Fornas Oscar (Orcid ID: 0000-0003-2017-9100)
Martínez-García Manuel (Orcid ID: 0000-0001-5056-1525)

Manuscript submitted to Environmental Microbiology

Benchmarking of Single-Virus Genomics: a new tool for uncovering the virosphere

Inmaculada Garcia-Heredia[1], Ananda Shankar Bhattacharjee[4], Oscar Fornas[2,3], Monica Lluesma Gomez[1], Joaquín Martínez Martínez[4], and Manuel Martinez-Garcia[1*].

1. Department of Physiology, Genetics, and Microbiology, University of Alicante, Carretera San Vicente del Raspeig, San Vicente del Raspeig, Alicante 03690, Spain.
2. Flow Cytometry Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology (BIST), Barcelona, Spain.
3. Flow Cytometry Unit, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra (UPF), Barcelona, Spain.
4. Bigelow Laboratory for Ocean Sciences, 60 Bigelow Drive, PO Box 380, East Boothbay, Maine 04544, USA.

*Corresponding author: m.martinez@ua.es

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.15375

Originality-significance statement

Single-virus genomics (SVG) have recently enabled the discovery of widespread and abundant uncultured viruses in nature by sequencing one virus at a time directly collected from the environment. In this study we show a unique genomic dataset of single viral-amplified genomes obtained by different SVG procedures representing a step further to investigate, optimize and standardize novel technologies, which is critical in microbial ecology to expand the knowledge of the uncultured realm.

Summary

Metagenomics and single-cell genomics have enabled the discovery of relevant uncultured microbes. Recently, single-virus genomics (SVG), although still in an incipient stage, has opened new avenues in viral ecology by allowing the sequencing of one single virus at a time. The investigation of methodological alternatives and optimization of existing procedures for SVG is paramount to deliver high-quality genomic data. We report a sequencing dataset of viral single-amplified genomes (vSAGs) from cultured and uncultured viruses obtained by applying different conditions in each SVG step, from viral preservation and novel whole-genome amplification (WGA) to sequencing platforms and genome assembly. Sequencing data showed that cryopreservation and mild fixation were compatible with WGA, although fresh samples delivered better genome quality data. The novel TruPrime WGA, based on primase-polymerase features, and WGA-X employing a thermostable phi29 polymerase, were proven to be with sufficient sensitivity in SVG. The Oxford Nanopore (ON) sequencing platform did not provide a significant improvement of vSAG assembly compared to Illumina alone. Finally, the SPAdes assembler

performed the best. Overall, our results represent a valuable genomic dataset that will help to standardized and advance new tools in viral ecology.

Introduction

Viruses are important biological agents with global consequences beyond their impact as pathogens in individual host organisms (Suttle, 2007; Abedon, 2008). However, the tremendous morphological and genomic diversity of viruses (Rohwer, 2003; Edwards and Rohwer, 2005; Paez-Espino *et al.*, 2016) leads to challenges in the discovery and study of environmental viruses. A suite of complementary culture-independent methods, such as metagenomics and binning, have been developed in recent years to reveal the uncultured virosphere (Suttle, 2007; Mizuno *et al.*, 2013; Kang and Cho, 2014; Hugerth *et al.*, 2015; Coutinho *et al.*, 2017, 2019; López-Pérez *et al.*, 2017; Parks *et al.*, 2017). The *in silico* mining of microbial single amplified genomes (SAGs) obtained by single-cell genomics (SCGs) (Stepanauskas and Sieracki, 2007; Yoon *et al.*, 2011; Lasken, 2012; Martinez-Garcia *et al.*, 2012; Stepanauskas, 2012; López-Escardó *et al.*, 2017; Mangot *et al.*, 2017; Tara Oceans Coordinators *et al.*, 2018; Sieracki *et al.*, 2019) is another culture-independent method that has allowed the discovery of new viruses and the investigation of host-virus interactions (Yoon *et al.*, 2011; Dhillon and Li, 2015; Labonté *et al.*, 2015). More recently, single-virus genomics (SVG) (Allen *et al.*, 2011), although still in an incipient stage, has arisen as a complementary approach to investigate the uncultured viriosphere by recovering and sequencing one virus at a time. The methodological steps of SVG are as follows: 1) flow cytometry sorting of fluorescently stained viruses, 2) capsid lysis, 3) whole-genome amplification (WGA) of viral genetic material and 4) DNA sequencing. To date, a handful of SVG-based studies of marine and human-related environments have demonstrated the power of this method to elucidate viral communities (Allen *et al.*, 2011; Martinez-Hernandez *et al.*, 2017;

Stepanauskas *et al.*, 2017; Wilson *et al.*, 2017; de la Cruz Peña *et al.*, 2018; Martinez-Hernandez, Fornas, *et al.*, 2019; Martinez-Hernandez, Garcia-Heredia, *et al.*, 2019). For instance, SVG revealed the marine virus vSAG 37-F6 to potentially represent the most abundant viral population in the surface ocean viriosphere (Martinez-Hernandez *et al.*, 2017), which likely infects *Pelagibacter* spp. and has been overlooked for years in metagenomic studies. The SVG approach also represents an inflection point for the capturing genomes of large-capsid-size dsDNA viruses, such as giant viruses (Martínez *et al.*, 2014; Wilson *et al.*, 2017)

While sample preservation, cell lysis, and WGA protocols have been tested for SCGs (Woyke *et al.*, 2011; Clingenpeel *et al.*, 2014; Chen *et al.*, 2018), more thorough testing in each of the fundamental methodological SVG steps is needed to achieve the full potential of this approach (Figure 1). To evaluate the efficiency of the tested approaches in recovering viral genomes, we employed sorted single viral particles of the *Escherichia coli* T4 virus as a model (Jesaitis and Goebel, 1955; Jesaitis, 1957; Miller *et al.*, 2003), and we tested some of the approaches on marine environmental viruses. Based on our findings, we provide specific SVG methodological recommendations and discuss some of the limitations and biases of this novel methodology that will help to expand our knowledge in the uncultured virosphere.

Results and Discussion

First, accurate flow cytometric detection and sorting of viruses fluorescently stained with SYBR dye is critical for the overall success of SVG. When working with biological samples expected to contain high levels of cellular debris or extracellular DNA (e.g., wastewater, stool samples or a bacterial culture such as the pure T4 culture used herein), we strongly recommend the implementation of a DNase digestion step

(see Methods for details) prior to viral DNA staining to reduce noise and the potential masking effect of free fragmented DNA from the actual signal of the stained viral particles (Figure 2A), thus ensuring that extracellular DNA is not co-sorted with viral particles. Vesicles that have a similar virus size are also abundant in environmental samples (Biller *et al.*, 2014) and sometimes vesicles contain sufficient DNA to be visible with SYBR dyes and thus could be confounded with viruses. However, it has been demonstrated that only a very small proportion of vesicles in environmental samples (<0.01–1%) package enough DNA material to be visible by SYBR staining (Biller *et al.*, 2017). Alternatively, a double stain with a lypophylic dye (e.g. FM4-64, red fluorescence) targeting the vesicle membrane combined with SYBR dye (green fluorescence) would aid to unequivocally distinguish vesicles (double positive stain) from non-enveloped viruses, which are dominant in aquatic and many other environments. The preservation of viral samples is crucial to prevent biases introduced by biological changes or degradation associated with the storage time and conditions. Our goal here was to compare three commonly used viral sample preservation methods(Brussaard, 2004; Martínez *et al.*, 2014; Martinez-Hernandez *et al.*, 2017) (fresh unfixed, mild fixation with 0.1% glutaraldehyde and cryopreservation; see Methods section for more details) to determine their relative impact on flow cytometry detection and genome recovery. As shown in Figure 2B-D, in all three methods, the gating of positively stained viral particles of bacteriophage T4 showed a similar resolution. From each of the preservation methods, a total of 668 single viral particles were sorted by flow cytometry and subjected to whole-genome amplification (WGA). In this study, we tested only multiple-displacement amplification-based methods since they surpass other methods amplifying minute quantities of DNA (de Bourcy *et al.*, 2014). Previous works have shown the sensitivity of multiple-

displacement amplification (MDA) in SVG studies of cultured and environmental viruses (Allen *et al.*, 2011; de Bourcy *et al.*, 2014; Martinez-Hernandez *et al.*, 2017; Stepanauskas *et al.*, 2017; Wilson *et al.*, 2017; de la Cruz Peña *et al.*, 2018; Haro-Moreno *et al.*, 2019; Martinez-Hernandez, Fornas, *et al.*, 2019; Martinez-Hernandez, Garcia-Heredia, *et al.*, 2019). Specifically, we explored the novel TruPrime WGA technology based on primase-polymerase features combined with phi29 DNA polymerase (Picher *et al.*, 2016) (see methods for details). To date, the sensitivity of TruPrime WGA technology has only been proven in sorted single eukaryotic cells (Picher *et al.*, 2016). From the sorted single T4 viruses subjected to each preservation method (n=668), we obtained a positive WGA rate (i.e. frequency of successful whole genome amplified single viruses out of total sorted single viruses per 384-well plate) of between 17-26% of viral single amplified genomes (vSAGs) (Table 1). At random, 10 vSAGs from each preservation method were sequenced with Illumina technology (Table 1). The mapping of quality filtered reads against the T4 reference genome confirmed that this novel technology showed a sufficient sensitivity for the whole-genome amplification of single sorted viral particles. Regarding the effects of the preservation methods on genomic quality data (Figure 3), we could conclude that among the three analysed methods, there were significant differences (Tukey test, Supplementary Table 1) between the fresh unfixed protocol and the other two protocols (aldehyde fixation and cryopreservation), with the first performing the best regarding genome recovery. Thus, the data indicated that when possible, for SVG, it is preferable to perform sorting from fresh unfixed samples. Alternatively, if preservation is needed, GlyTE cryopreservation seems to deliver slightly better sequencing data. This trend was also shown when the average coverage of T4 genome was compared (Supplementary Figure 2), although differences were not statistically robust (Supplementary Table 4). Perhaps

glutaraldehyde cross-linking may prevent or reduce the efficiency of DNA amplification and genome recovery compared to samples unfixed or preserved with non-aldehyde chemicals (Das *et al.*, 2014; Wilson *et al.*, 2017). Complementarily, we proved that TruePrime technology was also able to amplify the whole genome of single sorted *E. coli* cells (n=668) (see genome recovery and comparison with MDA in Supplementary Figure 1).

We further aimed to assess the performance of the novel MDA-based method WGA-X, which uses a more thermostable polymerase phi29 at 45ºC (Stepanauskas *et al.*, 2017). For this purpose, we carried out the WGA-X method with single viruses obtained from the same marine viral sample and batch of sorted viruses originally used in a previous marine study in which conventional MDA was employed (Martinez-Hernandez *et al.*, 2017). The results indicated that, as described for single cells (Stepanauskas *et al.*, 2017), the amplification kinetic reactions were also faster in SVG (<3 h), while when the classical phi 29 polymerase was used, the reaction took >7 h to reach a plateau (Martinez-Hernandez *et al.*, 2017). Furthermore, this novel enzyme yielded more amplified DNA from sorted single viruses (up to 1-fold more). The Illumina sequencing results and genome analyses of 4 (or 3) single environmental viruses (42-C9, 42-H22, 42-I14, 42-N18) selected at random corroborated the observation that the vSAGs were indeed marine viruses and confirmed that the novel enzyme was also able to amplify sorted single environmental viruses (Supplementary Data 1). Sequencing data indicated that the novel phi29 polymerase delivers assembled viral contigs of the same average length and quality as conventional MDA (Martinez-Hernandez *et al.*, 2017). Considering that this new phi 29 polymerase shows an overall higher performance, especially in the presence of a high GC content, and exhibits faster kinetics, we recommend the use of this enzyme as a standard for future single-virus

genomic surveys. Thus, based on our findings, both the novel TruePrime WGA and WGA-X methods are likely compatible for most SVG surveys. Further efforts will be required to determine which method performs the best for each type of environmental and biological sample, but since both are based on phi 29 polymerase kinetics, we do not anticipate major differences in terms of genome recovery.

In previous SVG studies, Illumina technology was used to sequence environmental vSAGs. New sequencing platforms such as the PacBio and Oxford Nanopore platforms deliver longer reads (>10 kb), which can facilitate assembly (Zhang *et al.*, 2012; Jain *et al.*, 2018). Here, 3 vSAGs that were previously sequenced by Illumina sequencing (one from the T4 virus and two single environmental viruses) were additionally sequenced on the Nanopore MinIon platform and assembled with the Canu program (version 1.7) (Li *et al.*, 2016; Koren *et al.*, 2017). The quality of the raw reads was not as expected, displaying a rather low Q-value (≈9), and the assembly dataset consisted of contigs that were highly chimeric, with repetition artefacts from the reference viral genome (Table 2). Hybrid assembly with Illumina data was also carried out to complement and improve the Nanopore assembly data, which resulted in non-chimeric data, although genome recovery was not significantly improved compared to the use of Illumina sequencing data alone (Table 2). Thus, when performing the *de novo* assembly of uncultured single viruses, we strongly suggest the use of SVG to perform hybrid assembly.

Finally, we sought to investigate the performance of three common *de novo* assemblers for the vSAGs obtained from phage T4: SPAdes (Bankevich *et al.*, 2012), IDBA (Peng *et al.*, 2012) and Megahit (Li *et al.*, 2015). In previous publications (Martinez-Hernandez *et al.*, 2017; de la Cruz Peña *et al.*, 2018), the authors showed that for most vSAGs, a long contig is generated that contains most of the recovered viral

genome, in addition to other shorter fragments (typically < 1-2 kb) from reagent contaminants, which are removed downstream. Thus, among the different possible genome assembly parameters to be considered, we selected the value of the longest contigs as a proxy for assessing each of the assemblers tested (SPAdes, Idba and Megahit, Figure 4). The data were analysed both individually for each single virus (Figure 4A) and jointly considering the mean and distribution of all analysed single viruses (Figure 4B). Both approaches indicated SPAdes as the best assembler, and this difference was significant based on one-way ANOVA and Tukey's test (Supplementary Table 2). Altogether, while still in its infancy, SVG is a remarkably valuable tool for the recovery of reference environmental viral genomes, bridging the gap between genomic and metagenomic studies.

Experimental Procedures

Virus samples

Stocks of *E. coli* T4 phage (DSM 163) were produced by infecting *E. coli* strain BL21 grown in LB medium at 37ºC once culture absorbance reached 0.3, as previously described (Wyckoff, 1948). Once cell lysis was apparent, the culture was centrifuged at 6,000 g for 15 min, and the supernatant filtered through 0.22 um syringe polyethersulfone (PES) membrane filters ref. SLGP033RS, Millipore, MA, USA) to purify the phages. The presence of bacteriophage T4 was confirmed by nucleic-acid staining and epifluorescence microscopy (Patel *et al.*, 2007) before flow cytometry analyses and sorting. T4 phage stocks were stored at 4ºC until further being used.

A 50 mL surface seawater sample was collected at the Blanes Bay Microbial Observatory (BBMO) in the north-western Mediterranean Sea (41°40'13.5"N 2°48'00.6"E; 2.7 miles offshore) on 15 April 2015 (chlorophyll a concentration 0.32 µg

l^{-1} and temperature 14.6 ºC) for sorting environmental virus particles. Seawater sample was immediately filtered upon collection through 0.22 µm syringe polyethersulfone (PES) membrane filters (ref. SLGP033RS, Millipore, MA, USA).

Sample preservation

Standard flow cytometry (FCM) protocols for discrimination and enumeration of virus-like particles (VLPs) use samples fixed with ≥0.5% glutaraldehyde (final concentration) (Brussaard, 2004). Non-preserved samples degrade relatively fast. Preservation allows long-term storage and consistent quantitative discrimination and sorting of VLPs (Martínez, et al 2014; De Corte et al. 2019) However, glutaraldehyde cross-linking may prevent or reduce the efficiency of DNA amplification and genome recovery compared to samples unfixed or preserved with non-aldehyde chemicals (Das *et al.*, 2014; Wilson *et al.*, 2017)

In this work, we tested the effect of three preservation methods. One milliliter aliquots of bacteriophage T4 were treated with 2U/ml of TurboDNAse I (Invitrogen) at 37ºC for 1 h to remove/reduce host and virus free DNA. Then, TurboDNAse I was inactivated with the Inactivation Buffer (10% Sample Volume) following strictly manufacture´s protocol The samples were then either: a) kept fresh (i.e., no added fixative) at 4ºC; b) fixed with either glutaraldehyde (0.1% final concentration) (Martínez et al 2014); or c) cryoprotectant glycerol Tris-EDTA buffer (GlyTE, 5% glycerol and 1×TE (10mM Tris-HCl, 1mM EDTA, pH 8.0) final concentrations) and stored at -80ºC, until ready for sorting. A non-infected ampicillin-treated (100 µg/ml) lysed *E. coli* BL21 was treated with DNase and preserved as described above to serve as non-virus controls. Controls helped discriminating non-viral FCM background signal to ensure its exclusion from sorting gates.

Viral nucleic acids staining for fluorescence detection by flow cytometry

Bacteriophage T4 samples were stained with SYBR Gold (Invitrogen). Glutaraldehyde-fixed samples were stained following the protocol in (Martinez-Hernandez *et al.*, 2017). Bacteriophage T4 fresh and GlyTE-cryopreserved samples were concentrated down to 50 µl and washed with 0.02 µm-filtered 1×TE buffer using Nanosep 10 kDa ultracentrifugal columns (OMEGA, Pall Life Sciences) prior to being incubated with SYBR Gold (4× final concentration) at room temperature in the dark for 20 min. Excess SYBR Gold in solution was removed by washing the sample three times with 500 µl of sterile TE buffer in the Nanosep 10 kDa column as previously described (Martinez-Hernandez *et al.*, 2017) prior to FCM analysis. Non-virus control were stained following the same procedures for their respective sample type.

Virus particles sorting

Fluorescence activated virus sorting (FAVS) was carried out with BD Influx high-resolution cell sorter (BD Biosciences, San Jose, CA). Specific instrument setup and DNA-free conditions were used as previously described (Martinez-Hernandez *et al.*, 2017) at the Flow Cytometry Core Facility University of Pompeu Fabra-Centre for Genomic Regulation (Barcelona, Spain). Virus particles were sorted into individual wells in 384-well plates (4titude Ltd) pre-loaded with 0.6 µl sterile 0.02 µm-filtered 1×TE buffer and UV-decontaminated for 20 min under 254 nm UV wavelength light. The sorted plate layout included 338 wells each containing a single virus particle, 3 wells containing 50 sorted virus particles each, and 43 no-drop wells to serve as negative controls.

Viral whole genome amplification

We tested three commercially available multiple displacement amplification (MDA) based whole genome amplification (WGA) enzymes commonly used for single-cell genomics. Decontamination of reagents prior to setting up the reactions was done as previously described (Martinez-Hernandez *et al.*, 2017). After combining the reagents, minus the fluorescence dye SYTO9 (Invitrogen), the master mix was UV-decontaminated for an additional 15-180 min at 4ºC in a UVP Ultraviolet CL-1000 Crosslinker described in detail (Rinke *et al.*, 2014). After which SYTO9 was added to the mixture. WGA methods were as follows:

1.- Phi29 polymerase

Sorted bacteriophage T4 and environmental virus particles were lysed with a KOH pH14 solution to release their nucleic acids. Alkaline lysis was done by adding 0.7 µl of lysis buffer DLB (0.4 M KOH, 10 mM EDTA and 100 mM dithiothreitol) to each well in the sorted plate and incubating for 5 min at 4ºC. The lysis reaction was stopped by adding 0.7 µL of Stop solution (Qiagen, ref. 1032393) or 1M Tris-HCL, pH 4 per well. DNA amplification was done by MDA in 10 µl final volume reactions containing 0.26 µl of Phi29 DNA polymerase (ref. M0269L; 10 U/ µl; New England Biolab), 1 µl of Phi29 10× reaction buffer (ref. M0269L; New England Biolab), 1 µl of hexamers (0.5 mM; IDT), 0.1 µl of DTT (1 M; Sigma), 0.4 µl of dNTPs (10 mM each; ref. N0447L, New England Biolab), 0.002 µl of SYTO 9 (Invitrogen) and 5.2 µl of sterile UV-treated (16 h) de-ionized water. Finally, 0.6 ng of genomic lambda DNA (ref. N3011S, New England Biolab) was added to four corner wells of the sorted plates to serve as MDA positive controls. MDA reactions were incubated at 30ºC for 16 h in a CLARIOstar (BMG Labtech) or a FilterMax F5 (Molecular Devices) fluorimeter plate readers. The reaction kinetics were monitored real time. The MDA reaction was stopped by heat-

inactivation of the phi 29 polymerase at 65ºC for 10 min and the MDA product was diluted 50-fold in sterile TE buffer.

2.- Equiphi29 polymerase (WGA-X)

Virus particles were lysed using thermal shock or pH14 alkaline lysis as described above. Based on previous results, pH10 alkaline solution was not tested here. Final volume for WGA-X reaction was 10 µl. Final concentration of whole-genome amplification reactions was as follows: 0.2 U μl^{-1} Equiphi29 polymerase (Thermo Fisher Scientific) (Picher *et al.*, 2016), 1X Equiphi29 reaction buffer (Thermo Fisher Scientific), 0.4mM each dNTP (New England BioLabs), 10mM dithiothreitol (Thermo Fisher Scientific), 40µM random heptamers with two 3'-terminal phosphorothioated nucleotide bonds (Integrated DNA Technologies), and 1µM SYTO-9 (Thermo Fisher Scientific). These reactions were performed at 45ºC for 3-4 h in plate readers and inactivated by incubation at 75ºC during 15 min.

3.- TruPrime method

The TruePrime scWGA kit (Sygnis® Expedeon cat. num: 351100) (Picher *et al.*, 2016) combines Phi29 polymerase with the *Thermus thermophilus* PrimPol primase, which displays a potent primase activity and synthesizes new primers on the displaced strands that are later extended by the high-fidelity Phi29 DNA polymerase. Virus particle lysis was carried out by adding 0.6 µl of buffer L2 in the kit followed by incubating for 5 min at room temperature plus one additional minute at 95ºC. The lysis reaction was stop with 0.6 µl of buffer N. Whole-genome amplification reactions (10 µl final volume)were set as follows:1.2 µl of Reaction buffer, 1.2 µl of dNTPs, 1.2 µl PrimPol primase, 0.168 µl of Phi29 polymerase, 6.06 µl of sterile UV--treated (16 h) de-ionized

water and 0.0024 of SYTO-9. Amplification was performed and monitored for 16h at 30ºC in plate readers. The reactions were inactivated by a 10 min incubation at 65ºC.

Sequencing and bioinformatics analyses

Selected bacteriophage T4 and environmental vSAGs were sequenced by Illumina technology using the Nextera XT DNA library (ref. FC-131-1024, Illumina) in a MiSeq sequencer (2x250, pair-end) according to manufacturer´s protocol. In addition, the following vSAGs were also sequenced with Oxford Nanopore MinION: environmental marine vSAGs 37-F6 and 42-C9 and bacteriophage T4 vSAG 142-4-J7 (Figure 1). The original vSAG 142-4-J7 MDA product was re-amplified with Equiphi29 polymerase to generate sufficient template for Nanopore sequencing, which requires at least a total of 0.5-1 µg of DNA. Nanopore sequencing was performed at the Genomics Center of FISABIO (Valencia, Spain) following manufacture´s condition with MinIon platform (version as of date May 2018).

Prior to assembling, illumina reads were quality filtered with Trimmomatic v0.3210 (Bolger, et al 2014) using the following parameters: phred33 LEADING:0 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. Illumina reads assembly was compared for assemblers: a) SPAdes version 3.10 (Bankevich *et al.*, 2012) by applying the following parameters: --sc, -k 21,33,55,77,99,127, --careful; b) IDBA with default parameters; and c) Megahit also with default parameters. For SPAdes we employed the specific options “-sc” and “-careful” that minimizes the effect of low or uneven coverage obtained from MDA and reduces the number of mismatches and short indels.

Nanopore reads were assembled using Canu 1.7 (Koren *et al.*, 2017) with -trim-assemble option to use as input raw reads from MinIon. Hybrid assemblies were

done with SPAdes v 3.10 using sequences belonged to Nanopore and cleaned with Canu 1.7 using –correct option and reads from Illumina and filtered with trimmomatic v 0.36. Parameters applied for hybrid assembly with SPAdes v 3.10 were: -nanopore to introduce the corrected sequences from Canu 1.7 and -k 21,33,55,77,99,127 (Figure 1). Percentage of recovered genome with reads was calculated with BWA program, Samtools, Bedtools, and bbmap packages. Prokka and BlastP against NR database were used to annotate the assembled genomes.

Statistical analyses

ANOVA analysis and Tukey posteriori test was made using R package Vegan (Oksanen *et al.*, 2017), first we check that our samples accomplish all the ANOVA requirements (Independency of the samples, Normalization of the levels and Variance Homogeneity) and gglopt2 package (Wickham, 2009) was used to represent the results graphically.

Data Records

Shotgun sequences generated on the Illumina and ONT platforms are publicly available through NCBI Bioprojects numbers PRJNA611681, PRJNA611684 y PRJNA611689.

Acknowledgements

This work has been supported by Gordon and Betty Moore Foundation (grant 5334) and Spanish Ministry of Economy and Competitiveness (refs CGL2013-40564-R, RTI2018-094248-B-I00 and SAF2013-49267-EXP). Work at CRG, BIST and UPF was in part funded by the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' and the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Maria de Maeztu 2016-2019'.

The authors declare no competing interests.

References

Abedon, S.T. ed. (2008) Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses, 1. ed. Cambridge: Cambridge Univ. Press.

Allen, L.Z., Ishoey, T., Novotny, M.A., McLean, J.S., Lasken, R.S., and Williamson, S.J. (2011) Single Virus Genomics: A New Tool for Virus Discovery. *PLoS ONE* **6**: e17722.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**: 455–77.

Biller, S.J., McDaniel, L.D., Breitbart, M., Rogers, E., Paul, J.H., and Chisholm, S.W. (2017) Membrane vesicles in sea water: heterogeneous DNA content and implications for viral abundance estimates. *ISME J* **11**: 394–404.

Biller, S.J., Schubotz, F., Roggensack, S.E., Thompson, A.W., Summons, R.E., and Chisholm, S.W. (2014) Bacterial Vesicles in Marine Ecosystems. *Science* **343**: 183–186.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

de Bourcy, C.F.A., De Vlaminck, I., Kanbar, J.N., Wang, J., Gawad, C., and Quake, S.R. (2014) A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods. *PLoS ONE* **9**: e105585.

Brussaard, C.P.D. (2004) Optimization of Procedures for Counting Viruses by Flow Cytometry. *Applied and Environmental Microbiology* **70**: 1506–1513.

Chen, D., Zhen, H., Qiu, Y., Liu, P., Zeng, P., Xia, J., et al. (2018) Comparison of single cell sequencing data between two whole genome amplification methods on two sequencing platforms. *Sci Rep* **8**: 4963.

Clingenpeel, S., Schwientek, P., Hugenholtz, P., and Woyke, T. (2014) Effects of sample treatments on genome recovery via single-cell genomics. *ISME J* **8**: 2546–2549.

Coutinho, F.H., Rosselli, R., and Rodríguez-Valera, F. (2019) Trends of Microdiversity Reveal Depth-Dependent Evolutionary Strategies of Viruses in the Mediterranean. *mSystems* **4**: e00554-19, /msystems/4/6/msys.00554-19.atom.

Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D., et al. (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**: 15955.

de la Cruz Peña, M., Martinez-Hernandez, F., Garcia-Heredia, I., Lluesma Gomez, M., Fornas, Ò., and Martinez-Garcia, M. (2018) Deciphering the Human Virome with Single-Virus Genomics and Metagenomics. *Viruses* **10**: 113.

Das, K., Fernando, M.R., Basiaga, S., Wigginton, S.M., and Williams, T. (2014) Effects of a novel cell stabilizing reagent on DNA amplification by PCR as compared to traditional stabilizing reagents. *Acta Histochemica* **116**: 55–60.

De Corte, D., Martínez, J.M., Cretoiu, M.S., Takaki, Y., Nunoura, T., Sintes, E., et al. (2019) Viral Communities in the Global Deep Ocean Conveyor Belt Assessed by Targeted Viromics. *Front Microbiol* **10**: 1801.

Dhillon, V. and Li, X. (2015) Single-Cell Genome Sequencing for Viral-Host Interactions. *J Comput Sci Syst Biol 2015* **8**:.

Edwards, R.A. and Rohwer, F. (2005) Viral metagenomics. *Nat Rev Microbiol* **3**: 504–510.

Haro-Moreno, J.M., Rodriguez-Valera, F., Rosselli, R., Martinez-Hernandez, F., Roda-Garcia, J.J., Gomez, M.L., et al. (2019) Ecogenomics of the SAR11 clade. *Environ Microbiol* 1462-2920.14896.

Hugerth, L., Larsson, J., Alneberg, J., Lindh, M., Legrand, C., Pinhassi, J., and Andersson, A. (2015) Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biol* **16**: 279.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338–345.

Jesaitis, M.A. (1957) The nucleic acids of T2, T4, and T6 bacteriophages. *J Exp Med* **106**: 233–246.

Jesaitis, M.A. and Goebel, W.F. (1955) Lysis of T4 phage by the specific lipocarbohydrate of phase II Shigella sonnei. *J Exp Med* **102**: 733–752.

Kang, I. and Cho, J.-C. (2014) Depth-Specific Distribution of the SAR116 Phages Revealed by Virome Binning. *Journal of Microbiology and Biotechnology* **24**: 592–596.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Res* **27**: 722–736.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., et al. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J* **9**: 2386–2399.

Lasken, R.S. (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nat Rev Microbiol* **10**: 631–640.

Li, C., Chng, K.R., Boey, E.J.H., Ng, A.H.Q., Wilm, A., and Nagarajan, N. (2016) INC-Seq: accurate single molecule reads using nanopore sequencing. *GigaSci* **5**: 34.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674–1676.

López-Escardó, D., Grau-Bové, X., Guillaumet-Adkins, A., Gut, M., Sieracki, M.E., and Ruiz-Trillo, I. (2017) Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. *Sci Rep* **7**: 11025.

López-Pérez, M., Haro-Moreno, J.M., Gonzalez-Serrano, R., Parras-Moltó, M., and Rodriguez-Valera, F. (2017) Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters. *PLoS Genet* **13**: e1007018.

Mangot, J.-F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S., et al. (2017) Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* **7**: 41498.

Martínez, J.M., Swan, B.K., and Wilson, W.H. (2014) Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J* **8**: 1079–1088.

Martinez-Garcia, M., Swan, B.K., Poulton, N.J., Gomez, M.L., Masland, D., Sieracki, M.E., and Stepanauskas, R. (2012) High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. *ISME J* **6**: 113–123.

Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., de la Cruz Peña, M.J., Martínez, J.M., et al. (2017) Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat Commun* **8**: 15892.

Martinez-Hernandez, F., Fornas, Ò., Lluesma Gomez, M., Garcia-Heredia, I., Maestre-Carballa, L., López-Pérez, M., et al. (2019) Single-cell genomics uncover Pelagibacter as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. *ISME J* **13**: 232–236.

Martinez-Hernandez, F., Garcia-Heredia, I., Lluesma Gomez, M., Maestre-Carballa, L., Martínez Martínez, J., and Martinez-Garcia, M. (2019) Droplet Digital PCR for Estimating Absolute Abundances of Widespread Pelagibacter Viruses. *Front Microbiol* **10**: 1226.

Miller, E.S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Rüger, W. (2003) Bacteriophage T4 genome. *Microbiol Mol Biol Rev* **67**: 86–156, table of contents.

Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E., and Ghai, R. (2013) Expanding the Marine Virosphere Using Metagenomics. *PLoS Genet* **9**: e1003987.

Oksanen, J., Blanchet, G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2017) vegan: Community Ecology Package.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., et al. (2016) Uncovering Earth's virome. *Nature* **536**: 425–430.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B.J., Evans, P.N., et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**: 1533–1542.

Patel, A., Noble, R.T., Steele, J.A., Schwalbach, M.S., Hewson, I., and Fuhrman, J.A. (2007) Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. **2**: 269–277.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Picher, Á.J., Budeus, B., Wafzig, O., Krüger, C., García-Gómez, S., Martínez-Jiménez, M.I., et al. (2016) TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. *Nat Commun* **7**: 13296.

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., et al. (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nature protocols* **9**: 1038–48.

Rohwer, F. (2003) Global Phage Diversity. *Cell* **113**: 141.

Sieracki, M.E., Poulton, N.J., Jaillon, O., Wincker, P., de Vargas, C., Rubinat-Ripoll, L., et al. (2019) Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci Rep* **9**: 6025.

Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology* **15**: 613–620.

Stepanauskas, R., Fergusson, E.A., Brown, J., Poulton, N.J., Tupper, B., Labonté, J.M., et al. (2017) Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun* **8**: 84.

Stepanauskas, R. and Sieracki, M.E. (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* **104**: 9052–9057.

Suttle, C.A. (2007) Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801–812.

Tara Oceans Coordinators, Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., et al. (2018) Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat Commun* **9**: 310.

Wickham, H. (2009) ggplot2: Elegant Graphics for Data Analysis., New York: Springer-Verlag.

Wilson, W.H., Gilg, I.C., Moniruzzaman, M., Field, E.K., Koren, S., LeCleir, G.R., et al. (2017) Genomic exploration of individual giant ocean viruses. *ISME J* **11**: 1736–1745.

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., et al. (2011) Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. *PLoS ONE* **6**: e26161.

Wyckoff, R.W.G. (1948) The electron microscopy of developing bacteriophage. *Biochimica et Biophysica Acta* **2**: 246–253.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., et al. (2011) Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. *Science* **332**: 714–717.

Zhang, X., Davenport, K.W., Gu, W., Daligault, H.E., Munk, A.C., Tashima, H., et al. (2012) Improving genome assemblies by sequencing PCR products with PacBio. *BioTechniques* **53**: 61–62.

Figure Legends

Figure 1: Scheme of samples and tested methodological steps for each one of SVG steps. MDA: Multiple Displacement Amplification, **WGA:** Whole genome amplification, **WGA-X**: Whole Genome Amplification-X, TruePrime method: **Sygnis**

Figure 2: Flow cytometry analyses and sorting of bacteriophage T4. Biplots of flow cytometry showing side scatter (SSC-H) and fluorescence signal (FITC-H, relative units). **a:** control culture of *E. col*i cells alone without infection of T4 phage. Culture was treated with Ampicillin (2.5µl/mll) and DNase (2U/ml) to determine the flourescence noise due to *E. col*i DNA released from lysed cells. **b:** Infected culture treated with DNase treatment (2U/ml). No preservation method was applied. **c:** Infected culture preserved with 0.1% glutaraldehyde and treated with DNase (2U/ml). **d:** Treatment with 10% GlyTE and DNase (2U/ml). **P1**: bacteriophage T4 gate used for virus sorting. **P2:** noise from debris

Figure 3: Comparison of preservation methods for SVG. Box plot comparing the percentage of bases covered of bacteriophage T4 genome calculated using a random sampling of 10 sequenced vSAGs data.

Figure 4: Comparison of assembly programs for sequenced vSAGs. a: Bar plot showing the length (bp) of the longest contig obtained in each single-virus for each assembler. **b:** Box plot comparing the mean of length (bp) of the longest contig obtained for each assembler.

Table 1. Summary data of methods used in each Single Virus and Single Cell

single virus name	type	preservation method	amplification method	sequencing method	assembly method	Average coverage	Bases recovered (%)
37-F6	Marine vSAG	Fresh	MDA	Illumina	SPAdes/Idba/Megahit		
37-F7	Marine vSAG	Fresh	MDA	Nanopore	Canu+ SPAdes		
42-C9	Marine vSAG	Fresh	WGA-X	Illumina	SPAdes/Idba/Megahit		
42-C10	Marine vSAG	Fresh	WGA-X	Nanopore	Canu+ SPAdes		
42-H22	Marine vSAG	Fresh	WGA-X	Illumina	SPAdes/Idba/Megahit		
42-I14	Marine vSAG	Fresh	WGA-X	Illumina	SPAdes/Idba/Megahit		
42-N18	Marine vSAG	Fresh	WGA-X	Illumina	SPAdes/Idba/Megahit		
139_2_C4	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	1045.19	21.06
139-2-A4	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	983.12	10.43
139-2-A6	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	0.15	6.86
139-2-C2	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	1070.6	8.95
139-2-E2	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	0.81	7.16
139-2-E6	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	1158.82	11.94
139-2-G8	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	1359.22	9.77
139-2-I4	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	0.14	6.33
139-2-M6	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	0.43	11
139-2-O3	phage T4 vSAG	0.1% Glut	TruePrime	Illumina	SPAdes/Idba/Megahit	0.12	6.11
141_4_A10	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	880.42	19.09
141_4_A2	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	1535.24	20.82
141_4_A4	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	1272.49	24.24
141_4_B9	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	1028.75	25.74
141-4-A11	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	477.24	20.35
141-4-A5	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	1435.84	16.71
141-4-A8	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	1241.21	13.63
141-4-I2	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	854.72	18.29
141-4-M8	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	866.13	19
4-N8	phage T4 vSAG	GlyTE	TruePrime	Illumina	SPAdes/Idba/Megahit	847.91	17.21
142_4_J7	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1259.56	61.23
142_4_J8	phage T4 vSAG	Fresh	TruePrime	Nanopore	Canu+ SPAdes		
142_4_A8	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1028.98	31.32
142_4_E7	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	999.35	31.17
142_4_F10	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1412.23	35.7
142_4_J8	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1338.5	70.66
142_4_M10	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1275.52	31.42
142-4-E8	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1704.79	16.49
142-4-K2	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	1147.62	24.72
142-4-L2	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	0.16	8.01
142-4-M8	phage T4 vSAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	0.4	21.36
105_2_C15	E. coli SAG	Fresh	MDA	Illumina	SPAdes/Idba/Megahit	2.68	5.23
105_2_C16	E. coli SAG	Fresh	MDA	Nanopore	Canu+ SPAdes		
105_1_E8	E. coli SAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	2.44	1.14
105_1_M5	E. coli SAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	3.97	1.22
105_1_O3	E. coli SAG	Fresh	TruePrime	Illumina	SPAdes/Idba/Megahit	4.64	0.85
105_2_I39	E. coli SAG	Fresh	MDA	Illumina	SPAdes/Idba/Megahit	1.29	3.93
147_1_E8	E. coli SAG	Fresh	MDA	Illumina	SPAdes/Idba/Megahit	2.82	30.68

Table 2. Data of short- and long-read sequencing of single cell and viruses

Name	Type	Sequencing method	Assembly method	Chimeric contigs	Longest contig	contig > 500	% contamination
37-F6	single virus	Illumina	SPAdes	No	7344	6	0.00
		Nanopore	Canu 1.7	Yes	29664*	32	0.00
			Canu 1.7+SPAdes	No	2908	13	0.00
142_4_J7	single virus	Illumina	SPAdes	No	13642	44	2.27
		Nanopore	Canu 1.7	Yes	42857*	51	0.00
			Canu 1.7+SPAdes	No	13642	37	2.70
42_C9	single virus	Illumina	SPAdes	No	20058	16	31.25
		Nanopore	Canu 1.7	Yes	35846*	24	0.00
			Canu 1.7+SPAdes	No	8649	261	1.15
105_2_C15	single cell	Illumina	SPAdes	No	41666	53	0.00
		Nanopore	Canu 1.7	Yes	63341*	19	0.00
			Canu 1.7+SPAdes	No	53650	27	0.00

* longest contig with repetitions

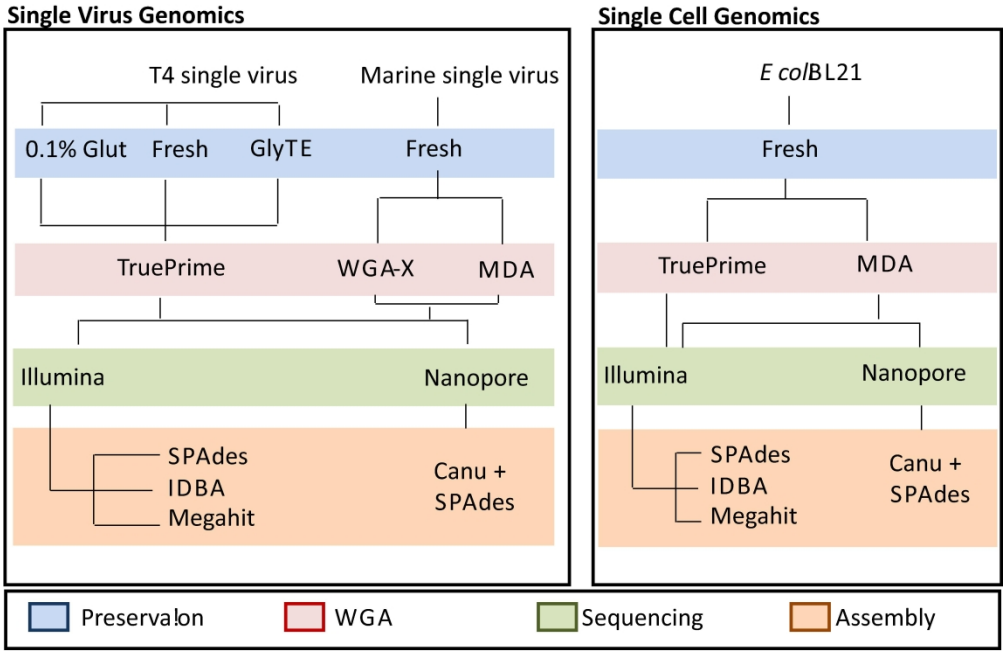

Figure 1: Scheme of samples and tested methodological steps for each one of SVG steps. MDA: Multiple Displacement Amplification, WGA: Whole genome amplification, WGA-X: Whole Genome Amplification-X, Truprime: Sygnis

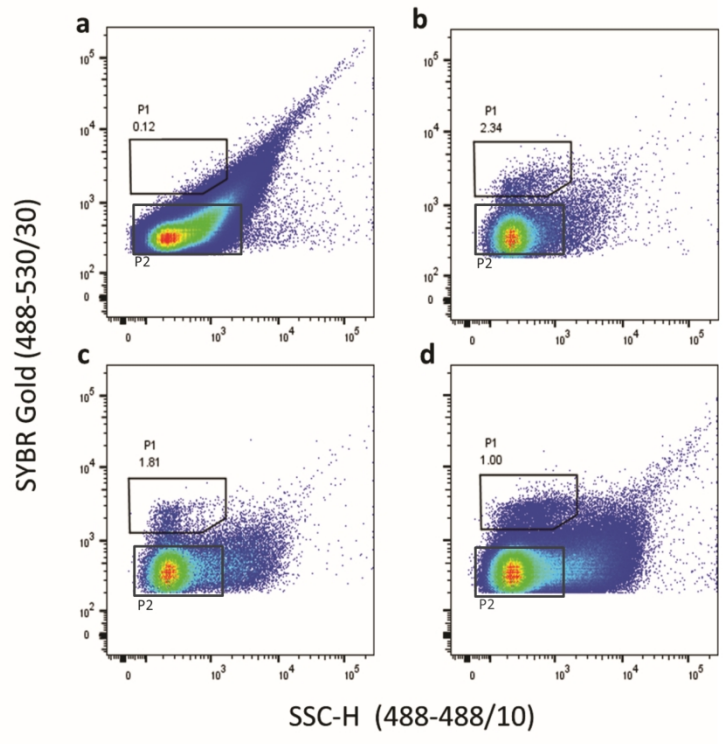

Figure 2: Flow cytometry analyses and sorting of bacteriophage T4. Biplots of flow cytometry showing side scatter (SSC-H) and fluorescence signal (FITC-H, relative units). a: control culture of E. coli cells alone without infection of T4 phage. Culture was treated with Ampicillin (2.5µl/mll) and DNase (2U/ml) to determine the flourescence noise due to E. coli DNA released from lysed cells. b: Infected culture treated with DNase treatment (2U/ml). No preservation method was applied. c: Infected culture preserved with 0.1% glutaraldehyde and treated with DNase (2U/ml). d: Treatment with 10% GlyTE and DNase (2U/ml). P1: bacteriophage T4 gate used for virus sorting. P2: noise from debris

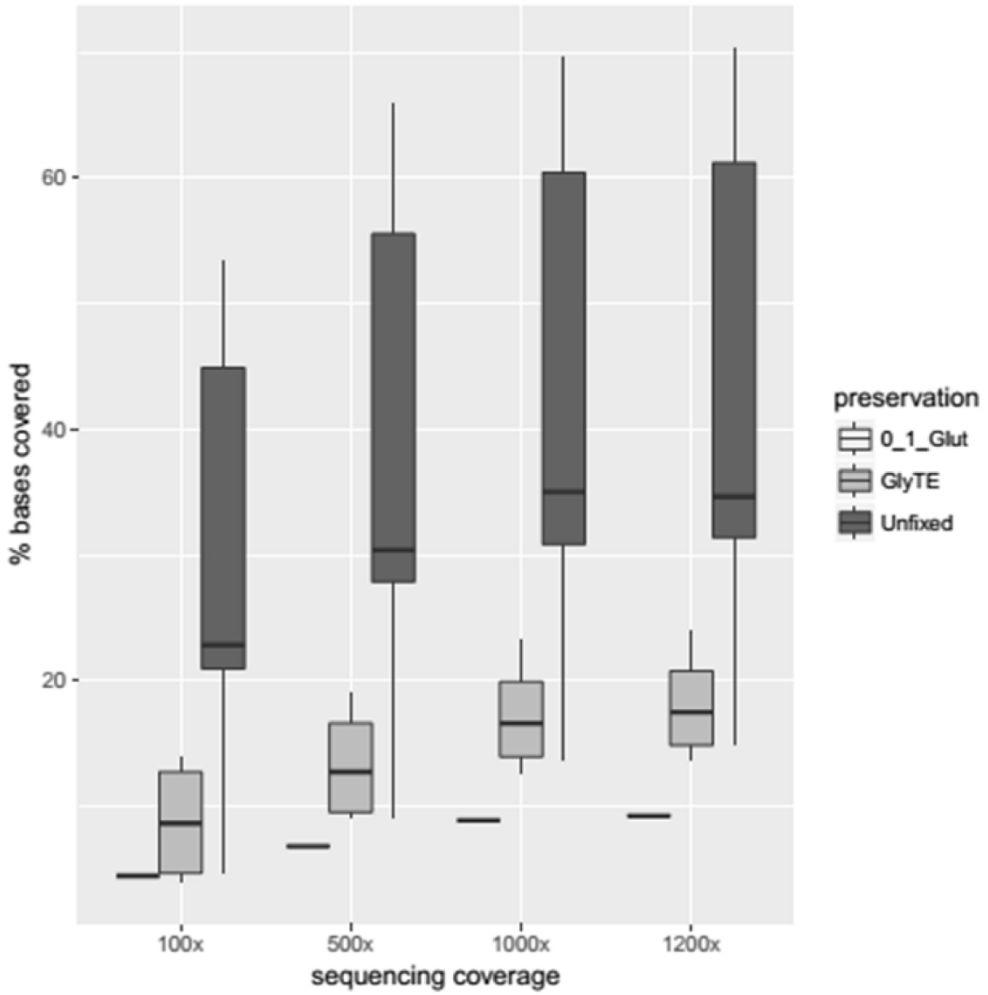

Figure 3: Comparison of preservation methods for SVG. Box plot comparing the percentage of bases covered of bacteriophage T4 genome calculated using a random sampling of 10 sequenced vSAGs data.

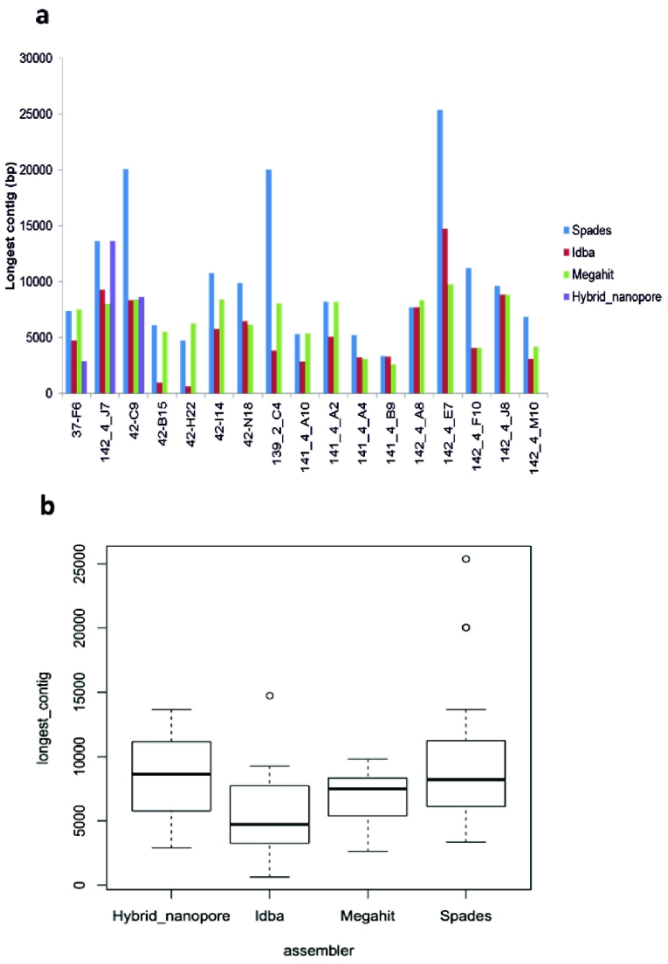

Figure 4: Comparison of assembly programs for sequenced vSAGs. a: Bar plot showing the length (bp) of the longest contig obtained in each single-virus for each assembler. b: Box plot comparing the mean of length (bp) of the longest contig obtained for each assembler.