Viruses Hot Paper

Deutsche Ausgabe: DOI: 10.1002/ange.201507047
Internationale Ausgabe: DOI: 10.1002/anie.201507047

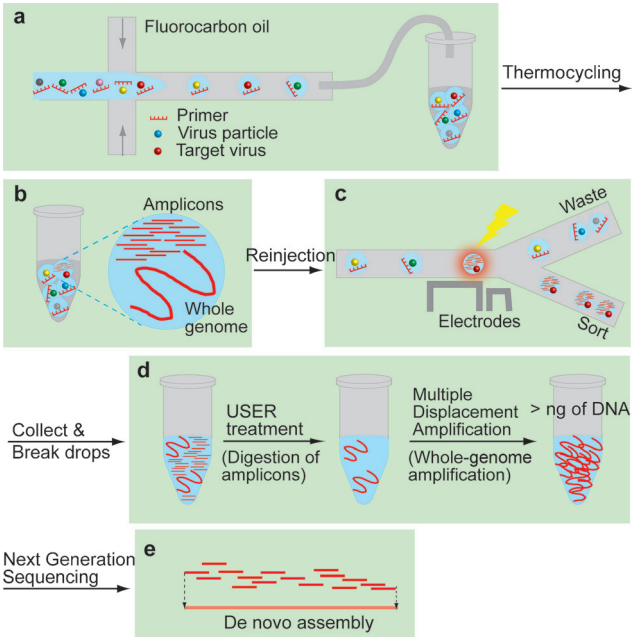
Whole-Genome Sequencing of a Single Viral Species from a Highly Heterogeneous Sample

*Hee-Sun Han, Paul G. Cantalupo, Assaf Rotem, Shelley K. Cockrell, Martial Carbonnaux, James M. Pipas, and David A. Weitz**

Abstract: *Metagenomic studies suggest that only a small fraction of the viruses that exist in nature have been identified and studied. Characterization of unknown viral genomes is hindered by the many genomes populating any virus sample. A new method is reported that integrates drop-based microfluidics and computational analysis to enable the purification of any single viral species from a complex mixed virus sample and the retrieval of complete genome sequences. By using this platform, the genome sequence of a 5243 bp dsDNA virus that was spiked into wastewater was retrieved with greater than 96% sequence coverage and more than 99.8% sequence identity. This method holds great potential for virus discovery since it allows enrichment and sequencing of previously undescribed viruses as well as known viruses.*

Viruses are the most abundant biological entities on Earth and significantly affect living organisms by causing diseases and shaping their immune systems. Despite their ubiquity and influence, less than 0.01% of existing viruses have been sequenced.[1] Establishment of an extensive virus database is crucial to identify potential emerging infectious diseases[2] and to improve our understanding of virus diversity, ecology, adaptation, and evolution. The major roadblock to characterizing unknown viral genomes is the lack of technologies enabling efficient enrichment of various types of viruses. Enrichment of a target viral species is required for the most common virus samples such as environmental samples, which generally harbor diverse viral populations,[3] or clinical samples, where the amount of viral genomes is often lower than the amount of host genomes and the virions are localized to a small subset of cells in the tissue. An enrichment step is particularly crucial for viral genome sequencing because other abundant DNAs in the sample, such as genomic fragments of host DNA, are often much larger than viral genomes and dominate the sequence space even with a small number of copies. Traditional enrichment methods for viruses include cell culture,[4] immunoscreening[5] followed by sequence-independent PCR,[6] and differential hybridization.[7] All of these methods are labor-intensive, inefficient, and more importantly, only applicable to a limited subset of viruses. Recently, a flow cytometry method was developed to disperse single virions into microwells and obtain their individual genome sequences.[8] However, this method does not employ a selection strategy. A selection strategy allows efficient usage of sequencing power and enables the sequencing of rare viruses with a reasonable sequencing cost and time.

Herein, we report the development of a platform to isolate and sequence any single viral species from a large genetic space of viral sequences. Our platform integrates drop-based microfluidics for high-throughput single-virus assays and sorting (Scheme 1 a–c), molecular biology techniques for whole-genome amplification of the selected viruses (Scheme 1 d), and a computational workflow for de novo assembly of viral genome sequences (Scheme 1 e).



Scheme 1. Identification of viral genomes from an environmental sample. a) Viruses are encapsulated into drops. b) After thermocycling, amplicons are generated in the drops containing the target virus. c) Drops are sorted based on their fluorescence intensities. d) The enriched virus solution is treated with USER to digest amplicons, followed by whole-genome amplification. e) The amplified products are sequenced and assembled by using our computational workflow.

Drop-based microfluidics offers unprecedented advantages, allowing high-throughput screening of single cells or viruses with high sensitivity and minimal time.[9] By combin-

[*] Dr. H.-S. Han, Dr. A. Rotem, M. Carbonnaux, Prof. D. A. Weitz
Department of Physics, School of Engineering and Applied Sciences
Harvard University, Cambridge, MA 02138 (USA)
E-mail: weitz@seas.harvard.edu
P. G. Cantalupo, Dr. S. K. Cockrell, Prof. J. M. Pipas
Department of Biological Sciences
University of Pittsburgh, Pittsburgh, PA 15260 (USA)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201507047.

ing droplet digital PCR (ddPCR) with microfluidic sorting techniques, we selectively enrich viruses of interest from an environmental sample. A drop maker is used to encapsulate an aqueous mixture containing a virus sample, PCR buffer, primers, dUTP/dATP/dCTP/dGTP, and SYBR Green I. Encapsulation is performed so that the majority of drops contain no more than one virion. All virus types can be efficiently encapsulated into drops following Poisson statistics since the size of viruses (5 nm–3 μm) is considerably smaller than the size of the drops (25–100 μm), and virions are well-dispersed in wastewater samples (Figure S3). With primers specific to a target virus, amplicons are generated selectively in a drop containing a target virion after thermocycling. Primers can be designed for both known viruses and unknown viruses (Figure S4, see the Supporting Information for details). In the PCR mixture, dTTP is replaced with dUTP so that amplicons generated during PCR contain dUTP in the place of dTTP. A dsDNA intercalating dye, SYBR Green I, is added to the mixture to stain amplicons and identify the drops containing a target virion.

After thermocycling, we inject the drops into the microfluidic sorter and select the drops displaying an enhanced fluorescence signal.[9c] The selected virus solution is then treated with uracil-specific excision reagent (USER), which generates a nucleotide gap at the location of a uracil. During USER treatment, amplicons that contain dUTP are selectively digested into small pieces while the viral genomes that do not contain dUTP remain intact. Selective digestion of amplicons is important since they are over-represented in the enriched virus solution and would impede the characterization of the complete viral genome if left undigested.

Multiple displacement amplification (MDA) is performed on the USER-treated virus solution to generate more than 1 ng of clonal copies, an amount sufficient for DNA sequencing.[10] To suppress the amplification of high-molecular-weight DNA contaminants in commercial MDA reagents,[11] we treat the reagents with UV light[12] and perform MDA reaction in a reduced reaction volume[13] of 4 μL.

The amplified DNA products are sequenced by using an Illumina Platform and the sequence reads are assembled to recover the genome sequence of the sorted viruses. A computational workflow was developed for sequence data cleaning, de novo assembly, and selection of the target virus contig (a contiguous assembled piece of DNA sequence). Low-quality reads and human reads are removed and the remaining reads are assembled into contigs by using CLC assembler (CLC bio). Meta-assembly is then performed by using a "homebuilt" MATLAB code since CLC assembler occasionally yields contigs with 13–50 bp sequence overlaps. The genome sequence of the enriched viruses is determined by selecting the contig with the highest relative abundance among the high-quality contigs that do not originate from known organisms.

Previous metagenomic studies have revealed that wastewater harbors diverse viral species, including previously undescribed viruses with unknown viral sequences that outnumber known viral sequences by a factor of 10–1000.[3] This indicates that raw sewage is an excellent source of previously undescribed viruses. To assess our platform for sequencing viral genomes from wastewater, we spiked a well-characterized virus, SV40, into wastewater, selected SV40 virions by using the microfluidic platform, sequenced the enriched samples, and performed de novo assembly. We then investigated whether the assembled genome sequence aligns with the SV40 genome and how much of the SV40 genome is covered by this sequence.

For enrichment of SV40, SV40/wastewater mixtures are encapsulated into 8 pL drops (Figure 1a). After thermocycling, a small fraction of the drops displays a high level of

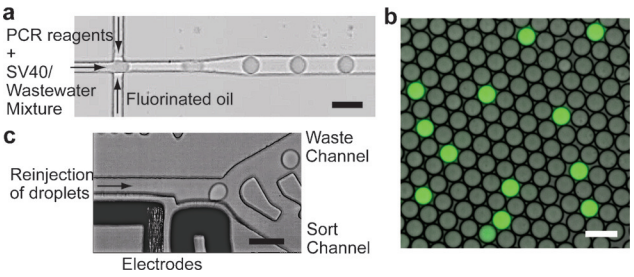


Figure 1. a) A partial image of the drop-making device. b) A fluorescence image of drops after PCR. The fluorescence signal is from intercalation of SYBR green I into amplicons. c) A partial image of the sorting device. Drops are sorted based on their fluorescence intensities. Scale bars: 50 μm.

fluorescence under excitation at 470 nm, thereby indicating the presence of SV40 virions in those drops (Figure 1b). These drops are selected by using the microfluidic sorter (Figure 1c). Amplicons in the collected solution are digested with USER (Figure 2a), followed by whole-genome amplification with ϕ29 DNA polymerase. Restriction analysis allows us to quickly examine whether the amplified DNA contains the SV40 genome sequence without sequencing. The restriction enzyme PvuII cleaves SV40 at three sites to produce DNA fragments of 1446, 1790, and 1997 bp (Figure 2b). The digestion test shows that at least 100 positive drops are required to obtain a detectable amount of the SV40 genome sequence after MDA (Figure 2c). With fewer drops, the amplified DNA is not cleaved by PvuII and migrates as large DNA during gel electrophoresis.

To further evaluate the effectiveness of our purification and amplification method, the amplified DNA was sequenced and analyzed. First, the percentage of SV40 mapped reads to total sequence reads was compared for 3% SV40/wastewater mixture and the selected SV40 samples (Sample 1, Table S1). Without sorting, the amplified sequences contain 0.004% of SV40 reads. The low percentage of SV40 reads is due to high-molecular-weight DNA in wastewater. During MDA, larger DNA is preferentially amplified over shorter DNA; the sequences of high-molecular-weight contaminants are thus over-represented in the amplified product. This issue is particularly prominent and problematic in viral genome sequencing becasue viral genomes are generally much smaller than other abundant DNA in the sample. The percentage of SV40 reads increases to 94.0% and 98.6% for samples of 100 and 1000 selected drops, respectively. The sequencing results on the selected SV40 samples from other sets of sorting

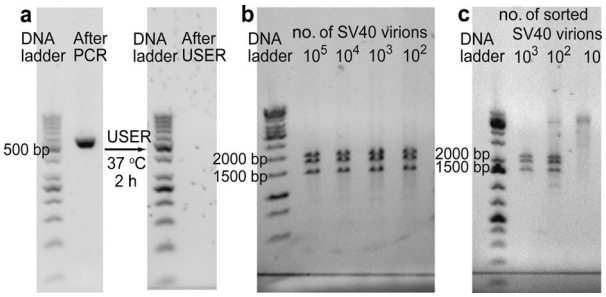


Figure 2. Gel electrophoresis was used to analyze the DNA fragments present in the samples at different stages of the process. a) After PCR, 519 bp amplicons are produced in the drops containing SV40 virions. These amplicons incorporate dUTP and are digested after USER treatment. b, c) Restriction analysis was used to verify successful amplification of the SV40 genome. b) After PvuII treatment, the genomic copies of SV40 are cleaved into three fragments of 1446, 1790, and 1997 bp. c) At least 100 PCR-positive drops are required to obtain a detectable amount of the SV40 genome sequence by MDA.

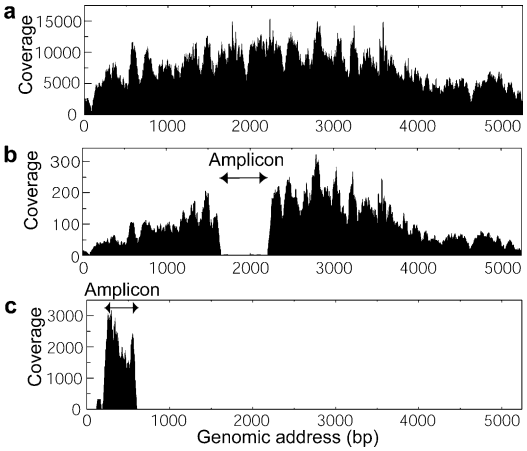


Figure 3. a) A representative sequence-coverage graph for the enriched SV40 samples. The sequence reads cover the entire SV40 genome. b) A sequence-coverage graph for the sample lacking a large portion of the amplicon sequence (1761–2259). c) A sequence-coverage graph for the sample where the 259 bp (237–495) amplicons are not digested before MDA. Without USER treatment, only amplicon sequences are recovered after sequencing.

experiments also show high percentages of SV40 reads (Table S1 in the Supporting Information). This confirms that our microfluidic platform effectively enriches target virions from a complex virus mixture. We also investigated whether the SV40 reads cover the complete genome of SV40. The reference mapping results show that the complete genome sequence of SV40 is recovered after collecting as little as 100 positive drops (Figure 3a and Figure S5). In two out of eleven samples, a portion of the SV40 genome corresponding to the amplicon sequence is absent (Figure 3b and Figure S5). This absence may be due to the digested amplicon fragments, which can anneal to the SV40 genome and hinder DNA strand extension by ϕ29. The sequence mapping results highlight the importance of the amplicon digestion step for whole-genome amplification. When the sorted virions are amplified without USER treatment, the amplified sequences only cover sequences from the amplicons themselves (Figure 3c).

The sequence reads are assembled into contigs by using our computational workflow. The contigs are analyzed by a BLAST search to identify their origin and sequence homology to the reference genome. In all cases, the contig with the highest relative abundance aligns with the SV40 genome (Table S2). When the sequence reads cover the complete SV40 genome, the generated SV40 contigs cover more than 97% of the genome sequence (Figure 3a and Table 1). The missing sequences are in the 72 bp repeat region of the SV40 genome. In our study, sequencing was performed with 50 bp single-end read, which makes it challenging to distinguish repeating sequences larger than 50 bp. Therefore, the sequence coverage can be further improved by performing paired-end sequencing with longer reads. We obtain SV40 contigs that cover more than 96% of the genome sequence even from the samples lacking a large portion of the amplicon sequence (Figure 3b and Table 1). To correct the absence of the amplicon sequence in these contigs, we characterized the amplicon sequence by performing Sanger sequencing on the sorted drops and manually inserted that sequence based on the sequence overlap. Our assembly technique is highly robust towards contamination. The samples amplified with the MDA reagents that were incompletely decontaminated showed only 8.9% and 0.74% SV40 reads; however, the SV40 contigs from these samples cover more than 95.8% of the genome with 100% sequence identity (Table 1). Sequence analysis on non-SV40 reads shows that over 99% of the contaminant sequences are from human, bacteria, and mouse, which are common contaminant sources in laboratory environments (Figure S6). For de novo assembly of unknown viral genomes, raw sequencing data will be computationally cleaned by removing the common contaminant sequences. The high-purity sequence reads will ensure accurate de novo assembly. Our assembly result indicates that a cheaper

Table 1: Summary of the SV40 contigs generated from the representative enriched virus samples.

Sample	Sorted drops	SV40 reads/ Total reads [%]	Contig length [bp]	Sequence Coverage [%]	Sequence Identity [bp]	Missing Sequence[c] [locus]
1–1[a]	100	94.0	5066	96.6	5066	124–207 1648–1740
1–2	1000	98.6	5162	98.5	5162	125–205
2–1	1000	95.8	5134	97.9	5134	124–232
2–2	5000	98.8	5187	98.9	5186	151–207
3–1[b]	100	8.9	5159	98.4	5158	124–207
3–2[a,b]	1000	0.74	5016	95.7	5016	99–232 1648–1740

[a] The sequence reads from these samples lack a large portion of the amplicon sequence (1761–2259 bp). [b] MDA reagents used for these samples were incompletely decontaminated. [c] Two 72 bp repeats reside in 107–250 bp.

sequencing option can also be used for sequencing. De novo assembly was successful when the sequence coverage was higher than 50× (Table S1). Considering that the size of viral genomes is a few kb to hundreds of kb, the Ion Torrent Personal Genome Machine, which reads 10 million bases per run, can be used for sequencing a small number of virus samples.

In summary, by using this platform, we were able to retrieve more than 97% of the target genome sequence after collecting as little as 100 virions. The total mass of DNA in 100 SV40 virions is 5×10^{-16} g. This value is orders of magnitude less than the amount of DNA in a single bacterium, which has the smallest genome among cells used for single-cell sequencing.

Even though the platform was developed for DNA viruses, RNA viruses can also be enriched and sequenced. For RNA virus sequencing, the cDNA copies of genomes are encapsulated into drops instead of the genome itself. The genomic cDNA copies are synthesized by reverse-transcription with either oligo-dT or random hexamers. We synthesized genomic cDNA copies for norovirus and showed that various segments of the genome can be amplified from these cDNA copies (Table S4 and Figure S7). The genomic cDNA copies are then encapsulated into drops. After thermocycling, a few drops display a high level of fluorescence under excitation at 470 nm (Figure S8). The PCR-positive drops can then be selected and processed following the same protocol as for DNA viruses.

Finally, we demonstrated the potential of our platform for the sequencing of unknown viruses by designing primers specific to a previously undescribed virus related to human rhinovirus and performing ddPCR. From metagenomic analysis on a wastewater sample,[3] we identified contigs showing 50–65% sequence homology to human rhinovirus. We designed primers for each contig (Table S5) and performed PCR on the reverse-transcribed wastewater sample. The size of amplicons generated from PCR agrees with the size predicted from the contig sequences (Figure S9). This result confirms the validity of our method for designing primers for unknown viruses. We then performed ddPCR with the designed primers and showed that a small fraction of drops display a high level of fluorescence signal under excitation at 470 nm (Figure S10). These positive drops can be selected by using our microfluidic sorter and the target viral genome can be sequenced by following the same protocol as for the known viruses.

Our platform enables efficient isolation of single viral species from a mixture of other viruses and DNA contaminants. This eliminates the requirement for cell culture to prepare a homogeneous virus solution and allows genome sequencing of uncultivable viruses from an environmental sample. It thus holds great potential for virus discovery and the establishment of an extensive genomic database. A comprehensive virus database will facilitate the interpretation of metagenomic data by providing reference genomes; lead to a better understanding of virus diversity, ecology, adaptation, and evolution; and enable the prediction of emerging infectious diseases caused by viruses. Our work represents an important step towards exploration of the viral universe. We expect extensions of this work to enable faster sorting for the isolation of rare viruses and the sequencing of individual viruses will further enhance the power of this new platform.

Acknowledgements

The research work of the authors is supported by the U.S. National Institute of Health grants, R21-AI101291 (D.A.W. and J.M.P.) and by Defense Advanced Research Projects Agency, HR0011-11-C-0093 (D.A.W. and J.M.P.).

Keywords: genome sequencing · high-throughput screening · microemulsions · microfluidics · viruses

How to cite: *Angew. Chem. Int. Ed.* **2015**, *54*, 13985–13988
Angew. Chem. **2015**, *127*, 14191–14194

[1] S. J. Anthony, J. H. Epstein, K. A. Murray, I. Navarrete-Macias, C. M. Zambrana-Torrelio, A. Solovyov, R. Ojeda-Flores, N. C. Arrigo, A. Islam, S. Ali Khan, P. Hosseini, T. L. Bogich, K. J. Olival, M. D. Sanchez-Leon, W. B. Karesh, T. Goldstein, S. P. Luby, S. S. Morse, J. A. K. Mazet, P. Daszak, W. I. Lipkin, *mBio* **2013**, *4*, e00598-13.
[2] K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, P. Daszak, *Nature* **2008**, *451*, 990–993.
[3] P. G. Cantalupo, B. Calgua, G. Zhao, A. Hundesa, A. D. Wier, J. P. Katz, M. Grabe, R. W. Hendrix, R. Girones, D. Wang, J. M. Pipas, *mBio* **2011**, *2*, e00180-11.
[4] D. S. Leland, C. C. Ginocchio, *Clin. Microbiol. Rev.* **2007**, *20*, 49–78.
[5] Q. Choo, G. Kuo, A. Weiner, L. Overby, D. Bradley, M. Houghton, *Science* **1989**, *244*, 359–362.
[6] a) G. R. Reyes, J. P. Kim, *Mol. Cell. Probes* **1991**, *5*, 473–481; b) P. Froussard, *Nucleic Acids Res.* **1992**, *20*, 2900.
[7] N. Lisitsyn, N. Lisitsyn, M. Wigler, *Science* **1993**, *259*, 946–951.
[8] L. Z. Allen, T. Ishoey, M. A. Novotny, J. S. McLean, R. S. Lasken, S. J. Williamson, *PLoS One* **2011**, *6*, e17722.
[9] a) E. Brouzes, M. Medkova, N. Savenelli, D. Marran, M. Twardowski, J. B. Hutchison, J. M. Rothberg, D. R. Link, N. Perrimon, M. L. Samuels, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 14195–14200; b) J. J. Agresti, E. Antipov, A. R. Abate, K. Ahn, A. C. Rowat, J.-C. Baret, M. Marquez, A. M. Klibanov, A. D. Griffiths, D. A. Weitz, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4004–4009; c) L. Mazutis, J. Gilbert, W. L. Ung, D. A. Weitz, A. D. Griffiths, J. A. Heyman, *Nat. Protoc.* **2013**, *8*, 870–891; d) B. L. Wang, A. Ghaderi, H. Zhou, J. Agresti, D. A. Weitz, G. R. Fink, G. Stephanopoulos, *Nat. Biotechnol.* **2014**, *32*, 473–478.
[10] a) K. Zhang, A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm, G. M. Church, *Nat. Biotechnol.* **2006**, *24*, 680–686; b) Y. Marcy, T. Ishoey, R. S. Lasken, T. B. Stockwell, B. P. Walenz, A. L. Halpern, K. Y. Beeson, S. M. D. Goldberg, S. R. Quake, *PLoS Genet.* **2007**, *3*, e155; c) S. Rodrigue, R. R. Malmstrom, A. M. Berlin, B. W. Birren, M. R. Henn, S. W. Chisholm, *PLoS One* **2009**, *4*, e6864.
[11] P. C. Blainey, S. R. Quake, *Nucleic Acids Res.* **2011**, *39*, e19.
[12] T. Woyke, A. Sczyrba, J. Lee, C. Rinke, D. Tighe, S. Clingenpeel, R. Malmstrom, R. Stepanauskas, J.-F. Cheng, *PLoS One* **2011**, *6*, e26161.
[13] C. A. Hutchison, H. O. Smith, C. Pfannkoch, J. C. Venter, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17332–17336.

Received: July 29, 2015
Published online: August 28, 2015

