High-coverage, massively parallel sequencing of single-cell genomes with CAP-seq

Mengzhen Li[a,#], Xuanpei Zhai[a,#], Jie Li[a,#], Shiyan Li[b,f], Yanan Du[a], Jian Zhang[a], Rong Zhang[a], Yuan Luo[c], Wu Wei[b,*] and Yifan Liu[a,d,e,*]

[a] *School of Physical Science and Technology, ShanghaiTech University, Shanghai 201210, China.*

[b] *Lingang Laboratory, Shanghai 200031, China.*

[c] *State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China.*

[d] *State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China.*

[e] *Shanghai Clinical Research and Trial Center, Shanghai 201210, China.*

[f] *CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, 200031, China.*

* Corresponding author. Email: wuwei@lglab.ac.cn (W.W.), liuyf6@shanghaitech.edu.cn (Y.L.)

# These authors contribute equally to this work.

Keywords: single-cell, whole genome sequencing, microbe, microfluidics, metagenomics

Abstract

Microorganisms dominate Earth's ecosystems in both abundance and diversity. Studying these organisms relies on their genome information, but obtaining high-quality genomes has long been challenging due to technical limitations in genomic sequencing. Traditional genome sequencing methods are limited by the need to culture isolated strains, excluding unculturable taxa and restricting the study of complex communities. Metagenomics bypasses this but lacks single-cell resolution and often misses rare, critical species. Droplet-based single-cell genomics offers high-throughput genome library preparation but faces challenges like requiring complex microfluidic setups and low genome coverage. To address these challenges, we introduce CAP-seq, a high-throughput single-cell genomic sequencing method with markedly improved genome coverage and ease of use. CAP-seq employs semi-permeable compartments that allow reagent exchange while retaining large DNA fragments, enabling efficient genome processing. This innovation results in higher-quality single amplified genomes (SAGs) and significantly improves resolution. In validation tests with simple and complex microbial communities, CAP-seq yielded high-quality SAGs with over 50% genome coverage, capturing rare taxa more accurately and providing detailed insights into strain-level variation. CAP-seq thus offers a scalable, high-resolution solution for microbial genomic analysis, overcoming the limitations of droplet-based single-cell genomics and enhancing the study of complex microbial ecosystems.

Main

Our planet hosts a vast array of organisms, with microorganisms (primarily bacteria, archaea, and fungi) dominating in both abundance and diversity[1-3]. Current estimates suggest the existence of up to a trillion microbial species, most of which remain uncharacterized[3-5]. This remarkable genomic diversity constitutes a rich repository of genetic material[6]. For decades, researchers have been sequencing microbial genomes to understand their biology and community dynamics[7,8] (e.g., human microbiota), while leveraging the resulting sequence libraries[9] to discover novel drug candidates[10,11], molecular tools[12], and biosynthetic pathways[13,14].

Classical genome sequencing of microorganisms relies on pure culture of isolated strains to amplify their DNA[15], but this method excludes unculturable taxa, which can occupy a significant portion of a given community[3]. It is also labor-intensive, limiting its throughput and suitability for profiling complex ecosystems. To overcome these issues, cultivation-free methods like metagenomics have been widely adopted[16,17]. However, these techniques compromise single-cell or strain-level resolution, depending on *in-silico* binning to assemble metagenomic "genomes"[18-20]. In addition, these methods are often biased towards abundant groups during the library preparation[21,22] and binning steps[23], missing low-abundance taxa that are critical to community function[24,25].

To tackle the challenges in community-wide approaches, efforts have been devoted to single-cell genomics[26]. This technique isolates individual cells and then prepares their genomes independently through processes such as cell sorting, lysis, and whole-genome amplification[26]. While this approach offers valuable insights, it remains limited by resource availability, resulting in low throughput. Recently, the workflow has been implemented in droplet microfluidics to enable thousands of single-cell genome libraries to be prepared simultaneously[27,28]. However, this requires multiple (e.g., five) advanced microfluidic modules and experimentations, posing significant technical challenges for non-experts and thus limiting broad use. In addition, droplets act as isolated environments, preventing purification and buffer exchange[29]. It is thus difficult to optimize sequential reactions, such as lysis, genome amplification, fragmentation, and

indexing, in the same droplet. As a result, the quality of single amplified genomes (SAGs) is often low, with typical genome coverage under 5%, restricting our understanding of microbial communities.

Here, we introduce single-cell genomic sequencing using compartments with adjusted permeability (CAP-seq), a high-throughput method that achieves high-coverage SAGs with only two microfluidic steps. As outlined in Fig. 1a, CAP-seq begins by encapsulating individual microbial cells into semi-permeable compartments, referred to as CAPs (the first microfluidic step). These compartments are designed with an engineered hydrogel shell that allows the selective transport of smaller molecules[30] (e.g., enzymes, dNTPs, buffers) while retaining large DNA fragments (>600 bp), Fig. 1b. This unique feature enables efficient washing and reagent exchange throughout the process. In the subsequent steps, various assays are conducted on the CAP-encapsulated cells, all in bulk solutions, including cell lysis, whole-genome amplification, and DNA fragmentation.

Once the genomic DNA is processed, the CAPs, now containing single-cell genomes, are co-encapsulated with DNA barcode beads in droplets (second microfluidic step). This step assigns each genome a unique cell barcode. Afterward, the CAPs are extracted from the droplets, washed, and dissolved to release the barcoded DNA fragments (~1 kb). These fragments are then further amplified and prepared for nanopore sequencing. Finally, the sequenced reads are categorized into individual SAGs based on their cell barcodes, yielding high-coverage genomes with significantly improved throughput and resolution. CAP-seq thus overcomes limitations in droplet single-cell genomics, providing a streamlined, scalable approach for studying microbial communities with enhanced genome recovery.

We validated the CAP-seq workflow using a 1:1 mixture of *Escherichia coli* and *Bacillus subtilis* cells. Around 30,000 cells were processed. The resulting library was sequenced by the Oxford nanopore sequencer, yielding 84.5 million raw reads. Quality control and our in-house binning pipeline identified 48.5 million reads with full-length and correct cell barcodes, which were sorted into 5,484 groups of ≥1,000 reads, including 1,607 groups

with ≥10,000 reads (Fig. 1c). After subsequent barcode trimming, the read lengths ranged from 300 to 1,500 nt, with an average length of 898 nt (Fig. 1d). Organism-specific mapping (Fig. 1e) revealed that 56.4% of barcode groups were assigned to *E. coli* (98.8% reads on average mapped to the *E. coli* genome), and 39.1% were mapped to *B. subtilis* (average of 98.5% mapped reads). A small fraction (4.5%) of groups were mixed. These results validate the CAP-seq workflow, demonstrating its capability for high-throughput, long-read sequencing of single microbial cells.

Next, we tested CAP-seq with a more complex mock microbial community composed of four bacterial species with varied Gram polarities at uneven ratios (Fig. 2a). Sequencing around 30,000 cells and bioinformatic binning resulted in 3,490 barcode groups containing over 1,000 reads, 91% of which had a purity (percent reads mapped to a dominant genome) of over 95%, and 82.6% with over 99% purity (Fig. 2b). CAP-seq's species profile closely matched the community composition, capturing rare species more accurately than metagenomic sequencing, which was biased towards the more abundant species (Fig. 2c). Specifically, *S. aureus* and *K. pneumoniae*, which accounted for only 0.3% of the metagenomic reads, were accurately represented by CAP-seq, with 173 (5.4%) and 13 (0.4%) SAGs assigned to each species, respectively. These results highlight CAP-seq's ability to provide an unbiased microbial population profile, capturing rare taxa with single-cell accuracy, making it more effective for profiling complex microbial communities.

After mapping them to reference genomes, we calculated the genome completeness of 3,177 high-quality SAGs. Since genome completeness is depth-dependent, we plotted coverage versus sequencing depth (Fig. 2d). At 10X depth, most SAGs achieved over 50% genome coverage, regardless of Gram status. This demonstrates CAP-seq's ability to generate high-completeness SAGs with moderate sequencing efforts. The improved coverage is likely due to the CAPs' semi-permeable nature, enabling efficient washing, buffer exchange, and sustained reagent supply, which enhance key biochemical steps, especially whole-genome amplification, during library preparation.

We further examined the *E. coli* groups, which contained two distinct strains: BL21 and MACH1 (Fig. 2a). These strains were engineered to carry specific plasmid-encoded genes: BL21 harbored AmpR, while MACH1 carried sfGFP and CmR. We profiled these genes across 3,210 *E. coli* SAGs and observed distinct patterns in their relative abundance (Fig. 2e). The abundance of sfGFP and CmR were highly correlated, being prevalent in SAGs where AmpR was absent. In the remaining SAGs, only AmpR was detected, with neither sfGFP nor CmR present. These distinct gene abundance patterns enabled us to distinguish between the two strains. As a result, 10 SAGs were assigned to MACH1 and the rest to BL21. These findings confirm CAP-seq's ability to resolve intraspecific heterogeneity with single-cell resolution, enabling new avenues of research, such as tracking mobile genetic elements across individual cells and mining genes directly from SAGs.

In conclusion, CAP-seq enables the massively parallel sequencing of single-cell genomes. Compared to existing single-cell genomic methods, CAP-seq offers significantly improved coverage of SAGs and greater ease of use. We expect that CAP-seq will have broad applications across various fields related to microorganisms, paving the way for microbial single-cell genomics and genome-resolved metagenomics.

Author Contributions

Y. Liu conceived the idea of CAP-seq; Y. Liu, J. Li, and R. Zhang developed the workflow; M. Li performed the experiments presented in this work; Y. Du, J. Zhang, and Y. Luo assisted in the method development; X. Zhai and S. Li conducted the bioinformatics; Y. Liu and X. Zhai analyzed the data; Y. Liu and W. W supervised the research; Y. Liu prepared the figures and wrote the manuscript.

Competing Interests

The authors declare no competing interest.

Figures

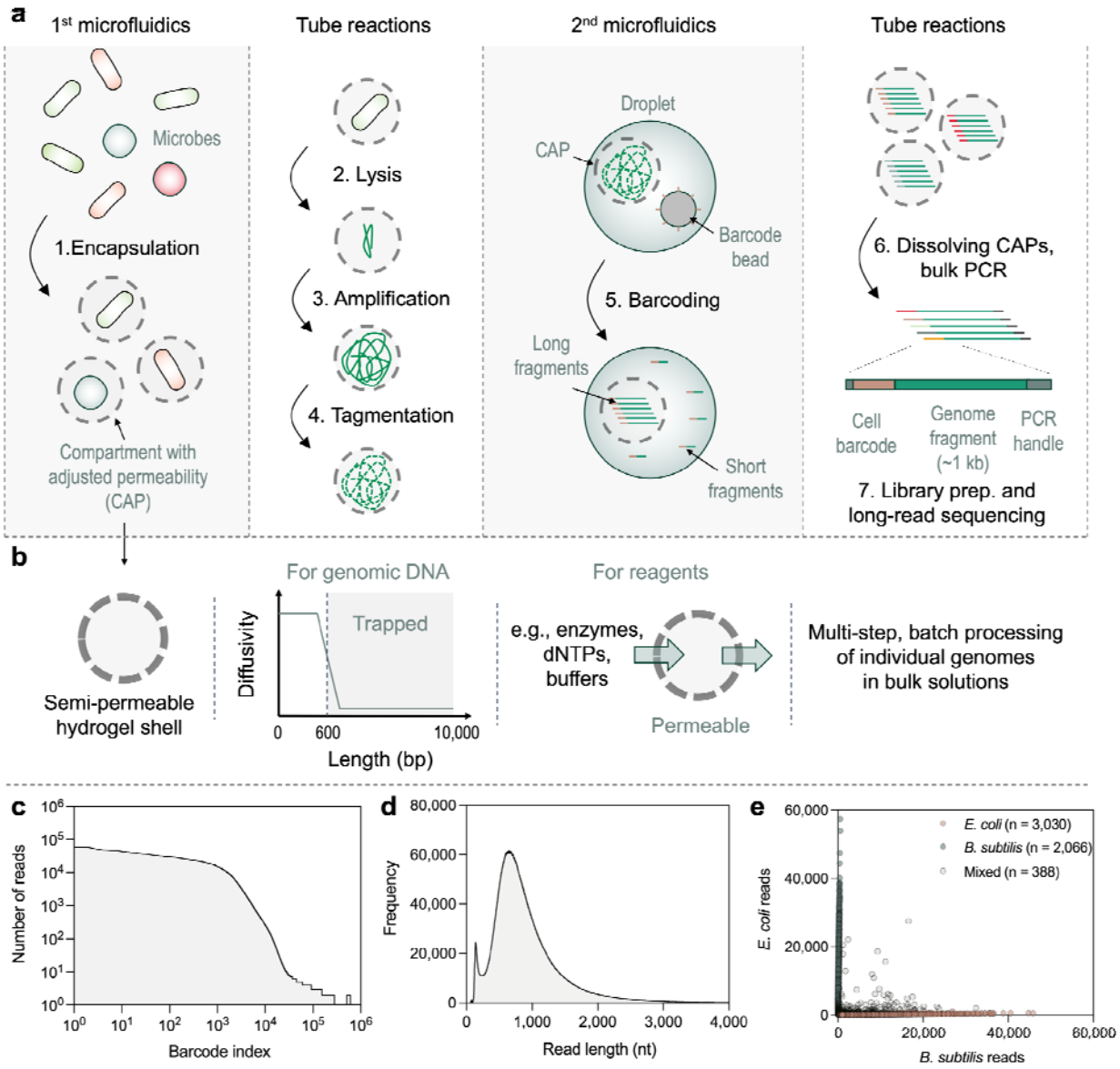

Figure 1. CAP-seq enables high-throughput and long-read sequencing of single-cell genomes. (a) Conceptual workflow of CAP-seq. The workflow generates amplified and barcoded single-cell genomic fragments suitable for nanopore sequencing, with only two microfluidic steps required. (b) Key features of CAP, a kind of semi-permeable hydrogel compartments. Through permeability engineering, the CAPs retain long DNA fragments (> 600 bp) while being permeable to other reagents used in the workflow. This allows for repeated washing and buffer exchange, enabling multistep processing of the individual genomes being retained. (c-d) Validation of CAP-seq using a 1:1 mixture of *Escherichia coli* and *Bacillus subtilis* cells. (c) Read distribution across different

barcode groups. (d) The length profile of all the reads after quality control, binning, and barcode trimming. (e) Scatter plot depicts the number of *E. coli* and *B. subtilis* reads associated with each barcode group.

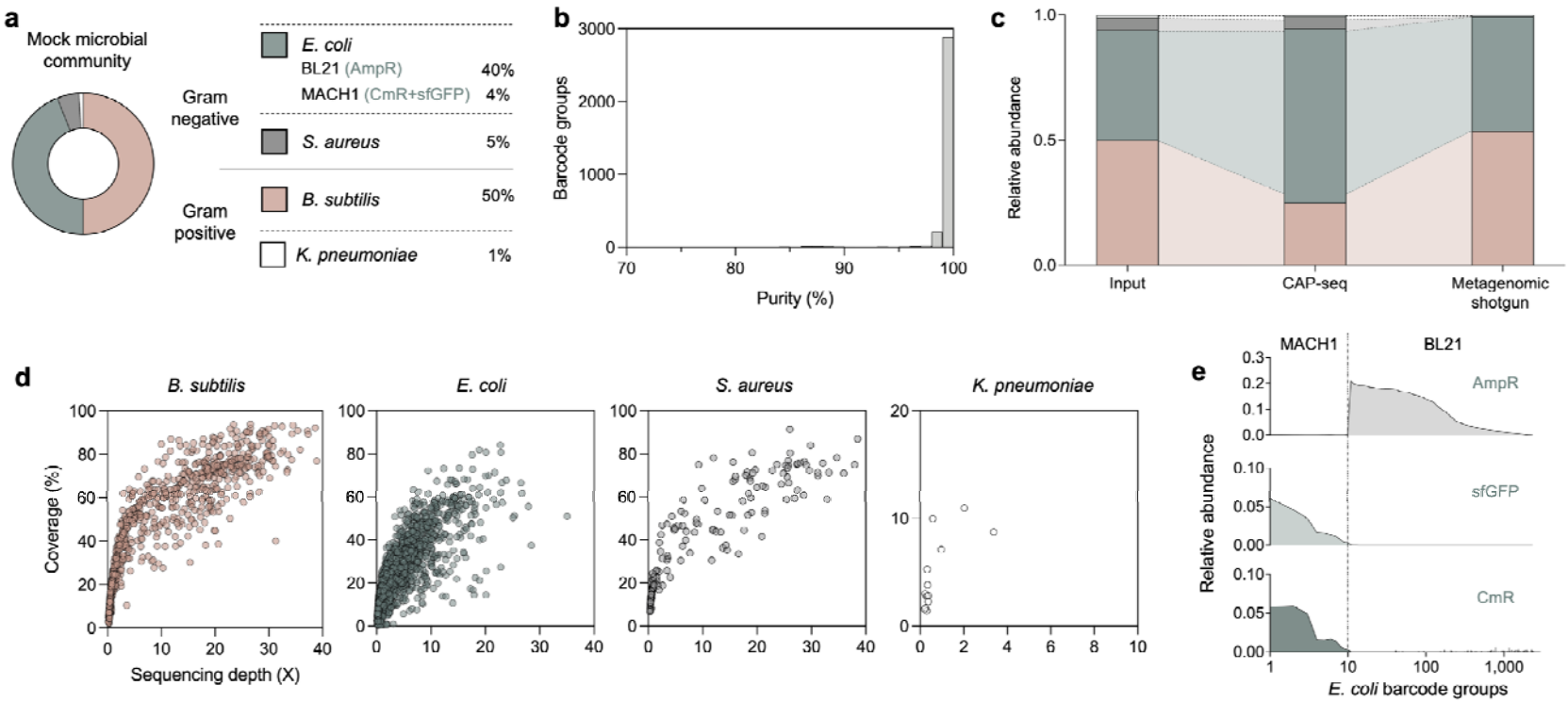

Figure 2. Cap-seq provides an unbiased profile of rare microbial populations and generates high-coverage single amplified genomes. (a) A mock microbial community to examine CAP-seq. (b) Purity distribution of all SAGs from the mock community sample, with 91% SAGs exceeding 95%. (c) Relative abundance of microbial species derived from CAP-seq and metagenomic shotgun sequencing. Compared with metagenomics, CAP-seq is more capable of capturing rare taxa, with single-cell accuracy. (d) Genome coverage of 3,490 SAGs against the sequencing depth, separately plotted by species. (e) Relative gene abundance of AmpR, sfGFP, and CmR across 3,210 *E. coli* SAGs.

References

1. Jiao, J.Y. et al. Microbial dark matter coming to light: challenges and opportunities. *National Science Review* **8**(2021).
2. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437 (2013).
3. Yarza, P. et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* **12**, 635-645 (2014).
4. Dykhuizen, D.E. Santa Rosalia revisited: Why are there so many species of bacteria? *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **73**, 25-33 (1998).
5. Parte, A.C., Carbasse, J.S., Meier-Kolthoff, J.P., Reimer, L.C. & Göker, M. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. *International Journal of Systematic and Evolutionary Microbiology* **70**, 5607-5612 (2020).
6. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research* **20**, 947-959 (2010).
7. Qin, J.J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-70 (2010).
8. Methé, B.A. et al. A framework for human microbiome research. *Nature* **486**, 215-221 (2012).
9. Lok, C. Mining the Microbial Dark Matter. *Nature* **522**, 270-273 (2015).
10. Medema, M.H., de Rond, T. & Moore, B.S. Mining genomes to illuminate the specialized chemistry of life. *Nature Reviews Genetics* **22**, 553-571 (2021).
11. Pavlopoulos, G.A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594-602 (2023).
12. Altae-Tran, H. et al. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* **382**, eadi1910 (2023).
13. Wilkinson, B. & Micklefield, J. Mining and engineering natural-product biosynthetic pathways. *Nature Chemical Biology* **3**, 379-386 (2007).
14. Chiang, C.Y., Ohashi, M. & Tang, Y. Deciphering chemical logic of fungal natural product biosynthesis through heterologous expression and genome mining. *Natural Product Reports* **40**, 89-127 (2023).
15. Goig, G.A. et al. Whole-genome sequencing of directly from clinical samples for high-resolution genomic epidemiology and drug resistance surveillance: an observational study. *Lancet Microbe* **1**, e175-e183 (2020).
16. Liu, Y.X. et al. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & Cell* **12**, 315-330 (2021).
17. Ustick, L.J. et al. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* **372**, 287-291 (2021).
18. Nissen, J.N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* **39**, 555-560 (2021).

19. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nature Methods* **20**, 1170-1173 (2023).
20. Yue, Y. et al. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* **21**, 334 (2020).
21. Poulsen, C.S., Ekstrom, C.T., Aarestrup, F.M. & Pamp, S.J. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology Spectrum* **10**(2022).
22. Nearing, J.T., Comeau, A.M. & Langille, M.G.I. Identifying biases and their potential solutions in human microbiome studies. *Microbiome* **9**(2021).
23. Sun, Z. et al. Challenges in benchmarking metagenomic profilers. *Nature Methods* **18**, 618-626 (2021).
24. Jin, H. et al. Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome. *Gut Microbes* **14**(2022).
25. Pust, M.M. & Tümmler, B. Bacterial low-abundant taxa are key determinants of a healthy airway metagenome in the early years of human life. *Computational and Structural Biotechnology Journal* **20**, 175-186 (2022).
26. Llorens-Rico, V., Simcock, J.A., Huys, G.R.B. & Raes, J. Single-cell approaches in human microbiome research. *Cell* **185**, 2725-2738 (2022).
27. Zheng, W.S. et al. High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* **376**, eabm1483 (2022).
28. Lan, F., Demaree, B., Ahmed, N. & Abate, A.R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nature Biotechnology* **35**, 640-646 (2017).
29. Moragues, T. et al. Droplet-based microfluidics. *Nature Reviews Methods Primers* **3**, 32 (2023).
30. Li, L.Y. et al. Permeability-Engineered Compartmentalization Enables In Vitro Reconstitution of Sustained Synthetic Biology Systems. *Advanced Science* **9**(2022).