DNA-spiking in viral metagenome sequencing: A new method with low bias

Geert Cremers [1] , **Lavinia Gambelli** [1] , **Theo van Alen** [1] , **Laura van Niftrik** [1] , **Huub Op den Camp** Corresp. [1]

[1] Microbiology, Radboud University Nijmegen

Corresponding Author: Huub Op den Camp
Email address: h.opdencamp@science.ru.nl

With the emergence of Next Generation Sequencing, major advances were made with regard to identifying viruses in natural environments. However, bioinformatical research on viruses is still limited because of the low amounts of viral DNA that can be obtained for analysis. To overcome this limitation, DNA is often amplified with multiple displacement amplification (MDA), which causes an unavoidable bias. Here, we describe a DNA-spiking method to avoid the bias that is created when using amplification of DNA before metagenome sequencing. To obtain sufficient DNA for sequencing, a bacterial 16S rRNA gene was amplified and the obtained DNA was spiked to a DNA sample containing DNA from a bacteriophage population before sequencing using Ion Torrent technology. After sequencing, the 16S rRNA gene reads DNA was removed by mapping to the Silva database. The new DNA-spiking method was compared with the MDA technique. When MDA was applied, the overall GC content of the reads showed a bias towards lower GC%, indicating a change in composition of the DNA sample. Assemblies using all available reads from both MDA and the DNA-spiked samples resulted in six viral genomes. All six genomes could be almost completely retrieved (97.9%-100%) when mapping the reads from the DNA-spiked sample to those six genomes . In contrast, 6.3%-77.7% of three viral genomes were covered by reads obtained using the MDA amplification method and only three were nearly fully covered (97.4%-100%). The new method provides a simple and inexpensive protocol with very low bias in sequencing of metagenomes for which low amounts of DNA are available.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2912v1 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

DNA-spiking in viral metagenome sequencing: A new method with low bias

Geert Cremers, Lavinia Gambelli, Theo van Alen, Laura van Niftrik & Huub J.M. Op den Camp*

Department of Microbiology, Institute for Water and Wetland Research, Faculty of Science, Radboud University, Heyendaalseweg 135, NL 6525 AJ Nijmegen, the Netherlands.

*Corresponding author:

h.opdencamp@science.ru.nl

Heyendaalseweg 135, NL 6525 AJ Nijmegen, The Netherlands

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2912v1 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

Key words: virus, bacteriophage, metagenome, metavirome, DNA spiking, multiple displacement amplification

Abstract

With the emergence of Next Generation Sequencing, major advances were made with regard to identifying viruses in natural environments. However, bioinformatical research on viruses is still limited because of the low amounts of viral DNA that can be obtained for analysis. To overcome this limitation, DNA is often amplified with multiple displacement amplification (MDA), which causes an unavoidable bias. Here, we describe a DNA-spiking method to avoid the bias that is created when using amplification of DNA before metagenome sequencing. To obtain sufficient DNA for sequencing, a bacterial 16S rRNA gene was amplified and the obtained DNA was spiked to a DNA sample containing DNA from a bacteriophage population before sequencing using Ion Torrent technology. After sequencing, the 16S rRNA gene reads DNA was removed by mapping to the Silva database. The new DNA-spiking method was compared with the MDA technique. When MDA was applied, the overall GC content of the reads showed a bias towards lower GC%, indicating a change in composition of the DNA sample. Assemblies using all available reads from both MDA and the DNA-spiked samples resulted in six viral genomes. All six genomes could be almost completely retrieved (97.9%-100%) when mapping the reads from the DNA-spiked sample to those six genomes. In contrast, 6.3%-77.7% of three viral genomes were covered by reads obtained using the MDA amplification method and only three were nearly fully covered (97.4%-100%). The new method provides a simple and inexpensive protocol with very low bias in sequencing of metagenomes for which low amounts of DNA are available.

Introduction

Microbial research has been mainly culture-based since the work of Pasteur and Koch. This has led to great improvement of our knowledge of the microbial and viral world. However, our knowledge is probably still only the tip of the iceberg, as most of the microorganisms cannot be cultured (Rosario & Breitbart, 2011). In recent years, there has been a greater focus on the hidden bacterial and viral ‘black matter’ since the development of next generation sequencing (NGS) techniques which allow the determination of the microbial community without the need for cultivation. Without the necessity of cultivation prior to sequencing, organisms that cannot be cultured under artificial conditions are now being sequenced in increasing numbers. This is especially true for bacteria; sequencing viral black matter from environmental samples is still hampered by a variety of factors. Besides the obvious problem that not all viruses are DNA viruses (Steward et al., 2013), there is also the challenge of the low quantity of DNA that can be retrieved from viruses. Although viruses outnumber bacteria 5 to 25 times in numbers (Fuhrman, 1999; Clokie et al., 2011), the fact that viruses have on average a significantly smaller genome size means that the viral DNA yield from any given sample is significantly lower compared to bacterial DNA (Brum & Sullivan, 2015). Because of this, DNA is often extracted using DNA extraction methods optimized for virions gathered from a large amount (20-200 L) of sample (Breitbart et al., 2002;Thurber et al., 2009; Duhaime, 2012; Steward et al., 2013).

Since sampling large amounts is not always possible and because there is loss of DNA in every step of DNA isolation protocols, the final yield of viral DNA remains low. Therefore, the low amount of DNA must be amplified before sequencing. Several methods are available (Duhaime et al., 2012; Brum & Sullivan, 2015), with the Multiple Displacement Amplification (MDA) method being the most widely used. However, MDA has a few drawbacks. Because it is amplification, it unavoidably introduces a bias (Kim & Bae, 2011; Marine et al., 2014) and furthermore information of relative abundance in the original sample is lost (Yilmaz et al., 2010). In practical terms, there is a high potential of cross contamination throughout the lab, which is particularly problematic in a laboratory where viral work is the main area of research.

In view of the growing interest in the impact of viruses on the ecology and the natural world, there is a growing need for an easy method or technology that is bias-limited and preferably easy, inexpensive and readily available. However, the recurring problem on most sequencing platforms is the relative large amount of DNA needed for the preparation of DNA fragments for sequencing compared to the typically low yield of DNA from viral population samples. As discussed above, two ways to obtain more DNA is to collect more samples or apply amplification. A theoretical third option for viruses would be to artificially raise the DNA concentration by adding DNA that is naturally non-occurring in any virus. Prime candidate would be the bacterial 16S ribosomal gene as it has not been found in any known virus up to date. In this report we describe a method for metagenome sequencing with the Ion Torrent Personal Genome Machine in which we spiked low amounts (≈ 0.1 ng) of viral DNA from a bacteriophage population with 16S ribosomal DNA and compare this with traditional MDA amplification.

Material and methods

Sample collection and DNA extraction

The bacteriophage population used for sequencing was obtained from a *Methylomirabilis oxyfera* bioreactor enrichment culture (Gambelli et al., 2016). Bioreactor material was collected over a period of about three months, stored at 4˚C and viral particles were obtained as described before (Gambelli et al., 2016). Briefly, the aggregated microbial biomass was disrupted to free the bacteriophages and viral particles were precipitated using PEG8000 (Guo et al., 2012). Free bacteriophages present in the bioreactor supernatant medium and not within the bacterial aggregates were precipitated by iron chloride flocculation (Cunningham et al., 2015).

The two samples obtained by iron chloride flocculation and PEG 8000 precipitation were pooled together and bacteriophages were concentrated by ultracentrifugation (Optima XE90, Beckman-Coulter; Rotor: Type 90 Ti, Beckman-Coulter) at 77,000x g at 4˚C for 1h. The pellet was resuspended in 1 ml of supernatant and the total DNA was extracted according to the protocol published by Thurber et al. (2009). Using the Qubit dsDNA HS assay kit (Thermo Scientific, Waltham, USA), the extracted DNA was quantified at 0.2 ng DNA.

Library preparation and sequencing

MDA method

For the MDA method, 0.1 ng of viral metagenome DNA was amplified using the Illustra GenomePhi HY DNA amplification kit (GE Healthcare, Piscataway, NJ, USA) as per manufacturer's protocol. The DNA was cleaned using GeneJET plasmid Miniprep kit (Fermentas, Amherst, USA) according to manufacturer’s protocol, except for step one, in which 200 µl of DEPC was used instead of lysis-buffer. This first amplification round yielded 15 ng of DNA (referred as 1 x MDA sample), after which 10 ng of DNA was used for a second amplification round (referred as 2 x MDA sample) which resulted in a final yield of 5.4 µg DNA. Sequence library preparation was started with 5 ng of 1 x MDA and with 100 ng 2 x MDA amplified DNA. Both samples were sheared for 6 cycles (1 min on, 1 min off) on the Bioruptor® Standard (Diagenode Liege, Belgium). After shearing, the samples were cleaned using a 1:1 volume ratio with AMPure XP beads (Beckman Coultier, High Wycombe, UK) and further prepared for sequencing as per manufacturer's protocol (IonXpress Plus gDNA fragment library preparation Rev C.0, Life, Carlsbad, USA).

DNA-spiking method

For the DNA-spiking method, 0.1 ng of viral metagenome DNA was spiked with 43.2 ng of amplified 16S rRNA DNA from "*Candidatus* Kuenenia stuttgartiensis" (GenBank CT573071) (referred as DNA-spiked sample) and sheared using the Bioruptor for 6 cycles (1 min on, 1 min off) and prepared according to manufacturer's protocol (IonXpress Plus gDNA fragment library preparation Rev C.0, Life).

The amplicons of the 16S rRNA gene were obtained by PCR of an in-house 16S rRNA gene clone of “*Ca.* K. stuttgartiensis*”* using primers pla46 (5'-GGATTAGGCATGCAAGTC-'3) and 630R (5'-CAKAAAGGAGGTGATCC-'3) with the following settings: 5 min at 94˚C, followed by 35 cycles of 40 s at 96˚C, 40 s at 49˚C and 1 min at 72˚C and finalised with an elongation step of 5 min at 72˚C. After amplification, the sample was purified from non 16S ribosomal DNA by excision and re-extraction of the DNA from a 0.9% gel (v/w) using the GeneJET gel extraction kit (Thermo Scientific, Waltham, USA) according to manufacturer’s protocol. The final concentration (10.8 ng/µl) was measured using the Qubit dsDNA HS assay kit (Thermo Scientific, Waltham, MA, USA).

Both MDA and DNA-spiking method samples were sequenced using the Personal Genome Machine Ion Torrent (Thermo Scientific, Waltham, MA, USA) as per manufacturer’s protocol. 1 x MDA was sequenced twice, once on a 314v2 chip and once on a 318v2 chip, and reads were combined. Sample 2 x MDA and the DNA-spiked sample were run on a 318v2 chip. All samples were constructed using the Ion PGM™ Sequencing 400 Kit and Ion PGM™ Template OT2 400 kit and sequenced with 850 flow cycles. The raw sequence data were submitted to the European Nucleotide Archive and received accession number PRJEB20134 (http://www.ebi.ac.uk/ena/data/view/PRJEB20134).

Bioinformatics

Trimming

After sequencing, reads from the DNA-spiked sample were trimmed with default quality settings, size 25 to 375 bp. The 1 x MDA sample was sequenced twice and reads from the first run were trimmed with default quality settings, size 25 to 325 bp. Reads from the second 1 x MDA sample run were trimmed with default quality settings, size 25 to 400 bp and 15 bp on 5' end. Reads from the 2 x MDA sample were trimmed with default quality settings, size 25 to 375 bp, using CLC genomics workbench v. 8 (CLCbio, Aarhus, Denmark).

Recovery of viral contigs

To remove genomic DNA from the most abundant microorganism in the bioreactor, the trimmed reads from the DNA-spiked sample were mapped against the genome of "*Candidatus* Methylomirabilis oxyfera" (Ettwig et al., 2010) (length 0.5, similarity 0.85) and the unmapped reads were assembled (length 0.5, similarity 0.95, word size 22, bubble size 276), using CLC genomics workbench v. 8 (CLCbio, Aarhus, Denmark).

The obtained contigs were subsequently mapped against the SILVA database 16S rRNA v119 (Yarza et al., 2008) (length 0.5, similarity 0.7%) and contigs that mapped to "*Ca.* K. stuttgartiensis" were removed from the database, resulting in 4088 remaining contigs. These contigs were checked with ESOM (Ultsch & Moerchen, 2005) (default settings) and from this, seven clustering contigs with a high coverage were obtained and reassembled with SPAdes (v.3.5.0) (Bankevich et al., 2012) using the 'trusted-contigs' and 'careful' settings for those seven contigs.

The reads that were used in the SPAdes re-assembly, were obtained by mapping the spiked DNA to the SILVA database 16S rRNA v119 (length 0.5, similarity 0.7%). The reads that did not map were used.

Reassembly with SPAdes created 2094 contigs, 14 of which were bigger than 5000 bp. From this set of contigs five putative viral genomes could be extracted (197 kbp, 86 kbp, 71 kbp, 41 kbp and 17 kbp).

Assembly of the combined MDA data (Length fraction = 0.5, similarity fraction = 0.9, minimum contig length = 1,500) resulted in 689 contigs, ranging from 130,897 to 1,504 nucleotides. With the use of ESOM, one more putative viral genome was identified (42 kbp).

Differential coverage

For differential coverage, reads from the DNA-spiked sample were mapped against the SILVA database 16S rRNA v119 (length 0.5, similarity 0.7%). Unmapped reads were combined with the reads from the second 1 x MDA run and assembled (length= 0.5, similarity 0.9, word=35, bubble size 271) with. Subsequent mapping of each read set (length 0.5, similarity 0.8%) was performed against the assembled contigs. The depth of both sets was plotted against one another.

Horizontal coverage

To assess how much of each virus was present in each set, the trimmed reads from the three sets were mapped against the six putative viral genomes (length 0.5, similarity 0.95%). The number of mapped reads and total length was normalised to the size of the dataset and the length of virus, respectively.

Results

Viral DNA extracted from bioreactor biomass was sequenced following two different approaches: MDA amplification (two samples) and non-amplified DNA spiked with bacterial 16S rRNA gene DNA. This resulted in three datasets comprising of a total of 774,366 trimmed reads for 1 x MDA and 187,178 trimmed reads for 2 x MDA. After the reads from the non-amplified spiked DNA were trimmed and mapped against the SILVA database, the final number of reads left was 529,481. After trimming, GC graphs were created showing the GC distribution of each dataset (Fig. 1). Clearly visible is the shift in GC content from non-amplified spiked DNA to amplified DNA (1 x MDA, 2 x MDA). The non-amplified set, DNA-spiking method (Fig. 1A), shows a large peak at 63-64% GC content and a shoulder at around 42%. For the MDA method, after the first run of amplification, a large increase of the shoulder at 42% GC content is visible (Fig. 1B) which becomes even more pronounced after another round of amplification (Fig. 1C). The large peak at 63-64% GC content shifts to 57% after one round of amplification. The second round of MDA does not lower the GC content of this peak, but the relative amount of sequences is lowered within the sample.

A total of six different viral sequences (putative genomes) could be assembled using the reads from all three datasets by a combination of various methods (see Materials and methods). The DNA-spiking method resulted in five viral genomes while the MDA set only resulted in one. The length of the viral genomes ranged from 197 kbp to 17 kbp with GC contents from 67% to 35%. Five of the viral genomes (197 kbp, 86 kbp, 71 kbp, 42 kbp and 17 kbp) contained the same sequence on each end of the contig indicating a full circular genome. Fig. 2 shows the comparison of the total amount of virus genomes that could be obtained with the individual datasets. Fig. 2a shows that all six virus genomes could be recovered in nearly complete length from the DNA-spiked dataset. In contrast, MDA amplification clearly lowers the percentage of viral genomes that can be recovered. Fig. 2b shows the percentage of reads from each viral metagenome sample mapping to the six virus genomes. With the MDA amplified samples, the percentages of reads mapping is low with a comparable relative abundance. The non-amplified sample not only shows a lot more difference in relative abundance, but the total amount of mapped reads comprises over half of the original dataset. When the MDA method is used, the number of mapped reads drops to lower than 3%. Fig. 2c shows the average coverage of each viral genome after mapping the reads of the individual datasets. The data are comparable to Fig. 2b. Using the DNA-spiked dataset, high coverage (> 35) is found for five out of six viral genomes. The MDA datasets show very low coverage and only give better results for the 42 kbp virus.

Assembly of the reads from the DNA-spiked sample and the 1 x MDA sample resulted in a total of 1644 contigs. Differential coverage of these contigs, comparing the two datasets is shown in Fig. 3. From the figure it is clear that the depth for each set of contigs differs completely since two similar sets would result in contigs mapping on a diagonal straight line. The two datasets show an almost perpendicular distribution. When looking at the location of contigs with a high and low GC content it becomes clear that high GC% contigs (green) are present within the DNA-spiked dataset but are low in the amplified sample while low GC% contigs (pink) are much more abundant in the 1 x MDA dataset. The figure also demonstrates the wide range in sequencing depth when no amplification was applied (maximum around 380) in clear contrast to amplification, as the sample coverage was lowered tenfold, with a maximum around 38.

Discussion

Here we describe a new method for sequencing low amount of DNA (viral DNA) and compare this to a traditional MDA amplification method. Like in previous reports (Kim & Bae, 2011; Marine et al., 2014), our experiments show the bias that MDA amplification of metavirome or metagenome DNA introduces into the dataset. In the extreme case of applying MDA twice, the dataset is completely changed with major consequences for results obtained, which as such may not reflect the sampled ecosystem. This is emphasising once more the caution that has to be taken when utilizing the MDA technique for environmental DNA samples. Where annotation of the contigs obtained from the unamplified DNA-spiked sample showed a majority of virus related genes, the MDA datasets showed a general shift towards bacterial genes (data not shown). In the MDA datasets, the genes match most closely to WS6-like bacteria, a group of unusually small bacteria with small genomes (~0.1 Mbp) and (in this case) low GC content (Speth et al, 2016). Considering the small size of these bacteria it is very likely they were not removed during the filtration process used for selecting viral particles.

Our results illustrate that specifically DNA with a low GC content is favoured in MDA, which can lead to a considerable shift of GC content causing a severe underestimation of the quantity of viruses with a higher GC percentage in the sample. Viruses with higher GC content can be easily overlooked.

Differential binning of spiked DNA and the 1 x MDA not only shows a remarkable difference between both samples, but also demonstrates the loss of information about the abundance, as the range of this abundance of the different viruses shifts from 380 to 38. This shift in depth means that using MDA as an abundance indicator would have been impossible in this example and as a consequence we used the abundance of the spiked DNA in recovery of the viral genomes.

A nice side effect of the spiking method is the lowering of contamination risk with DNA from the lab. By choosing the 16S rRNA gene of a microorganism which is very uncommon to the laboratory for spiking, contamination can easily be recognized and filtered out. One could even use eukaryotic DNA encoding the 18S rRNA gene in a prokaryotic-based lab and vice versa.

In this report we described an extremely low input method (only 0.1 ng of DNA needed) for viral metagenome sequencing that is unbiased, inexpensive, easy and readily available for any lab with sequence facilities and that can possibly be extended for other non-16S/18S containing DNA like plasmids, or other Next Generation Sequencing platforms like MinIon or Illumina.

Conclusions

When dealing with low quantities of DNA for Next Generation Sequencing, multiple displacement amplification (MDA) of the DNA from the sample might not be the method of choice to obtain enough starting material. We propose to use the DNA-spiking method, where one adds DNA with a known sequence, as a more valid alternative. The reads resembling the added DNA can be easily discarded afterwards. This new method is not only technically feasible, but also results in a very low bias in the dataset compared to the MDA method. The obvious bias of the MDA method has to be taken into consideration since it may cause major shifts in the DNA profile of an ecosystem.

Acknowledgements

We thank Simon Guerrero for supplying the bioreactor material. LG is supported by ERC Advanced grant (ERC-339880), GC by a Gravitation grant of The Netherlands Organisation for Fundamental Research (SIAM024002002). HOdC is supported by ERC Advanced grant (ERC-669371).

References

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, and Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:455-477. 10.1089/cmb.2012.0021

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, and Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99:14250-14255. 10.1073/pnas.202488399

Brum JR, and Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. Nature Reviews Microbiology 13:147-159. 10.1038/nrmicro3404

Clokie MR, Millard AD, Letarov AV, and Heaphy S. 2011. Phages in nature. Bacteriophage 1:31-45. 10.4161/bact.1.1.14942

Cunningham BR, Brum JR, Schwenck SM, Sullivan MB, and John SG. 2015. An inexpensive, accurate, and precise wet-mount method for enumerating aquatic viruses. Applied and Environmental Microbiology 81:2995-3000. 10.1128/Aem.03642-14

Duhaime MB, Deng L, Poulos BT, and Sullivan MB. 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. Environmental Microbiology 14:2526-2537. 10.1111/j.1462-2920.2012.02791.x

Ettwig KF, Butler MK, Le Paslier D, Pelletier E, Mangenot S, Kuypers MM, Schreiber F, Dutilh BE, Zedelius J, de Beer D, Gloerich J, Wessels HJ, van Alen T, Luesken F, Wu ML, van de Pas-Schoonen KT, Op den Camp HJM, Janssen-Megens EM, Francoijs KJ, Stunnenberg H, Weissenbach J, Jetten MSM, and Strous M. 2010. Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. Nature 464:543-548. 10.1038/nature08883

Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399:541-548. Doi 10.1038/21119

Gambelli L, Cremers G, Mesman R, Guerrero S, Dutilh BE, Jetten MSM, Op den Camp HJM, and van Niftrik L. 2016. Ultrastructure and viral metagenome of bacteriophages from an anaerobic methane oxidizing Methylomirabilis bioreactor enrichment culture. Frontiers in Microbiology 7:1740. 10.3389/fmicb.2016.01740

Guo P, El-Gohary Y, Prasadan K, Shiota C, Xiao XW, Wiersch J, Paredes J, Tulachan S, and Gittes GK. 2012. Rapid and simplified purification of recombinant adeno-associated virus. Journal of Virological Methods 183:139-146. 10.1016/j.jviromet.2012.04.004

Kim KH, and Bae JW. 2011. Amplification methods bias metagenomic Libraries of uncultured single-stranded and double-stranded DNA viruses. Applied and Environmental Microbiology 77:7663-7668. 10.1128/Aem.00289-11

Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, and Wommack KE. 2014. Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. Microbiome 2. Artn 310.1186/2049-2618-2-3

Rosario K, and Breitbart M. 2011. Exploring the viral world through metagenomics. Current Opinion in Virology 1:289-297. 10.1016/j.coviro.2011.06.004

Speth DR, In 't Zandt MH, Guerrero-Cruz S, Dutilh BE, and Jetten MS. 2016. Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. Nature Communications 7:11172. 10.1038/ncomms11172

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, and Poisson G. 2013. Are we missing half of the viruses in the ocean? ISME Journal 7:672-679. 10.1038/ismej.2012.121

Thurber RV, Haynes M, Breitbart M, Wegley L, and Rohwer F. 2009. Laboratory procedures to generate viral metagenomes. Nature Protocols 4:470-483. 10.1038/nprot.2009.10

Ultsch A, Moerchen F 2005. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM, Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46

Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glockner FO, and Rossello-Mora R. 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and Applied Microbiology. 31:241-250. 10.1016/j.syapm.2008.07.001

Yilmaz S, Allgaier M, and Hugenholtz P. 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nature Methods 7:943-944. 10.1038/nmeth1210-943

Legends to the figures.

Figure 1: Distribution of reads obtained from Ion Torrent sequencing using three different sample preparation methods based on their individual GC content in % of the total number of reads in one sample. **A.** DNA-spiking method; **B.** 1 x MDA method; **C.** 2 x MDA method. GC graphs were created using CLCgenomics workbench.

Figure 2: Comparison of the viral reads from the individual datasets mapping to the six assembled virus genomes (Green, DNA-spiked sample; Blue, 1 x MDA; Red, 2 x MDA). **A.** Horizontal coverage of the viral genomes with reads from the individual datasets. **B.** Distribution of the reads from the individual datasets over the assembled viral genomes. **C.** Depth (vertical coverage) of the viral genomes with reads from the individual datasets as a measure of abundance.

Figure 3: Differential coverage of the viral contigs assembled using a combination of the DNA-spiked sample and the 1 x MDA sample with each individual read sets. Each circle represents a contig present after assembly and the placement in the plot shows the abundance of the contig for each read set. Two similar read sets would result in a diagonal straight line. GC content of the different contigs is indicated as depicted in the colour scale. Three outliers caused by the MDA amplification method are not shown in the plot.

Figure 1(on next page)

Figure 1: Distributionof reads obtained from Ion Torrent sequencing using three different samplepreparation methods based on their individual GC content in % of the totalnumber of reads in one sample.

A. DNA-spiking method; **B.** 1 x MDA method; **C.** 2 x MDA method. GC graphs were created using CLCgenomics workbench.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2912v1 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

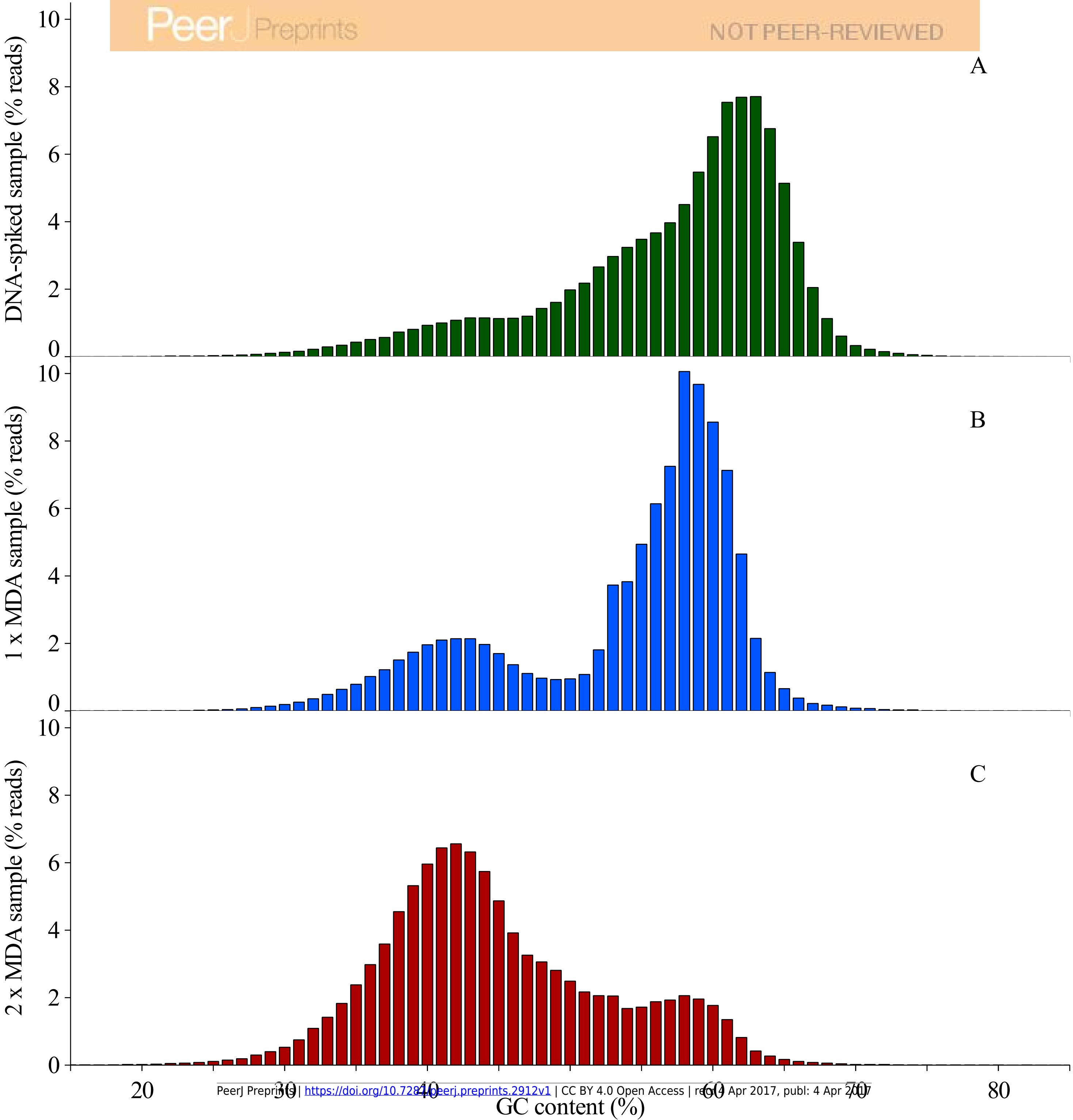

A
B
C
DNA-spiked sample (% reads)
1 x MDA sample (% reads)
2 x MDA sample (% reads)
0
2
4
6
8
10
20
30
40
60
70
80
GC content (%)

Figure 2(on next page)

Figure 2: Comparisonof the viral reads from the individual datasets mapping to the six assembled virusgenomes.

(Green, DNA-spiked sample; Blue, 1 x MDA; Red, 2 x MDA). **A.** Horizontal coverage of the viral genomes with reads from the individual datasets. **B.** Distribution of the reads from the individual datasets over the assembled viral genomes. **C.** Depth (vertical coverage) of the viral genomes with reads from the individual datasets as a measure of abundance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2912v1 | CC BY 4.0 Open Access | rec: 4 Apr 2017, publ: 4 Apr 2017

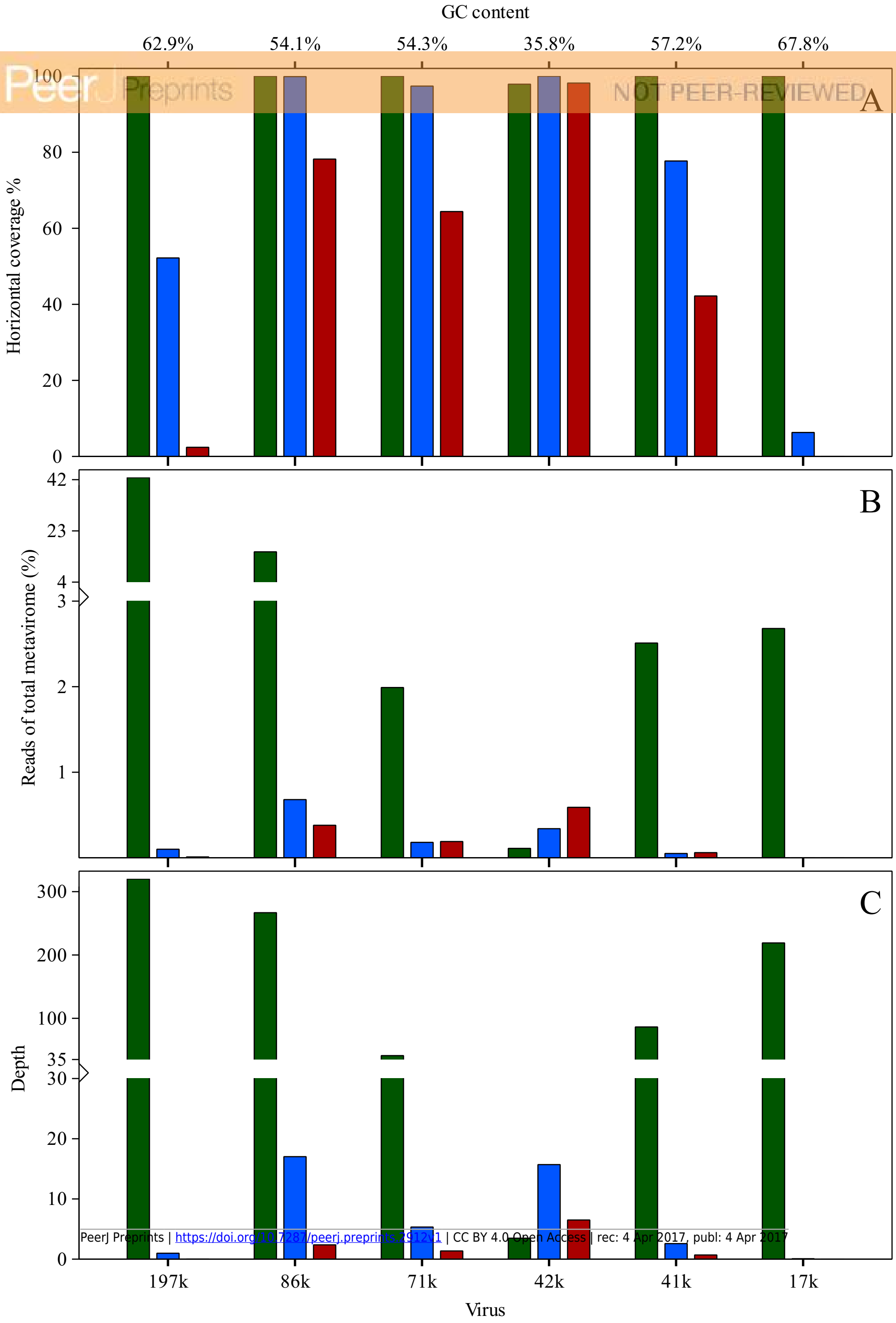
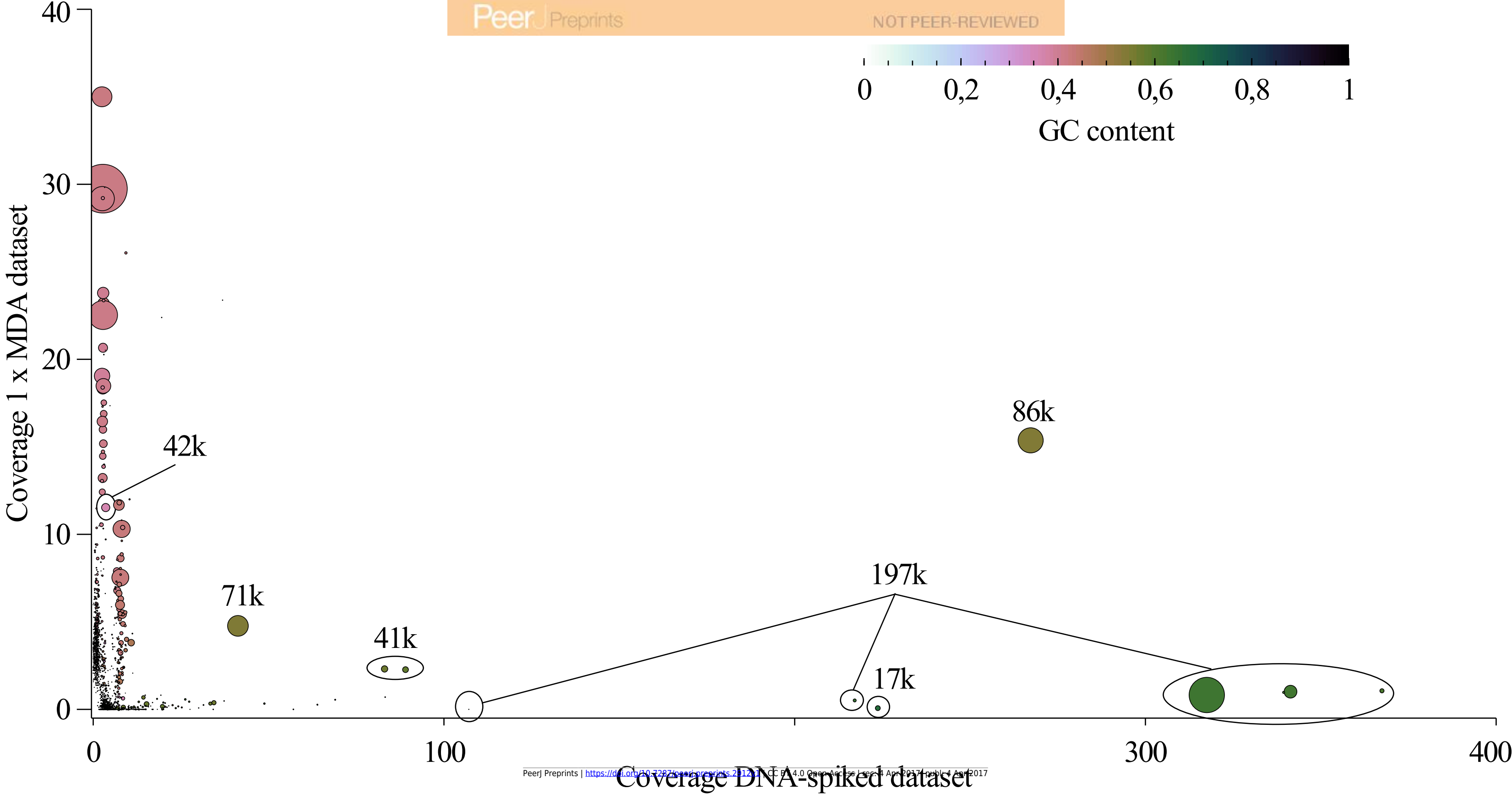

GC content
62.9%
54.1%
54.3%
35.8%
57.2%
67.8%
A
Horizontal coverage %
0
20
40
60
80
100
B
Reads of total metavirome (%)
1
2
3
4
23
42
C
Depth
0
10
20
30
35
100
200
300
197k
86k
71k
42k
41k
17k
Virus

Figure 3(on next page)

Figure3: Differential coverage of the viral contigs assembled using a combination of theDNA-spiked sample and the 1 x MDA sample with each individual read sets.

Each circle represents a contig present after assembly and the placement in the plot shows the abundance of the contig for each read set. Two similar read sets would result in a diagonal straight line. GC content of the different contigs is indicated as depicted in the colour scale. Three outliers caused by the MDA amplification method are not shown in the plot.


Coverage 1 x MDA dataset
40
30
20
10
0
Coverage DNA-spiked dataset
0
100
300
400
0
0,2
0,4
0,6
0,8
1
GC content
42k
71k
41k
86k
197k
17k