Sequencing ultra-rare targets with compound nucleic acid cytometry

Chen Sun[1], Kai-Chun Chang[1] and Adam R. Abate[1,2,3*]
[1] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA
[2] California Institute for Quantitative Biosciences, University of California San Francisco, San Francisco, CA 94158, USA
[3] Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
[*] Corresponding author: adam@abatelab.org

Abstract
Targeted sequencing enables sensitive and cost-effective analysis of rare targets by focusing resources on regions of interest. However, existing target enrichment methods are limited in enrichment power and target molecule length. Here, we present a novel method that uses compound nucleic acid cytometry to achieve millions-fold enrichment of a target. This enrichment requires minimal prior target information and allows rapid recovery of long molecules. We demonstrate the power of this approach by sequencing HIV proviruses in infected individuals. Our method is valuable to rare target analysis in various research and clinical applications, such as identifying cancer-associated mutations or characterizing target viruses in infected 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 persistence and finding a cure, 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 human genetic disease, variants may be present at fractions of a percent[8-10]. In instances like these, sequencing all DNA is wasteful because only a fraction of reads correspond to the region of interest. The most common target enrichment strategies are based on PCR amplification or hybridization capture[1,11,12]. PCR methods recover only the amplified portion and miss information beyond the 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 novel microbe or genetic lesion sequencing.

Nucleic acid cytometry (NAC) is a conceptually novel approach to target enrichment based on droplet microfluidics[17]. The overarching principle is to identify and physically isolate molecules by hydrodynamic sorting. Identification is accomplished using droplet PCR, while isolation is accomplished by sorting positive droplets[17-20]. The approach is akin to querying a diverse mixture of sequences for keyword subsequences of interest, and isolating all molecules containing the keyword. The critical factor in NAC enrichment is the sensitivity for recovering the target of interest. Sensitivity, in turn, is limited by the number of droplet PCR reactions 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 amount[5,8,21,22]. This enrichment is insufficient for applications with ultra-rare targets often present 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 possible using the next best technology[8,15,21]. To demonstrate the approach, we use it to isolate and sequence single provirus genomes from HIV infected individuals. No other assay or physical enrichment approach has the sensitivity to recover and sequence such rare single virus genomes. Compound NAC provides a general platform for recovering long, ultra-rare molecules with minimal prior sequence information.

Overview
Nucleic acid cytometry isolates molecules of interest from a mixed population based on specific sequence biomarkers[17]. This is achieved by combining droplet TaqMan PCR identification and microfluidic droplet sorting to physically isolate molecules based on the TaqMan signal (Fig. 1a). The DNA mixture is partitioned at limiting dilution such that individual droplets rarely contain more than one target. The purity of the target sequence is

$\frac{N_T}{N_T+N_O}$ before sorting, and $\frac{N_T}{\frac{N_T+N_O}{D}(N_T+fD)}$ after, where N_T is the number of target molecules or positively sorted drops, N_O the total number of off-target molecules in the sample, D the total number of droplets and f the false positive droplet rate of the assay. Thus, the enrichment factor is determined by $\frac{1}{\frac{N_T}{D}+f}$. Because the false positive rate is generally small (~10^{-5}) and difficult to reduce[23], the best way to increase enrichment factor is to encapsulate the sample into more droplets, which thus delivers fewer co-encapsulated off target molecules per sorted positive droplet. However, the number of droplets that can be sorted is limited by the sorting to ~10 million[8,17]. Consequently, the maximum practical enrichment that can be achieved per NAC round is ~10^4.

Like hybridization capture, NAC does not fragment the original target molecules. However, in contrast to hybridization capture, NAC can recover long intact targets (>100 kbp) and operate on samples with a wide range of DNA concentrations[8]. These features allow NAC to be performed repeatedly on a sample such that the enrichments compound (Fig. 1b). In such a strategy, the overall enrichment is $\frac{1}{\frac{N_T}{D_1}+f_1} \times \frac{1}{\frac{N_T}{D_2}+f_2}$, when using D_1 drops in the first round and D_2 drops in the second. Compound enrichment allows marked increases to enrichment compared to sorting more drops in a single round. For example, for a total of ~10^7 drops screened, one round typically achieves ~10^3 enrichment of 10,000 target molecules, while two consecutive rounds achieve ~10^6. Achieving 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 (Fig. 1c).

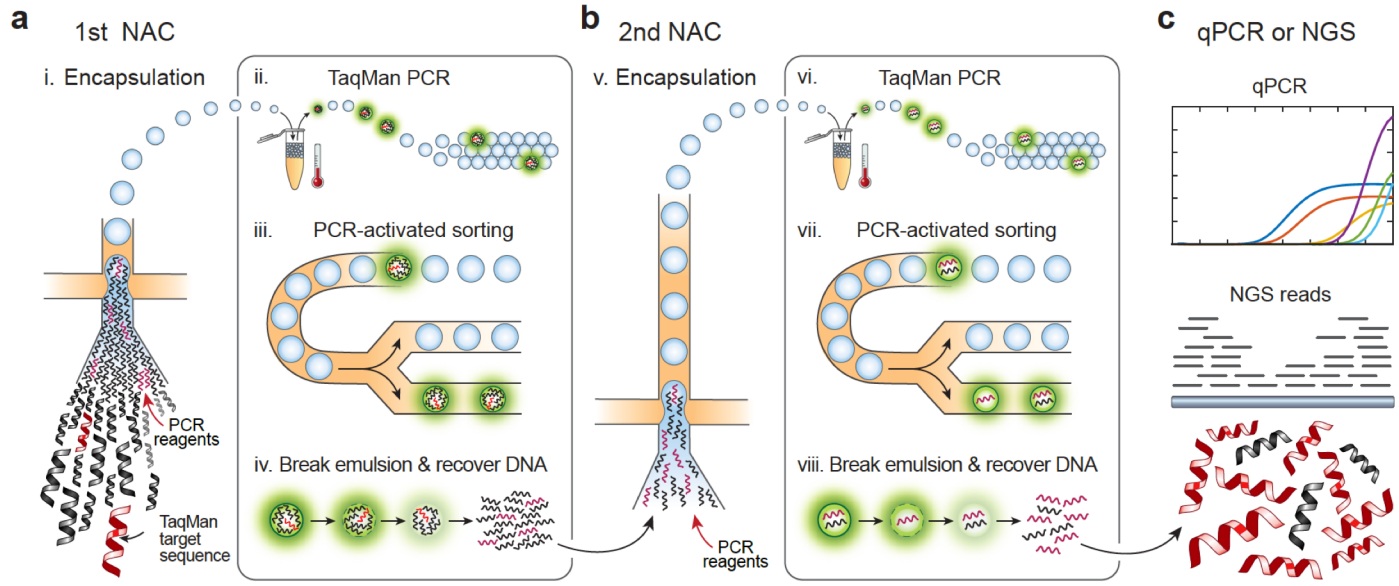

Figure1. Schematic of the workflow for compound NAC. (a) A heterogeneous sample of DNA is processed with one NAC run for target enrichment. TaqMan assay is designed to target specific sequence of interest. The NAC consists of (i) droplet encapsulation of DNA and TaqMan PCR reagents; (ii) in-droplet PCR to generate fluorescent drop when target sequence is present; (iii) droplet sorting to select PCR positive drops; (iv) breaking sorted drops to recover DNA. **(b)** The 1st NAC sorted DNA is processed with another round of NAC for 2nd target enrichment, consisting of (v) droplet encapsulation of DNA and 2nd TaqMan PCR reagents; (vi) in-droplet PCR to generate fluorescent drop when target sequence is present; (vii) droplet sorting to select PCR positive drops; (viii) breaking sorted drops to recover DNA. **(c)** Compound enriched DNA is analyzed by qPCR to quantify the enrichment factor, or NGS to study the specific sequence of interest.

Microfluidic workflow for compound enrichment

NAC relies crucially on high throughput microfluidic techniques to perform, analyze, and sort millions of PCR reactions. Droplet encapsulation loads sample DNA with TaqMan reagents in ~45 μm droplets (Fig. 2a). Our device runs at ~2.5 kHz, partitioning a 150 μL reaction into ~3 million drops in ~20 min. The drops are

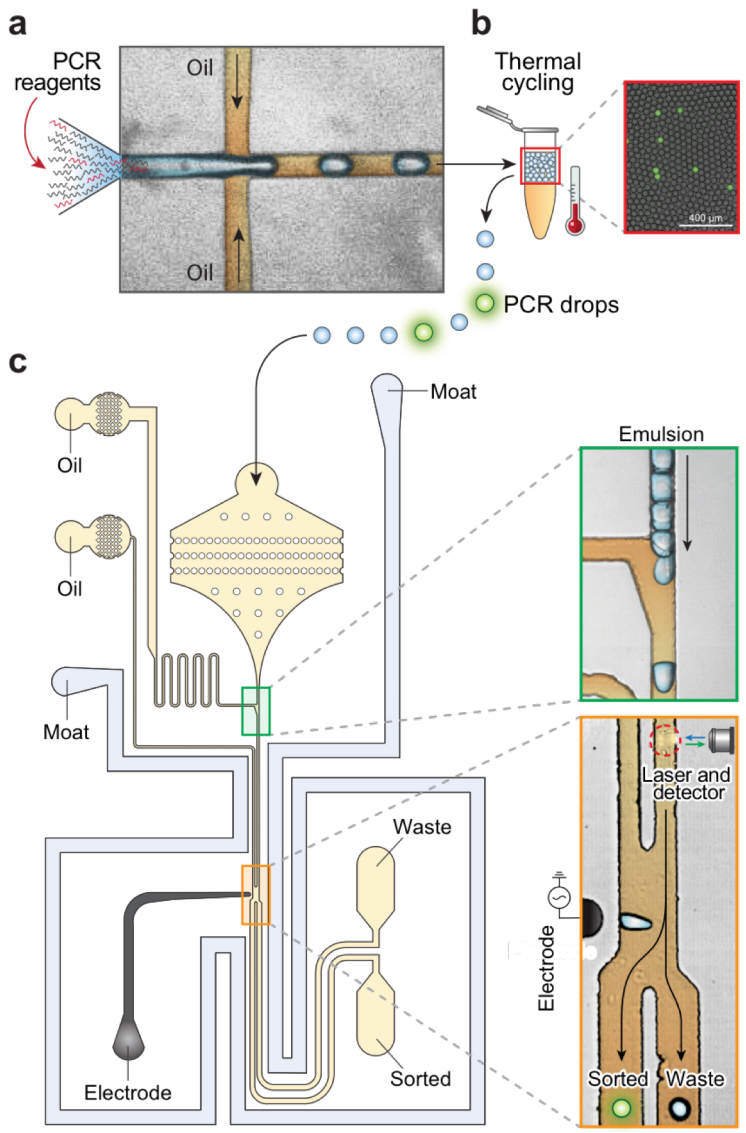

Figure 2. NAC uses droplet microfluidics to enrich target sequences. This is accomplished by **(a)** droplet encapsulation of DNA and PCR reagents; **(b)** in-drop PCR to generate fluorescent drop when target sequence is present. Fluorescence image shows representative drops after thermocylcing; and **(c)** droplet sorting to select drops containing target sequence. scale bars: 400um.

thermocycled, generating TaqMan fluorescence when the target is present in a drop (Fig. 2b). The drops are then analyzed and sorted using a laser-induced fluorescence detector and dielectrophoretic droplet deflector[24,25] (Fig. 2c). We operate this integrated device at ~400 Hz to ensure accurate sorting and efficient positive droplet recovery, screening ~3 million drops in ~2 hr. The sorted target molecules are recovered by demulsifying the droplets with perfluoro-octanol[25] and diluted into new TaqMan reagents for the next round of NAC.

Compound enrichment of ΦX 174 virus DNA

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 (Fig. 3a). Both sets reliably detect ΦX 174 DNA (Supplementary Fig. 1). To obtain an optimal enrichment of four hundred, we set the target concentration such that in the first round 0.24% droplets are positive (Fig. 3b(i)). The recovered DNA is diluted into fresh reaction buffer again to achieve a 400-fold enrichment, and subjected to another round of NAC (Fig. 3b(ii)). Because the method is nondestructive, the number of positive drops should be equal for both rounds, but sample loss during preparation for the second round, results in slightly fewer total positives (Fig. 3b(iii)). To confirm 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 (concentrated) while that for lambda shifts to higher cycles (diluted), illustrating enrichment (Fig. 3c(i)). For two rounds compounded, these shifts are greater (Fig. 3c(ii)). To quantify the enrichments, we calculate the enrichment factor e based on the cross-threshold values of the qPCR curves[18]. For one round of sorting ~3 million droplets, the actual enrichment is ~150. For two rounds of sorting ~6 million droplets, the enrichment is ~16,000. To achieve this enrichment in one round would require sorting ~300 million droplets, totaling 16 mL of PCR reagent, and a week of nonstop sorting.

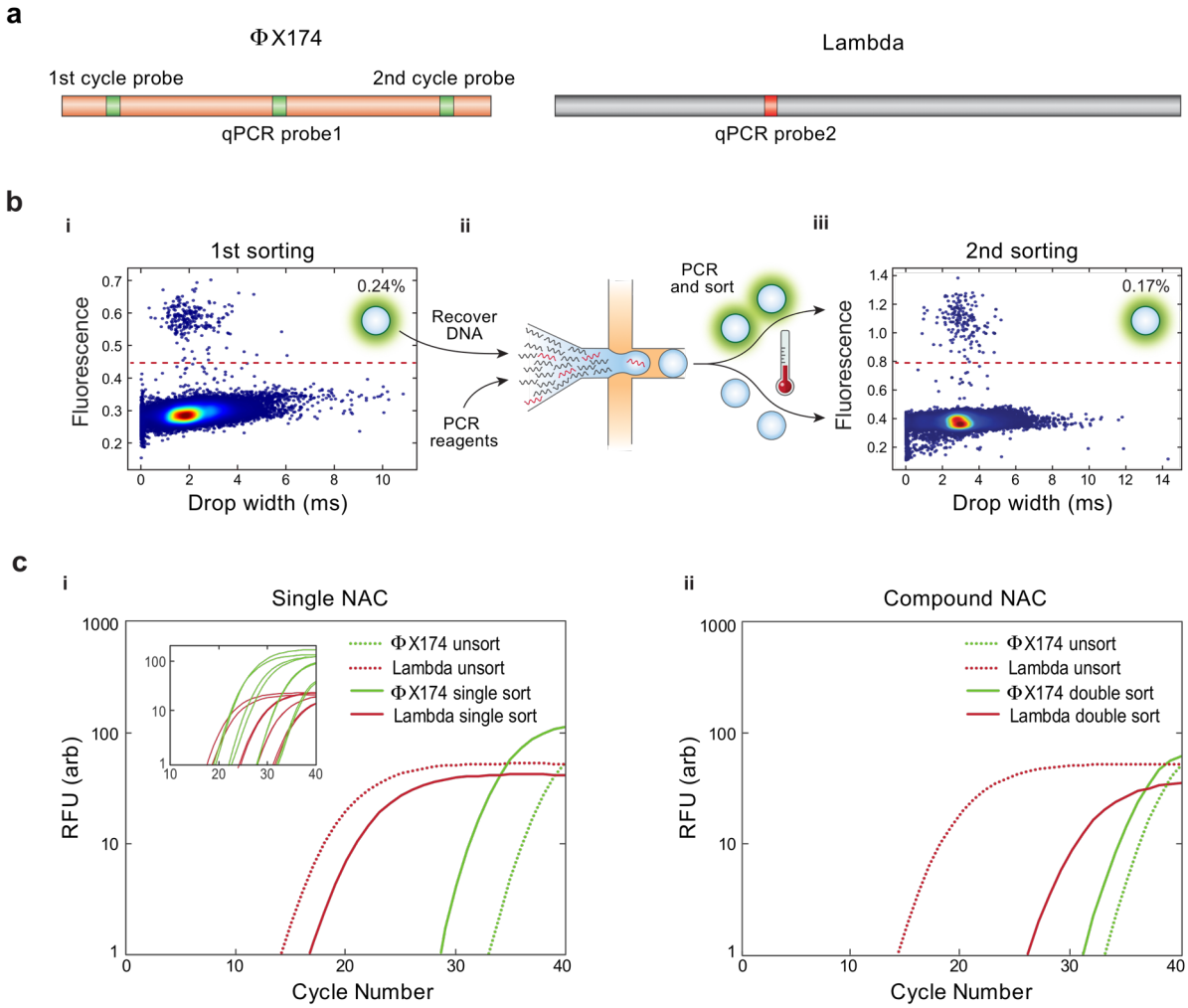

Figure 3. (a) Taqman assays targeting different regions are used in each NAC cycle for detection of droplets containing ΦX 174 genome. Multiplexed qPCR targeting ΦX 174 genome (green) and lambda genome (red) is used to quantify enrichment efficiency. **(b)** Droplets are measured for their fluorescence to determine the sorting. (i) 0.24% drops are PCR positive (distributed above the red dashed line of a representative scatterplot of fluorescence intensity) and isolated in the 1st NAC. (ii) DNA from these drops are recovered and processed to 2nd NAC. (iii) 0.17% drops are PCR positive and isolated in the 2nd NAC. **(c)** qPCR quantifies the enrichment of ΦX 174 DNA from lambda DNA in single or double sorting. Based on the curve shifts, we estimate that (i) the ratio of ΦX 174 -to- lambda DNA increases by ~150 fold by 1st sorting, and (ii) a total of ~16,000 by two rounds of NAC. Standard amplification curves for ΦX 174 DNA and Lambda DNA are shown in the inset.

Single genome sequencing of ultra-rare 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 thus requires potent enrichment. The only effective strategy presently available is terminal dilution PCR in well plates[3,4,27]. This brute force approach aliquots thousands of cells in hundreds of microwells, using long ranged, multi-primer 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 in a single contig with the complete virus genome. Without this dual information, specific proviruses cannot be related to host insertion sites and thereby the combination associated to disease behavior[3,28]. Consequently, other strategies must be employed to infer viral genome and host junction relationships[27,28].

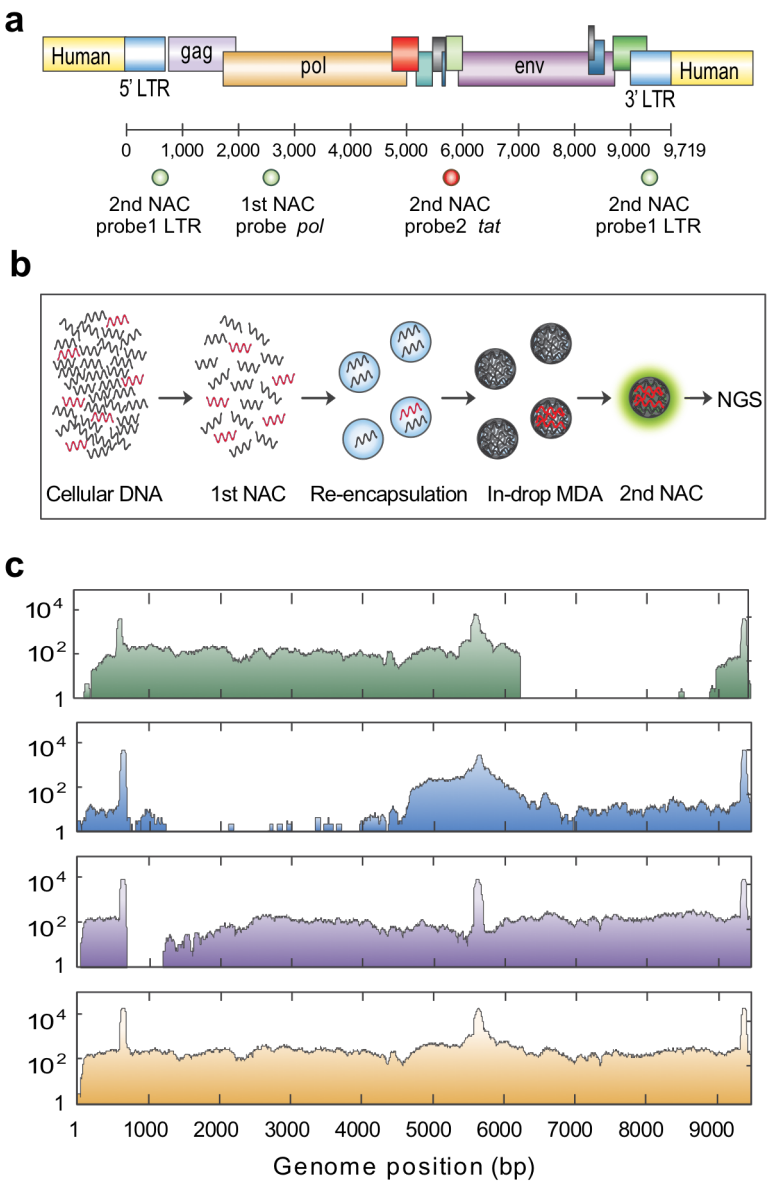

Figure 4. (a) Taqman assay targeting *pol* gene (FAM) is used in 1st NAC, and targeting LTR (FAM) and *tat* (Cy5) is used in 2nd NAC to identify droplets containing HIV genome. **(b)** HIV sequences are enriched from cellular gDNA by two cycles of NAC for sequencing. 1st NAC sorted DNA is amplified by in-droplet MDA before 2nd sorting to generate enough materials for NGS. **(c)** HIV genome and integration site are detected at single provirus resolution in 3 representative sorted drops. Pooled sequencing reads of the 3 drops cover the full-length HIV genome. Coverage maps are generated by mapping sequencing reads to the HIV reference (HXB2). The peaks at *tat* and LTRs are abundant TaqMan PCR amplicons used in the 2nd droplet detection.

Due to the potent enrichment enabled by compound NAC and its ability to recover intact DNA fragments spanning tens of kilobases, performing 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 further dilute the sample with non-HIV cell gDNA at a 1:30 ratio. This dilution models the concentration of latent infection. Due to the extreme rarity of HIV DNA in this sample, we use a multiplexed TaqMan PCR targeting conserved regions of HIV in the second cycle for accurate HIV detection and isolation before sequencing (Fig. 4a). The DNA mixture is encapsulated in droplets and ones containing HIV genomes are isolated. By incorporating in-droplet whole genome amplification before the second round[29], each sorted droplet yields ~3 pg DNA, just enough for sequencing (Fig. 4b). This novel workflow affords ultra-high enrichment power and single drop sequencing, allowing recovery of single provirus genomes (Fig.4c, first to third row). The sorted drop contains the HIV genome and its adjacent host junctions. We thus identify the integration site in host gene ARIH2 by extracting virus-human chimeric reads from the sequencing data. Assembling reads from 3 droplets provides coverage of the full-length viral genome and integration site (Fig.4c, 4th row). In total, sequencing after two rounds detects ~6 million times more proviral reads compared to the initial sample. These results illustrate that compound NAC enables sequencing of extremely rare HIV proviruses, and that the enriched molecules retain complete information on the genetic context of the insertion.

Discussion

By leveraging droplet microfluidics, NAC enables enrichment of target molecules containing specific sequence biomarkers. By processing the sample repeatedly, feeding the output of the first round into the input of the second, the enrichment compounds, allowing recovery of ultra-rare target molecules. In single nucleic acid cytometry, the maximum enrichment is determined by the false positive droplet rate. Partitioning the sample into more droplets enhances the enrichment in a linear fashion, but this does not allow the marked increases required to recover ultra-rare targets. Moreover, such brute force also increases cost and processing time and becomes impractical beyond enrichments of 30,000[8,21]. Our method eliminates the need for large numbers of droplets and exponentially increases the maximum enrichment limit to 10^9 fold, allowing highly specific target recovery at high throughput.

Compound NAC opens new avenues in target enrichment due to unique features. First, it allows the isolation of long molecules using minimal sequence information. Second, integration of two rounds of NAC enables time and cost-effective target enrichment of millions-fold and is extendable to further enrichment by engaging additional rounds. For these reasons, compound enrichment is not only efficient in isolating and sequencing ultra-rare HIV proviruses, as we demonstrate here, but also provides a general strategy for rare genomic targets, including other retrovirus infections, human genetic mutations, gene clusters in metagenomic samples, and rare viral genomes in environmental or clinical samples. Furthermore, compound NAC can be applied to any nucleic acid detectable by PCR and, thus, is applicable to RNA by adding a reverse transcription step, which would allow isolation and sequencing of rare RNAs or transcripts that are important in disease studies requiring monitoring gene expression changes, or detecting fusion genes and low-abundance variants.

Methods

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 transferred to a silicon wafer (University Wafer) using negative photoresist (MicroChem, SU-8 2025) by UV photolithography. 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 ºC oven for 2 hr. 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 ºC 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 ºC overnight for hydrophobicity.

Droplet TaqMan PCR

ΦX 174 virion DNA and Lambda DNA (New England BioLabs) were added to PCR reagents containing 1X 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 stability of droplets during thermal cycling[20]. 1,2-propanediol was used as a PCR enhancer when low temperature was used for denaturation[30]. Two syringes backfilled with HFE-7500 fluorinated oil (3M, catalog no. 98-0212-2928-5) were loaded with (1) TaqMan PCR reaction mix, (2) HFE-7500 oil with 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 PCR mix and 800 μl/h for oil phase) sustained by computer programmed syringe pumps (New Era). Monodispersed droplets (diameter ~40 μm) were generated and collected to 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 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: 2 min 30 s at 86 ºC; 35 cycles of 30 s at 86 ºC, 1 min 30 s at 60 ºC and 30 s at 72 ºC; and a final extension of 5 min at 72 ºC. A low denaturation temperature of 86 ºC 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 10% (v/v) solution of perfluoro-octanol (Sigma-Aldrich, catalog no. 370533) and addition of 10 μl 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 enrichment factor.

Dielectrophoretic sorting

The thermocycled drops were transferred to a 1 ml syringe and reinjected to a microfluidic dielectrophoretic (DEP) sorter (Fig. 2) at 50 μl/h[8,18]. 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 unsorted drops flowed to waste. The salt water electrodes and moat shielding were filled with 2M 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 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 pre-set 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 2nd 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 addition of 20 μl 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 2nd 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" 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 reaction was set as: 1X 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 Supplementary Table S1. The TaqMan assays were tested for specificity and linearity by constructing a serial dilution of 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 were 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 HIV *pol* specific TaqMan probe and biotinylated primers. All primer and TaqMan probe sequences for HIV are listed in Supplementary Table S2. The sorted emulsions were broken using perfluoro-1-octanol and the aqueous was diluted in 5 μl 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 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 supernatant to a fresh tube. The 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 to a 1 ml syringe and incubated at 30 ºC for 20 h. After incubation, MDA droplets and 2nd TaqMan PCR reagents targeting HIV LTR and *tat* were injected into a microfluidic merger device[24]. PCR reagent drops were formed on chip and merged with MDA drops pairwise. Merging was achieved at a salt electrode connected to a cold cathode fluorescent inverter and 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 5% (w/w) PEG-PFPE surfactant for thermal cycling: 3 min at 86 ºC; 35 cycles of 30 s at 86 ºC, 90 s at 60 ºC and 30 s at 72 ºC; and finally, 5 min at 72 ºC. After PCR, the drops were reinjected into a DEP sorter for the 2nd 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 DI H_2O was added to dissolve the sorted DNA. The dissolved DNA was then tagmented using 0.6 μl TD Tagmentation buffer and 0.3 μl ATM Tagmentation enzyme from Nextera DNA Library Prep Kit (Illumina, catalog no. FC-121-1030) for 5min at 55 °C. 1 μl NT buffer were added to neutralize the tagmentation. The tagmented DNA was then mixed with PCR solution containing 1.5 μl NPM PCR master mix, 0.5 μl of each index primers i5 and i7 from Nextera Index Kit (Illumina, catalog no. FC-121-1011) and 1.5 μl 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 Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific) and 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. Sequencing reads were mapped to the HIV reference genome (HXB2) using Bowtie 2[31]. Genomic coverage as a function of genome position was generated using SAMtools[32]. The non-HIV regions of chimeric reads were extracted using extractSoftclipped (https://github.com/dpryan79/SE-MEI) and analyzed by a web base tool for integration sites (https://indra.mullins.microbiol.washington.edu/integrationsites/).

Conflicts of interest

The authors declare that they have no competing financial interests.

Acknowledgements

This work was supported by the Chan Zuckerberg Biohub, the National Institutes of Health (NIH) (Grant No. R01-EB019453-01, R01-HG008978-01 and DP2- AR068129-01), the National Science Foundation CAREER Award DBI-1253293.

References

1 Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**, 111-118, doi:10.1038/nmeth.1419 (2010).

2 Finzi, D. *et al.* Latent infection of CD4(+) T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Medicine* **5**, 512-517 (1999).

3 Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *Journal of Clinical Investigation* **129**, 988-998, doi:10.1172/jci124291 (2019).

4 Hiener, B. *et al.* Identification of Genetically Intact HIV-1 Proviruses in Specific CD4(+) T Cells from Effectively Treated Participants. *Cell Reports* **21**, 813-822, doi:10.1016/j.celrep.2017.09.081 (2017).

5 Xu, P. *et al.* Microfluidic automated plasmid library enrichment for biosynthetic gene cluster discovery. *Nucleic Acids Research* **48**, e48-e48, doi:10.1093/nar/gkaa131 (2020).

6 Suenaga, H. Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ. Microbiol.* **14**, 13-22, doi:10.1111/j.1462-2920.2011.02438.x (2012).

7 Sharon, I. & Banfield, J. F. Genomes from Metagenomics. *Science* **342**, 1057-1058, doi:10.1126/science.1247023 (2013).

8 Eastburn, D. J. *et al.* Microfluidic droplet enrichment for targeted sequencing. *Nucleic Acids Research* **43**, doi:10.1093/nar/gkv297 (2015).

9 Wei, X. M. *et al.* Identification of Sequence Variants in Genetic Disease-Causing Genes Using Targeted Next-Generation Sequencing. *Plos One* **6**, doi:10.1371/journal.pone.0029500 (2011).

10 Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* **18**, doi:10.1186/s13059-017-1212-4 (2017).

11 Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nature Reviews Microbiology* **15**, 183-192, doi:10.1038/nrmicro.2016.182 (2017).
12 Mertes, F. *et al.* Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics* **10**, 374-386, doi:10.1093/bfgp/elr033 (2011).
13 Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9296-9301, doi:10.1073/pnas.0803240105 (2008).
14 Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* **27**, 1025-U1094, doi:10.1038/nbt.1583 (2009).
15 Iwase, S. C. *et al.* HIV-1 DNA-capture-seq is a useful tool for the comprehensive characterization of HIV-1 provirus. *Scientific Reports* **9**, doi:10.1038/s41598-019-48681-5 (2019).
16 Bodi, K. *et al.* Comparison of commercially available target enrichment methods for next-generation sequencing. *J Biomol Tech* **24**, 73-86, doi:10.7171/jbt.13-2402-002 (2013).
17 Clark, I. C. & Abate, A. R. Finding a helix in a haystack: nucleic acid cytometry with droplet microfluidics. *Lab on a Chip* **17**, 2032-2045, doi:10.1039/c7lc00241f (2017).
18 Lim, S. W., Tran, T. M. & Abate, A. R. PCR-Activated Cell Sorting for Cultivation-Free Enrichment and Sequencing of Rare Microbes. *Plos One* **10**, doi:10.1371/journal.pone.0113549 (2015).
19 Lance, S. T., Sukovich, D. J., Stedman, K. M. & Abate, A. R. Peering below the diffraction limit: robust and specific sorting of viruses with flow cytometry. *Virology Journal* **13**, doi:10.1186/s12985-016-0655-7 (2016).
20 Sukovich, D. J., Lance, S. T. & Abate, A. R. Sequence specific sorting of DNA molecules with FACS using 3dPCR. *Scientific Reports* **7**, doi:10.1038/srep39385 (2017).
21 Han, H. S. *et al.* Whole-Genome Sequencing of a Single Viral Species from a Highly Heterogeneous Sample. *Angewandte Chemie-International Edition* **54**, 13985-13988, doi:10.1002/anie.201507047 (2015).
22 Tao, Y. *et al.* Artifact-Free Quantification and Sequencing of Rare Recombinant Viruses by Using Drop-Based Microfluidics. *Chembiochem* **16**, 2167-2171, doi:10.1002/cbic.201500384 (2015).
23 Rowlands, V. *et al.* Optimisation of robust singleplex and multiplex droplet digital PCR assays for high confidence mutation detection in circulating tumour DNA. *Scientific Reports* **9**, doi:10.1038/s41598-019-49043-x (2019).
24 Sciambi, A. & Abate, A. R. Generating electric fields in PDMS microfluidic devices with salt water electrodes. *Lab on a Chip* **14**, 2605-2609, doi:10.1039/c4lc00078a (2014).
25 Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols* **8**, 870-891, doi:10.1038/nprot.2013.046 (2013).
26 Bruner, K. M. *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120-+, doi:10.1038/s41586-019-0898-8 (2019).
27 Patro, S. C. *et al.* Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 25891-25899, doi:10.1073/pnas.1910334116 (2019).
28 Wiegand, A. *et al.* Single-cell analysis of HIV-1 transcriptional activity reveals expression of proviruses in expanded clones during ART. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E3659-E3668, doi:10.1073/pnas.1617961114 (2017).
29 Sidore, A. M., Lan, F., Lim, S. W. & Abate, A. R. Enhanced sequencing coverage with digital droplet multiple displacement amplification. *Nucleic Acids Research* **44**, doi:10.1093/nar/gkv1493 (2016).
30 Mousavian, Z., Sadeghi, H. M. M., Sabzghabaee, A. M. & Moazen, F. Polymerase chain reaction amplification of a GC rich region by adding 1,2 propanediol. *Adv Biomed Res* **3**, 65-65, doi:10.4103/2277-9175.125846 (2014).

31 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354, doi:10.1038/nmeth.1923 (2012).
32 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).