. Microbiol.* **78**, 6309–6320 (2012).
16. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12450–12455 (2013).
17. Kim, Y., Aw, T. G., Teal, T. K. & Rose, J. B. Metagenomic Investigation of Viral Communities in Ballast Water. *Environmental Science and Technology* **49**, 8396–8407 (2015).
18. Ly, M. *et al.* Transmission of viruses via our microbiomes. *Microbiome* **4**, 64 (2016).
19. Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11941–11946 (2015).
20. Roux, S. *et al.* Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**, (2012).
21. Weynberg, K. D., Wood-Charlson, E. M., Suttle, C. A. & van Oppen, M. J. H. Generating viral metagenomes from the coral holobiont. *Front. Microbiol.* **5**, 1–11 (2014).
22. Hannigan, G. D. *et al.* The human skin double-stranded DNA virome: Topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *MBio* **6**, (2015).
23. de Cárcer, D. A., López-Bueno, A., Alonso-Lobo, J. M., Quesada, A. & Alcamí, A. Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the Antarctic Peninsula. *FEMS Microbiol. Ecol.* **92**, 1–10 (2016).
24. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).
25. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
26. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).
27. Zolfo, M., Tett, A., Jousson, O., Donati, C. & Segata, N. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. *Nucleic Acids Res.* gkw837 (2016).
28. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* **18**, 181 (2017).

29. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).

30. Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat. Commun.* **10**, 752 (2019).

DATA AVAILABILITY STATEMENT

The raw reads analyzed in this study are available using accession numbers provided in **Table S1** and **Table S2**.

CODE AVAILABILITY STATEMENT

Code and documentation are available at http://segatalab.cibio.unitn.it/tools/viromeqc

AUTHOR CONTRIBUTIONS

Study conception and design: M.Z. and N.S.; Methodology and analysis: M.Z., F. P, F.A., A.T., F.B. and N.S.; Public datasets collection and curation: M.Z. and P.M.. All authors contributed to the writing of the final manuscript.

ACKNOWLEDGEMENTS

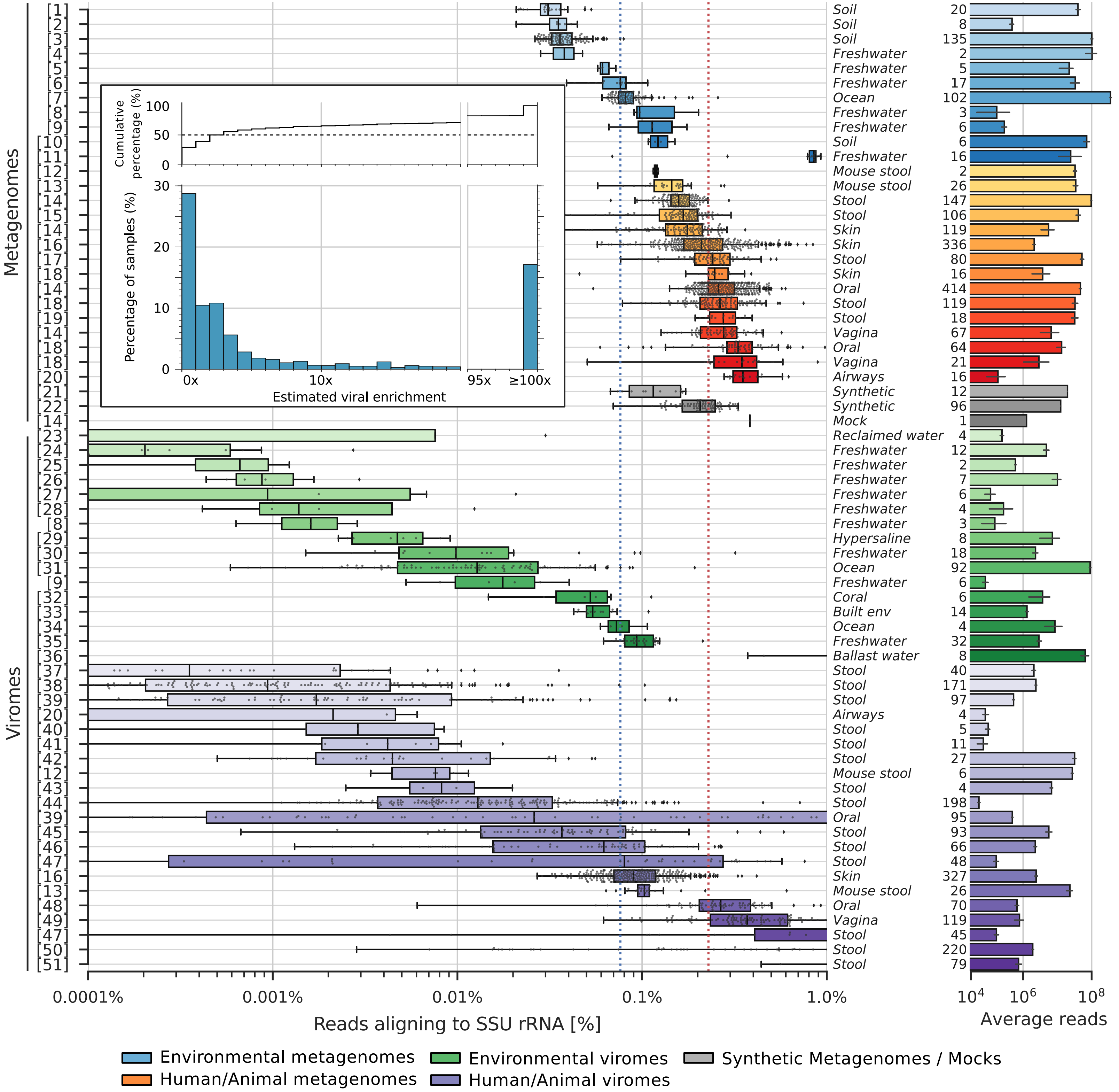
This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 716575) to N.S. The work was also supported by MIUR ‘‘Futuro in Ricerca’’ RBFR13EWWI_001 and by the European Union (H2020-SFS-2018-1 project MASTER-818368 and H2020-SC1-BHC project ONCOBIOME-825410) to N.S.

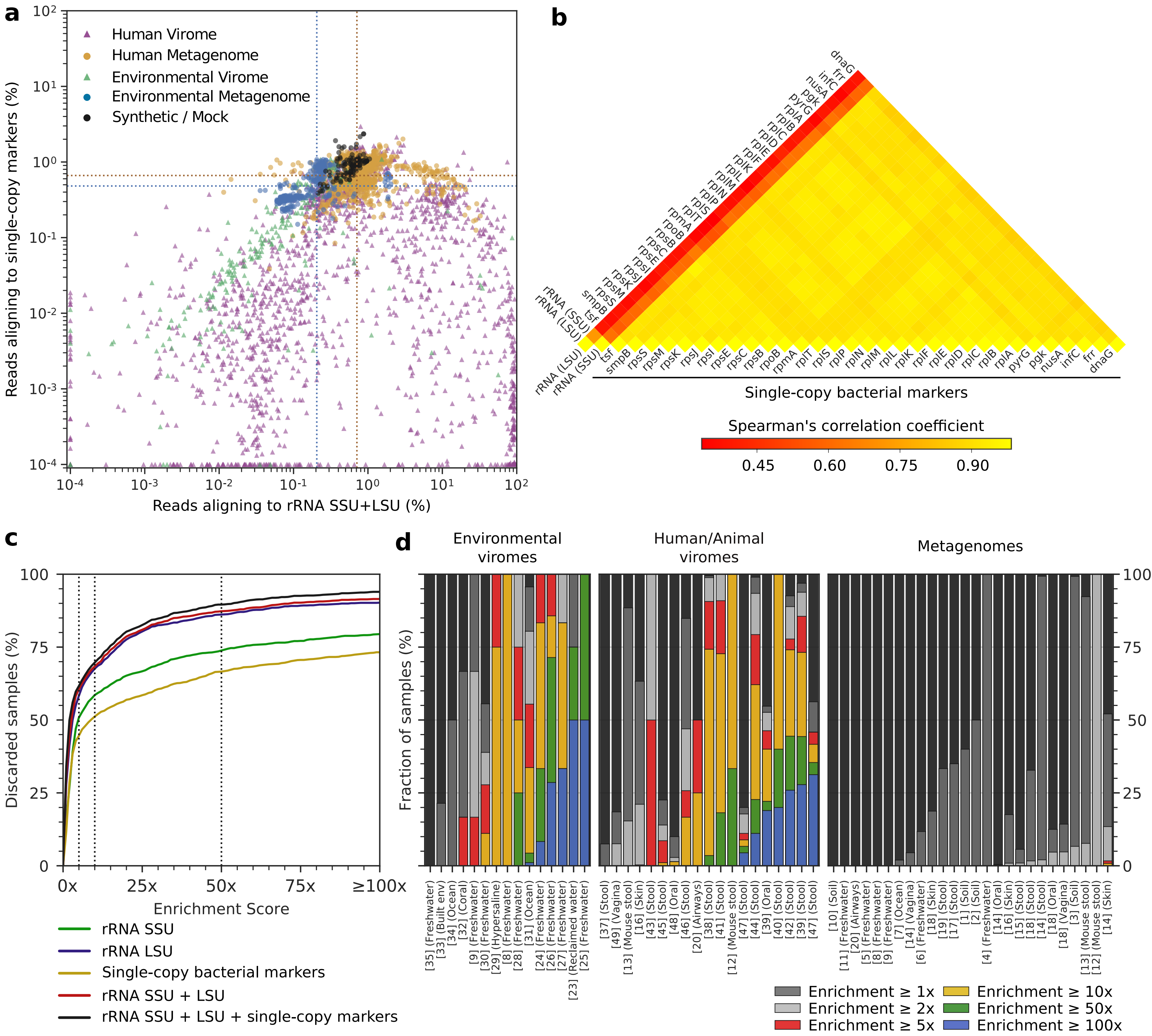
COMPETING INTERESTS

The authors declare no competing interests.

Figure 1. **Survey of viral enrichment rates on 1,977 samples from 35 studies estimated as percentage of reads aligning to the small subunit rRNA gene.** The vertical dotted lines indicate the median of median SSU rRNA abundances in human/animal (red dotted line) and environmental (blue dotted line) unenriched metagenomes, as a reference. The two medians are used to calculate the enrichment rate of each virome with respect to the reference metagenomes. The frequency of enrichment levels of all the 1,977 viromes that passed quality-control is represented in the inset histogram. The histogram on the right side shows the number of reads (bar height) and the number of samples (to the left of the bar) in each dataset. Datasets are grouped by type (environmental or Human/animal). Datasets within the same group are ordered by median abundance. References to each dataset are provided in (**Supplementary Tables 1** and **2**). Error bars in the right barplot show the 95% confidence intervals. Boxes show the quartiles of each dataset, the central line in each box indicates the median, while whiskers extend to show data points within 1.5 IQR range. Data-points, including outliers, are overlaid to the boxes

Figure 2. **Combined quantification of ribosomal genes and genes coding for universal proteins identifies the cross-study set of 101 samples with >100x VLP enrichment.** **(a)** The retrieved viromes were mapped against rRNA small and large subunits reference sequences (x-axis), and against 31 single-copy bacterial markers (y-axis). The scatter plot shows the percentage of aligned reads on 1,751 human and animal viromes (red) and 226 environmental viromes (blue). The dotted lines indicate the median abundances in the corresponding metagenomes. **(b)** Spearman's correlation coefficients between the 31 single-copy markers and the small and large subunits of the rRNA gene. **(c)** Fraction of the discarded viromes at different enrichment thresholds (dashed lines) and using different components to calculate the enrichment. The proposed threshold (rRNA SSU + LSU + single-copy markers) is drawn in black. **(d)** Enrichment scores of samples within each dataset grouped by dataset type together with metagenomes used as controls. The enrichment score is based on both SSU and LSU rRNAs and single-copy markers. References to each dataset are provided in **Supplementary Tables 1** and **2**


Metagenomes
Viromes
[1] Soil 20
[2] Soil 8
[3] Soil 135
[4] Freshwater 2
[5] Freshwater 5
[6] Freshwater 17
[7] Ocean 102
[8] Freshwater 3
[9] Freshwater 6
[10] Soil 6
[11] Freshwater 16
[12] Mouse stool 2
[13] Mouse stool 26
[14] Stool 147
[15] Stool 106
[14] Skin 119
[16] Skin 336
[17] Stool 80
[18] Skin 16
[14] Oral 414
[18] Stool 119
[19] Stool 18
[14] Vagina 67
[18] Oral 64
[18] Vagina 21
[20] Airways 16
[21] Synthetic 12
[22] Synthetic 96
[14] Mock 1
[23] Reclaimed water 4
[24] Freshwater 12
[25] Freshwater 2
[26] Freshwater 7
[27] Freshwater 6
[28] Freshwater 4
[8] Freshwater 3
[29] Hypersaline 8
[30] Freshwater 18
[31] Ocean 92
[9] Freshwater 6
[32] Coral 6
[33] Built env 14
[34] Ocean 4
[35] Freshwater 32
[36] Ballast water 8
[37] Stool 40
[38] Stool 171
[39] Stool 97
[20] Airways 4
[40] Stool 5
[41] Stool 11
[42] Stool 27
[12] Mouse stool 6
[43] Stool 4
[44] Stool 198
[39] Oral 95
[45] Stool 93
[46] Stool 66
[47] Stool 48
[16] Skin 327
[13] Mouse stool 26
[48] Oral 70
[49] Vagina 119
[47] Stool 45
[50] Stool 220
[51] Stool 79
Cumulative percentage (%)
100
50
0
Percentage of samples (%)
30
20
10
0
0x
10x
95x
≥100x
Estimated viral enrichment
0.0001%
0.001%
0.01%
0.1%
1.0%
Reads aligning to SSU rRNA [%]
10^4
10^6
10^8
Average reads
Environmental metagenomes
Human/Animal metagenomes
Environmental viromes
Human/Animal viromes
Synthetic Metagenomes / Mocks


a
Human Virome
Human Metagenome
Environmental Virome
Environmental Metagenome
Synthetic / Mock
Reads aligning to single-copy markers (%)
Reads aligning to rRNA SSU+LSU (%)
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
b
rRNA (LSU)
rRNA (SSU)
tsf
smpB
rpsS
rpsM
rpsK
rpsJ
rpsI
rpsE
rpsC
rpsB
rpoB
rpmA
rplT
rplS
rplP
rplN
rplM
rplL
rplK
rplF
rplE
rplD
rplC
rplB
rplA
pyrG
pgk
nusA
infC
frr
dnaG
Single-copy bacterial markers
Spearman's correlation coefficient
0.45
0.60
0.75
0.90
c
Discarded samples (%)
0
25
50
75
100
0x
25x
50x
75x
≥100x
Enrichment Score
rRNA SSU
rRNA LSU
Single-copy bacterial markers
rRNA SSU + LSU
rRNA SSU + LSU + single-copy markers
d
Environmental viromes
Human/Animal viromes
Metagenomes
Fraction of samples (%)
0
25
50
75
100
[35] (Freshwater)
[33] (Built env)
[34] (Ocean)
[32] (Coral)
[9] (Freshwater)
[30] (Freshwater)
[29] (Hypersaline)
[8] (Freshwater)
[28] (Freshwater)
[31] (Ocean)
[24] (Freshwater)
[26] (Freshwater)
[27] (Freshwater)
[23] (Reclaimed water)
[25] (Freshwater)
[37] (Stool)
[49] (Vagina)
[13] (Mouse stool)
[16] (Skin)
[43] (Stool)
[45] (Stool)
[48] (Oral)
[46] (Stool)
[20] (Airways)
[38] (Stool)
[41] (Stool)
[12] (Mouse stool)
[47] (Stool)
[44] (Stool)
[39] (Oral)
[40] (Stool)
[42] (Stool)
[39] (Stool)
[47] (Stool)
[10] (Soil)
[11] (Freshwater)
[20] (Airways)
[5] (Freshwater)
[8] (Freshwater)
[9] (Freshwater)
[7] (Ocean)
[14] (Vagina)
[6] (Freshwater)
[18] (Skin)
[19] (Stool)
[17] (Stool)
[1] (Soil)
[2] (Soil)
[4] (Freshwater)
[14] (Oral)
[16] (Skin)
[15] (Stool)
[18] (Stool)
[14] (Stool)
[18] (Oral)
[18] (Vagina)
[3] (Soil)
[13] (Mouse stool)
[12] (Mouse stool)
[14] (Skin)
Enrichment ≥ 1x
Enrichment ≥ 2x
Enrichment ≥ 5x
Enrichment ≥ 10x
Enrichment ≥ 50x
Enrichment ≥ 100x