Benchmarking viromics: An *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity

Simon Roux [Corresp., 1] , Joanne B Emerson [1] , Emiley A Eloe-Fadrosh [2] , Matthew B Sullivan [Corresp. 1, 3]

[1] Department of Microbiology, Ohio State University, Columbus, Ohio, United States
[2] Joint Genome Institute, Department of Energy, Walnut Creek, California, United States
[3] Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, Ohio, United States

Corresponding Authors: Simon Roux, Matthew B Sullivan
Email address: sroux@lbl.gov, mbsulli@gmail.com

Background. Viral metagenomics (viromics) is increasingly used to obtain uncultivated viral genomes, evaluate community diversity, and assess ecological hypotheses. While viromic experimental methods are relatively mature and widely accepted by the research community, robust bioinformatics standards remain to be established. Here we used *in silico* mock viral communities to evaluate the viromic sequence-to-ecological-inference pipeline, including (*i*) read pre-processing and metagenome assembly, (*ii*) thresholds applied to estimate viral relative abundances based on read mapping to assembled contigs, and (*iii*) normalization methods applied to the matrix of viral relative abundances for alpha and beta diversity estimates. **Results.** Tools specifically designed for metagenomes, specifically metaSPAdes, MEGAHIT, and IDBA-UD, were the most effective at assembling viromes. Read pre-processing, such as partitioning, had virtually no impact on assembly output, but may be useful when hardware is limited. Viral populations with 2–5x coverage typically assembled well, whereas lesser coverage led to fragmented assembly. Strain heterogeneity within populations hampered assembly, especially when strains were closely related (average nucleotide identity, or ANI ≥ 97%) and when the most abundant strain represented < 50% of the population. Viral community composition assessments based on read recruitment were generally accurate when the following thresholds for detection were applied: (*i*) ≥ 10kb contig lengths to define populations, (*ii*) coverage defined from reads mapping at ≥ 90% identity, and (*ii*) ≥ 75% of contig length with ≥ 1x coverage. Finally, although data are limited to the most abundant viruses in a community, alpha and beta diversity patterns were robustly estimated (±10%) when comparing samples of similar sequencing depth, but more divergent (up to 80%) when sequencing depth was uneven across the dataset. In the latter cases, the use of normalization methods specifically developed for metagenomes provided the best

estimates. **Conclusions.** These simulations provide benchmarks for selecting analysis cut-offs and establish that an optimized sample-to-ecological-inference viromics pipeline is robust for making ecological inferences from natural viral communities. Continued development to better accessing RNA, rare, and/or diverse viral populations and improved reference viral genome availability will alleviate many of viromics remaining limitations.

1   **Benchmarking viromics: An *in silico* evaluation of metagenome-enabled estimates of viral**
2   **community composition and diversity**

4   Simon Roux[1,†,*], Joanne B. Emerson[1], Emiley A. Eloe-Fadrosh[2], Matthew B. Sullivan[1,3,*]
5
6   [1] Department of Microbiology, The Ohio State University, Columbus, OH, USA
7   [2] United States Department of Energy Joint Genome Institute, Lawrence Berkeley National
8   Laboratory, Walnut Creek, CA, USA
9   [3] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University,
10   Columbus OH, USA
11
12   [†] Current address: United States Department of Energy Joint Genome Institute, Lawrence
13   Berkeley National Laboratory, Walnut Creek, CA, USA
14   [*] to whom correspondence should be addressed

17   **Abstract**

18   **Background**. Viral metagenomics (viromics) is increasingly used to obtain uncultivated viral
19   genomes, evaluate community diversity, and assess ecological hypotheses. While viromic
20   experimental methods are relatively mature and widely accepted by the research community,
21   robust bioinformatics standards remain to be established. Here we used *in silico* mock viral
22   communities to evaluate the viromic sequence-to-ecological-inference pipeline, including (*i*)
23   read pre-processing and metagenome assembly, (*ii*) thresholds applied to estimate viral relative
24   abundances based on read mapping to assembled contigs, and (*iii*) normalization methods
25   applied to the matrix of viral relative abundances for alpha and beta diversity estimates.
26   **Results.** Tools specifically designed for metagenomes, specifically metaSPAdes, MEGAHIT,
27   and IDBA-UD, were the most effective at assembling viromes. Read pre-processing, such as
28   partitioning, had virtually no impact on assembly output, but may be useful when hardware is
29   limited. Viral populations with 2–5x coverage typically assembled well, whereas lesser coverage
30   led to fragmented assembly. Strain heterogeneity within populations hampered assembly,
31   especially when strains were closely related (average nucleotide identity, or ANI ≥ 97%) and
32   when the most abundant strain represented < 50% of the population. Viral community
33   composition assessments based on read recruitment were generally accurate when the following
34   thresholds for detection were applied: (*i*) ≥ 10kb contig lengths to define populations, (*ii*)
35   coverage defined from reads mapping at ≥ 90% identity, and (*iii*) ≥ 75% of contig length with ≥
36   1x coverage. Finally, although data are limited to the most abundant viruses in a community,
37   alpha and beta diversity patterns were robustly estimated (±10%) when comparing samples of
38   similar sequencing depth, but more divergent (up to 80%) when sequencing depth was uneven
39   across the dataset. In the latter cases, the use of normalization methods specifically developed for
40   metagenomes provided the best estimates.
41   **Conclusions.** These simulations provide benchmarks for selecting analysis cut-offs and establish
42   that an optimized sample-to-ecological-inference viromics pipeline is robust for making
43   ecological inferences from natural viral communities. Continued development to better accessing
44   RNA, rare, and/or diverse viral populations and improved reference viral genome availability
45   will alleviate many of viromics remaining limitations.

Background

46
47   Microbial communities and their associated viruses are abundant, diverse, and play key roles
48   in Earth's ecosystems and processes (Falkowski, Fenchel & Delong, 2008; Cobián Güemes et
49   al., 2016). However, because most microbes and viruses remain uncultivated, and because
50   viruses do not harbor a universal marker gene, viral ecology studies remain challenging to
51   perform (Brum & Sullivan, 2015; Solden, Lloyd & Wrighton, 2016). Viral metagenomics
52   (viromics) is a uniquely powerful tool for high-throughput analysis of uncultivated viruses
53   (Brum & Sullivan, 2015; Cobián Güemes et al., 2016). Initial viromics studies, despite being
54   limited to gene-level analyses, revealed the large diversity of viral-encoded genes (Edwards &
55   Rohwer, 2005; Schoenfeld et al., 2008), provided first estimates of richness and functional
56   diversity across natural viral communities (Hurwitz, Hallam & Sullivan, 2013; Hurwitz, Brum &
57   Sullivan, 2015), and suggested the existence of biome-specific viral communities distributed
58   worldwide (Rodriguez-Brito et al., 2010; Roux et al., 2012).

59   Thanks to recent improvements in high-throughput sequencing technologies and genome
60   assembly, viromes now also provide the opportunity to assemble large genomes fragments (and
61   even complete genomes) of uncultivated viruses (reviewed in Brum and Sullivan, 2015; Rose et
62   al., 2016). Historically, *in silico* benchmarks of the assembly process for microbial metagenomes
63   indicated that accurate bacterial and archaeal genomes (complete or partial) could be recovered
64   for relatively abundant lineages given sufficient sequencing depth, but revealed potential issues
65   including misassemblies deriving from the presence of very closely related organisms
66   (Mavromatis et al., 2007; Mende et al., 2012; Greenwald et al., 2017; Sczyrba et al., 2017). Viral
67   community datasets are typically processed using the same methodologies, and viral-specific
68   benchmarks came to a similar conclusion: viral genomes can be assembled from metagenomes,
69   but the presence of co-existing viruses with highly similar regions in their genome can lead to
70   reduced contig size and/or chimeric contigs (Aguirre de Cárcer, Angly & Alcamí, 2014;
71   Vázquez-Castellanos et al., 2014; García-López, Vázquez-Castellanos & Moya, 2015; Martinez-
72   Hernandez et al., 2017; White, Wang & Hall, 2017). However, new metagenome assembly
73   softwares (e.g. metaSPAdes (Nurk et al., 2017)) and methods for read filtering and/or
74   partitioning prior to assembly (e.g. khmer (Crusoe et al., 2015)) that might improve assembly
75   quality have yet to be evaluated with viral data.

76   For bacteria and archaea, advances in genome binning and genome validation approaches (e.g.,
77   (Parks et al., 2015)) have significantly improved the recovery of accurately reconstructed
78   genomes from increasingly complex environments (Wrighton et al., 2012; Sharon et al., 2013;
79   Waldor et al., 2015; Sangwan, Xia & Gilbert, 2016; Sczyrba et al., 2017). These methods rely on
80   single-copy marker genes to assess genome bin completeness and "contamination" (i.e. multiple
81   genomes in the same genome bin), two metrics critical to guide the optimization of genome
82   binning parameters and curate the final dataset (Parks et al., 2015; Bowers et al., 2017).
83   Unfortunately, because of the absence of universal single-copy viral marker gene, viral genome
84   bins are much more challenging to interpret and analyze. Since viral genomes are also smaller
85   than microbial ones and thus more frequently assembled in a single contig, viromics studies
86   usually rely on the assembled contigs without applying any genome binning step.

87   For ecological analyses, a community abundance matrix of microbial OTU counts across
88   samples is the typical starting point, and this "OTU table" is often derived from 16S rRNA gene
89   abundances in amplicon sequencing datasets or metagenomes (Hill et al., 2003; Roesch et al.,
90   2007; Fulthorpe et al., 2008; Fierer et al., 2011; Logares et al., 2014). Even for these relatively
91   established microbial ecological analyses, appropriate normalization methods that account for

92   different sequencing throughput across samples are still debated, and rarely are results compared
93   across multiple normalization methods to establish best practices (Doll et al., 2013; Paulson et
94   al., 2013; McMurdie & Holmes, 2014). This microbial ecology pipeline also needs adjustment
95   when applied to viruses because viruses lack a universal marker gene, precluding amplicon-
96   based viral population abundance estimates at the community scale (although amplicon-based
97   studies have been successful for ecological analyses of specific viral lineages, e.g., (Filée et al.,
98   2005; Goldsmith et al., 2011; Chow & Fuhrman, 2012)). Notably, comparative genomic and
99   ecological analysis of model systems enabled the identification of sequence-discrete populations,
100 which represent stable ecotypes in natural viral communities (Marston & Amrich, 2009; Gregory
101 et al., 2016; Marston & Martiny, 2016). Thus, in the absence of a universal viral marker gene,
102 these genome-based populations have been proposed to be used as a viral population units (akin
103 to a microbial operational taxonomic unit, OTU) in ecological analysis (Brum et al., 2015).
104 Pragmatically, viral populations are derived from *de novo* metagenomic assemblies, with
105 abundances estimated by metagenomic read recruitment. Ecological analyses of these contig-
106 derived abundance matrices still have to be comprehensively evaluated, although one bias
107 specific to this approach has already been identified: counting each assembled contig as a
108 separate OTU can lead to over-estimates of the number of different viruses in the community
109 (Aziz et al., 2015; García-López, Vázquez-Castellanos & Moya, 2015).
110   Here we used 14 *in silico* simulated viral metagenomes to (*i*) compare the assembly results
111 across different reads pre-processing methods and assemblers, both in terms of the overall
112 genomes recovery and the number and type of errors observed, (*ii*) assess potential biases and
113 identify optimal thresholds for identification and quantification of viral populations from
114 metagenomic contigs, and (*iii*) determine if virome populations abundance matrices can provide
115 reliable estimates of alpha diversity (i.e. diversity within a community) and beta diversity (i.e.
116 differentiation between communities), even in cases where sequencing depth vary widely (up to
117 two orders of magnitude) between samples.
118
119 **Methods**
120
121   **Mock community design**
122   Viral genomes were randomly selected among the complete genomes of viruses infecting
123 bacteria or archaea in the NCBI RefSeq database (v69, 2015-02). For each mock community, the
124 total number of viruses randomly selected (between 500 and 1,000, Table S1, Fig. S1A), as well
125 as the parameter of the power law distribution used to model relative abundances (between 1 and
126 50) were varied (Fig. S1B-D). To create patterns of beta diversity across samples, the 50 most
127 abundant viruses were homogenized within each of four sample groups, *i.e.*, samples within a
128 group shared 30 to 50 of their most abundant viruses, and samples between groups did not share
129 any of their most abundant 50 viruses. This led to a clear beta diversity pattern with the mock
130 communities clustering into four groups (Fig. S1E & F, a PerMANOVA was performed in R
131 with the package vegan (Oksanen et al., 2008) to verify that the sample groups were significantly
132 different).
133
134   **Virome simulations**
135   To simulate virome sequencing for each mock community, the number of reads derived from
136 each genome was first calculated based on the relative abundance of the genome in the mock
137 community and the total number of reads sequenced in the virome (10 millions paired-end reads

138 in the initial viromes, 1 million and 100,000 paired-end reads for the subsets at 10% and 1%
139 respectively). Then, NeSSM (Jia et al., 2013) was used to generate random reads (2 x 100bp) at
140 the prescribed abundances with simulated Illumina HiSeq errors.
141
142 **Reads processing**
143 Reads generated by NeSSM were first quality-controlled with Trimmomatic (Bolger, Lohse &
144 Usadel, 2014) with a minimum base quality threshold of 30 evaluated on sliding windows of 4
145 bases, and minimum read length of 50. We opted not to evaluate different error correction
146 softwares or to compare raw reads to quality-controlled (QC) reads, as previous studies have
147 already provided such benchmarks for genomic assembly, which should be applicable to
148 metagenomic assembly as well (e.g., (Yang, Chockalingam & Aluru, 2013)).
149 All sets of additionally pre-processed reads were generated from these QC reads using khmer
150 v1.4.1 (Crusoe et al., 2015), following the online protocols (http://khmer-
151 protocols.readthedocs.io/, Fig. S2). First, a dataset of digitally normalized reads was generated,
152 i.e. a dataset in which all reads with median k-mer abundance higher than a specified threshold
153 were eliminated. This was done in two steps by normalizing k-mer coverage first to 20x then to
154 5x (script "normalize-by-median", dataset "Digital normalization"). The script "do-partition"
155 was then used to partition these digitally normalized datasets, i.e. separate reads that did not
156 connect to each other in the k-mer graphs in different bins (dataset "Partitioned reads
157 (normalized)"). These reads partitions were then re-inflated, i.e., the original abundance of reads
158 was restored to its value prior to digital normalization, with the script "sweep-reads" (dataset
159 "Partitioned reads (inflated)"). Finally, three sets of reads were generated by trimming all low-
160 abundance k-mers for highly covered reads, i.e., highly covered reads (in this case, $\geq$ 20x) were
161 truncated at the first occurrence of a k-mer below a given abundance cutoff (here $\leq$ 2x, $\leq$ 5x, and
162 $\leq$ 20x for the three datasets "Low k-mer filter (2x)", "Low k-mer filter (5x)", "Low k-mer filter
163 (20x)", respectively). This was done with the script "filter-abund", with option "variable-
164 coverage" as recommended for metagenomes.
165
166 **Assembly and comparison to input genomes**
167 The different read sets were assembled with five different assembly software tools, using
168 metagenomic-optimized parameters (when available, Fig. S2). IDBA-UD v.1.1.1 (Peng et al.,
169 2012) was used with the option "pre-correction" and from fasta reads (converted from fastq reads
170 with the tool "fq2fa"). MetaSPAdes assemblies (Nurk et al., 2017) were computed from the
171 software version 3.10.0, with the option "metagenomic" (all other options default). MEGAHIT
172 assemblies (Li et al., 2016) were computed from version v1.0.6 with presets "meta" (all other
173 options default). MetaVelvet assemblies (Namiki et al., 2012) were computed with software
174 version 1.2.07 with the "discard_chimera" option selected, default parameters otherwise. Omega
175 assemblies (Haider et al., 2014) were computed with software version 1.0.2 and minimum
176 overlap length of 60. Each assembler was applied to each read pool from each sample (7 read
177 pools x 14 samples = 98 assemblies, Fig. S2), retaining all contigs $\geq$ 500 bp for each assembly
178 (Table S4).
179 Contigs were compared to the input genomes with nucmer (Delcher, Salzberg & Phillippy,
180 2003)(default options). When $\geq$95% of a contig's length matched an input genome at $\geq$ 90%
181 nucleotide identity, that contig was considered to be a genuine assembly of the input genome.
182 Otherwise, if a contig was similar to multiple genomes but to none over $\geq$ 95% of its length, it
183 was considered a chimera. Circular contigs were detected based on identical 5' and 3' ends, as in

184   (Roux et al., 2014). A circular contig with a length corresponding to ≥ 95% of the original
185   genome length was considered a genuine complete genome assembly, while circular contigs
186   covering less than 95% of the original genomes were considered false positives (i.e., incomplete
187   contigs incorrectly predicted as complete genome assemblies). R was used to conduct t-test when
188   comparing rate of chimeric contigs across assemblers and reads pre-processing methods, using
189   the assembly of QC reads with MEGAHIT as the control (the set of contigs with the lower
190   number of chimeras).
191
192   **Generation of the non-redundant pool of population contigs and coverage estimation**
193   Based on the previous benchmarks, the assemblies obtained with metaSPAdes from the QC
194   reads were considered to be the most optimal assemblies and were used in all subsequent
195   benchmarking analyses. Contigs from all samples were clustered with nucmer (Delcher, Salzberg
196   & Phillippy, 2003) at ≥ 95% ANI across ≥ 80% of their lengths, as in (Brum et al., 2015;
197   Gregory et al., 2016), to generate a pool of non-redundant "population contigs". QC reads from
198   each sample were then mapped to these population contigs with bbmap (http://bit.ly/bbMap),
199   with ambiguous mapping assigned to contigs at random (option ambiguous=random). A custom
200   python script was then used to estimate the number of reads and coverage of each contig.
201
202   **Alpha and beta diversity estimates**
203   The abundance of each population contig in a given sample was estimated based on the
204   number of reads mapping to that contig, normalized by the contig length (to account for
205   differences in contig / genome size). Beyond the raw read counts (normalized by contig length),
206   five abundance matrices were generated with different library size normalization methods as
207   follow (summarized in Fig. S2):

- "Normalized": counts were divided by the total library size, i.e., the total number of QC reads in the sample, as used for example in (Brum et al., 2015). This approach is also known as "total-sum scaling".
- "MGSeq": counts were normalized through cumulative-sum scaling with the metagenomeSeq R package (Paulson et al., 2013). This method was specifically designed for metagenomes in which communities are under-sampled (as is the case in most viral metagenome studies), and will divide counts by a cumulative sum of count to a given percentile (as opposed to dividing by total counts as in "Normalized"). This will minimize the effects of the few highly abundant viruses potentially dominating the community, and introducing biases in relative abundances (Paulson et al., 2013).
- "EdgeR": counts were normalized using scaling factors for libraries designed to minimize the log-fold change between samples for most of the populations, computed with the edgeR R package (Robinson, McCarthy & Smyth, 2009). This method was initially developed for count-based expression data and assumes that the relative abundances of most features (here populations) will not vary between two samples.
- "DeSeq": as with EdgeR, counts were normalized to minimize variations between samples for most populations but with a different underlying model, computed with the DESeq R package (Anders & Huber, 2010). As with EdgeR, this method was initially developed for the detection of differentially expressed features in sequence count data analysis.
- "Rarefied": new counts were generated based on rarefied sets of reads, i.e. quality-controlled reads are subsampled (without replacement) to the smallest number of quality-

230   controlled reads across all samples. Thus, all of the libraries are artificially set to the
231   same size, however some data are "wasted" in the process, i.e., for the more deeply
232   sequenced samples, some observations will not be included in the rarefied counts
233   (McMurdie & Holmes, 2014).
234 Each abundance matrix was then used to calculate alpha and beta diversity indices, namely the
235 Shannon index, Simpson index, and pairwise Bray-Curtis dissimilarities between samples with a
236 custom perl script. R was used to generate all plots using the ggplot2 package (Wickham, 2009),
237 as well as the NMDS and PerMANOVA analyses, computed with the vegan package (Oksanen
238 et al., 2008). For alpha diversity, we opted to only test indices reflecting community structure
239 (Shannon and Simpson indexes) and not indices predicting sample richness (e.g. Chao estimators
240 (Chao, 1984)), since the latter have been highlighted as not suitable for cases in which rare
241 members of the community are not adequately sampled (Haegeman et al., 2013).
242

243 **Under-sequencing and strain heterogeneity benchmarks**
244 To evaluate the impact of under-sequencing on alpha and beta diversity estimates, the same
245 pipeline (assembly with metaSPAdes from QC Reads, selection of population contigs, and
246 estimation of alpha and beta diversity) was applied to datasets in which seven of the 14 samples
247 were under-sequenced. Two levels of under-sequencing were tested, one in which under-
248 sequenced samples were set at 10% of the initial library size (i.e. 1,000,000 reads) and another at
249 1% of the initial library size (100,000 reads, Table S1).
250 To evaluate the impact of strain heterogeneity (within-population genomic diversity) on
251 assembly success, a custom perl script was used to simulate strain variations as observed on
252 natural populations of T4-like cyanophages (Gregory et al., 2016), i.e. a set of potentially
253 mutated positions were determined for each new simulated strain gathering all intergenic
254 positions, all third codons positions in protein-coding genes, and all positions in two randomly
255 selected genes (to simulate genes undergoing diversifying selections). These simulations were
256 based on the mock community "Sample_1", for which every genome was transformed into a
257 population composed of a set of related strains.
258 For each population, three parameters selected randomly and independently:
259 • The total number of strains was set at 10, 50, or 100 strains simulated
260 • The strain divergence, controlled by a "mutation rate", i.e. the ratios of positions mutated
261   within the set of positions identified as "potentially mutated" (see above). The other
262   positions in the genome, not selected as potentially mutated, were mutated at a rate 100
263   times lower. This "mutation rate" was set at 5%, 10%, or 20%. This led to ANI between
264   the generated strains and the original reference genomes of 97-100%, 95-97%, and 90-
265   95%, respectively.
266 • The relative abundance of individual strains within the population, sampled from a
267   power-law distribution. The shape of the distribution was controlled by the power-law
268   parameter, set at 0.1, 1, 10, 100, or 1000. This led to the dominant (i.e. most abundant)
269   strain representing from 1% to 100% of the population.
270 For each population, reads were then simulated with NeSSM (Jia et al., 2013), with the total
271 reads generated for each population calculated based on the input coverage (as for previous
272 simulations), and the number of reads generated from each strain calculated from the strains
273 relative abundance. Reads were then processed as previously, i.e. quality-controlled, partitioned,
274 or filtered, and assembled with the 5 assemblers tested using the same options as for the
275 simulated viromes. Finally, the size of the largest contig recovered for each population was

276 compared to the size of the largest contig recovered for the same genome without strain
277 heterogeneity, to evaluate the impact of strain heterogeneity independently from differences in
278 assembly efficiency between coverage levels, reads processing methods, and assemblers.
279
280 **Results & Discussion**
281 *Mock communities design*
282 A set of 14 viral communities was designed to provide a gradient of alpha diversity and clear
283 beta diversity patterns (Fig. S1, Table S1 & S2). These communities were composed of 500 to
284 1,000 genomes (randomly sampled within bacteriophages and archaeal viruses available in NCBI
285 RefSeq v69), with the relative abundance of individual genomes based on power law
286 distributions with varying exponents. These simulations are thus designed to reflect a diverse
287 viral community, as is usually observed in environmental samples (e.g., oceans, lakes, soils, or
288 human gut), but would not correspond to viral communities dominated by a single type of virus,
289 e.g., clinical samples associated with a specific host or epidemiological samples targeting a
290 specific type of virus. Beyond differences in alpha diversity, these communities were also
291 designed to organize into four "ecological" clusters, i.e., four groups of mock communities
292 sharing more genomes within than between groups (Fig. S1). Thus, this simulated dataset allows
293 us to evaluate the ability of virome-based population ecology approaches to recover absolute
294 values of alpha diversity, as well as trends in alpha diversity and beta diversity patterns across
295 samples.
296 Virome reads were simulated *in silico* with NeSSM (Jia et al., 2013) for each mock community
297 (10,000,000 paired-end Illumina HiSeq reads, 2 x 100bp). Since the number of reads derived
298 from each genome was based on its prescribed relative abundance in the community, 29.1% to
299 75.2% of the viral genomes in each mock community did not get "sequenced" at all (*i.e.*, did not
300 yield any reads). This was by design to mimic the lack of sampling for rare viruses by current
301 sequencing efforts of environmental samples.
302
303 *Testing the capacity and accuracy of assembly tools*
304 Given metagenomic sequence data from these 14 mock communities, we first evaluated
305 currently available assembly algorithms. To this end, five assemblers (IDBA-UD (Peng et al.,
306 2012), MEGAHIT (Li et al., 2016), MetaVelvet (Namiki et al., 2012), Omega (Haider et al.,
307 2014), and metaSPAdes (Nurk et al., 2017), all adapted to assemble metagenomic data) were
308 compared to assess their ability to accurately assemble genomes of bacterial and archaeal viruses
309 from viromes (Fig. S2). As expected, each of the assemblers successfully assembled highly
310 covered genomes (10x or higher) and failed to assemble most low-coverage genomes (2x and
311 lower, Fig. 1A, Fig. S3A). However, MetaVelvet and Omega required higher coverage to
312 assemble viral genomes (~5-10x), while IDBA-UD, MEGAHIT, and metaSPAdes routinely
313 assembled genomes at ~2-5x coverage (Fig. 1A, Fig. S3A). A similar trend was found when
314 observing genome recovery in a single contig (i.e., the percentage of a genome assembled in a
315 single contig, as opposed to the percentage of a genome assembled when cumulating all contigs).
316 Again, IDBA-UD, MEGAHIT, and metaSPAdes were more efficient than MetaVelvet and
317 Omega for assembling large genome fragments at lower read coverage (~2-20x), and
318 metaSPAdes was also better than IDBA-UD and MEGAHIT for assembling low-coverage
319 genomes in a single large contig (Fig. 1B, Fig. S3B).
320 When comparing individual genome assemblies across the three best assemblers (metaSPAdes,
321 IDBA-UD, and MEGAHIT), no clear differences could be observed in the genome recovery

322   (Fig. S4, correlation coefficients between assemblers > 0.99). However, the percentage of each
323   genome recovered in a single contig was more variable among assemblers (Fig. S4, correlations
324   coefficients: 0.88-0.98). This comparison did not indicate that one assembler would be
325   systematically better than another, but rather that the best assembly for a given genome could
326   come from any of these three assemblers.
327   Together these comparisons suggest that: (*i*) IDBA-UD, MEGAHIT, and metaSPAdes are
328   currently the best available choices for maximizing assembly of viral contigs from short-read
329   (100 bp) viromes (assembly accuracy discussed below), (*ii*) regardless of the choice of assembly
330   tool, low coverage genomes (< 2x) are under-assembled, and (*iii*) because assembly success
331   varies across genomes and assemblers, multiple tools should be compared to optimally assemble
332   desired target genomes from viromes. Overall, these results are consistent with microbial
333   metagenomic benchmarks, which also indicated that assemblers designed specifically for
334   metagenomes, especially metaSPAdes, MEGAHIT, and IDBA-UD, provided the best assemblies
335   (Sczyrba et al., 2017; Vollmers, Wiegand & Kaster, 2017).
336
337   *Impact of k-mer-based read filtering and partitioning on assembly*
338   Next, we evaluated how available read pre-processing approaches impacted genome assembly
339   (using approaches from the khmer package and summarized in Table S3 and Fig. S2) (Crusoe et
340   al., 2015). Briefly, beyond the reference dataset of quality controlled reads, the different methods
341   tested were (*i*) trimming of reads based on low-abundance k-mers, i.e., reads are truncated at the
342   first occurrence of a low-abundance k-mer likely originating from sequencing error, (*ii*) digital
343   normalization, i.e., the removal of redundant sequences to normalize genome coverage at or
344   under a specific value (here 5x), and (*iii*) read partitioning, i.e., separate assembly of the
345   disconnected components of the k-mer graph.
346   Overall, and compared with the effect of the different assembly algorithms, the read pre-
347   processing had a minimal impact on the assembly output (Fig. 1C & D, Fig. S3 C & D with
348   metaSPAdes; the same observations were made with different assemblers in Fig. S5). The main
349   effects observed were that (*i*) digital normalization (treatments "Digital normalization" and
350   "Partitioned reads (normalized)") led to sub-optimal assemblies, likely because differences in
351   coverage above 5x are useful for assemblers to distinguish between related genomes, and (*ii*)
352   trimming of low-abundance k-mers led to sub-optimal assemblies when the threshold used to
353   define low abundance k-mers was close to the threshold used to define "abundant" reads to be
354   trimmed (effect especially noticeable for the 20x filter, Fig. 1C & D). Conversely, partitioning
355   reads and keeping their coverage information (treatment "Partitioned reads (inflated)") or
356   trimming low-abundance k-mers from high coverage reads (with thresholds of 2x and 5x) had
357   little effect on the assembly output, except on low-coverage genomes (< 5x). These observations
358   are consistent with the initial expectations of khmer's performance (Crusoe et al., 2015),
359   although these simulations illustrate that digital normalization alone (i.e., without read
360   partitioning and restoration of original read coverage) can lead to a sub-optimal metagenomic
361   assembly.
362
363
364   *Errors and limitations of genome assembly from viromes*
365   Beyond the assembly of low-coverage genomes, which was found to be challenging for all
366   assemblers tested, other errors are known to occur during the *de novo* assembly of viromes.

367   First, chimeric contigs (i.e., contigs representing artificial constructs assembled from two or
368   more distinct genomes) were generated in each assembly, as previously noted (Aguirre de
369   Cárcer, Angly & Alcamí, 2014; Vázquez-Castellanos et al., 2014; García-López, Vázquez-
370   Castellanos & Moya, 2015). In our simulated data, these usually represented less than 2.5% of
371   the assembled datasets, and less than 5% of the large contigs (≥ 10kb), but these numbers varied
372   between assemblers and read curation methods (Fig. 2A&B). This low number of chimeric
373   contigs is in accordance with benchmarks of microbial metagenomes, and suggests that
374   metagenome assemblers in general can correctly reconstruct microbial and/or viral genomes
375   (Mende et al., 2012). For all assemblers, reads after digital normalization always yielded more
376   chimeric contigs, which confirmed that the digital normalization step led to sub-optimal
377   assemblies (p-value < 0.01). MEGAHIT systematically produced fewer chimeric contigs than
378   IDBA-UD and metaSPAdes, especially for large (≥ 10kb) contigs (Fig. 2B, p-value < 0.01).
379   Hence, although MEGAHIT did not assemble as many large genome fragments, the fragments
380   that were assembled contained fewer chimeras.
381   Next, we investigated whether finished and closed viral genomes assemblies could be robustly
382   identified as "circular" contigs, i.e., contigs with matching 5' and 3' ends, as previously
383   suggested (Roux et al., 2014). The ratio of false-positive circular contigs, i.e. circular contigs that
384   represented less than 95% of the original genome and thus likely arose from repeat regions
385   within a genome, was not modified by read pre-processing but was different among assemblers
386   (Fig. 2C). Specifically, 10 to 30% of the circular contigs generated by MEGAHIT and IDBA-UD
387   did not correspond to a complete genome, while metaSPAdes assemblies rarely included any
388   false positive (4 contigs, or <2%, for metaSPAdes assemblies of quality-controlled reads). This
389   suggests that metaSPAdes circular contigs are more likely to correspond to complete genomes
390   and that the "circularization" of a contig cannot be considered as proof of completeness for
391   MEGAHIT and IDBA-UD contigs.
392   Finally, we evaluated the impact of population strain heterogeneity – i.e., the co-existence of
393   closely related strains with distinct genomes from the same population – on virome assembly. In
394   microbial communities, strain heterogeneity is known to considerably hamper the assembly of
395   the corresponding genomes (Sharon et al., 2015; Martinez-Hernandez et al., 2017; Sczyrba et al.,
396   2017). Population genetic studies of natural viral communities are however challenged by the
397   paucity of cultivated systems that include multiple viral genomic representatives from a single
398   population. Pragmatically, this means that although strain heterogeneity has been observed for
399   specific model systems (Gregory et al., 2016; Marston & Martiny, 2016), community-wide strain
400   variations that would accurately reflect natural viral communities cannot be pulled from these
401   data. Hence, we opted to generate a mock community using the same populations and relative
402   abundances as Sample 1 above, but introduced some level of strain heterogeneity for each
403   population by varying a combination of 3 parameters: (*i*) the number of strains in the population,
404   either low (n=10), medium (n=50), or high (n=100), (*ii*) the diversity of these strains, presented
405   as the average ANI of strains compared to the consensus population genome, either low (90-
406   95%), medium (95-97%), or high (97-100%), and (*iii*) the evenness of the power-law distribution
407   of strain frequency in the population, either low (dominant variant represents 75-100% of the
408   population), medium (dominant variant 50-75%), or high (dominant variant < 25%). For each
409   genome, reads were thus not generated from the reference genome sequence as before, but from
410   a set of strains generated and sampled using a random combination of these 3 parameters. Then,
411   the same pipeline of read processing and assembly was applied, and the size of the largest contig

412   obtained for each population was compared to the size of the largest contig obtained in the
413   previous mock community assembly (i.e. without strain heterogeneity, Fig. 2D and Fig. S6).
414   An ANOVA was performed on the complete dataset (i.e. all combinations of assemblers and
415   read processing) to evaluate which component of strain heterogeneity impacted the assembly
416   process (see Methods). The three parameters (number of strains, strain diversity, and evenness of
417   strain distribution) significantly but differently impacted the assembly: population shape (i.e.,
418   strain distribution) was the main explanatory variable of suboptimal assemblies (F-value 149.8,
419   p-value<1e-16), strain diversity was also a strong driver of assembly failures (F-value 70.4, p-
420   value<1e-16), while the number of strains in the populations had a more marginal effect (F-value
421   2.8, p-value 0.06). Overall, when compared to the assemblies generated without strain
422   heterogeneity, contigs were shorter for populations with an even strain distribution (i.e. dominant
423   strain ≤ 50% of the population) and/or when strains were more similar to the consensus genome
424   (i.e. average ANI to consensus ≥ 97%) and to each other, with the combination of both leading to
425   the greater reduction in contig length (Fig. 2D). These results indicate that strain heterogeneity
426   within natural viral populations will likely be a key factor contributing to assembly success and
427   failure, and populations of evenly distributed closely related strains will be the most likely to fail
428   to assemble in virome studies. A similar trend was observed for microbial genomes in the
429   Critical Assessment of Metagenome Interpretation benchmarks, where the assembly of closely
430   related genomes (i.e. those with strain-level heterogeneity) was found to be challenging for all
431   assemblers tested, although the experimental design did not allow the evaluation of which level
432   and parameter of strain heterogeneity were most impactful (Sczyrba et al., 2017).
433
434   *Population identification and quantification*
435   In viral ecological studies, the next step after assembly often consists of identifying viral
436   populations (i.e. contigs representing individual populations) and quantifying their relative
437   abundances in each sample. We opted to use the contigs assembled with metaSPAdes from
438   quality-controlled reads, as they represented the largest contigs overall across the different
439   samples (despite ~1% chimerism). We pooled contigs generated from all samples into a single
440   non-redundant database (contigs were clustered at ≥ 95% of nucleotide identity across ≥ 80% of
441   the contig length, in accordance with population genome analysis (Gregory et al., 2016)).
442   Quality-controlled reads were then mapped to this database to estimate contig coverage across
443   the 14 samples. Two types of thresholds were evaluated in this mapping step: (*i*) minimum
444   nucleotide identity for a given read to be considered mapped to a given contig, and (*ii*) minimum
445   length of the contig covered to consider a contig as "detected" in a sample (Fig. S2). Reads not
446   meeting the threshold were removed from abundance counts, and contigs not meeting the
447   detection threshold in a given sample were given abundance values of zero for that sample in the
448   resulting coverage table.
449   Considering all non-redundant contigs ≥ 500 bp as different populations, we observed that
450   increasing the two thresholds (read mapping identity percentage and length of contig covered)
451   progressively decreased the sensitivity of the analysis (evaluated here as the percentage of
452   genomes recovered among genomes which were covered ≥ 1x in the sample, Fig. 3A) and the
453   false discovery rate (or FDR, which is the percentage of contigs recovered that were not part of
454   the initial community, i.e. these genomes did not provide any reads to the simulated
455   metagenome, Fig. 3B). However, because FDR decreased more precipitously than sensitivity,
456   there is an optimal combination of thresholds for which FDR can be minimized and sensitivity
457   maximized. In these simulations, that optimal threshold was ≥ 75% on the contig length coverage

458 associated with ≥ 90% nucleotide identity for the read mapping, which led to a 3% decrease in
459 sensitivity (compared to the most permissive thresholds), but only 13% FDR (compared to 49%
460 for the most permissive thresholds).

461 As noted by previous studies (Aziz et al., 2015; García-López, Vázquez-Castellanos & Moya,
462 2015), considering all non-redundant contigs as distinct populations strongly over-estimated the
463 total number of populations (on average, 2 to 3 contigs were counted for each individual
464 genome, Fig. 3C). Thus, we re-analyzed our dataset using only non-redundant contigs ≥ 10kb or
465 circular as was proposed previously, and as required for taxonomic classification by gene content
466 network-based analysis (Bolduc et al., 2017). Again, the optimal threshold combination was ≥
467 75% of the contig length covered and ≥ 90% read mapping identity (Fig. 3 panels D,E, & F).
468 However, while sensitivity declined slightly (~15%) compared to the dataset including all
469 contigs ≥ 500bp, FDR improved drastically to 0.2%, compared to 13% observed in the above
470 analyses. Further, by increasing the stringency of the population definition, the number of
471 contigs per genome that were counted as a population was 1.2 which is much closer to the
472 correct number of 1 contig per genomne. More generally, increasing this contig size threshold
473 quickly decreased the number of contig observed per genome, and most of the over-estimation
474 observed earlier seemed to arise from contigs < 5kb (Fig. S7).

475 In summary, we recommend that viral populations (as an operational taxonomic unit) be
476 defined and analyzed in viromes using contigs that are ≥ 10 kb or circular, and only considered
477 "detected" when the contig is covered over ≥ 75% of its length by read mapping at ≥ 90%
478 nucleotide identity. However, we also anticipate that the data from these sensitivity analyses will
479 help researchers tune these thresholds to match a given study's need for high sensitivity or low
480 FDR. Importantly though, these suggestions are specific to viromes, since microbial
481 metagenomic studies can rely on genome binning and universally conserved, single-copy marker
482 genes to estimate more robustly the global number and completeness of the different genomes
483 assembled (Sczyrba et al., 2017).

484

485 *Alpha and beta diversity estimation from virome-derived populations*
486 We next sought to evaluate how the variation in community structure of our 14 mock
487 community metagenomes impacted diversity estimations, and did so using our recommended
488 optimized population cut-offs for identifying populations and then estimating their abundances
489 by read mapping. These population count matrices (counting either base pairs or reads mapped to
490 each population contig) were used as input for alpha and beta diversity estimations and compared
491 across the dataset. Notably, these matrices included only a fraction (10-33%) of the original
492 genomes in the dataset, as rare viral genomes were not "sequenced", and low-coverage genomes
493 produced only small (< 10 kb) contigs (Fig. 4A).

494 Before calculating any index, the read counts were first normalized by the contig length, since
495 viral genome lengths can be highly variable (~2 orders of magnitude, (Angly et al., 2009)). Then,
496 to account for potential differences in library sizes, we compared five different methods: (*i*) a
497 simple normalization in which counts are divided by the library size, "Normalized" (*ii*) a method
498 specifically designed to account for under-sampling of metagenomes, from the metagenomeSeq
499 R package, "MGSeq" (*iii and iv*) two methods designed to minimize log-fold changes between
500 samples for most of the populations, from the edgeR R package, "edgeR", and the DESeq R
501 package, "DESeq", and (*v*) a rarefaction approach whereby all libraries get randomly down-
502 sampled without replacement to the size of the smallest library, "Rarefied" (Fig. S2).

503   For both Shannon and Simpson alpha diversity indices, the values calculated from normalized
504   count matrices were within 10% of the actual value calculated from the whole community (Fig.
505   4B & C). Hence, the recovery of abundant members of the community seems to be enough to
506   estimate alpha diversity values. Since both Shannon and Simpson indices are based on the
507   relative abundance of individual members of the community, the three methods that applied a
508   sample-wide correction factor (normalization by library size, MGSeq, EdgeR) all led to the same
509   estimations, while rarefied count matrices and DESeq, which can (slightly) modify relative
510   abundance of populations within communities, provided statistically indistinguishable estimates
511   (Fig. 4 B & C). Similarly, for beta diversity estimates, pairwise Bray-Curtis dissimilarities
512   between samples calculated from normalized counts matrices were highly similar to the
513   dissimilarities calculated from the whole communities for all normalization methods (within
514   15% of actual values, p-value≤0.001 for Mantel test comparing true and estimated dissimilarity
515   matrices, Fig. 4D). Thus, as long as the count matrices were normalized to account for different
516   contig lengths and library sizes, each of the five methods tested here provided reliable estimates
517   of alpha and beta diversity.
518
519   *Impact of under-sequencing and possible corrections*
520   Finally, to help guide researchers in making decisions about under-sequenced samples, we
521   evaluated how alpha and beta diversity estimates were impacted by such samples in a dataset.
522   Specifically, we performed the same computations (assembly with metaSPAdes from quality-
523   controlled reads, generation of a pool of dereplicated population contigs, mapping of quality-
524   controlled reads and computation of normalized count matrices), but we did so with a dataset in
525   which half of the samples were drastically under-sequenced either at 10% (subset_10) or 1%
526   (subset_1) of the original sequencing depth, respectively (Table S1, Fig. S2).
527   Not surprisingly, under-sequenced samples resulted in fewer genomes detected (t-test, p-value
528   < 1e-05, Fig. 4A). Using the same five normalization methods to account for these differences in
529   sequencing depth, we found that the diversity estimations were impacted. The subset_10 samples
530   resulted in Shannon and Simpson estimations that were close (within 16%) to the initial
531   estimates, but the diversity estimates in the subset_1 samples varied as much as 30% (Fig. 4 B &
532   C). Hence, although the different normalization methods tested here helped to compensate for
533   some degree of under-sequencing, none was able to recover the correct values of alpha diversity
534   when sequencing depth was highly variable and/or when some samples were significantly under-
535   sequenced.
536   Similarly, beta diversity patterns (evaluated as pairwise Bray Curtis dissimilarities) were not
537   estimated as accurately with the under-sequenced samples than with the initial samples:
538   dissimilarities estimated from subset_10 samples varied as much as 61% compared with the true
539   dissimilarities (mean: 5.9%), and the ones estimated from subset_1 samples varied as much as
540   77% (mean: 4.4%; Fig. 4 E & F). Rarefaction and MGSeq were the two normalization methods
541   most efficient at limiting these biases, as they led to maximum variations of 11.5% and 11.3%
542   for subset_10, and 10.9% and 52.7% for subset_1, respectively. Moreover, even with the
543   subset_1 samples, the results of an NMDS based on these normalized count matrices were still
544   strongly correlated with the results of an NMDS based on true relative abundances (Fig. S8,
545    $r^2>0.9$ for all normalization methods but "rarefied", for which the positions of two groups are
546   switched leading to a lower r^2 of 0.64). Hence, beta diversity trends can be recovered even when
547   sequencing depth was highly variable.

548   Although not formally evaluated through *in silico* benchmarks, it is very likely that microbial
549 metagenomes with highly uneven sequencing depth would be subjected to similar biases, and the
550 tools tested here would be expected to perform comparably on viral and microbial metagenomes,
551 since the input data (i.e., coverage matrix) is essentially identical. Hence, the information and
552 guidelines provided here can in all likelihood be considered relevant for microbial metagenomes
553 as well.
554
555   *Current limitations of the sample-to-ecological-inference pipeline*
556   Overall, these benchmarks confirmed that virome-derived abundance matrices can be used in
557 ecological studies, with two main caveats. First, absolute viral richness will likely be under-
558 estimated, because the assembly will only yield large contigs for abundant viral genotypes
559 without evenly distributed and/or closely related strains. Hence, absolute values of richness and
560 diversity should be interpreted with care, although once normalized, sample comparisons of
561 these richness and diversity metrics are generally robust to differences in community complexity
562 and sequencing depth. Second, because this approach relies on coverage as a proxy for relative
563 abundance, only quantitative (or near-quantitative) datasets can be used as input (Duhaime et al.,
564 2012). Notably, protocols to generate these quantitative viromes are currently available only for
565 dsDNA and/or ssDNA viruses (Duhaime et al., 2012; Roux et al., 2016), and still remain to be
566 developed for their RNA counterparts, although these RNA viruses might represent up to half of
567 the viral particles in some environments (Steward et al., 2013). Thus, when interpreting
568 viromics-based ecological studies, it is important to remember and clearly state that these reflect
569 only the sub-part of viral communities with (ds)DNA genomes.
570

Conclusions

571
572   Our comparative analysis of 14 simulated viromes showed that the genome-assembly-to-
573 ecological-inference viromics pipeline can efficiently and robustly identify abundant viruses and
574 recover trends in alpha and beta diversity. As viromics becomes routine in viral ecology, the
575 approaches underlined here (both the tools and thresholds used) offer an initial set of "best
576 practices" for data analysis.
577   Moving forward, increased library size and number associated with improved genome
578 recovery from metagenomes will undoubtedly lead to an unprecedented catalog of uncultivated
579 viral genomes (e.g. 125,000 released in a single study (Paez-Espino et al., 2016)). These will be
580 complemented by viral genomes obtained from other methods, such as single-virus sequencing,
581 which can access less dominant viruses and those with high strain heterogeneity (Martinez-
582 Hernandez et al., 2017). As standards emerge, such uncultivated viral genomes will migrate
583 toward specifically-designed databases (e.g. IMG/VR (Paez-Espino et al., 2016)), and viral
584 ecological studies will be greatly improved by these centralized reference genome data. Beyond
585 improved references (which will also need to include uncultivated RNA viruses), viromics will
586 need to advance from relative abundance estimations to absolute quantification of viral
587 populations, likely coupled with "ground-truthing" provided by quantitative, lineage-specific
588 molecular methods such as phageFISH, polonies, microarrays, or microfluidic PCR (Tadmor et
589 al., 2011; Allers et al., 2013; Martínez-García et al., 2014). Once in-hand, such approaches
590 should enable researchers to address long-standing questions in the viral ecology field, and more
591 fully bring viruses into predictive ecological models across Earth's ecosystems.

592 **List of abbreviations**
593 ANI: Average Nucleotide Identity
594 ANOVA: ANalysis Of Variance
595 FDR: False Discovery Rate
596 NMDS: Non-metric MultiDimensional Scaling
597 OTU: Operational Taxonomic Unit
598 QC: Quality-controlled (for reads)
599
600 **Declaration**
601
602 *Availability of data and material*
603 The datasets generated for this study are available at
604 http://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/Virome_pipeline_benchmark
605 The scripts used in this study are available at
606 https://bitbucket.org/MAVERICLab/benchmarking_viromics
607

608 *Acknowledgements*
609 High performance computing resources were provided by the Ohio Supercomputer Center, and
610 the National Energy Research Scientific Computing Center supported by the Office of Science of
611 the US Department of Energy.

612
613

614   **References**

615   Aguirre de Cárcer D., Angly FE., Alcamí A. 2014. Evaluation of viral genome assembly and
616   diversity estimation in deep metagenomes. *BMC Genomics 2014 15:1* 15:e368. DOI:
617   10.1186/1471-2164-15-989.

618   Allers E., Moraru C., Duhaime MB., Beneze E., Solonenko N., Canosa JB., Amann R., Sullivan
619   MB. 2013. Single-cell and population level viral infection dynamics revealed by
620   phageFISH, a method to visualize intracellular and free viruses. *Environmental*
621   *Microbiology* 15:2306–18. DOI: 10.1111/1462-2920.12100.

622   Anders S., Huber W. 2010. Differential expression analysis for sequence count data. *Genome*
623   *Biology* 11:R106. DOI: 10.1186/gb-2010-11-10-r106.

624   Angly FE., Willner D., Prieto-Davó A., Edwards RA., Schmieder R., Vega-Thurber R.,
625   Antonopoulos DA., Barott K., Cottrell MT., Desnues C., Dinsdale EA., Furlan M., Haynes
626   M., Henn MR., Hu Y., Kirchman DL., McDole T., McPherson JD., Meyer F., Miller RM.,
627   Mundt E., Naviaux RK., Rodriguez-Mueller B., Stevens R., Wegley L., Zhang L., Zhu B.,
628   Rohwer F. 2009. The GAAS metagenomic tool and its estimations of viral and microbial
629   average genome size in four major biomes. *PLoS computational biology* 5:e1000593. DOI:
630   10.1371/journal.pcbi.1000593.

631   Aziz RK., Dwivedi B., Akhter S., Breitbart M., Edwards RA. 2015. Multidimensional metrics
632   for estimating phage abundance , distribution , gene density , and sequence coverage in
633   metagenomes. *Frontiers in Microbiology* 6. DOI: 10.3389/fmicb.2015.00381.

634   Bolduc B., Jang H Bin., Doulcier G., You Z-Q., Roux S., Sullivan MB. 2017. vConTACT: an
635   iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ*
636   5:e3243. DOI: 10.7717/peerj.3243.

637   Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence
638   data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.

639   Bowers RM., Kyrpides NC., Stepanauskas R., Harmon-Smith M., Schulz F., Doud D., Reddy
640   TBK., Jarett J., Rivers AR., Eloe-Fadrosh EA., Tringe SG., Ivanova N., Copeland A., Clum
641   A., Becraft ED., Malmstrom RR., Birren B., Schriml L., Podar M., Bork P., Weinstock
642   GM., Banfield JF., Garrity GM., Hugenholtz P., Parks DH., Tyson GW., Rinke C.,
643   Dodsworth J a., Yooseph S., Sutton G., Yilmaz P., Glockner FO., Meyer F., Gilbert JA.,
644   Nelson WC., Hallam SJ., Jungbluth SP., Ettema TJG., Tighe S., Konstantinidis KT., Liu W-
645   T., Baker BJ., Rattei T., Eisen JA., Hedlund BP., Mcmahon KD., Fierer N., Knight R., Finn
646   RD., Karsch-Mizrachi I., Eren A., Woyke T. 2017. Minimum information about a single
647   amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria
648   and archaea. *Nature Biotechnology* in press.

649 Brum J., Ignacio-Espinoza J., Roux S., Doulcier G., Acinas SG., Alberti A., Chaffron S., Cruaud
650   C., de Vargas C., Gasol JM., Gorsky G., Gregory AC., Ogata H., Pesant S., Poulos BT.,
651   Schwenck SM., Speich S., Dimier C., Kandels-Lewis S., Picheral M., Searson S.,
652   Coordinators TO., Bork P., Bowler C., Sunagawa S., Wincker P., Karsenti E., Sullivan MB.
653   2015. Patterns and ecological drivers of ocean viral communities. *Science* 348:1261498-1–
654   10. DOI: 10.1126/science.1261498.

655 Brum JR., Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms
656   marine virology. *Nature reviews. Microbiology* 13:1–13. DOI: 10.1038/nrmicro3404.

657 Chao A. 1984. Nonparametric Estimation of the Number of Classes in a Population.
658   *Scandinavian Journal of Statistics* 11:265–70.

659 Chow CET., Fuhrman JA. 2012. Seasonality and monthly dynamics of marine myovirus
660   communities. *Environmental Microbiology* 14:2171–2183. DOI: 10.1111/j.1462-
661   2920.2012.02744.x.

662 Cobián Güemes AG., Youle M., Cantú VA., Felts B., Nulton J., Rohwer F. 2016. Viruses as
663   Winners in the Game of Life. *Annual Review of Virology* 3:197–214. DOI:
664   10.1146/annurev-virology-100114-054952.

665 Crusoe MR., Alameldin HF., Awad S., Boucher E., Caldwell A., Cartwright R., Charbonneau A.,
666   Constantinides B., Edvenson G., Fay S., Fenton J., Fenzl T., Fish J., Garcia-Gutierrez L.,
667   Garland P., Gluck J., González I., Guermond S., Guo J., Gupta A., Herr JR., Howe A., Hyer
668   A., Härpfer A., Irber L., Kidd R., Lin D., Lippi J., Mansour T., McA'Nulty P., McDonald
669   E., Mizzi J., Murray KD., Nahum JR., Nanlohy K., Nederbragt AJ., Ortiz-Zuazaga H., Ory
670   J., Pell J., Pepe-Ranney C., Russ ZN., Schwarz E., Scott C., Seaman J., Sievert S., Simpson
671   J., Skennerton CT., Spencer J., Srinivasan R., Standage D., Stapleton JA., Steinman SR.,
672   Stein J., Taylor B., Trimble W., Wiencko HL., Wright M., Wyss B., Zhang Q., Zyme E.,
673   Brown CT. 2015. The khmer software package: enabling efficient nucleotide sequence
674   analysis. *F1000Research* 4:900. DOI: 10.12688/f1000research.6924.1.

675 Delcher AL., Salzberg SL., Phillippy AM. 2003. Using MUMmer to identify similar regions in
676   large sequence sets. *Current protocols in bioinformatics* 00:10.3:10.3.1–10.3.18. DOI:
677   10.1002/0471250953.bi1003s00.

678 Doll HM., Armitage DW., Daly R a., Emerson JB., Goltsman DSA., Yelton AP., Kerekes J.,
679   Firestone MK., Potts MD. 2013. Utilizing novel diversity estimators to quantify multiple
680   dimensions of microbial biodiversity across domains. *BMC microbiology* 13:259. DOI:
681   10.1186/1471-2180-13-259.

682 Duhaime MB., Deng L., Poulos BT., Sullivan MB. 2012. Towards quantitative metagenomics of
683   wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and

684   optimization of the linker amplification method. *Environmental microbiology* 14:2526–37.
685   DOI: 10.1111/j.1462-2920.2012.02791.x.

686   Edwards RA., Rohwer F. 2005. Viral metagenomics. *Nature Reviews Microbiology* 3:504–510.

687   Falkowski PG., Fenchel T., Delong EF. 2008. The Microbial Engines That Drive Earth's
688   Biogeochemical Cycles. *Science* 320:1034–9. DOI: 10.1126/science.1153213.

689   Fierer N., McCain C., Meir P., Zimmerman M., Rapp JM., SIlman MR., Knight R. 2011.
690   Microbes do not follow the elevational diversity patterns of plants and animals. *Ecology*
691   92:2013–2019.

692   Filée J., Tétart F., Suttle CA., Krisch HM. 2005. Marine T4-type bacteriophages, a ubiquitous
693   component of the dark matter of the biosphere. *Proceedings of the National Academy of*
694   *Sciences of the United States of America* 102:12471–6. DOI: 10.1073/pnas.0503404102.

695   Fulthorpe RR., Roesch LFW., Riva A., Triplett EW. 2008. Distantly sampled soils carry few
696   species in common. *Isme Journal* 2:901–910. DOI: 10.1038/ismej.2008.55.

697   García-López R., Vázquez-Castellanos JF., Moya A. 2015. Fragmentation and Coverage
698   Variation in Viral Metagenome Assemblies, and Their Effect in Diversity Calculations.
699   *Frontiers in bioengineering and biotechnology* 3:141. DOI: 10.3389/fbioe.2015.00141.

700   Goldsmith DB., Crosti G., Dwivedi B., McDaniel LD., Varsani A., Suttle CA., Weinbauer MG.,
701   Sandaa R-AA., Breitbart M. 2011. Development of phoH as a novel signature gene for
702   assessing marine phage diversity. *Applied and environmental microbiology* 77:7730–9.
703   DOI: 10.1128/AEM.05531-11.

704   Greenwald WW., Klitgord N., Seguritan V., Yooseph S., Venter JC., Garner C., Nelson KE., Li
705   W. 2017. Utilization of defined microbial communities enables effective evaluation of
706   meta-genomic assemblies. *BMC Genomics* 18:296. DOI: 10.1186/s12864-017-3679-5.

707   Gregory AC., Solonenko SA., Ignacio-Espinoza JC., LaButti K., Copeland A., Sudek S.,
708   Maitland A., Chittick L., dos Santos F., Weitz JS., Worden AZ., Woyke T., Sullivan MB.
709   2016. Genomic differentiation among wild cyanophages despite widespread horizontal gene
710   transfer. *BMC Genomics* 17:930. DOI: 10.1186/s12864-016-3286-x.

711   Haegeman B., Hamelin J., Moriarty J., Neal P., Dushoff J., Weitz JS. 2013. Robust estimation of
712   microbial diversity in theory and in practice. *The ISME journal* 7:1092–101. DOI:
713   10.1038/ismej.2013.10.

714   Haider B., Ahn TH., Bushnell B., Chai J., Copeland A., Pan C. 2014. Omega: An Overlap-graph
715   de novo Assembler for Metagenomics. *Bioinformatics* 30:2717–2722. DOI:
716   10.1093/bioinformatics/btu395.

717 Hill TCJ., Walsh K a., Harris J a., Moffett BF. 2003. Using ecological diversity measures with
718   bacterial communities. *FEMS Microbiol Ecol* 43:1–11. DOI: FEM1 [pii]\n10.1111/j.1574-
719   6941.2003.tb01040.x.

720 Hurwitz BL., Brum JR., Sullivan MB. 2015. Depth-stratified functional and taxonomic niche
721   specialization in the "core" and "flexible" Pacific Ocean Virome. *The ISME Journal* 9:472–
722   84. DOI: 10.1038/ismej.2014.143.

723 Hurwitz BL., Hallam SJ., Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit
724   and dark ocean. *Genome biology* 14:R123. DOI: 10.1186/gb-2013-14-11-r123.

725 Jia B., Xuan L., Cai K., Hu Z., Ma L., Wei C. 2013. NeSSM: a Next-generation Sequencing
726   Simulator for Metagenomics. *PloS one* 8:e75448. DOI: 10.1371/journal.pone.0075448.

727 Li D., Luo R., Liu CM., Leung CM., Ting HF., Sadakane K., Yamashita H., Lam TW. 2016.
728   MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced
729   methodologies and community practices. *Methods* 102:3–11. DOI:
730   10.1016/j.ymeth.2016.02.020.

731 Logares R., Sunagawa S., Salazar G., Cornejo-Castillo FM., Ferrera I., Sarmento H., Hingamp
732   P., Ogata H., de Vargas C., Lima-Mendez G., Raes J., Poulain J., Jaillon O., Wincker P.,
733   Kandels-Lewis S., Karsenti E., Bork P., Acinas SG. 2014. Metagenomic 16S rDNA
734   Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and
735   structure of microbial communities. *Environmental Microbiology* 16:2659–2671. DOI:
736   10.1111/1462-2920.12250.

737 Marston MF., Amrich CG. 2009. Recombination and microdiversity in coastal marine
738   cyanophages. *Environmental microbiology* 11:2893–903. DOI: 10.1111/j.1462-
739   2920.2009.02037.x.

740 Marston MF., Martiny JBH. 2016. Genomic diversification of marine cyanophages into stable
741   ecotypes. *Environmental Microbiology* 18:4240–4253. DOI: 10.1111/1462-2920.13556.

742 Martínez-García M., Santos F., Moreno-Paz M., Parro V., Antón J. 2014. Unveiling viral–host
743   interactions within the "microbial dark matter." *Nature Communications* 5:1–8. DOI:
744   10.1038/ncomms5542.

745 Martinez-Hernandez F., Fornas O., Lluesma M., Bolduc B., Cruz M., Martinez Martinez J.,
746   Anton J., Gasol J., Rosselli R., Rodriguez-Valera R., Sullivan MB., Acinas SG., Martinez-
747   Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses.
748   *Nature communications* 8:15892. DOI: 10.1038/ncomms15892.

749 Mavromatis K., Ivanova N., Barry K., Shapiro H. 2007. Use of simulated data sets to evaluate
750   the fidelity of metagenomic processing methods. *Nature methods* 4:495–500. DOI:
751   10.1038/NMETH1043.

752   McMurdie PJ., Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is
753   Inadmissible. *PLoS Computational Biology* 10. DOI: 10.1371/journal.pcbi.1003531.

754   Mende DR., Waller AS., Sunagawa S., Järvelin AI., Chan MM., Arumugam M., Raes J., Bork P.
755   2012. Assessment of metagenomic assembly using simulated next generation sequencing
756   data. *PLoS ONE* 7. DOI: 10.1371/journal.pone.0031386.

757   Namiki T., Hachiya T., Tanaka H., Sakakibara Y. 2012. MetaVelvet: an extension of Velvet
758   assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids*
759   *research* 40:e155. DOI: 10.1093/nar/gks678.

760   Nurk S., Meleshko D., Korobeynikov A., Pevzner PA. 2017. metaSPAdes: a new versatile
761   metagenomic assembler. *Genome Research*. DOI: 10.1101/gr.213959.116.

762   Oksanen J., Kindt R., Legendre P., O'Hara B., Simpson GL., Solymos P., Stevens MHH.,
763   Wagner H. 2008. *The vegan Package*.

764   Paez-Espino D., Chen I-MA., Palaniappan K., Ratner A., Chu K., Szeto E., Pillay M., Huang J.,
765   Markowitz VM., Nielsen T., Huntemann M., K Reddy TB., Pavlopoulos GA., Sullivan
766   MB., Campbell BJ., Chen F., McMahon K., Hallam SJ., Denef V., Cavicchioli R., Caffrey
767   SM., Streit WR., Webster J., Handley KM., Salekdeh GH., Tsesmetzis N., Setubal JC.,
768   Pope PB., Liu W-T., Rivers AR., Ivanova NN., Kyrpides NC. 2016. IMG/VR: a database of
769   cultured and uncultured DNA Viruses and retroviruses. *Nucleic acids research*
770   45:gkw1030. DOI: 10.1093/nar/gkw1030.

771   Parks DH., Imelfort M., Skennerton CT., Hugenholtz P., Tyson GW. 2015. CheckM: assessing
772   the quality of microbial genomes recovered from. *Genome Research* 25:1043–55. DOI:
773   10.1101/gr.186072.114.

774   Paulson JN., Stine OC., Bravo HC., Pop M. 2013. Differential abundance analysis for microbial
775   marker-gene surveys. *Nature methods* 10:1200–2. DOI: 10.1038/nmeth.2658.

776   Peng Y., Leung HCM., Yiu SM., Chin FYL. 2012. IDBA-UD: a de novo assembler for single-
777   cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–
778   1428.

779   Robinson MD., McCarthy DJ., Smyth GK. 2009. edgeR: A Bioconductor package for
780   differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.
781   DOI: 10.1093/bioinformatics/btp616.

782   Rodriguez-Brito B., Li L., Wegley L., Furlan M., Angly F., Breitbart M., Buchanan J., Desnues
783   C., Dinsdale E., Edwards R., Felts B., Haynes M., Liu H., Lipson D., Mahaffy J., Martin-
784   Cuadrado AB., Mira A., Nulton J., Pašić L., Rayhawk S., Rodriguez-mueller J., Rodriguez-
785   Valera F., Salamon P., Srinagesh S., Thingstad TF., Tran T., Thurber RV., Willner D.,

786   Youle M., Rohwer F. 2010. Viral and microbial community dynamics in four aquatic
787   environments. *The ISME journal* 4:739–51. DOI: 10.1038/ismej.2010.1.

788   Roesch L., Fulthorpe R., Riva A., Casella G., Hadwin A., Kent A., Daroub S., Camargo F.,
789   Farmerie W., Triplett E. 2007. Pyrosequencing enumerates and contrasts soil microbial
790   diversity. *The ISME Journal* 1:283–290. DOI: 10.1038/ismej.2007.53.Pyrosequencing.

791   Rose R., Constantinides B., Tapinos A., Robertson DL., Prosperi M. 2016. Challenges in the
792   analysis of viral metagenomes. *Virus Evolution* 2:vew022. DOI: 10.1093/ve/vew022.

793   Roux S., Enault F., Robin A., Ravet V., Personnic S., Theil S., Colombet J., Sime-Ngando T.,
794   Debroas D. 2012. Assessing the Diversity and Specificity of Two Freshwater Viral
795   Communities through Metagenomics. *PLoS One* 7:e33641.

796   Roux S., Tournayre J., Mahul A., Debroas D., Enault F. 2014. Metavir 2: new tools for viral
797   metagenome comparison and assembled virome analysis. *BMC Bioinformatics* 15:1–12.

798   Roux S., Solonenko NE., Dang VT., Poulos BT., Schwenck SM., Goldsmith DB., Coleman ML.,
799   Breitbart M., Sullivan MB. 2016. Towards quantitative viromics for both double-stranded
800   and single-stranded DNA viruses. *PeerJ* 4:e2777. DOI: 10.7717/peerj.2777.

801   Sangwan N., Xia F., Gilbert JA. 2016. Recovering complete and draft population genomes from
802   metagenome datasets. *Microbiome* 4:8. DOI: 10.1186/s40168-016-0154-5.

803   Schoenfeld T., Patterson M., Richardson PM., Wommack KE., Young M., Mead D. 2008.
804   Assembly of viral metagenomes from yellowstone hot springs. *Applied and environmental*
805   *microbiology* 74:4164–4174.

806   Sczyrba A., Hofmann P., Belmann P., Koslicki D., Dröge J., Gregor I., Majda S., Fiedler J.,
807   Dahms E., Bremges A., Fritz A., Garrido-oter R., Jørgensen S., Shapiro N., Blood PD.,
808   Gurevich A., Hansen LH., Sørensen SJ., Chia BKH., Denis B., Froula JL., Wang Z., Egan
809   R., Kang DD., Singer W., Jain C., Strous M., Klingenberg H., Meinicke P., Barton M.,
810   Lingner T., Lin H., Liao Y., Silva GZ., Cuevas DA., Edwards RA., Saha S., Vitor C.,
811   Renard BY., Hill CM., Pop M., Goeker M., Kyrpides N., Woyke T., Vorholt J., Rubin EM.,
812   Darling AE., Rattei T., Alice C. 2017. Critical Assessment of Metagenome Interpretation –
813   a comprehensive benchmark of computational metagenomics software. *bioRxiv*. DOI:
814   10.1101/099127.

815   Sharon I., Morowitz MJ., Thomas BC., Costello EK., Relman DA., Banfield JF. 2013. Time
816   series community genomics analysis reveals rapid shifts in bacterial species, strains, and
817   phage during infant gut colonization. *Genome Research* 23:111–120. DOI:
818   10.1101/gr.142315.112.

819   Sharon I., Kertesz M., Hug LA., Pushkarev D., Blauwkamp TA., Castelle CJ., Amirebrahimi M.,
820   Thomas BC., Burstein D., Tringe SG., Williams KH., Banfield JF. 2015. Accurate, multi-kb

821   reads resolve complex populations and detect rare microorganisms. *Genome Research*
822   25:534–543. DOI: 10.1101/gr.183012.114.

823   Solden L., Lloyd K., Wrighton K. 2016. The bright side of microbial dark matter: Lessons
824   learned from the uncultivated majority. *Current Opinion in Microbiology* 31:217–226. DOI:
825   10.1016/j.mib.2016.04.020.

826   Steward GF., Culley AI., Mueller JA., Wood-Charlson EM., Belcaid M., Poisson G. 2013. Are
827   we missing half of the viruses in the ocean? *The ISME journal* 7:672–679. DOI:
828   10.1038/ismej.2012.121.

829   Tadmor AD., Ottesen EA., Leadbetter JR., Phillips R., Gallant JA., Cramer P., Weintraub H.,
830   Rich A., Maas S., Platas AA., Hawley DK., Yuzenkova Y., Severinov K., Yano T.,
831   Kawabata H., Ueda H., Suzuki T., Scott J., Anant S., Navaratnam N., Nicklen S., Coulson
832   AR., Hamilton CE., Mwangi MM., Dewell S., Papavasiliou FN., Gish W., Miller W., Myers
833   EW., Lipman DJ., Cox J., Mann M., Patt S., Schrey M., Rich A., Williams MA., Mulligan
834   RM. 2011. Probing Individual Environmental Bacteria for Viruses by Using Microfluidic
835   Digital PCR. *Science* 333:58–62. DOI: 10.1126/science.1200758.

836   Vázquez-Castellanos JF., García-López R., Pérez-Brocal V., Pignatelli M., Moya A. 2014.
837   Comparison of different assembly and annotation tools on analysis of simulated viral
838   metagenomic communities in the gut. *BMC genomics* 15:37. DOI: 10.1186/1471-2164-15-
839   37.

840   Vollmers J., Wiegand S., Kaster AK. 2017. Comparing and evaluating metagenome assembly
841   tools from a microbiologist's perspective - Not only size matters! *PLoS ONE* 12:1–31. DOI:
842   10.1371/journal.pone.0169662.

843   Waldor MK., Tyson G., Borenstein E., Ochman H., Moeller A., Finlay BB., Kong HH., Gordon
844   JI., Nelson KE., Dabbagh K., Smith H. 2015. Where Next for Microbiome Research? *PLoS*
845   *Biology* 13:1–9. DOI: 10.1371/journal.pbio.1002050.

846   White DJ., Wang J., Hall RJ. 2017. Assessing the Impact of Assemblers on Virus Detection in a
847   De Novo Metagenomic Analysis Pipeline. *Journal of Computational Biology*
848   24:cmb.2017.0008. DOI: 10.1089/cmb.2017.0008.

849   Wickham H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company.

850   Wrighton KC., Thomas BC., Sharon I., Miller CS., Castelle CJ., Verberkmoes NC., Wilkins MJ.,
851   Hettich RL., Lipton MS., Williams KH., Long PE., Banfield JF. 2012. Fermentation,
852   Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science*
853   337:1661–1666.

854   Yang X., Chockalingam SP., Aluru S. 2013. A survey of error-correction methods for next-
855   generation sequencing. *Briefings in Bioinformatics* 14:56–66. DOI: 10.1093/bib/bbs015.

856

Figure 1

Figure 1. Influence of assembly software and read curation on genome recovery.

All plots display the input coverage on the x-axis, and either the cumulated genome recovery across all contigs (A & C) or the highest genome recovery by a single contig (B & D) on the y-axis. Panels A & B display a comparison of assemblers applied to quality-controlled (QC) reads. Panels C & D present a comparison of read pre-processing methods, all assembled with metaSPAdes. Comparable plots for reads assembled with the other assemblers are available in Fig. S5.
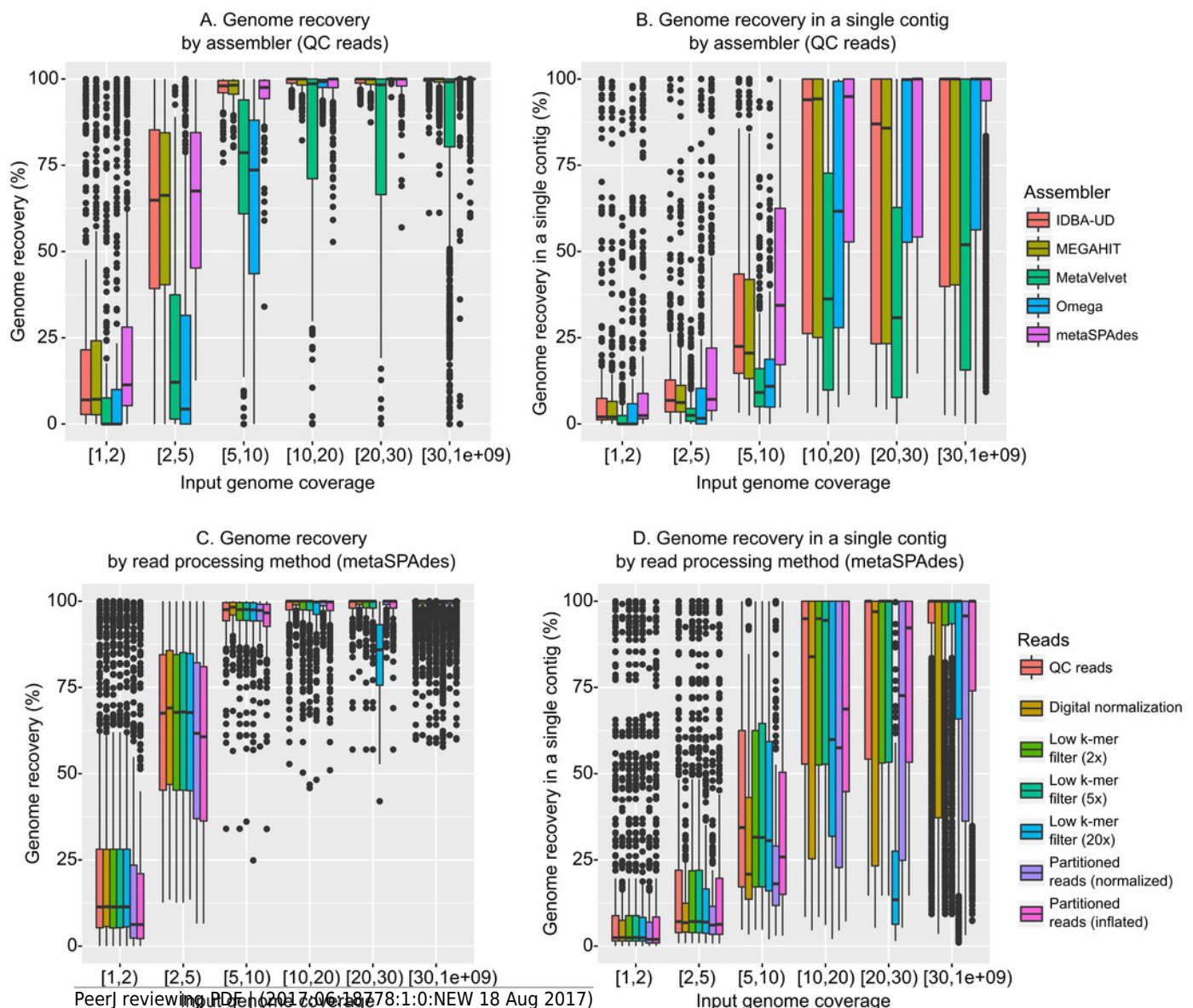
Figure 2

Figure 2. Types and frequency of errors observed in genome assembly from viral metagenomes.

A. Percentage of chimeric contigs (i.e. contigs originating from two distinct genomes) across all assembled sequences, by assembler (x-axis) and read curation method (colors). B. Percentage of chimeric contigs among large (≥10kb) contigs, by assembler (x-axis) and read curation method (colors). C. Percentage of false-positive circular contigs, i.e. contigs identified as circular (matching 5' and 3' ends) but representing 95% or less of the original genome, by assembler (x-axis) and read curation method (color). D. Impact of strain heterogeneity (i.e. presence of multiple strains from the same population) on the assembly efficiency. These tests were computed on one mock community (Sample_1), for which each reference genome was replaced with a set of related strains with varying divergence and relative abundances. The y-axis represents the ratio between the largest contig assembled for a genome when strain heterogeneity is introduced and the same parameter without strain heterogeneity (i.e. previous assemblies of the same Sample_1). Populations are grouped based on the two main parameters explaining assembly inefficiency: proportion of the most abundant strain in the population (bottom) and divergence of strains in the population (top). Data presented here include assemblies from QC reads with IDBA-UD, MEGAHIT, and metaSPAdes, while the full set of parameters and approaches tested are presented in Fig. S6.

A. Chimeric contigs ratio by assembler and read processing method

B. ≥10kb chimeric contigs ratio by assembler and read processing method

C. False positive circular contigs by assembler and read processing method

D. Impact of strain heterogeneity on assembly efficiency
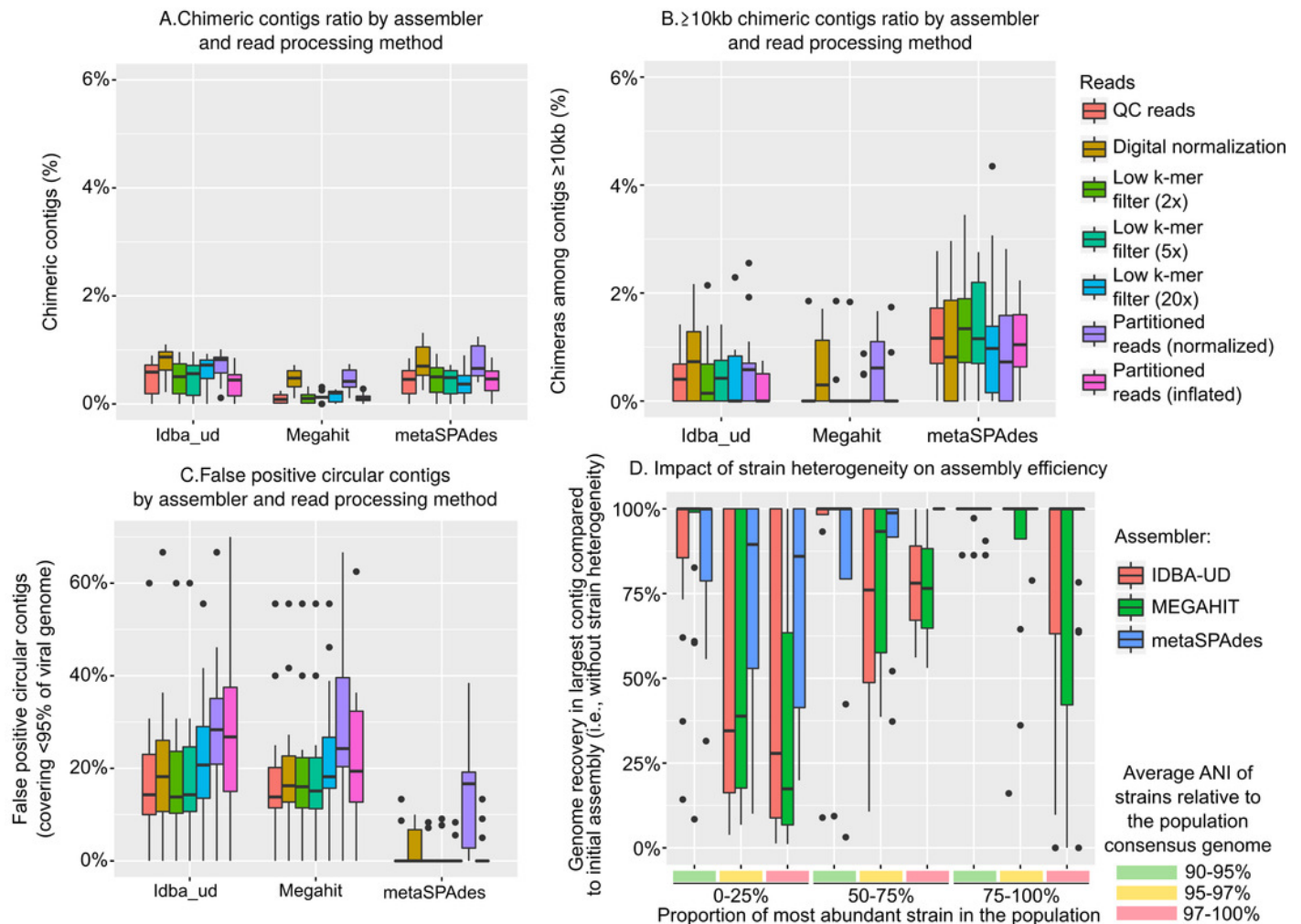
Figure 3

Figure 3. Impact of read mapping thresholds on accuracy of viral population detection.

Two parameters were investigated when parsing the mapping of individual virome reads to the population contigs pool: (*i*) the percentage of a contig covered by a sample to considered the contig as detected (x-axis), and (*ii*) the percentage of identity of reads mapping to the contig (color scale). Two pools of population contigs were tested: all non-redundant contigs of ≥ 500 bp (panels A, B, and C), and all non-redundant contigs ≥ 10 kb (panels D, E, and F). Three metrics were calculated to evaluate the impact of mapping reads thresholds. The detection sensitivity is estimated as the percentage of "expected" genomes (i.e., genomes covered ≥ 1x in the sample) that were detected through mapping to population contigs (panels A and D). The false-discovery rate corresponds to the percentage of contigs detected in a sample through mapping to population contigs, but were not associated with any genomes from the initial sample (i.e., these genomes did not provide any reads to the simulated virome, so these contigs should not be detected, panels B and E). Finally the average number of distinct population contigs detected is calculated for each individual genome initially covered ≥ 1x, and correspond to the number of times a single genome is "counted" (i.e. multiple contigs suggest multiple populations, even though it is really just one population, panels C and F).
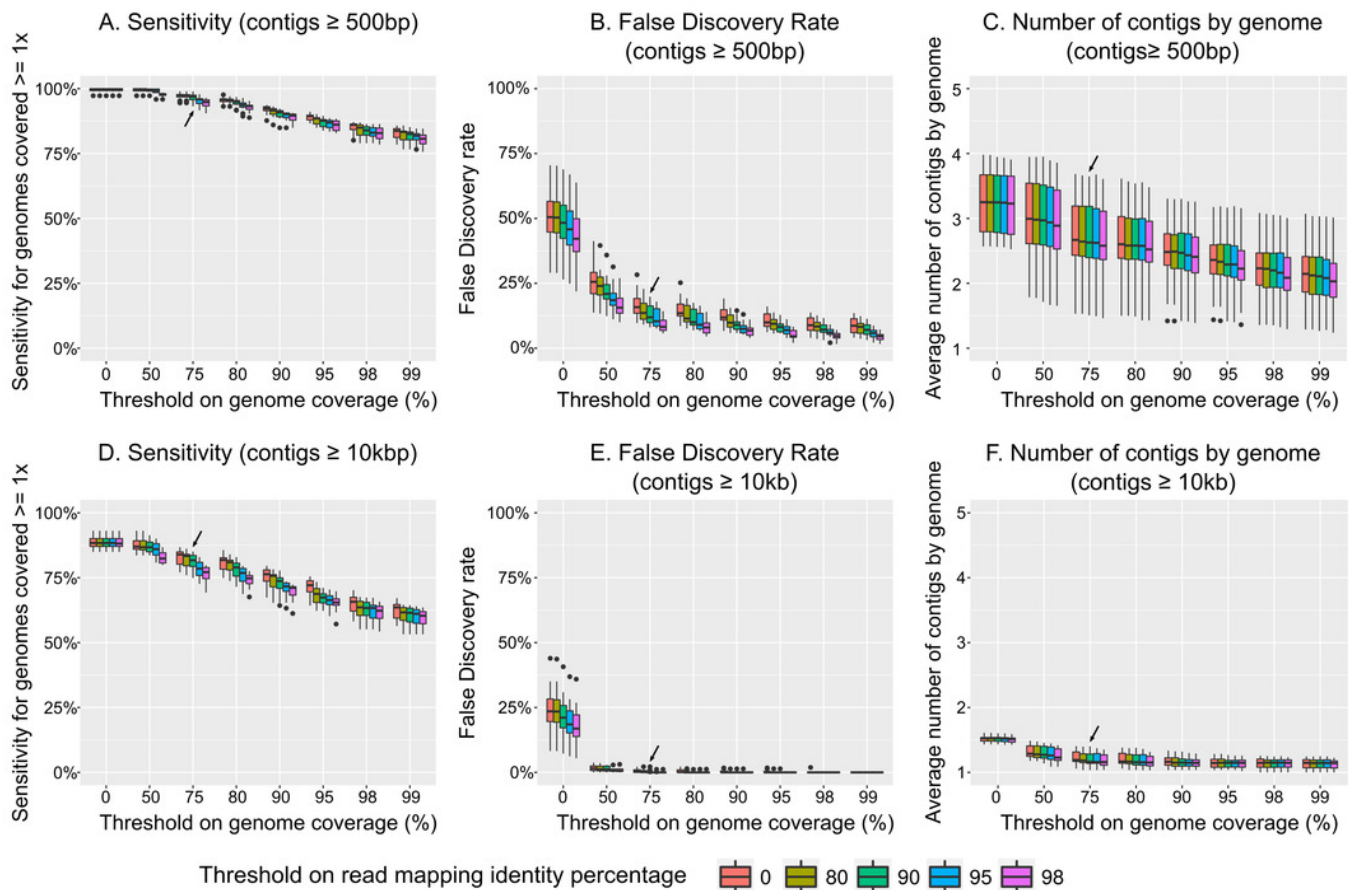
Figure 4

Figure 4. Estimation of alpha and beta diversity from virome-derived viral populations.

To evaluate the impact of varying sequencing depth, six viromes (highlighted in bold in panels A, B, and C), were sub-sampled at 10% (long dash) or 1% (short dash) of the original read number ("Initial" corresponds to the assemblies presented in Figure 1, 2, and 3, for which all viromes had the same initial number of reads). A. Number of genomes observed from the read mapping to viral populations. The actual number of genomes in the initial simulated community is indicated with black dots, while estimated based on viromes are colored in red. B. Comparison of Shannon diversity index from the true community composition (black dots) and estimated from the viromes (colored dots). The different estimations are based on 3 different normalization methods: counts divided by the total number of reads sequenced in the virome and the contig size ("Normalized"), counts after rarefying all viromes to the smallest dataset and normalized by contig size ("Rarefied"), and counts normalized via DESeq ("DESeq"). C. Comparison of Simpson diversity index from the true community composition and estimated from the viromes (color codes are the same as in panel B). D. Distribution of differences in Bray-Curtis dissimilarities between samples calculated from true community composition and the same dissimilarities estimated from the viromes analysis. The different normalization methods (x-axis) are as follows: counts divided by genome size ("Counts"), counts rarefied to the smallest dataset and normalized by contig size ("Rarefied"), counts divided by the total number of reads sequenced in the library and the contig size ("Normalized"), counts normalized by metagenomeSeq ("MGSeq"), EdgeR ("RPKM"), and DESeq ("DESeq"). E. Distribution of differences in Bray-Curtis dissimilarities between samples calculated from true community composition and the same dissimilarities estimated from virome analysis, including 6 samples sequenced at 10%. Methods are similar as in panel D. F. Distribution of differences in Bray-Curtis dissimilarities between samples calculated from true community composition and from virome analysis, including 6 samples

sequenced at 1%. Methods are similar as in panel D.