

1 **Detecting contamination in viromes using ViromeQC**

2

3 **To the editor**

4

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

16 Viruses affect the ecology and composition of microbial communities^{1,2}. Bacteriophages
17 (viruses of bacteria and archaea) are extremely abundant and diverse, and affect microbiomes in
18 several ways including transduction, which is an important mechanism of lateral gene transfer³.
19 Metagenomics can be used to characterize phage populations, but phage are so diverse, and
20 evolve so rapidly, that they are poorly represented in sequence databases. Also, there are no
21 universal viral genetic markers and the overall biomass of viruses, compared with other
22 microorganisms in a sample, is low. For these reasons, phage sequences are difficult to identify
23 in metagenomes although specific methods that are partly based on sequence characteristics of
24 known phages have been reported^{4,5}.

25 VLP purification can be used to enrich microbiome samples for viral nucleic acids⁶,
26 thereby improving virus detection. VLP protocols have various goals, ranging from untargeted
27 analyses of highly purified phage populations to targeted identification of rare sequences of viral
28 pathogens in diagnostic samples. These methods typically include filtration through small pore
29 size filters that retain bacteria, cesium chloride gradient purification, treatment with chloroform
30 to disrupt membranes, and exposure to nucleases to reduce free DNA and RNA concentration. If
31 the aim is to use metagenomics to detect known viral pathogens, a low-purity sample may suffice

32 because identification will be by alignment of sequence reads to viral databases. However, if the
33 aim is to detect unknown viruses or report all viruses in a sample, a high purity sample is
34 required. When coupled with untargeted shotgun sequencing⁷, VLP enrichment has underpinned
35 numerous studies in human^{8,9}, environmental^{10,11}, and built-environment settings¹², but there is
36 no single VLP enrichment protocol that is optimal for all sample types.

37 Regardless of the VLP protocol, non-viral genetic material remains after enrichment¹³.
38 These unwanted nucleic acids are contaminants, and their presence particularly confounds the *de*
39 *novo* discovery of phages in untargeted virome sequencing. If the VLP virome is pure, it is
40 possible to assemble reads into possibly fragmented viral genomes without using computational
41 prediction approaches, which are unavoidably affected by low-confidence calls and false
42 negatives^{4,5}. The fraction of viruses in the VLP sample is associated with improved *de-novo*
43 recovery of new viruses, but methods for evaluating VLP purity in samples have not been
44 systematically explored. Studies have assessed contamination of VLP-preparations by PCR-
45 amplification of prokaryotic 16S rRNA gene sequences before virome sequencing^{11,14-19}. Others
46 have mapped the NGS virome sequencing output against the 16S rRNA gene, or a different
47 marker^{9,20-24}.

48 However, these studies haven't provided a validated pipeline to quantify viral enrichment
49 in viromes or unenriched samples. Although efforts towards VLP-protocol optimization have
50 been reported²⁴, the largest meta-analysis of post-sequencing non-viral quantification to date
51 considered just 67 viromes¹³. As the use of VLP enrichment for virome sequencing is increasing,
52 we set out to evaluate non-viral contamination in >2,000 virome samples.

53 To assess the enrichment rates of publicly available viromes, we applied our method
54 (**Supplementary Methods**) on a collection of 2,050 VLP samples (**Supplementary Table 1**).
55 As controls, we included 2,189 metagenomes that were not enriched for viruses from the
56 *curatedMetagenomicData*²⁵ and the National Center for Biotechnology Information Shortread
57 Archive (NCBI-SRA)²⁶ repositories, as well as 108 publicly accessible synthetic
58 metagenomes^{27,28} and one mock community (**Supplementary Table 2**). After uniform
59 preprocessing to remove low-quality reads (**Supplementary Methods**), we computed the
60 percentage of raw reads in each sample that align to the small subunit ribosomal RNA gene (SSU
61 rRNA), which has never been found in a virus genome. This provided a proxy for non-viral
62 microbial sequence abundance¹³. We estimated the abundance of the bacterial and archaeal 16S

63 and micro-eukaryotic 18S ribosomal genes in all of the viromes and metagenomes. Unenriched
64 metagenomes provided a baseline estimation of the environment-specific rRNA gene abundance,
65 from which we calculated the relative enrichment of viromes with respect to the metagenomes.
66 Environmental and human/animal unenriched metagenomes had a median rRNA gene abundance
67 of 0.08% (n=320, interquartile-range=0.07%) and 0.25% (n=1,551, interquartile-range=0.1%)
68 (**Fig. 1**).

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

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

90 Four VLP datasets were highly enriched in rRNA genes with a median abundance > 10%
91 and peaks of 90% reads aligning to either the 16S/18S or 23S/28S rRNA gene subunits (datasets
92 36, 47, 50 and 51, see **Supplementary Table 1**). Conversely, the median rRNA gene abundance
93 observed in unenriched real and synthetic metagenomes never exceeded 1% (**Supplementary**

94 **Table 2**). The experimental design of these four studies pointed at the likely cause of
95 contamination because they involved DNA and RNA co-extraction, with DNA and retro-
96 transcribed cDNA sequenced together. We hypothesize that higher rRNA abundance was
97 observed due to incompletely depleted structural rRNA in the samples. In a further 25 RNA
98 viromes, we also found higher rRNA abundances than would be expected (4.18% median
99 abundance when considering both rRNA subunits, maximum of 67.5%, **Supplementary Table**
100 **4**). As it was not possible to evaluate the VLP enrichment efficiency by estimating rRNA
101 abundances for samples with atypically high levels of rRNA, we excluded datasets with more
102 than 10% median abundance of rRNA genes from the downstream analysis because viromes
103 with such high levels of rRNA genes are unlikely to be useful in downstream genome assembly
104 and analysis. In total, 307 samples were removed, all of which were from studies that sequenced
105 DNA and RNA together. Although protocols of this type cannot be evaluated with our approach,
106 they may be useful for some tasks such as sequence-based detection of known pathogens.

107 To improve virus enrichment estimates we next calculated the abundance of the large
108 ribosomal subunit rRNA gene (LSU-rRNA), encoding for prokaryotic 23S and eukaryotic 28S
109 rRNAs (**Fig. 2a**) and of 31 single-copy universal markers from bacterial and archaeal ribosomal
110 proteins²⁹ (**Supplementary Figure 2**). Because some ribosomal proteins are occasionally found
111 in viral genomes³⁰, it is plausible that this might result in assigning viral genomes as
112 contaminants. However, extensive mapping of these universal ribosomal markers against viral
113 repositories provided evidence that the rare inclusion of a marker gene in a viral genome is
114 unlikely to affect the results (**Supplementary Note 1, Supplementary Fig. 3, Supplementary**
115 **Table 5**), especially when considering all 31 single-copy universal markers. Although a few
116 samples (11.8%) still harbored high levels of rRNA genes (i.e., >5% abundance, **Supplementary**
117 **Fig. 4b, Supplementary Fig. 5**), the abundance quantification of rRNA genes (SSU and LSU)
118 and genes encoding single-copy proteins were in agreement for most viromes. In 75.3% of the
119 viromes, rRNA genes and single-copy marker abundances were either both below (67.1%) or
120 above (8%) the reference unenriched-metagenomes medians (**Supplementary Fig. 4**). The
121 abundance of the individual markers was highly correlated (**Fig. 2b**), as were the abundances of
122 SSU rRNA and single-copy markers (Spearman's coefficient = 0.72 when considering the
123 abundance of all 31 markers together). A weaker correlation was observed between LSU rRNA
124 and single-copy markers (**Fig. 2b**, Spearman's coefficient = 0.47). Although rRNA and single-

125 copy marker abundances were generally in agreement, we propose that a multi-marker approach
126 is required to accurately estimate viral enrichment. For example, one of the datasets we
127 examined⁹ had substantial amounts of LSU rRNA genes, but was found to be highly virus-
128 enriched if only SSU rRNA were quantified.

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

142 To highlight the importance of quality control in untargeted virome metagenomics, we
143 investigated the extent to which the viral enrichment score is connected with success in
144 computational identification of viral genomes from virome samples subjected to metagenomic
145 assembly. We assembled 1,445 untargeted virome samples and classified each of the resulting
146 2.09×10^7 contigs as viral or not-viral using VirSorter⁴ (**Supplementary Methods**). The
147 proportion of viral and potentially-viral contigs increased from an average of 7.9% to an average
148 of 31% for samples with viral enrichment-scores of 1–2-fold and 5–9-fold, respectively.
149 However, the proportion of predicted viral contigs did not substantially increase at higher
150 enrichment values (**Supplementary Fig. 6**). Indeed, in most samples enriched by a factor of 100-
151 fold or more, for which there are, at best, just traces of ribosomal genes from prokaryotes and
152 eukaryotes, fewer than 25% of the assembled nucleotides could be classified as “potentially
153 viral” (i.e., VirSorter *category 1, 2 or 3*), and fewer than 4% was ‘surely viral’ (i.e., *category 1*).
154 At such high enrichment rates, assembled contigs could all be considered viral, which means
155 there is a substantial false negative rate. This is likely due to viral genomes not displaying

156 enough similarity with known reference viruses, and to the limitation of contig-based viral
157 detection tools when analyzing contigs with relatively short length⁴. Conversely, 55 of the 475
158 lowly enriched samples (i.e. less than threefold) had more than 20% of the assembled nucleotides
159 labelled as potentially viral, which is inconsistent with the high abundance of prokaryotic
160 organisms with much longer genomes and could suggest the presence of false positives. Caution
161 is needed when interpreting the results of viral mining software and incorporating virome-
162 enrichment into untargeted virome analyses should improve downstream analyses.

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

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

177

178 ***Moreno Zolfo¹, Federica Pinto¹, Francesco Asnicar¹, Paolo Manghi¹, Adrian Tett¹, Frederic***
179 ***D. Bushman² & Nicola Segata^{1,*}***

180

181 ¹*Department CIBIO, University of Trento, Trento, Italy*

182 ²*Department of Microbiology, University of Pennsylvania Perelman School of Medicine,*
183 *Philadelphia, PA, USA.*

184

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

186

- 187 1. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The ‘Known Unknown’ of
188 the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
- 189 2. Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.*
190 **5**, 801 (2007).
- 191 3. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat.*
192 *Commun.* **1**, 147 (2010).
- 193 4. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
194 microbial genomic data. *PeerJ* **3**, e985 (2015).
- 195 5. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based
196 tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69
197 (2017).
- 198 6. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures
199 to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
- 200 7. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun
201 metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
- 202 8. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers.
203 *Nature* **466**, 334–338 (2010).
- 204 9. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on
205 microbiome diversity. *PeerJ* **6**, e4694 (2018).
- 206 10. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant
207 ocean viruses. *Nature* **537**, 689–693 (2016).
- 208 11. Watkins, S. C. *et al.* Assessment of a metaviromic dataset generated from nearshore Lake
209 Michigan. *Mar. Freshwater Res.* **67**, 1700–1708 (2016).
- 210 12. Rosario, K., Fierer, N., Miller, S., Luongo, J. & Breitbart, M. Diversity of DNA and RNA
211 Viruses in Indoor Air As Assessed via Metagenomic Sequencing. *Environmental Science*
212 *and Technology* **52**, 1014–1027 (2018).
- 213 13. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral
214 community functional potential from viral metagenomes may be hampered by
215 contamination with cellular sequences. *Open Biol.* **3**, 130160 (2013).
- 216 14. Minot, S. *et al.* The human gut virome : Inter-individual variation and dynamic response to
217 diet The human gut virome : Inter-individual variation and dynamic response to diet.

- 218 *Genome Res.* 1616–1625 (2011).
- 219 15. Emerson, J. B. *et al.* Dynamic viral populations in hypersaline systems as revealed by
220 metagenomic assembly. *Appl. Environ. Microbiol.* **78**, 6309–6320 (2012).
- 221 16. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. U. S. A.*
222 **110**, 12450–12455 (2013).
- 223 17. Kim, Y., Aw, T. G., Teal, T. K. & Rose, J. B. Metagenomic Investigation of Viral
224 Communities in Ballast Water. *Environmental Science and Technology* **49**, 8396–8407
225 (2015).
- 226 18. Ly, M. *et al.* Transmission of viruses via our microbiomes. *Microbiome* **4**, 64 (2016).
- 227 19. Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute
228 malnutrition. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11941–11946 (2015).
- 229 20. Roux, S. *et al.* Assessing the diversity and specificity of two freshwater viral communities
230 through metagenomics. *PLoS One* **7**, (2012).
- 231 21. Weynberg, K. D., Wood-Charlson, E. M., Suttle, C. A. & van Oppen, M. J. H. Generating
232 viral metagenomes from the coral holobiont. *Front. Microbiol.* **5**, 1–11 (2014).
- 233 22. Hannigan, G. D. *et al.* The human skin double-stranded DNA virome: Topographical and
234 temporal diversity, genetic enrichment, and dynamic associations with the host microbiome.
235 *MBio* **6**, (2015).
- 236 23. de Cárcer, D. A., López-Bueno, A., Alonso-Lobo, J. M., Quesada, A. & Alcamí, A.
237 Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the
238 Antarctic Peninsula. *FEMS Microbiol. Ecol.* **92**, 1–10 (2016).
- 239 24. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal
240 phageomes. *Microbiome* **6**, 68 (2018).
- 241 25. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat.*
242 *Methods* **14**, 1023–1024 (2017).
- 243 26. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database
244 Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).
- 245 27. Zolfo, M., Tett, A., Jousson, O., Donati, C. & Segata, N. MetaMLST: multi-locus strain-
246 level bacterial typing from metagenomic samples. *Nucleic Acids Res.* gkw837 (2016).
- 247 28. Quince, C. *et al.* DESMAN: a new tool for de novo extraction of strains from metagenomes.
248 *Genome Biol.* **18**, 181 (2017).

- 249 29. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with
250 AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).
- 251 30. Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal
252 proteins. *Nat. Commun.* **10**, 752 (2019).

253

254 **DATA AVAILABILITY STATEMENT**

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

257

258 **CODE AVAILABILITY STATEMENT**

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

260

261 **AUTHOR CONTRIBUTIONS**

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

265

266 **ACKNOWLEDGEMENTS**

267 This project has received funding from the European Research Council (ERC) under the
268 European Union’s Horizon 2020 research and innovation programme (grant agreement No
269 716575) to N.S. The work was also supported by MIUR ‘‘Futuro in Ricerca’’
270 RBFR13EWWI_001 and by the European Union (H2020-SFS-2018-1 project MASTER-818368
271 and H2020-SC1-BHC project ONCOBIOME-825410) to N.S.

272

273 **COMPETING INTERESTS**

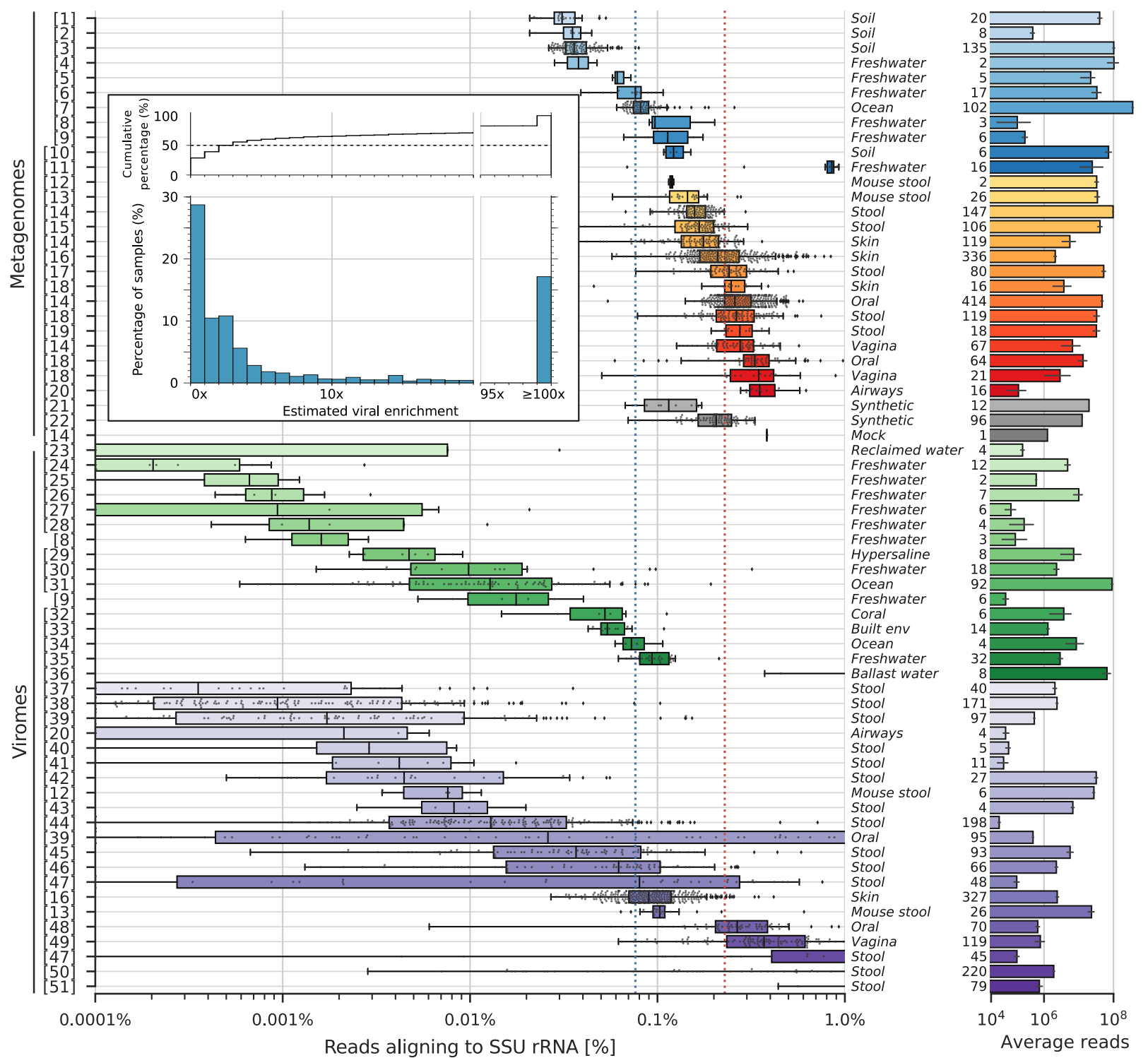
274 The authors declare no competing interests.

275

276

277 **Figure 1. Survey of viral enrichment rates on 1,977 samples from 35 studies estimated as**
278 **percentage of reads aligning to the small subunit rRNA gene.** The vertical dotted lines
279 indicate the median of median SSU rRNA abundances in human/animal (red dotted line) and
280 environmental (blue dotted line) unenriched metagenomes, as a reference. The two medians are
281 used to calculate the enrichment rate of each virome with respect to the reference metagenomes.
282 The frequency of enrichment levels of all the 1,977 viromes that passed quality-control is
283 represented in the inset histogram. The histogram on the right side shows the number of reads
284 (bar height) and the number of samples (to the left of the bar) in each dataset. Datasets are
285 grouped by type (environmental or Human/animal). Datasets within the same group are ordered
286 by median abundance. References to each dataset are provided in (**Supplementary Tables 1 and**
287 **2**). Error bars in the right barplot show the 95% confidence intervals. Boxes show the quartiles of
288 each dataset, the central line in each box indicates the median, while whiskers extend to show
289 data points within 1.5 IQR range. Data-points, including outliers, are overlaid to the boxes
290

291 **Figure 2. Combined quantification of ribosomal genes and genes coding for universal**
292 **proteins identifies the cross-study set of 101 samples with >100x VLP enrichment.** (a) The
293 retrieved viromes were mapped against rRNA small and large subunits reference sequences (x-
294 axis), and against 31 single-copy bacterial markers (y-axis). The scatter plot shows the
295 percentage of aligned reads on 1,751 human and animal viromes (red) and 226 environmental
296 viromes (blue). The dotted lines indicate the median abundances in the corresponding
297 metagenomes. (b) Spearman's correlation coefficients between the 31 single-copy markers and
298 the small and large subunits of the rRNA gene. (c) Fraction of the discarded viromes at different
299 enrichment thresholds (dashed lines) and using different components to calculate the enrichment.
300 The proposed threshold (rRNA SSU + LSU + single-copy markers) is drawn in black. (d)
301 Enrichment scores of samples within each dataset grouped by dataset type together with
302 metagenomes used as controls. The enrichment score is based on both SSU and LSU rRNAs and
303 single-copy markers. References to each dataset are provided in **Supplementary Tables 1 and 2**
304



■ Environmental metagenomes
 ■ Environmental viromes
 ■ Synthetic Metagenomes / Mocks
■ Human/Animal metagenomes
 ■ Human/Animal viromes

