analytical chemistry

pubs.acs.org/ac Article

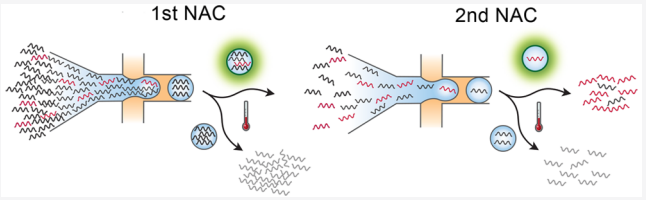
Sequencing Ultrarare Targets with Compound Nucleic Acid Cytometry

Chen Sun, Kai-Chun Chang, and Adam R. Abate*

Cite This: *Anal. Chem.* 2021, 93, 7422−7429 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Targeted sequencing enables sensitive and cost-effective analysis by focusing resources on molecules of interest. Existing methods, however, are limited in enrichment power and target capture length. Here, we present a novel method that uses compound nucleic acid cytometry to achieve million-fold enrichments of molecules >10 kbp in length using minimal prior target information. We demonstrate the approach by sequencing HIV proviruses in infected individuals. Our method is useful for rare target sequencing in research and clinical applications, including for identifying cancer-associated mutations or sequencing viruses infecting cells.



INTRODUCTION

Target enrichment focuses valuable sequencing on important molecules and is useful when the sample comprises a large background of uninteresting DNA.[1] For instance, characterizing HIV genomic diversity is important for understanding persistent infection, but under treatment viral DNA is outnumbered by human DNA by billions of times.[2−4] In metagenomic analyses, organisms of interest may be present at a few percent,[5−7] while in a human genetic disease, variants may be present at fractions of a percent.[8−10] In instances such as these, sequencing all DNA is wasteful because only a fraction of reads corresponds to the region of interest. The most common target enrichment strategies are based on polymerase chain reaction (PCR) amplification or hybridization capture.[1,11,12] PCR methods recover only the amplified portion and miss information beyond primers.[13,14] Hybridization capture recovers information extending beyond probes but can require hundreds of probes;[15,16] this necessitates considerable prior information for probe design, which is often unavailable, especially when little information is known about the region of interest, such as in the novel microbe or genetic lesion sequencing.[17,18]

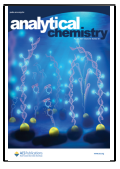
Nucleic acid cytometry (NAC) is a conceptually novel approach to target enrichment based on droplet microfluidics.[19] The overarching principle is to physically isolate molecules by hydrodynamic sorting. Target identification is accomplished using droplet PCR, while isolation is accomplished by sorting positive droplets.[19−22] The approach is akin to querying a diverse mixture for keyword subsequences and isolating all molecules containing the keyword. The critical factor in NAC enrichment is sensitivity for recovering the target of interest. Sensitivity, in turn, is limited by the number of droplet PCRs that can be sorted which, presently, is ∼10 million. Considering losses in DNA recovery and the need for sufficient material to perform sequencing, current enrichments are capped to ∼30,000, allowing NAC to maximally concentrate the target by this factor.[5,8,23,24] This enrichment is insufficient for applications with ultrarare targets below one in a million. To broaden the applicability of NAC, a strategy to increase enrichment power is needed.

In this study, we demonstrate the ability to perform NAC repeatedly on a sample to achieve compound enrichment over multiple rounds. The final enrichment is the product of each round, allowing a ∼6 million-fold enrichment over two rounds. This is ∼200-fold higher than enrichments with the next best technology, and it affords the ability of recovering single virus genomes.[8,15,23] To demonstrate the approach, we use it to isolate and sequence single HIV genomes from infected individuals. Because of their limited enrichment power, previous approaches lack the sensitivity to recover and sequence such rare single virus genomes. Compound NAC provides a general platform for recovering long, ultrarare molecules with minimal prior sequence information.

EXPERIMENTAL SECTION

Microfluidic Device Fabrication. The microfluidic devices were fabricated in polydimethylsiloxane (PDMS) using standard soft lithography. Photomasks designed by AutoCAD were printed on transparencies and the features on the photomask were transferred to a silicon wafer (University

Received: November 10, 2020
Accepted: March 19, 2021
Published: May 10, 2021



https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422−7429

Wafer) using a negative photoresist (MicroChem, SU-8 2025) by UV photolithography. A PDMS (Dow Corning, Sylgard 184) prepolymer mixture of polymer and cross-linker at a ratio of 10:1 was poured over the pattered silicon wafer and cured in a 65 °C oven for 2 h. The PDMS replica was peeled off and punched for inlets and outlets by a 0.75 mm biopsy core (World Precision Instruments). The PDMS slab was bound to a clean glass using an oxygen plasma cleaner (Harrick Plasma), followed by baking at 65 °C for 30 min to ensure strong bonding between the PDMS and glass. The microfluidic channels were treated with Aquapel (PPG Industries) and baked at 65 °C overnight for hydrophobicity.

Droplet TaqMan PCR. 50 fg of ΦX 174 virion DNA and 500 ng of lambda DNA (New England BioLabs) were added to 150 μL of PCR reagents containing 1× Platinum Multiplex PCR Master Mix (Life Technologies, catalog no. 4464269), 200 nM TaqMan probe (IDT), 1 μM forward primer and 1 μM reverse primer (IDT), 2.5% (w/w) Tween 20 (Fisher Scientific), 2.5% (w/w) poly(ethylene glycol) 6000 (Sigma-Aldrich), and 0.8 M 1,2-propanediol (Sigma-Aldrich). Tween 20 and poly(ethylene glycol) 6000 were used to increase the stability of droplets during thermal cycling.[22] 1,2-Propanediol was used as a PCR enhancer when low temperature was used for denaturation.[25] Two syringes backfilled with HFE-7500 fluorinated oil (3M, catalog no. 98-0212-2928-5) were loaded with (1) TaqMan PCR reaction mix and (2) HFE-7500 oil with a 2% (w/w) PEG−PFPE amphiphilic block copolymer surfactant (RNA Biotechnologies, catalog no. 008-FluoroSurfactant-1G). The aqueous phase and oil phase were injected into a flow-focus droplet maker at controlled flow rates (400 μL/h for the PCR mix and 800 μL/h for the oil phase) sustained by computer-programmed syringe pumps (New Era). ∼3 million monodispersed droplets (diameter ∼40 μm) were generated and collected in PCR tubes *via* polyethylene tubing. The bottom oil was then removed and replaced with FC-40 fluorinated oil (Sigma-Aldrich, catalog no. 51142-49-5) with a 5% (w/w) PEG−PFPE amphiphilic block copolymer surfactant for better droplet stability before putting the emulsion into a thermal cycler (Bio-Rad, T100 model). Thermal cycling was performed at the following conditions: 2 min 30 s at 86 °C; 35 cycles of 30 s at 86 °C, 1 min 30 s at 60 °C, and 30 s at 72 °C; and a final extension of 5 min at 72 °C. A low denaturation temperature of 86 °C was used to minimize DNA fragmentation. After PCR, a small aliquot of drops was visualized with an EVOS inverted fluorescence microscope. Another small aliquot of drops was taken and broken with a 10% (v/v) solution of perfluoro-octanol (Sigma-Aldrich, catalog no. 370533) and an addition of 10 μL of deionized (DI) water, followed by gentle vortexing for 5 s and centrifuging for 1 min at 500 rpm. The recovered DNA in water, denoted as "unsort", was saved for later measurement of the enrichment factor.

Dielectrophoretic Sorting. The thermocycled drops were transferred to a 1 mL syringe and reinjected to a microfluidic dielectrophoretic (DEP) sorter (Figure 2) at 50 μL/h.[8,20] The syringe was placed vertically so that the drops remained at the top and closely packed. Individual drops were separated after entering the sorter by a spacer oil of HFE-7500 with a flow rate of 950 μL/h. Another stream of HFE-7500 oil at 1000 μL/h was introduced at the sorting junction to drive the drops to waste collection when the DEP force was off. A syringe at −1000 μL/h was used to produce a negative pressure at the waste collection to further ensure that unsorted drops flowed to waste. The salt water electrodes and moat shielding were filled with 2 M NaCl solution. A laser of 100 mW, 532 nm was focused upstream of the sorting junction to excite droplet fluorescence. Photomultiplier tubes (PMTs, Thorlabs, PMM01 model) were focused on the same spot to measure the emission fluorescence. A data acquisition card (FPGA card) and a LabVIEW program (available at GitHub: https://github.com/AbateLab/sorter-code) (National Instruments) were used to collect PMT outputs and activate the salt electrode when the emission fluorescence intensity is higher than a preset threshold. A high-voltage amplifier (Trek) was used to amplify the electrode pulse to 0.8−1 kV for DEP sorting. The sorted drops were collected into a 1.5 mL Eppendorf DNA LoBind tube.

DNA Recovery and Second Round of Enrichment. DNA from sorted drops was recovered by breaking the emulsion with 10% (v/v) solution of perfluoro-octanol (Sigma-Aldrich, catalog no. 370533) and the addition of 20 μL of DI water, followed by gentle vortexing for 5 s and centrifugation for 1 min at 500 rpm. 2 μL of the recovered DNA, denoted as "single sort", was saved for later measurement of the enrichment factor by qPCR. The remaining 18 μL recovered DNA was processed with a second round of droplet TaqMan PCR and DEP sorting as described above. After sorting, the sorted drops were broken and the recovered DNA, denoted as "double sort", was used to measure the degree of enrichment.

Quantitative PCR Analysis of Sorted Droplets. We used a multiplex TaqMan PCR, with one FAM-based probe targeting ΦX 174 DNA and one Cy5-based probe targeting lambda DNA to quantify ΦX 174 and lambda DNA in "unsort", "single sort", and "double sort". The PCR was set as follows: 1× Platinum Multiplex PCR Master Mix, 200 nM TaqMan probes, 1 μM forward primers and 1 μM reverse primers (IDT), recovered DNA, and DNAse-free water to bring the volume to 25 μL. The PCR was performed in a QuantStudio 5 Real-Time PCR System (Thermo Fisher Scientific) using the following parameters: 95 °C for 2 min; 40 cycles of 95 °C for 30 s, 60 °C for 90 s, and 72 °C for 30 s. C_t values for each sample were obtained and used to compute the enrichment factor. All primer and TaqMan probe sequences are listed in Table S2. The TaqMan assays were tested for specificity and linearity by constructing a serial dilution of the ΦX 174 DNA with a fixed concentration of lambda DNA. We obtained two C_t values for ΦX 174 (FAM) and lambda (Cy5) for each of the "unsort", "single sort", and "double sort" samples to compute the enrichment factors for each round of sorting.

HIV-Associated DNA Sample Preparation. The HIV-infected cells were prepared by plating resting CD4 T cells from an ART-treated person at ∼1 infected cell per 5 wells (∼100 total cells per well), followed by stimulation and a period of *in vitro* culture to allow proliferation.[26] Non-HIV-infected Jurkat cells (ATCC TIB-152) were cultured following the provided protocol. DNA was extracted from clonally expanded cells (from one well of the culture plate) and from Jurkat cells using Quick-DNA Miniprep Plus Kits (Zymo research, catalog no. D4068) according to the manufacturer's instructions and mixed at a 1:30 ratio.

Compound Enrichment of Single HIV Genomes. The DNA mixture was processed with droplet TaqMan PCR and 1st DEP sorting as described above using the HIV *pol* specific TaqMan probe and biotinylated primers. All primer and TaqMan probe sequences for HIV are listed in Table S3. The

https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422−7429

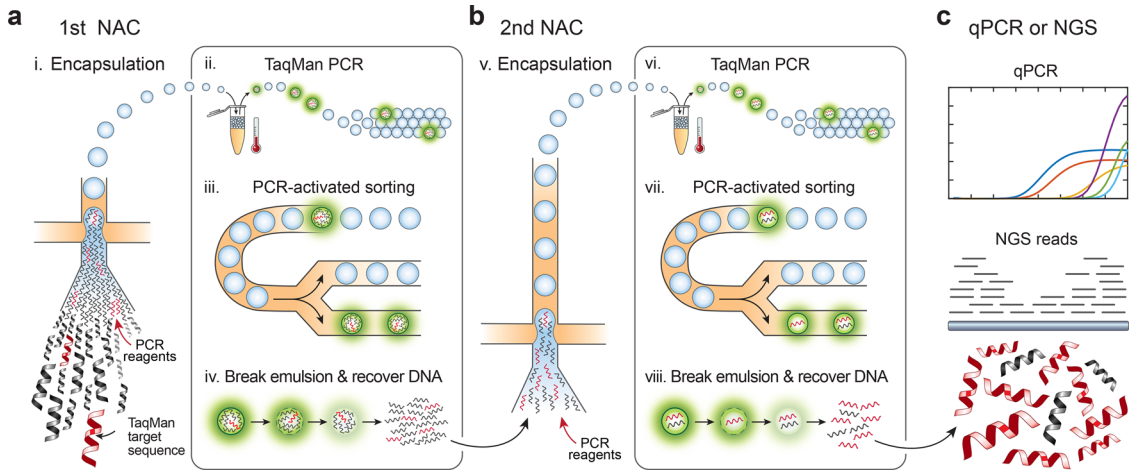


Figure 1. Schematic of compound NAC workflow. (a) A mixed DNA sample is sorted in the first NAC round using TaqMan targeting a desired sequence biomarker. Each NAC round comprises (i) DNA encapsulation with TaqMan reagents; (ii) in-droplet PCR to generate fluorescence when the target is present; (iii) sorting to select positive drops; and (iv) recovery of sorted DNA by droplet demulsification. (b) DNA recovered from the first NAC round is diluted and processed through another round consisting of the same steps (v–viii). (c) The double-enriched DNA is analyzed by qPCR to estimate enrichment and sequenced.

sorted emulsions were broken using perfluoro-1-octanol, and the aqueous fraction was diluted in 5 μL of H_2O. The aqueous layer containing sorted DNA was then added to streptavidin-conjugated magnetic beads (Dynabeads MyOne Streptavidin C, Thermo Fisher Scientific) and incubated for 15 min. D1 buffer from the REPLI-g single cell kit (Qiagen, catalog no. 150343) was added to denature the DNA. Biotinylated primers and amplicons were attached to the magnetic beads and removed after transferring the supernatant to a fresh tube. The multiple-displacement amplification (MDA) reaction mixture was then prepared with a REPLI-g single cell kit by following the manufacturer's protocol and emulsified by a flow-focus droplet maker (diameter ~20 μm) as described above. The emulsion was collected in a 1 mL syringe and incubated at 30 °C for 20 h. After incubation, MDA droplets and 2nd TaqMan PCR reagents were injected into a microfluidic merger device.[27] PCR reagent drops were formed on a chip and merged with MDA drops pairwise. Merging was achieved at a salt electrode connected to a cold cathode fluorescent inverter and a DC power supply (Mastech) to generate a ~2 kV AC signal from a 2 V input voltage. The merged drops (diameter: 40 μm) were collected to PCR tubes. The bottom oil layer was removed and replaced with FC-40 fluorinated oil with a 5% (w/w) PEG–PFPE surfactant for thermal cycling: 3 min at 86 °C; 35 cycles of 30 s at 86 °C, 90 s at 60 °C, and 30 s at 72 °C; and finally, 5 min at 72 °C. After PCR, the drops were reinjected into a DEP sorter for the second sorting round as described above.

Library Preparation and Sequencing from Sorted Droplets. The sorted single droplets with their carrier oil were collected into individual PCR tubes and dried out in a vacuum chamber. 1 μL of DI H_2O was added to dissolve the sorted DNA. The dissolved DNA was then tagmented using 0.6 μL of the TD Tagmentation buffer and 0.3 μL of the ATM Tagmentation enzyme from the Nextera DNA Library Prep Kit (Illumina, catalog no. FC-121-1030) for 5 min at 55 °C. 1 μL of the NT buffer was added to neutralize the tagmentation. The tagmented DNA was then mixed with a PCR solution containing 1.5 μL of the NPM PCR master mix, 0.5 μL of each index primers i5 and i7 from the Nextera Index Kit (Illumina, catalog no. FC-121-1011), and 1.5 μL of H_2O and placed on a thermal cycler with the following program: 3 min at 72 °C; 30 s at 95 °C; 20 cycles of 10 s at 95 °C, 30 s at 55 °C, and 30 s at 72 °C; and finally, 5 min at 72 °C. The DNA library was purified using a DNA Clean & Concentrator-5 kit (Zymo Research, catalog no. D4004), size-selected for 200–600 bp fragments using Agencourt AMPure XP beads (Beckman Coulter), and quantified using the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and the High Sensitivity DNA Bioanalyzer chip (Agilent). The library was sequenced using Illumina Miseq and ~1 million paired-end reads of 150 bp were used for each sorted droplet or unsorted starting sample. Sequencing reads were mapped to the HIV reference genome (HXB2) using Bowtie 2.[28] Genomic coverage as a function of genome position was generated using SAMtools.[29] HIV genome average coverage with or without duplicates removing by Picard (MarkDuplicates) (http://broadinstitute.github.io/picard/) was calculated using SAMtools.[29] TaqMan amplicon regions were excluded in the average coverage calculation. The non-HIV regions of chimeric reads were extracted using extractSoftclipped (https://github.com/dpryan79/SE-MEI) and analyzed by a web-based tool for integration sites (https://indra.mullins.microbiol.washington.edu/integrationsites/).

RESULTS AND DISCUSSION

NAC isolates molecules of interest from a mixed population based on specific sequence biomarkers.[19] This is achieved by combining the droplet TaqMan PCR identification and microfluidic droplet sorting to physically isolate molecules based on the TaqMan signal (Figure 1a). The DNA mixture is partitioned at the limiting dilution such that individual droplets rarely contain more than one target. The purity of the target sequence before sorting P_{before} and after sorting P_{after}

$$P_{before} = N_T/(N_T + N_O) \quad (1)$$

$$P_{after} = N_T D/[(N_T + N_O)(N_T + fD)] \quad (2)$$

where N_T is the number of target molecules (positively sorted drops), N_O is the total number of off-target molecules, D is the

https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422–7429

total number of droplets, and f is the assay false positive rate. Thus, the enrichment power E is (derivation in Supporting Information)

$$E = (N_T/D + f)^{-1} \quad (3)$$

Because the false positive rate is generally small ($\sim 10^{-4}$) and difficult to reduce,[30] the best way to increase the enrichment power is to encapsulate the sample into more droplets, which thus delivers fewer co-encapsulated off-target molecules per sorted positive. However, the number of droplets that can be sorted is limited to ~10 million.[8,19] It is possible to increase D in several ways, including implementing faster droplet sorting with gapped dividers, batched sorting, and novel sorting mechanisms.[31] However, while promising, none have yet been demonstrated for this purpose and doing so would introduce risk and require additional development. Consequently, the maximum practical enrichment that can be achieved per NAC round is $\sim 10^4$.

Similar to hybridization capture, NAC does not fragment the original target molecules. However, in contrast to hybridization capture, NAC can recover long intact targets (>100 kbp) present in a sample over a wide range of DNA concentrations.[8] These features allow it to be performed repeatedly on a sample such that compound enrichment is achieved (Figure 1b). In such a strategy, the overall enrichment with two rounds $E_{compound}$ is

$$E_{compound} = [(N_T/D_1 + f_1)(N_T/D_2 + f_2)]^{-1} \quad (4)$$

when using D_1 drops in the first round and D_2 drops in the second. Compound enrichment thus allows marked increases to enrichment compared to sorting more drops in a single round. For example, for a total of $\sim 10^7$ drops sorted, one round typically achieves $\sim 10^3$ enrichment of 10,000 target molecules, while two consecutive rounds achieve $\sim 10^6$. Obtaining such an enrichment with a single-round of NAC would require sorting over a billion droplets, which is impractical. The resultant concentrated DNA is intact and readily amenable to qPCR or sequencing analysis (Figure 1c).

Microfluidic Workflow for Compound Enrichment. NAC uses ultrahigh-throughput microfluidics to perform, analyze, and sort millions of PCRs. Flow focusing loads sample DNA with TaqMan reagents in ~45 μm droplets at ~2.5 kHz, partitioning the entire 150 μL reaction in ~3 million drops in ~20 min (Figure 2a). The drops are thermocycled, generating TaqMan fluorescence when the target is present (Figure 2b). The drops are analyzed and sorted using a laser-induced fluorescence detector and a DEP droplet deflector[27,32] (Figure 2c). We operate this integrated device at ~400 Hz to ensure accurate sorting and efficient positive recovery, screening ~3 million drops in ~2 h. The sorted target molecules are recovered by droplet demulsification with perfluoro-octanol[32] and diluted into new TaqMan reagents for the next round of NAC.

Compound Enrichment of the ΦX 174 Virus. To demonstrate the power of compound NAC, we apply it to enrich ΦX 174 viral genomes from a 10^7-fold greater background of lambda DNA. The TaqMan set used in each round detects a different region of the ΦX 174 genome, preventing amplicons carried over from the first round generating false positives in the second (Figure 3a). Both sets reliably detect ΦX 174 DNA (Figure S1). We set the target concentration such that ~0.3% droplets are expected to

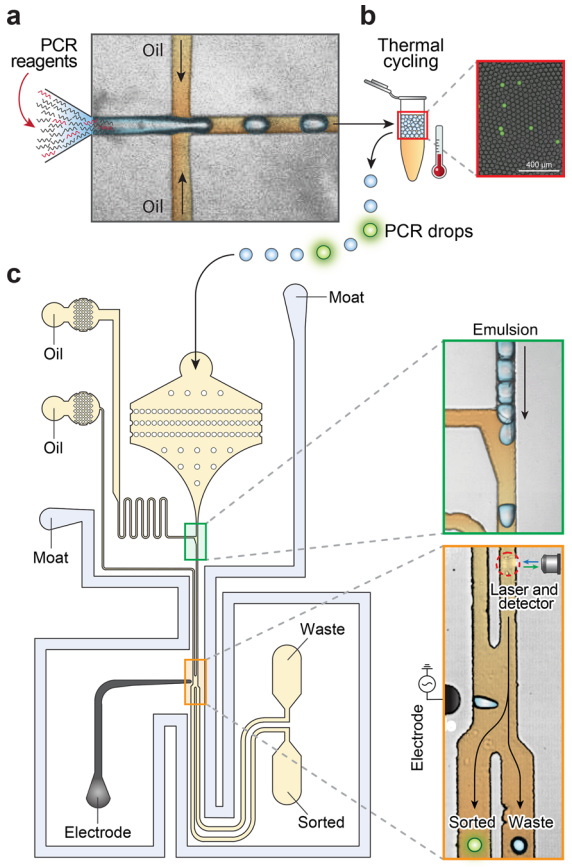


Figure 2. Microfluidic devices of NAC. (a) Droplet encapsulation of DNA and PCR reagents. (b) In-drop PCR to generate fluorescence when the target is present. The merged bright field/fluorescence image shows a representative sample post thermocycling. (c) Fluorescence-activated sorting selects droplets containing target sequences. Scale bars: 400 μm.

be positive and observe an actual positive rate of 0.24% in the first round (Figure 3b(i)), yielding an enrichment of ~400. This input concentration also generates less than 0.001% multiplets, in accordance with Poisson statistics. We screen a total of ~3 million droplets, collecting ~7000 positives. The recovered DNA is diluted into a fresh reaction buffer again to achieve another 400-fold enrichment and subjected to another round of NAC, collecting ~5000 positives (Figure 3b(ii)). Because the method is nondestructive (Figure S2), the number of positive droplets should be equal for both rounds, but sample loss during transfer and preparation for the second round results in a ~30% positive reduction (Figure 3b(iii)). To confirm the enrichment, we use qPCR to measure the fractions of ΦX 174 and lambda DNA in the sorted samples. After a single round, the qPCR curve for ΦX 174 shifts to lower cycles by ~5 (concentrated), while that for lambda shifts to higher cycles by ~3 (diluted), illustrating enrichment (Figure 3c(i)). For two rounds compounded, the shifts are greater (−2 for ΦX 174 and +12 for lambda) (Figure 3c(ii)). To quantify the enrichments, we calculate the enrichment factor E based on the cross-threshold values of the qPCR curves (Table S1).[20] For one round of sorting ~3 million droplets, the estimated enrichment is ~150. For two rounds of sorting comprising a total of ~6 million droplets, the enrichment is ~16,000. Achieving this enrichment in one round would require sorting

https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422−7429

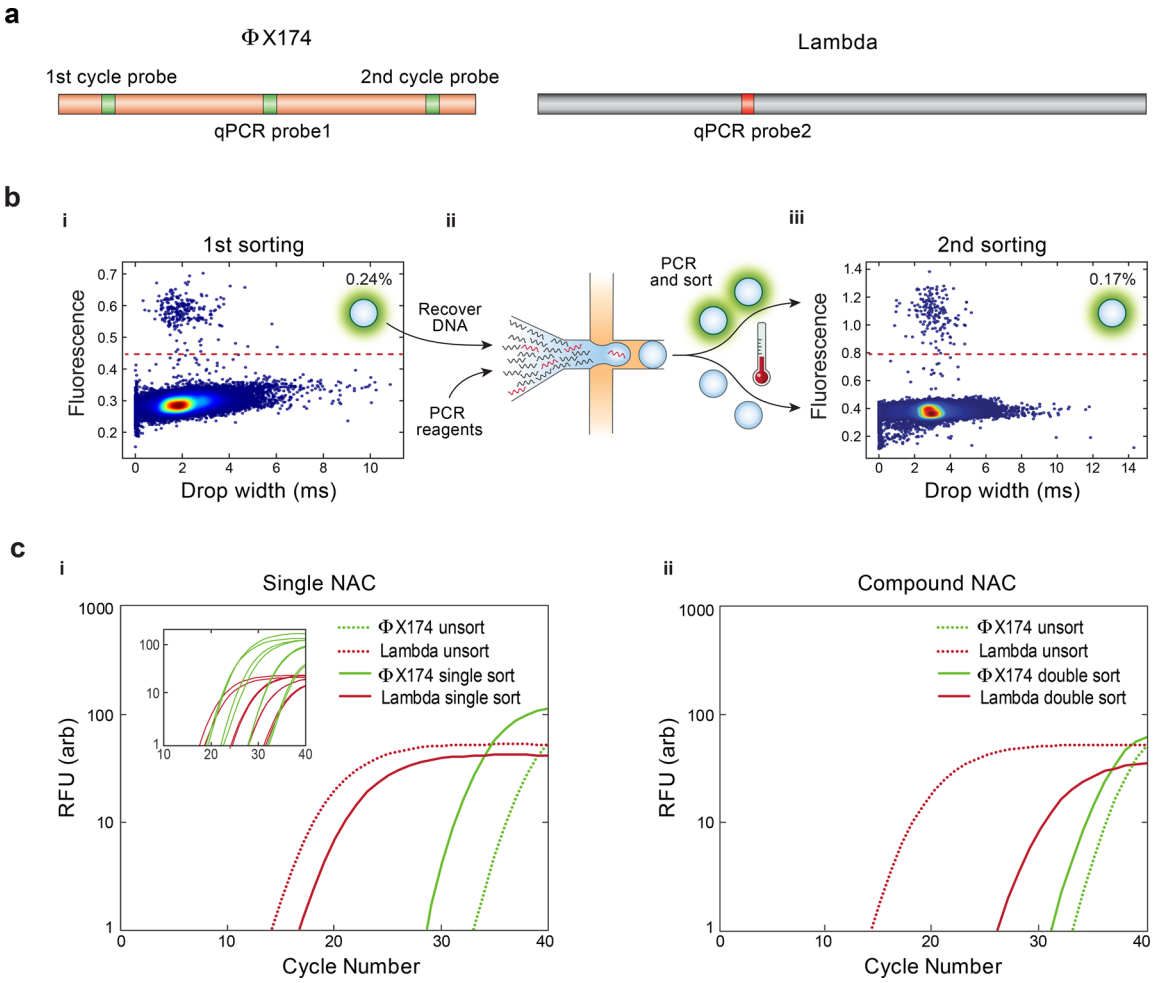


Figure 3. Enrichment of the ΦX 174 DNA from a background of lambda DNA with compound NAC. (a) TaqMan assays detect droplets containing ΦX 174 (green) and lambda (red) DNA. (b) The microfluidic sorter interrogates the droplets for fluorescence and sorts PCR positives. (i) Scatter plot of fluorescence *vs* size of drops from the first NAC round, with 0.24% positive. (ii) DNA from the first round is recovered, diluted, and processed again. (iii) Scatter plot of fluorescence *vs* size of drops from the second NAC round, with 0.17% positive. (c) qPCR plots for (i) single- and (ii) double-enriched DNA; based on curve shifts, single-round sorting enriches ΦX 174 by ∼150-fold and double-round sorting by ∼16,000-fold. Inset in (i) shows ΦX 174 and lambda standard curves.

∼300 million droplets, totaling 16 mL of PCR reagent, and a week of nonstop sorting.

Single Genome Sequencing of Ultrarare HIV Proviruses. During effective antiretroviral therapy, HIV persists in a latent state and circulates at extremely low levels, with human DNA outnumbering it by over a billion-fold.[2,3,26] Under such circumstances, unbiased sequencing would recover a minute fraction of one viral genome per human genome sequenced. To obtain comprehensive information on the genetics of HIV under such circumstances, potent enrichment of the virus is needed. Due to limited information on HIV genomes in a specific sample, capture probe design is challenging. In addition, hybridization methods require considerable input DNA and do not provide single provirus information. Long-range PCR methods are only applicable when specific primers can be designed to target known conserved sequences within the virus or host; this information is often not available for novel virus integrations and precludes the recovery of unknown integrations. The only effective strategy presently available is terminal dilution PCR in well plates.[3,4,33] This brute force approach aliquots thousands of cells in hundreds of microwells using long-ranged multiprimer amplification to obtain near full-length HIV genomes. However, in addition to often generating artifacts that can confound analysis, the approach does not obtain the crucial virus−host junction with the complete virus genome in a single contig. Without this dual information, specific proviruses cannot be related to host insertion sites and thereby the combination associated to disease behavior.[3,34] Consequently, other strategies must be employed to infer viral genome and host junction relationships.[33,34]

Due to the potent enrichment enabled by compound NAC and its ability to recover intact DNA fragments spanning integrated HIV genomes, such analyses are possible. To demonstrate this, we use compound NAC to isolate and sequence HIV proviruses from a patient-infected cell expansion.[26] This clonal lineage of infected cells contains one HIV provirus per ∼100 cells, each bearing the same provirus. To demonstrate the enrichment power of compound NAC, we dilute the sample with non-HIV cell gDNA at a 1:30 ratio. This dilution models the concentration of the latent infection. Due to the extreme rarity of HIV DNA in this sample, we use a multiplexed TaqMan PCR targeting multiple conserved regions of HIV in the second cycle for accurate

https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422−7429

detection and isolation before sequencing (Figure 4a). The DNA mixture is encapsulated in droplets and the ones

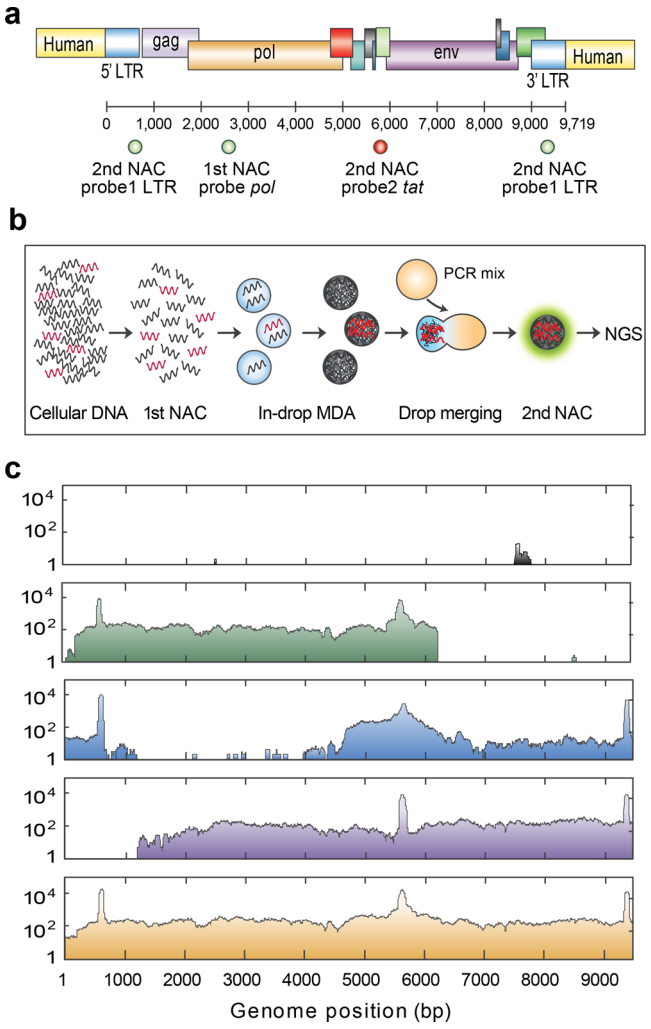


Figure 4. Compound enrichment and sequencing of HIV proviruses in infected individuals. (a) The first round of NAC uses a TaqMan assay targeting the *pol* gene (FAM). The second round of NAC uses a degenerate TaqMan assay targeting the long terminal repeat (LTR, FAM) and *tat* gene (Cy5). (b) Implementation of in-droplet MDA before the second round generates sufficient material for single droplet sequencing. The MDA droplet is merged with a droplet containing PCR reagents for TaqMan detection in the second round. (c) HIV genome and integration site coverage maps for nonenriched sample (top panel) and three sorted individual drops. Aggregating all data (bottom panel) assembles the full-length HIV genome including the human integration site. The peaks at LTR and *tat* are the TaqMan PCR amplicons.

containing HIV genomes are isolated. Microfluidic enrichment and pooled sequencing of DNA collected from many droplets have been used to characterize rare genome sequences, but isolation and sequencing of single viruses are impossible with this approach.[5,23] By incorporating in-droplet MDA before the second round,[35] each sorted droplet yields ~3 pg DNA, just enough for sequencing; this enables the linkage of a specific HIV genome to its integration site (Figure 4b). This novel workflow affords superior enrichment and single drop sequencing, allowing recovery of integrated provirus genomes (Figure 4c, first to third row). We thus identify the integration site in the host gene ARIH2 by extracting virus−human chimeric reads from the sequencing data. Shearing during PCR and DNA preparation results in a partial genome dropout. Thus, by assembling reads from three droplets, we obtain complete coverage of the full-length viral genome (Figure 4c, fourth row). In total, sequencing after two rounds detects ~6 million times more proviral reads compared to the initial sample, with an average coverage of ~200× for the targeted proviral genome and ~0.1× for the untargeted human genome. Discarding duplicates reduces the coverage of the targeted region to 60× but is not always necessary.[36] These results illustrate that compound NAC enables the sequencing of extremely rare HIV proviruses and that the enriched molecules retain information on the genetic context of the integration.

By leveraging droplet microfluidics, NAC enables enrichment of target molecules containing sequence biomarkers. By processing the sample repeatedly, feeding the output of the first round into the input of the second, compound enrichment is achieved, allowing the recovery of ultrarare targets. In single-round NAC, maximum enrichment is limited by the false positive droplet rate. Partitioning the sample into more droplets enhances enrichment in a linear fashion but does not allow the marked increases required for ultrarare targets. Moreover, such brute force also increases cost and processing time and becomes impractical beyond enrichments of 30,000.[8,23] Our method eliminates the need for large numbers of droplets and increases the maximum enrichment to 10^9 fold, allowing highly specific target recovery. A drawback of the multiround NAC is sample loss during transfer and processing. We observe ~30% loss between the two NAC rounds, which may result in the loss of low-abundant variants. Nevertheless, recovery of these variants by single-round NAC is essentially impossible, so compound enrichment is still the best approach for analysis of rare sequences.

CONCLUSIONS

Compound NAC allows the isolation of long molecules with minimal prior sequence information, opening new avenues in target enrichment. For example, million-fold enrichment of >100 kbp molecules is useful for a variety of ultrarare target applications, including characterizing novel human genetic mutations or natural product gene clusters in metagenomic samples.[5,8] Additionally, the approach is generalizable to other targets because it uses TaqMan PCR to define the sequence biomarker of capture and thus can be applied to any nucleic acid detectable by this assay, including RNA by adding a reverse transcription step; this would allow sequencing of fusion genes or low-abundant variants. Finally, as we have shown, implementation of in-droplet MDA allows sequencing of compound-enriched single molecules, making it a powerful tool for single virus genomics.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c04749.

Derivation of equations for enrichment power, images and plots showing that TaqMan sets reliably detect ΦX 174 DNA in compound NAC, gel analysis showing minimal DNA fragmentation during thermo cycling, C_t values from the qPCR plots for single and double enrichment, and primers and probes for the compound

https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422−7429

enrichment of ΦX 174 virus DNA and HIV provirus DNA (PDF)

AUTHOR INFORMATION

Corresponding Author

Adam R. Abate − *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, United States; California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, California 94158, United States; Chan Zuckerberg Biohub, San Francisco, California 94158, United States;* orcid.org/0000-0001-9614-4831; Email: adam@abatelab.org

Authors

Chen Sun − *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, United States;* orcid.org/0000-0003-1216-5091

Kai-Chun Chang − *Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158, United States;* orcid.org/0000-0001-7200-3532

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c04749

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank James I. Mullins at the University of Washington for providing HIV-infected cells. We also thank members of the Abate lab, in particular Leqian Liu, Cyrus Modavi, David J. Sukovich, and Samuel C. Kim, for helpful discussions. This work was supported by the Chan Zuckerberg Biohub and the National Institutes of Health (NIH) (grant nos. R01-EB019453-01 and R01-HG008978-01).

REFERENCES

(1) Mamanova, L.; Coffey, A. J.; Scott, C. E.; Kozarewa, I.; Turner, E. H.; Kumar, A.; Howard, E.; Shendure, J.; Turner, D. J. *Nat. Methods* **2010**, *7*, 111−118.

(2) Finzi, D.; Blankson, J.; Siliciano, J. D.; Margolick, J. B.; Chadwick, K.; Pierson, T.; Smith, K.; Lisziewicz, J.; Lori, F.; Flexner, C.; Quinn, T. C.; Chaisson, R. E.; Rosenberg, E.; Walker, B.; Gange, S.; Gallant, J.; Siliciano, R. F. *Nat. Med.* **1999**, *5*, 512−517.

(3) Einkauf, K. B.; Lee, G. Q.; Gao, C.; Sharaf, R.; Sun, X.; Hua, S.; Chen, S. M. Y.; Jiang, C.; Lian, X.; Chowdhury, F. Z.; Rosenberg, E. S.; Chun, T.-W.; Li, J. Z.; Yu, X. G.; Lichterfeld, M. *J. Clin. Invest.* **2019**, *129*, 988−998.

(4) Hiener, B.; Horsburgh, B. A.; Eden, J.-S.; Barton, K.; Schlub, T. E.; Lee, E.; von Stockenstrom, S.; Odevall, L.; Milush, J. M.; Liegler, T.; Sinclair, E.; Hoh, R.; Boritz, E. A.; Douek, D.; Fromentin, R.; Chomont, N.; Deeks, S. G.; Hecht, F. M.; Palmer, S. *Cell Rep.* **2017**, *21*, 813−822.

(5) Xu, P.; Modavi, C.; Demaree, B.; Twigg, F.; Liang, B.; Sun, C.; Zhang, W.; Abate, A. R. *Nucleic Acids Res.* **2020**, *48*, No. e48.

(6) Suenaga, H. *Environ. Microbiol.* **2012**, *14*, 13−22.

(7) Sharon, I.; Banfield, J. F. *Science* **2013**, *342*, 1057−1058.

(8) Eastburn, D. J.; Huang, Y.; Pellegrino, M.; Sciambi, A.; Ptacek, L. J.; Abate, A. R. *Nucleic Acids Res.* **2015**, *43*, No. e86.

(9) Wei, X.; Ju, X. C.; Yi, X.; Zhu, Q.; Qu, N.; Liu, T. F.; Chen, Y.; Jiang, H.; Yang, G. H.; Zhen, R.; Lan, Z. Z.; Qi, M.; Wang, J. M.; Yang, Y.; Chu, Y. X.; Li, X. Y.; Guang, Y. F.; Huang, J. *PLoS One* **2011**, *6*, No. e29500.

(10) Bomba, L.; Walter, K.; Soranzo, N. *Genome Biol.* **2017**, *18*, 77.

(11) Houldcroft, C. J.; Beale, M. A.; Breuer, J. *Nat. Rev. Microbiol.* **2017**, *15*, 183−192.

(12) Mertes, F.; ElSharawy, A.; Sauer, S.; van Helvoort, J. M. L. M.; van der Zaag, P. J.; Franke, A.; Nilsson, M.; Lehrach, H.; Brookes, A. J. *J. Brief. Funct. Genom.* **2011**, *10*, 374−386.

(13) Krishnakumar, S.; Zheng, J.; Wilhelmy, J.; Faham, M.; Mindrinos, M.; Davis, R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9296−9301.

(14) Tewhey, R.; Warner, J. B.; Nakano, M.; Libby, B.; Medkova, M.; David, P. H.; Kotsopoulos, S. K.; Samuels, M. L.; Hutchison, J. B.; Larson, J. W.; Topol, E. J.; Weiner, M. P.; Harismendy, O.; Olson, J.; Link, D. R.; Frazer, K. A. *Nat. Biotechnol.* **2009**, *27*, 1025−1031.

(15) Iwase, S. C.; Miyazato, P.; Katsuya, H.; Islam, S.; Yang, B. T. J.; Ito, J.; Matsuo, M.; Takeuchi, H.; Ishida, T.; Matsuda, K.; Maeda, K.; Satou, Y. *Sci. Rep.* **2019**, *9*, 12326.

(16) Bodi, K.; Perera, A. G.; Adams, P. S.; Bintzler, D.; Dewar, K.; Grove, D. S.; Kieleczawa, J.; Lyons, R. H.; Neubert, T. A.; Noll, A. C.; Singh, S.; Steen, R.; Zianni, M. *J. Biomol. Tech.* **2013**, *24*, 73−86.

(17) Riedl, J.; Ding, Y.; Fleming, A. M.; Burrows, C. J. *Nat. Commun.* **2015**, *6*, 8807.

(18) Petti, C. A. *Clin. Infect. Dis.* **2007**, *44*, 1108.

(19) Clark, I. C.; Abate, A. R. *Lab Chip* **2017**, *17*, 2032−2045.

(20) Lim, S. W.; Tran, T. M.; Abate, A. R. *PLoS One* **2015**, *10*, No. e0113549.

(21) Lance, S. T.; Sukovich, D. J.; Stedman, K. M.; Abate, A. R. *Virol. J.* **2016**, *13*, 201.

(22) Sukovich, D. J.; Lance, S. T.; Abate, A. R. *Sci. Rep.* **2017**, *7*, 39385.

(23) Han, H.-S.; Cantalupo, P. G.; Rotem, A.; Cockrell, S. K.; Carbonnaux, M.; Pipas, J. M.; Weitz, D. A. *Angew. Chem., Int. Ed.* **2015**, *54*, 13985−13988.

(24) Tao, Y.; Rotem, A.; Zhang, H.; Cockrell, S. K.; Koehler, S. A.; Chang, C. B.; Ung, L. W.; Cantalupo, P. G.; Ren, Y.; Lin, J. S.; Feldman, A. B.; Wobus, C. E.; Pipas, J. M.; Weitz, D. A. *ChemBioChem* **2015**, *16*, 2167−2171.

(25) Mousavian, Z.; Sadeghi, H. M. M.; Sabzghabaee, A. M.; Moazen, F. *Adv. Biomed. Res.* **2014**, *3*, 65.

(26) Bruner, K. M.; Wang, Z.; Simonetti, F. R.; Bender, A. M.; Kwon, K. J.; Sengupta, S.; Fray, E. J.; Beg, S. A.; Antar, A. A. R.; Jenike, K. M.; Bertagnolli, L. N.; Capoferri, A. A.; Kufera, J. T.; Timmons, A.; Nobles, C.; Gregg, J.; Wada, N.; Ho, Y.-C.; Zhang, H.; Margolick, J. B.; Blankson, J. N.; Deeks, S. G.; Bushman, F. D.; Siliciano, J. D.; Laird, G. M.; Siliciano, R. F. *Nature* **2019**, *566*, 120−125.

(27) Sciambi, A.; Abate, A. R. *Lab Chip* **2014**, *14*, 2605−2609.

(28) Langmead, B.; Salzberg, S. L. *Nat. Methods* **2012**, *9*, 357−359.

(29) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data P. *Bioinformatics* **2009**, *25*, 2078−2079.

(30) Rowlands, V.; Rutkowski, A. J.; Meuse, E.; Carr, T. H.; Harrington, E. A.; Barrett, J. C. *Sci. Rep.* **2019**, *9*, 12620.

(31) Xi, H.-D.; Zheng, H.; Guo, W.; Gañán-Calvo, A. M.; Ai, Y.; Tsao, C.-W.; Zhou, J.; Li, W.; Huang, Y.; Nguyen, N.-T.; Tan, S. H. *Lab Chip* **2017**, *17*, 751−771.

(32) Mazutis, L.; Gilbert, J.; Ung, W. L.; Weitz, D. A.; Griffiths, A. D.; Heyman, J. A. *Nat. Protoc.* **2013**, *8*, 870−891.

(33) Patro, S. C.; Brandt, L. D.; Bale, M. J.; Halvas, E. K.; Joseph, K. W.; Shao, W.; Wu, X.; Guo, S.; Murrell, B.; Wiegand, A.; Spindler, J.; Raley, C.; Hautman, C.; Sobolewski, M.; Fennessey, C. M.; Hu, W.-S.; Luke, B.; Hasson, J. M.; Niyongabo, A.; Capoferri, A. A.; Keele, B. F.; Milush, J.; Hoh, R.; Deeks, S. G.; Maldarelli, F.; Hughes, S. H.; Coffin, J. M.; Rausch, J. W.; Mellors, J. W.; Kearney, M. F. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 25891−25899.

(34) Wiegand, A.; Spindler, J.; Hong, F. F.; Shao, W.; Cyktor, J. C.; Cillo, A. R.; Halvas, E. K.; Coffin, J. M.; Mellors, J. W.; Kearney, M. F. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E3659−E3668.

(35) Sidore, A. M.; Lan, F.; Lim, S. W.; Abate, A. R. *Nucleic Acids Res.* **2016**, *44*, No. e66.

(36) Ebbert, M. T. W.; Wadsworth, M. E.; Staley, L. A.; Hoyt, K. L.; Pickett, B.; Miller, J.; Duce, J.; Kauwe, J. S. K.; Ridge, P. G. *BMC Bioinf.* **2016**, *17*, 239.

https://doi.org/10.1021/acs.analchem.0c04749
Anal. Chem. 2021, 93, 7422−7429