ADVANCED OPTICAL MATERIALS
www.advopticalmat.de

Computational Design of Synthetic Optical Barcodes in Microdroplets

Fumiko Kawasaki, Takahiro Mimori, Yuka Mori, Hiroyuki Aburatani, Nozomu Yachie, Issei Sato, and Sadao Ota**

Barcodes are useful for identifying objects across time, space, and information modalities. However, materializing and decoding optical and multimodal barcodes on microscopic objects remains difficult despite the increasing need for multiplexed cell analysis. Here, a computational design of randomly combinatorial is presented, yet decodable barcodes in microdroplets. The design is based on a novel Real2Sim2Real framework: it first collects experimental images of optically distinct microparticles, then simulates massive combinatorial images by randomly assembling the imaged particles to train a neural network-based decoder. It is demonstrated that the decoder, even though trained via simulation, accurately identifies the randomly assembled particles in real hydrogel microdroplets. It also shows that the microdroplets with an additional DNA barcoding functionality are applicable to individually link independently measured microscopic images and transcriptome profiles of pooled single cells.

1. Introduction

Barcodes are optically readable forms useful for identifying an object often used to track individual objects across physically and temporally disconnected measurements. In a macroscopic world, barcodes are designed computationally, printed, or displayed to be physically assigned to objects of interest and tracked by common optical devices such as barcode readers and cameras. Due to the potential of barcoding for characterizing microscale objects, intensive efforts were made to create various functional microscale barcodes on computed pre-designs.[1] However, it is demanding to materialize a large number of pre-designed optical and multimodal barcodes separately and assign them individually to randomly distributed microscopic objects such as biological cells and microdroplets.

A promising approach for generating microscopic barcodes that have sufficient variations and are readily assigned to target objects is using a random combination of elemental coding objects as an object identifier.[2] However, such randomly bottom-up-synthesized barcodes are often difficult to decode in practice, in contrast to the pre-designed and printed or top-down-fabricated barcodes, which are easy to identify and distinct from each other by design. For example, while segmenting elements has been an indispensable step in conventional image-based decoding,[2c,3] the elemental objects often overlap in real images such that the segmentation-based analysis can be easily ruined. One may think that a data-driven approach is promising for learning the required image recognition tasks considering the success in machine learning-based cell segmentations.[4] However, this approach is also hindered by the inherent difficulty in experimentally preparing training data of massive combinations and spatial locations with ground truth barcode labels. Moreover, existing approaches for random combinatorial barcoding have relied on the direct attachment of the elemental coding objects to the cell membrane or their uptake by cells, which may cause unexpected cellular responses and increase variability in the number of particles to be tagged.

Here we present a simulation-based design of combinatorial barcoding material adaptable to the real world. Specifically, we demonstrate to design optical and multimodal materials that barcode target objects by compartmentalizing them with multiple elemental image barcoding beads (iBBs); a novel Real2Sim2Real machine learning approach allows us to accurately decode the iBB combinations without segmenting each iBB. In this design, we first experimentally synthesize a small variety of iBBs with well-resolved spectral properties and obtain microscopic images of each (**Figure 1**. top left). Using the real images of the beads, we then computationally generate numerous images containing multiple beads with massive combinations and arbitrary spatial locations in spherical spaces (Figure 1. top right). These

F. Kawasaki, T. Mimori, Y. Mori
Center for Advanced Intelligence Project
RIKEN
Chuo-ku, Tokyo 103-0027, Japan
H. Aburatani, S. Ota
Research Center for Advanced Science and Technology
The University of Tokyo
Meguro-ku, Tokyo 153-8904, Japan
E-mail: sadaota@solab.rcast.u-tokyo.ac.jp
N. Yachie
School of Biomedical Engineering
The University of British Columbia
Vancouver 1200-1874, Canada
I. Sato
Graduate School of Information Science and Technology
The University of Tokyo
Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: sato@g.ecc.u-tokyo.ac.jp
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adom.202302564

DOI: 10.1002/adom.202302564

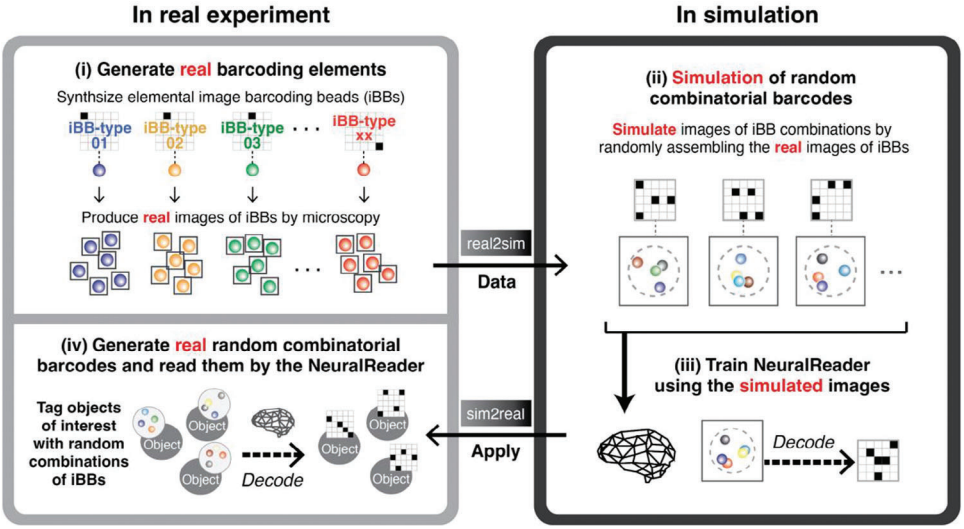

Figure 1. Computational design of optical barcoding materials based on the randomly assembled combination of image barcoding beads. We realized randomly combinatorial, yet decodable optical barcoding materials via a Real2Sim2Real machine learning framework: i) We first materialized distinct types of image-barcoding beads (iBBs) separately and captured their images by optical microscopy. ii) Using the experimentally collected images of the iBBs, we simulated numerous images of spherical barcoding units, each of which contained an arbitrary combination of iBBs at arbitrary spatial locations, and iii) then train an end-to-end algorithm named NeuralReader to decipher barcodes from the realistic simulated images. iv) Lastly, the trained NeuralReader allowed us to decode iBB combinations which were experimentally synthesized and assigned to each object of interest upon generation.

simulated images are used to train a neural network-based barcode reader, which we call NeuralReader, to decode the combinatorial codes in an end-to-end manner (Figure 1. bottom right). We demonstrate that NeuralReader accurately decodes combinatorial barcodes, which are materialized in the real world by encapsulating iBBs inside hydrogel capsules together with intact micro-objects to be tagged (Figure 1. bottom left).

2. Experimental Generation of Image Barcoding Materials

We first generated elemental image barcoding beads (iBBs), wherein we adopted fluorescence spectra as simple and robust optically readable signatures. Microscale beads made of alginate hydrogels were created by emulsifying an alginate solution with a porous glass membrane equipped with an external pressure system (**Figure** 2A,B), followed by a calcium ion-mediated alginate gelation.[5] Covalent conjugation of three fluorescent molecules (i.e., cascade blue, AlexaFluor 568, AlexaFluor 647) to the alginate beads at designed ratios results in the generation of twenty-five distinct image barcodes (Figure 2B; Table S1, Supporting Information); each optical signature of iBBs (iBB types) was distinctly identified by fluorescence intensity (Figure 2C). In our design, a combination of multiple iBBs serves as an optically trackable identifier for a larger number of objects when residing with them (image-code, Figure 1). To materialize the combinatorial barcodes with iBBs, we utilized hydrogel microdroplets (gel units) to tag single cells, given the wide applicability of microdroplets for biological and biochemical assays.[6] More concretely, we generated monodisperse water-in-oil hydrogel microdroplets encapsulating multiple iBBs with single cells at random combinations by microfluidics, followed by extracting the hydrogels from oil to aqueous media for the down-stream uses (Figure 2D,E).

3. Computational Design of Optical Barcodes Based on a Real2Sim2Real Machine Learning

Herein while the possible variation of barcodes rapidly grows along with the number of iBBs per unit (Figure 2F), it is challenging to identify densely and randomly encapsulated iBBs from optical images since aggregated and/or out-of-focus beads become more likely to cause overlaps of signals (Figure 2G). Indeed, a conventional method based on step-by-step particle segmentation, intensity quantification, and classification becomes error-prone as the iBBs in units become dense (herein we performed such analysis and call it a baseline, Figure S1, Supporting Information). On the other hand, an alternative image recognition approach of employing supervised machine learning is difficult since its training requires a massive amount of experimental data of arbitrary iBB combinations and their spatial configurations with proper labels.

To overcome the difficulties in identifying iBB combinations, we propose a computational scheme for the barcode design based on a Real2Sim2Real machine learning approach (Figure 1), and thereby realized an end-to-end neural network-based image decoder named NeuralReader (Figure 2H right; Figure S2, Supporting Information). Notably, the Real2Sim2Real approach allows us to address the challenge of obtaining enough training data by simulating various realistic images of randomly assembled iBBs by combining seed experimental images of each iBB. More specifically, we first image sparsely populated iBBs for each type on a microscope (Figure 2H left), segment 3D patches of isolated iBBs from the z-stack images, and prepare a library of patches after the quality filtering on the imaged beads (Figure 2H

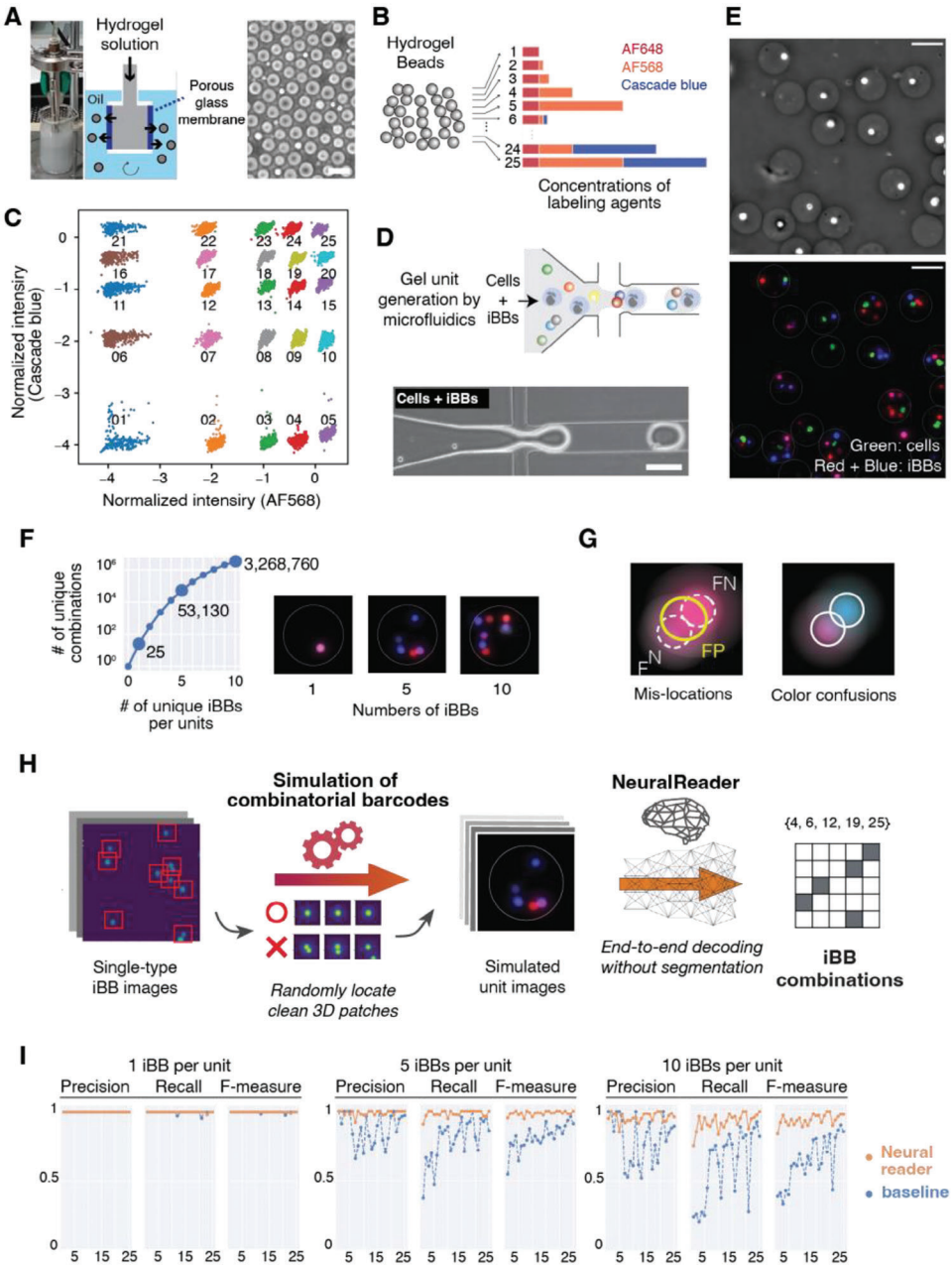

Figure 2. Generation of random combinatorial barcodes with image barcoding beads (iBBs) and end-to-end, segmentation-free decoding of iBB combinations by NeuralReader trained via a real-based simulation scheme. A) Large-scale emulsification of a hydrogel (alginate) solution. B) A schematic diagram of generating twenty-five iBB types via fluorescence functionalization at designed ratios. C) A scatter plot of fluorescence intensities obtained for 25 iBB types. The intensities of Alexa Fluor 568 signals and Cascade blue signals were scaled and normalized by intensities of reference Alexa Fluor 647 signals. D) A snapshot of a microfluidics chip during ID-coded unit generation. The scale bar is for 100 μm. E) An example phase contrast image (left) and a z-projected fluorescence image of the ID-coded units taken using a microscope. White circles drawn in the fluorescence image are for visualizing how each unit is segmented (right). Cells were stained with CellMask Green Plasma Membrane Stain. The scale bars are for 75 μm. F) The number of distinct iBB combinations and an example of simulated optical units are shown for the case with one, five, and ten unique iBBs per unit. G) Illustration of major failure modes in iBB detection with the baseline approach due to an aggregation of beads. On the left panel, FP and FN denote false positives and negatives, respectively. H) Schematics of the proposed end-to-end neural network for decoding iBB combinations (NeuralReader) and a data simulation scheme for training the network. In the simulation, z-stack unit images with arbitrary iBB combinations were synthesized by using 3D patches of isolated iBBs extracted from experimentally imaged sparse single-type iBBs. The whole protocol requires no human annotations for iBB types and locations. I) The performance of iBB detections with NeuralReader and the baseline approach. The precision, recall, and F-measure of 25-type iBB detections are shown for the three simulated datasets wherein iBBs per unit are one, five, and ten, respectively. Although both approaches were nearly perfect on single-iBB units, NeuralReader was much more reliable in the cases with denser iBBs, which were required for uniquely distinguishing more than thousands of randomly combinatorial barcoding units.

middle; Figure S1, Supporting Information). For the training of NeuralReader, we prepare a set of labeled images with various iBB combinations by randomly selecting patches from the library and overlaying those at arbitrary locations. This simulation-based training strategy is counted as a domain randomization technique,[7] in which a real-world adaptable object recognizer is solely trained by an extensive simulation of items at various possible placements and environmental perturbations. We further advanced the technique to incorporate the real iBB images for the simulation seeds and consider z-stack and overlapped signals, which are realistic characters confusing microscopic image analysis.

The architecture of NeuralReader is shown in Figure S2 (Supporting Information), wherein we employ multiple customized 3D convolutions[8] and fully connected layers followed by sigmoid activations to output twenty-five zero-to-one scores, which are rounded to predict a combination of iBBs. The input of NeuralReader is a concatenation of z-stack images taken through the three windows of the fluorescence spectrum around a unit and an additional mask channel representing the inner space of the unit. We trained NeuralReader by minimizing a modified cross-entropy loss for the prediction of iBB combinations with respect to a synthetic dataset of 10 000 unit images, wherein each unit contained one to fifteen iBBs. We evaluated the performance of NeuralReader in detecting iBBs on three datasets of simulated unit images encapsulating one, five, and ten iBBs, respectively, compared with the baseline approach (Figure 2I; Figure S3, Supporting Information). Notably, the average precision/recall of NeuralReader for units densely encapsulating iBBs (five and ten) were 0.99/0.98 and 0.96/0.93, respectively, clearly superior to the corresponding measures of the baseline method, i.e., 0.88/0.80 and 0.80/0.62.

4. DNA-Functionalized iBBs for Multimodal Data Integration

The combinatorial design allows us to generate a large number of unique variations from a small variety of components. By taking this advantage, we further create a large number of multimodal identifiers from a small variety of microparticles functionalized with optical codes and corresponding DNA-barcodes (**Figure** 3A left; Figure S4, Supporting Information). When encapsulated with single cells inside droplets, the combinations of DNA-functionalized iBBs become multimodal single-cell identifiers dually readable by imaging and sequencing, which we named image-DNA coding: ID-coding (Figure 3A right). Here we demonstrate the parallel multi-modal analysis of suspended single cells in a workflow using ID-coding. After generating ID-coded gel units of single cells, we imaged the pooled ID-coded gel units (Figure 3B top), and finally sequenced each unit by adapting a next-generation sequencing (NGS)-based single cell analysis (Figure 3B middle). An iBB combination read by NeuralReader serves as an identifier assigned to each single-cell image data, and an iBB combination represented by corresponding DNA-barcodes serves as an identifier assigned to each single-cell sequencing data. Matching the imaged and sequenced identifiers of ID-coded units enables linking the two completely independent measurements, optical imaging, and NGS, by single cell-basis in silico (Figure 3C). In the step of reading the DNA barcodes of iBBs together with single-cell transcriptome, we carried out "single unit" RNA sequencing library preparation by modifying a previously reported technique for droplet-based single-cell RNA (scRNA) sequencing (drop-seq),[9] where a whole unit is encapsulated inside a droplet together with a sequencing bead (Figure 3D). In the step of identifying the iBB combination of a sequenced unit, we defined positively detected iBB types when significant unique molecular identifiers (UMIs) were counted for the corresponding DNA-barcode compared with a background UMI distribution of the unit (Figure 3E,F) similarly to the reported barcode identification strategy.[10]

5. Single-Cell Image and RNA Profiling of ID-Coded Cells

As a demonstration of the multimodal cell barcoding ability of ID-coding, we experimented on linking single-cell images and sequencing data by using a mixed population of K562 and HEK293 cell lines (**Figure** 4A; Figure S5–S6, Supporting Information). In the experiment, we first imaged units containing K562 cells and HEK293 cells separately for assigning the ground truth labels to each unit (image-labels), then pooled all units to perform the droplet-based single-unit sequencing (Figures S5A and S6A,B, Supporting Information). By image pre-processing (Figure S1, Supporting Information), 5913 units with cellular images were detected (image-units). By sequencing readout analysis, 384 units were detected with transcriptome data (seq-units). Transcriptome profiles of the seq-units appeared in two clusters in a space projected with UMAP,[11] and we assigned cell type labels (RNA-labels) to units in the clusters by using a k-nearest neighbors algorithm using reference gene expression data of K562 and HEK293 cell lines (Figure 4B; Figure S7, Supporting Information). We then decoded iBB combinations in image data using NeuralReader. The average numbers of unique iBBs identified per unit was 4.2 in all NeuralReader outputs (Figure S6D,E, Supporting Information). We finally linked the images and sequencing data of a single unit on the basis of matching between NeuralReader outputs and iBB combinations identified by counting UMIs of each DNA barcode in the single-unit sequencing data. Herein, we linked sequencing data and images when a DNA barcode combination assigned to a single unit matches a uniquely identified iBB combination using NeuralReader within two-edit distance, wherein each edit step altered an iBB in the combination to one of its neighboring types on the 5×5 matrix reflecting the layout of fluorescent intensities (Figure S8, Supporting Information). The consistency of the links was evaluated by cross-checking the RNA-labels of the seq-units and the ground truth labels of the linked image units, which were highly concordant as seen in the image-labels projected on the UMAP space (Figure 4C). Representatives of linked iBB combinations, single-cell images, and gene expression profiles are shown in Figure 4D,E for K562 and HEK293 cells, respectively: label-free brightfield cellular images found one-to-one correspondence with scRNA records. As to the performance of ID-coding, our quantitative evaluation considers a trade-off between the number of linked units and the accuracy (i.e., label consistency between image data and sequencing data) with several different linking conditions: the threshold for the minimum number of unique iBBs per unit and the maximum allowed edit distance between iBB combinatorial patterns.

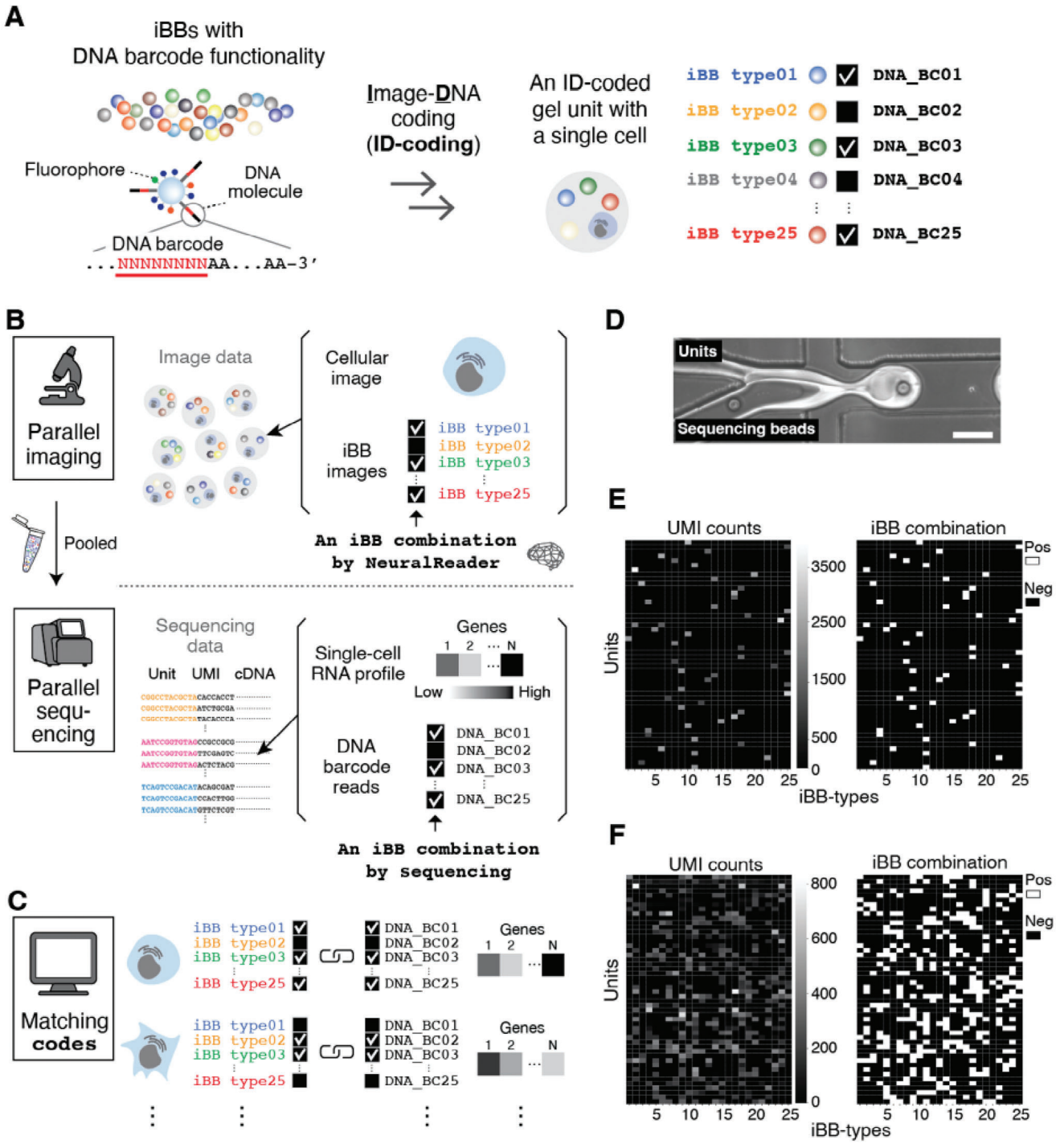

Figure 3. Scalable synthesis of image and DNA (ID)-codes and their application for parallel multimodal single-cell analysis. A) The conceptual illustration of iBBs with DNA barcode functionality and that of ID-coding. B,C) A pooled workflow consisting of generation, imaging, and sequencing of ID-coded single cells. D) A snapshot of a microfluidic chip during "single unit" sequencing library preparation. The scale bar is for 100 µm. E,F) UMI counts and identified seq-codes (iBB combinations) of 50 sequenced ID-coded units in an experiment with (E) sparsely encapsulated iBBs (average 1.1 iBBs/unit) and (F) densely encapsulated iBBs (average 6.0 iBBs/unit).

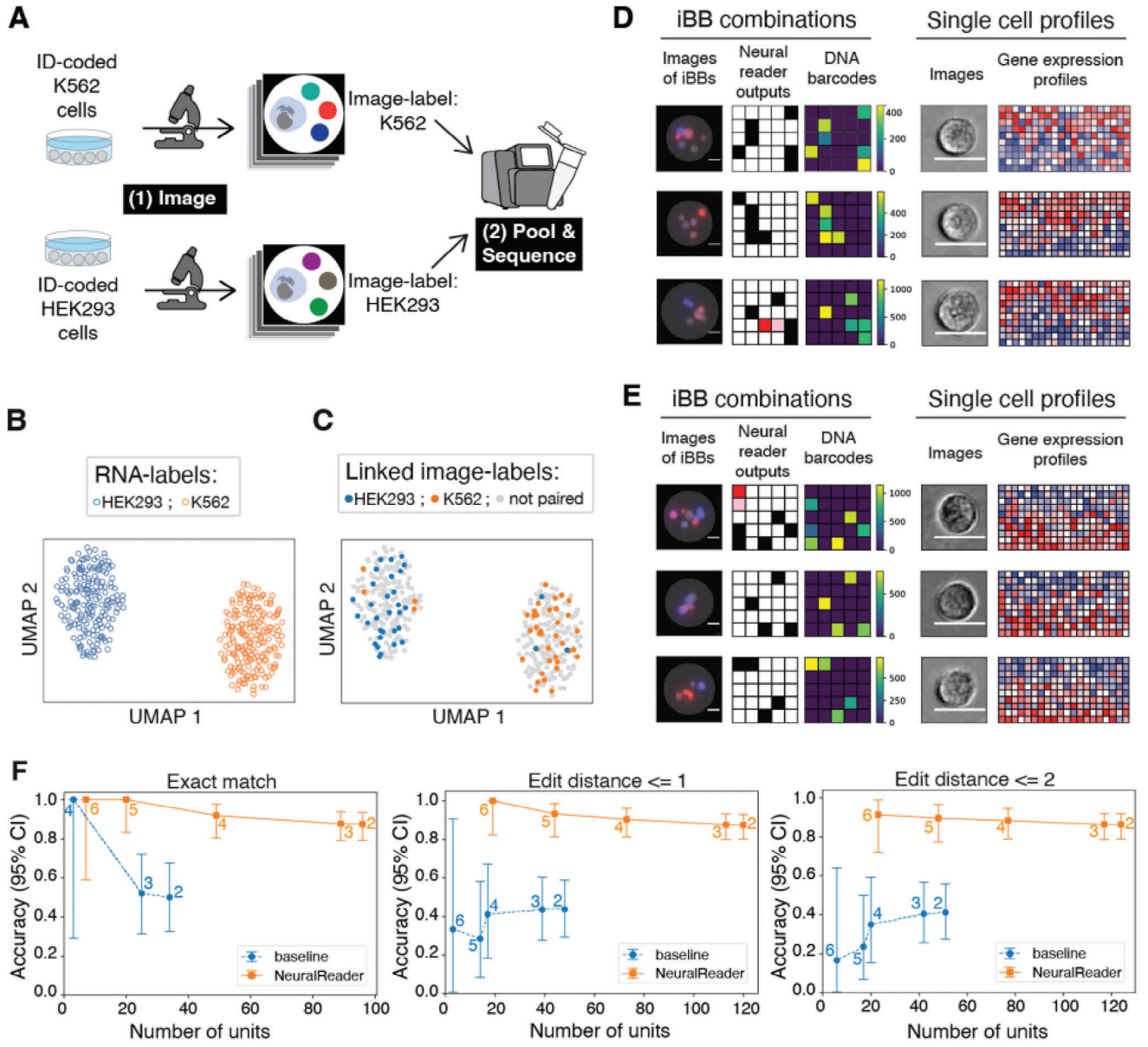

Figure 4. Single-cell images and RNA profiles linked by ID-coding. A) An experimental design. B,C) RNA profiles of seq-units in UMAP coordinates, with cell type labels created based on reference scRNA data (B), and with the ground truth labels of the linked image-units (C). The ground truth labels of the linked image-units by ID-codes show high consistency with the RNA-labels, where image-units encapsulating at least four unique iBBs and uniquely matched to the corresponding seq-units within one edit distance were used for linking. D,E) iBB images (layered phase contrast images and z-projected fluorescent images), iBB combination decoded by NeuralReader, UMI counts of DNA barcodes of individual iBBs, brightfield cell images, and cellular RNA profiles used for the cell-type labeling from individual linked units whose cellular RNA profiles were characterized as "K562" (D) and "HEK293" (E) individually. The decoded iBB combinations are presented in 5 × 5 matrices in accordance with the intensity levels of AF568 and cascade blue fluorescence (Figure 2C). In the matrix showing NeuralReader output, a black, red, and pink cell represents an iBB type detected both by NeuralReader and DNA barcode analysis, only by NeuralReader, and only by DNA barcode analysis, respectively. The expression levels of 200 genes displayed in the RNA profiles are z-score normalized, clipped to the range from −2 (blue) to 2 (red), and ordered according to Wilcoxon's rank-sum test scores from top left to bottom right. The scale bar is for 20 μm. F) Tradeoffs between the number of linked units and the accuracy of linked units are shown with respect to linking conditions: the maximum edit distance allowed for linking (0 to 2) and the minimum number of unique iBBs per gel unit (2 to 6). 95% confidence intervals of the accuracy were calculated using the Clopper-Pearson method. Analysis results of an independently repeated experiment are presented in Figure S9 (Supporting Information).

When we used at least two distinct iBBs per unit and allowed up to an edit distance of two (Figure S8, Supporting Information), the number of linked units was 124 with 86.3% accuracy (95% confidence intervals: 79.0 to 91.8%). As we expected, the accuracy of linked units can be further improved by tightening the linking conditions (Figure 4F). Lastly, our evaluation confirmed that NeuralReader relying on simulation-based training is effective for real data and essential for accurate linking, as the number of linked cases and the accuracy of linking were simultaneously improved from the baseline method with most of the linking conditions (Figure 4F).

Lastly, we present the scalability of our workflow of linking units across image and sequencing spaces in an experiment using approximately 3×10^4 ID-coded units. To ensure sufficient code variations required for unit distinction, we loaded six iBBs per unit on average and only used units with at least five iBBs (i.e., minimum of 53130 possible distinct patterns, Figure 2F) in the linking process. The number of unique unit optical identifiers decoded by NeuralReader was 29501 (Figure S10A, Supporting Information), demonstrating the high capacity of our image coding system at this scale. For the multi-modal analysis, we increased the rate of recovering the ID-coded units after the droplet-based sequencing library preparation by loading a higher number of sequencing beads in this step and concatenating sequencing data based on ID-codes (Figure S10B,C, Supporting Information). With this customized protocol of library preparation, we obtained 3155 sequenced units with at least five iBBs, where 1357 (43.0%) were linked with their corresponding image data (Figure S10B,C, Supporting Information), showing the practical per-experiment-linking efficiency for scalable applications. In addition, the scalability can be further improved by increasing the color diversity of iBBs and changing the imaging platform to one with more color channels and higher resolution to the depth direction.

6. Conclusion

In summary, we present the computational design of randomly combinatorial synthetic barcodes decodable with the neural net-based, end-to-end identifier, and their real-world implementation. In the Real2Sim2Real framework introduced here, we computationally synthesized large-scale images of the iBB combinations using experimental images of each iBB and thereby trained the identifier. The approach circumvents the requirement of a sufficient set of properly labeled, experimentally obtained training data for machine learning of the real world. The identifier, NeuralReader, is proven powerful even in a case when barcode elements are difficult to identify using conventional segmentation methods. In general, when elemental information can be obtained experimentally in advance, we anticipate that the Real2Sim2Real approach is widely applicable to perform elemental decomposition from measurement data obtained as linear superposition of elemental information, regardless of the dimension of the measurement data.

By exploiting the large information capacity of image spaces for object identification, the multimodal barcoding strategy allows the suspended droplets, hydrogels, and cells at a large scale to become trackable across different time, instruments, and information modalities. The microdroplet barcoding thus provides a solution for one of the fundamental limits of suspended pooled units, the loss of object identity across different measurements. In ID-coding, we demonstrated the potential of this solution by enabling the parallel characterization of pooled individual cells both by optical measurement and a single-cell sequencing method; such multi-modal profiling is currently of significant interest for high-resolution biology.[12] We note that a strategy of optically reading DNA barcodes assigned to each cell under microscopy (e.g., spatial transcriptomics or optical pooled screening) is powerful and widely adopted in combination with optical phenotyping methods.[12c,13] Using the same strategy, multimodal analysis of pooled suspended cells using a microchamber array has enabled the linking of images and sequencing data of single cells with high accuracy up to >99%.[12f] Yet, the decoding of the DNA barcodes in these strategies requires the sequential optical imaging of each biochemical reaction involving an iterative exchange of reagent solutions; this requirement increases the cost and time for the experiments and basically limits its application to immobile samples. In contrast, our strategy employs multi-modal identifiers which can be read by a single round of imaging without the iterative reagent exchange and molecular profiling on standard sequencers, which significantly simplifies the decoding process. With our multi-modal identifiers, we demonstrated linking images and sequencing data with a practical accuracy from 86.3% to nearly perfect (Figure 4F). Moreover, our strategy will work seamlessly with droplet-centric assays, which provide a unique, isolated environment for high-throughput cell evaluations in tandem with single-cell sequencing and thus has become powerful tools in drug discovery and gene perturbation analyses.[14] The application is not limited to suspension cells such as immunological cells and blood cells, but also to adherent cells detached from the original environments and suspended in droplets.[15] We thus expect it to find wide-ranging applications in biology and medicine. Finally, the combinatorial design for ID-coding is flexible and versatile such that it can be modified for multiplexed and multimodal observations other than sequencing.[16] Therefore, we envision that such barcoded suspended materials potentially broaden applications of pooled technologies such as flow cytometry and droplet microfluidics, widely in biology, biotechnology, and chemistry.

7. Experimental Section

Oligonucleotides and Sources of Reagents: DNA oligos were purchased from Integrated DNA Technologies, Inc. or Fasmac Co., Ltd. Sequences of primers and DNA oligos are listed in Table S3 (Supporting Information), and sources of reagents and equipment details are listed in Table S4 (Supporting Information).

PDMS Chip Design and Fabrication: Microfluidic chips were designed using AutoCAD software (Autodesk, Figure S11, Supporting Information). Preparation of polydimethylsiloxane (PDMS) microfluidic chips was outsourced from YODAKA Co. Ltd. PDMS chips were hydrophobidized by flowing Aquapel (PPG Industries) through PDMS chips followed by washing and drying.

iBB Preparation: Sodium alginate ULV-L3G (KIMIKA Co.) was functionalized with maleimide functional groups by incubating with EDC (Sigma–Aldrich) and N-β-maleimidopropionic acid hydrazide (BMPH, ThermoFisherScientific) in the MES buffer (pH 5.5). The maleimide-functionalized alginate was further incubated with a thiol-modified DNA oligomer (iBB_universal_ODN, Table S3, Supporting Information), which was pre-activated with BondBreaker TCEP solution

(ThermoFisherScientific). Then, alginate hydrogel beads were generated. Briefly, a 2% (w/v) aqueous solution of sodium alginate IL-6 (KIMIKA Co.) and DNA-bearing alginate was mixed with an equal amount of EDTA-Ca buffer (50 mM CaCl2, 50 mM EDTA, pH 7.2). The solution was emulsified in Droplet Generator oil for EvaGreen (Biorad) using an SPG micro kit (SPG) equipped with 5 µm filter at 8–9 kPa. Acetic acid was added to the oil phase at a final concentration of 0.05% (v/v) to initiate gelation of alginate.[5] Hydrogel beads were extracted to beads-wash buffer (10 mM Tris-HCl, 137 mM NaCl, 2.7 mM KCl, 1.8 mM CaCl2, 0.1% (v/v) Triton X-100, pH 7.5) by addition of perfluoro octanol (FUJIFILM Wako Chemicals, at a final concentration of 20%, v/v) to the oil phase, and the aqueous phase was washed with 20% (v/v) perfluoro octanol in HFE7200 (3 M), and hexane supplemented with 1% (w/v) of Span-80 (Sigma–Aldrich). Fluorescent labeling of the hydrogel beads was achieved by incubating the beads with EDC and various mixtures of fluorescent labeling agents (Table S1, Supporting Information) in an MES buffer (pH 5.5) supplemSupporting Informationented with calcium chloride. DNA labeling of the fluorescently labeled beads was achieved by incubating the beads with idBB_barcode_ODNs (Tables S2 and S3, Supporting Information). At each step of the reaction, beads were washed with beads-wash buffer.

Cells: HEK293 cells were purchased from JCRB CellBank (JCRB9068); K562 cells were purchased from JCRB CellBank (JCRB0019). Murine NIH/3T3 cells were purchased from JCRB CellBank (JCRB0615). K562 cells were grown in RPMI1640 medium (Sigma–Aldrich) supplemented with 10%FBS (Sigma–Aldrich) and 1x Antibiotic Antimycotic (Thermo Fisher Scientific). HEK293 cells were grown in Minimum Essential Medium Eagle (Sigma–Aldrich) supplemented with 10% Horse serum (Thermo Fisher Scientific) and 1x Antibiotic Antimycotic to reach 60–80% confluence. NIH/3T3 cells (used to estimate the rate of doublet errors during NGS library preparation, Figure S6a, Supporting Information) were grown in Dulbecco's Modified Eagles' Medium low glucose (Sigma–Aldrich) supplemented with 10% Newborn Calf serum (Thermo Fisher Scientific) and 1x Antibiotic Antimycotic to reach 60–80% confluence. HEK293 cells and NIH/3T3 cells were treated with TrypLE Express Enzyme (Thermo Fisher Scientific) for 4 min, harvested by 10 volumes of DPBS. The harvested cell pellets were resuspended in a solution of CellMask Green Plasma Membrane Stain (Thermo Fisher Scientific) at x1000 dilution in PBS in a 15 mL low absorption tube, and the cell suspension was incubated at 37 °C for 10 min. The suspension was then pelleted. The pellets were resuspended in PBS, passed through cell strainers, and pelleted again before use.

Combinatorial ID-Coded Unit Preparation: Cell pellets were resuspended in the agarose buffer [0.75% (w/v) Agarose Ultra-low Gelling Temperature (Sigma–Aldrich), 10 mM Tris-HCl, 137 mM NaCl, 2.7 mM KCl, and 1.8 mM CaCl2]. The suspension was passed through pluriStrainer (20 µm), mixed with iBBs pellets (typically, seven to ten times to the number of cells), and loaded into a 1 mL plastic syringe. HFE7500 supplemented with 2% 008-fluorosurfactant (RAN Biotechnologies) was loaded into a 10 mL plastic syringe. Syringes were connected to the droplet generation chip (Figure S11, Supporting Information). The liquid transfer was carried out by syringe pumps at a flow rate of 7 µL min^{-1} (for cell/iBB suspension), and 25 µL min^{-1} (for oil), which resulted in monodisperse droplets. After droplet generation, water-in-oil droplets were cooled at 4 °C at 500 rpm shaking for 10 min to solidify agarose gel. To the droplets, imaging buffer [FluoroBrite DMEM (ThermoFisherScientific) supplemented with 2 mM L-glutamine (Wako)] was slowly added, and hydrogel units were extracted into the aqueous layer by addition of perfluoro octanol (final concentration 20% v/v) to the oil layer, and the aqueous layer was washed with 20% perfluoro octanol in HFE7200, and hexane supplemented with 1% (w/v) of Span 80.

Imaging: Combinatorial ID-coded units in the imaging buffer were gently loaded on a glass-bottom 6 well plate (MatTek Corporation) covered by a round-shaped cover glass (Matsunami Corporation). Unit images were recorded on InCellAnalyzer6000 (GE Healthcare) at x10 magnification (the objective lens: Nikon 10X/0.45 Plan Apo CFI/60). The pixel size was 0.65 µm. For each view, z-stack images (5-um pitch) at five channels (DAPI, FITC, dsRed, Cy5, brightfield) as well as a phase contrast image at the central z-position.

Single-Unit Sequencing Library Preparation: Combinatorial ID-coded units were collected and re-suspended in a unit suspension beads-resuspension buffer (10 mM Tris-HCl, 137 mM NaCl, 2.7 mM KCl, 1.8 mM CaCl2, pH 7.5) supplemented with 0.1% (w/v) BSA. Sequencing beads (Macosko-2011-10 V Plus/Barcoded-Seq B, Chemgenes Corporation) were suspended in a lysis buffer. Droplet generation was carried out using a PDMS chip at a flow rate of 4 µL min^{-1} (for unit suspension), 8 µL min^{-1} (for sequencing beads suspension), and 25 µL min^{-1} (for Droplet Generator oil for EvaGreen). At a high loading concentration of sequencing beads (typically >500 beads μL^{-1}), droplet generation was carried out by an air pressure-regulated liquid transfer system (Fluigent). Droplets were incubated at 50 °C for 3 min, cooled down to room temperature, then broken by the addition of perfluoro octanol. The downstream sequencing library preparation was carried out by referring to protocols described in Drop-Seq Laboratory Protocol ver. 3.1.[17] The sequencing library for DNA barcodes was prepared by referencing CITE-seq & Cell Hashing Protocol Version 2019-02-1.[18]

Sequencing: All DNA barcode libraries for seq-code and cDNA libraries were analyzed on TapeStation2200 (Agilent Technologies), quantified using KAPA Library Quantification Kits (Roche), and sequenced on Mi-seq (Illumina). The cDNA and DNA barcode libraries were sequenced with a read length pair of 25 bp/126 bp, and 25 bp/50 bp, individually. A custom sequencing primer (Custom Read 1 primer; Table S3, Supporting Information) for the first read was added to the primer mix solution.

Obtaining Images of Individual iBB Types: Individual iBB types were re-suspended in a solution of 0.75% agarose in 10 mM Tris-HCl, 137 mM NaCl, 2.7 mM KCl, 1.8 mM CaCl2, 0.1% (v/v) BSA at a concentration of 1 × 105 beads mL^{-1} and loaded on a glass-bottom well plate individually. The well was cooled at 4 °C for 10 min and brought back to room temperature. iBB images were obtained on InCellAnalyzer6000. For each view, z-stack images (15 images in a 5 µm pitch) at three channels (DAPI, dsRed, and Cy5) were recorded.

Identification of Barcode Sequences on Elemental iBBs by Sanger Sequencing: DNA barcodes attached to individual elemental iBBs were amplified using T24V (Table S3, Supporting Information) and an indexed PCR primer in a TruSeq small RNA library preparation kit. PCR products were purified by FastGene Gel/PCR Extraction Kit. Sanger sequencing was outsourced (Eurofins Genomics K.K. or Fasmac Co., Ltd.).

Generating DNA Barcode Libraries of Elemental iBBs by NGS: Equally pooled elemental iBBs were resuspended in a solution of 0.75% agarose in 10 mM Tris-HCl, 137 mM NaCl, 2.7 mM KCl, 1.8 mM CaCl2, 0.1% (v/v) BSA in a concentration of 6.25×10^5 beads mL^{-1}. The calculated average number of iBBs per 90 µm droplets in this condition was 0.25, where most droplets contain either single iBB or no iBB. DNA barcode library preparation was performed as described in single-unit sequencing library preparation.

Computational Methods Overview: The data analysis workflow consists of three parts: image analysis, sequencing analysis, and their linking, as shown in Figure S1 (Supporting Information).

Image Data Analysis: Unit Detection From Phase-Contrast Images: The location and size of the hydrogel units of each view were identified in the phase-contrast (PC) image at the central z-position as follows: First, the intensity of the raw PC image was standardized with a linear transformation and clipped to take values between 0.2 and 0.5. From this image, a Sobel filter and a Gaussian blur were used to detect the edges of the units, and Otsu's thresholding method was applied for binarization. After removing small objects with an area of fewer than 20^2π pixels, the central position and the radius of each unit were identified with the Hough circle transform, wherein the radius was searched between 50 and 85 pixels, and the accumulation threshold was set to 0.9.

Image Data Analysis: Cell Detection From Brightfield Images: For experimental runs with HEK293 and K562 cells, the cell regions were identified from the z-stack brightfield (BF) images, where each image was processed with a Sobel filter, binarized with triangle thresholding, then applied a morphological opening to remove small holes and noises. Each connected component in this binary stacked image was labeled as a distinct cell region. The focal depth of the region was determined by peak detection from the layer-wise means of the gradient root mean squares (GradRMS) in the

region's bounding box. Finally, a watershed algorithm was used in the focal layer of the region to separate adhered or closed multiple cells and assign single-cell labels.

Image Data Analysis: iBB Segmentation and Fluorescent Intensity Quantification: iBBs were segmented by image-processing and their intensities were quantified from the z-stack fluorescent images as follows: First, each z-layer of the Cy5 fluorescence images was binarized with a thresholding method, then disks with up to 10-pixel radius were detected by Hough transform to identify the center XY coordinates of the iBBs in the layer, where the largest ones were retained in case of overlap. A layer-wise fluorescent intensity of an iBB in each channel was quantified as the mean intensities in a 4-pixel square around the bead center. To integrate the detected beads across z-layers, a graph was created in which a pair of disks residing in adjacent z-layers and overlapping in XY coordinates were connected. Then each connected component of the graph was identified as an iBB, where its Z coordinate was determined with a peak detection from the layer-wise Cy5 intensities of the bead.

Image Data Analysis: iBB Classification for Baseline Decoder: A classification method of iBB types was developed for the segmented iBBs using their fluorescent intensities. The classified iBB types were used in a unit as a baseline result of predicting iBB combinations in the unit (Figure S1, Supporting Information). The classification of the segmented iBBs was performed as follows: a reference intensity profile was first prepared for each of the 25 iBB types, wherein the beads were segmented and quantified their intensities by the aforementioned steps (Figure S1, Supporting Information). For classifying iBBs with unknown types in a new experimental run, an alignment of the iBBs' fluorescent intensities between the run and the reference was required due to the variability of intensity levels across experiments. This alignment was implicitly performed with a Bayesian inference of a generative model for the iBBs' intensities as follows:

$$\begin{aligned} &Y_{ic} = Y_{ic}^{(sig)} + Y_{ic}^{(bg)},\ Y_{ic}^{(sig)} | X_{ic} \sim Pois\left(X_{ic}\right),\ Y_{ic}^{(bg)} | B_c \sim Pois\left(B_c\right) \\ &X_{ic} = r_c S_i \sum_{k=1}^{K} Z_{ik} U_{kc},\ B_c \sim Gam\left(a_{B_c}^0, b_{B_c}^0\right),\ S_i \sim Gam\left(a_S^0, b_S^0\right) \\ &Z_{ik} | \pi \sim Cat\left(\pi\right),\ \pi \sim Dir\left(\alpha_\pi^0\right),\ U_{kc} \sim Gam\left(a_{U_{kc}}^0, b_{U_{kc}}^0\right) \end{aligned} \quad (1)$$

where, Pois, Gam, Cat, and Dir denote the Poisson, the Gamma, the Categorical, and the Dirichlet distribution, respectively, and variables Y_{ic}, $Y_{ic}^{(sig)}$, $Y_{ic}^{(bg)}$, X_{ic}, B_c, r_c, S_i, Z_{ik}, U_{kc} and π denote an observed intensity of iBB i in channel c, a signal intensity, a background intensity, the true signal intensity, the true background intensity, a relative channel intensity, a signal intensity varied with iBB, a class probability for iBB type k, the relative intensity of iBB type k in channel c and a ratio of the twenty-five iBB types, respectively. To infer the posterior probability of each iBB type, i.e., $P(Z_i \mid Y)$, a variational approximation[19] was employed to the posterior distribution of the hidden variables with a full factorization assumption as follows:

$$\begin{aligned} &Q\left(Y^{(sig)}, Y^{(bg)}, S, Z, U, B, \pi\right) = \left[\prod_i \left(\prod_c Q\left(Y_{ic}^{(sig)}\right) Q\left(Y_{ic}^{(bg)}\right) Q\left(S_{ic}\right)\right)\right. \\ &\left. Q\left(Z_i\right)\right] \left[\prod_c Q\left(B_c\right) \prod_k Q\left(U_{kc}\right)\right] Q\left(\pi\right) \end{aligned} \quad (2)$$

where Binomial distributions for $Q(Y_{ic}^{(sig)})$ and $Q(Y_{ic}^{(bg)})$, Gamma distributions for $Q(S_{ic})$,$Q(B_c)$ and $Q(U_{kc})$, and a Dirichlet distribution for $Q(\pi)$ were assumed. The Q distributions were updated up to 500 iterations or until the difference of the evidence lower bound (ELBO) from the previous step reached less than 10^{-3}.

Image Data Analysis: Simulating Image Barcoding Units Encapsulating iBBs: Image barcoding units that encapsulated various combinations of 25 iBB types were stimulated by combining experimentally obtained images of 25 iBBs. As illustrated in Figure 2H, 3D patches of isolated iBB images were first collected for each of 25 iBB types, then randomly selected, and located these patches within the spheres to compose the images of spherical units. To collect a set of image patches of isolated iBBs, 25 views of 15 z-stack images were experimentally prepared with a size of 2048 × 2048 pixels for each of the 25 iBB types. For each view, iBBs were segmented by image-processing (Figure 2H left; Figure S1, Supporting Information), which utilized the Hough circle transform, and detected the coordinates, radii, and channel intensities of the beads in the images, where 7322 patches in total with a size of 15 × 80 × 80 voxels were collected. To ensure only a single bead was located at the center of the patches and no other beads were contained, 6127 clean patches were selected in total (Figure 2H middle) as follows: For each patch, a max-normalized image of the patch was created by taking the maximum intensity projections in the Z-direction and normalized intensity values, ensuring the sum of the intensity equaled 1; for each iBB-type, an averaged image was generated using the pixel-wise mean of the max-normalized images; patches were retained only if the root-mean-squared difference between the max-normalized image and the averaged image of the same iBB-type was below a threshold value of 5×10^{-3}. For the selected patches, the mean and the standard deviation of the bead radius were 7.30 and 1.01 pixels, respectively. By using a set of the clean patches of iBB images, z-stack images of barcoding units were simulated with a size of 19 × 170 × 170 voxels, each containing a spherical unit encapsulating arbitrary combinations of iBBs. The iBB patches were positioned at random x, y, and z coordinates within the unit. To account for the physical exclusion volume of the beads, each bead center's location from the internal pixels of the spherical unit was iteratively sampled, ensuring no collisions with the sphere's surface or previously sampled beads. To determine the channel-wise intensities of the image, an estimated background intensities were factored and adjusted for background intensities of overlapping patches at each voxel. The background values of a patch were estimated as the ten percentiles of the voxel intensities, and these were subtracted prior to the addition.

As a training dataset for NeuralReader, 10 000 z-stack images were computationally generated, where a spherical unit region with a radius of 55 to 75 pixels in the XY scale was set at the center of a stack, and one to fifteen iBBs were placed inside of the unit. In addition, to train the network that correctly ignores unrelated beads, whose fluorescence occasionally leaks into the unit, iBBs were generated outside of the unit with the number following a Poisson distribution with a 30% rate of the number of inside ones. For the evaluation of iBB detections presented in Figure 2I and Figure S12 (Supporting Information), ten datasets were generated. Each of these datasets consisted of z-stack images of units encapsulating one to ten iBBs and the radius of the units was 65 pixels in the XY scale. To ensure that at least 1000 iBBs were included in each dataset, 1000 images for a single iBB dataset, 500 images for datasets with 2 to 4 iBBs, and 200 images for datasets with 5 to 10 iBBs were generated.

Image Data Analysis: Decoding iBB Combinations with NeuralReader: To accurately decode iBB combinations from z-stack images of the units, an end-to-end neural network (NeuralReader) based on customized 3D convolutional layers was developed, as shown in Figure S2 (Supporting Information). The Keras Framework[20] with Tensorflow[21] backend was used for the implementation. The input of NeuralReader was a z-stack image with 19 × 170 × 170 voxels with three fluorescence channels (Cy5, dsRed, and DAPI) and a binary mask image that indicates the region of the units. The network outputs existence probabilities of 25 iBB types in a unit.

NeuralReader on the 10 000 simulated images for 300 epochs was trained, using a batch size of 12. A modified binary cross-entropy (BCE) loss function was used, where 1% confusion was applied for the prediction probabilities between neighboring classes. This was achieved by multiplying the corresponding matrix before computing the average BCE for each iBB type. The Adam optimizer[22] with a base learning rate of 0.001 was employed and applied a learning rate scheduling[23] with a warm-up period of five epochs and a multi-step decay by multiplying the rate by 0.1 at the end of the 150th and 250th epochs. During training, 15% of the data was set aside for validation purposes. Additionally, data augmentation techniques were employed: each input z-stack image was subjected to a random selection of transformations including flipping along the X-, Y-, and Z-axis; rotation in the XY-plane; shift within 1 pixel in the Z direction; and shifts within 10 pixels in the X and Y directions. To inform the

basis of architectural decisions for NeuralReader, a summary of prediction performance was presented, as assessed in the validation split, across different architectures (Table S5, Supporting Information). This includes the use of a convolutional layer (Conv3D or MMConv3D), the inclusion of a confusion matrix in the loss function, and the application of the data augmentation. The analysis indicated that each of these components notably enhanced performance.

At the prediction with NeuralReader for a new dataset, input intensities of the dataset were calibrated by a linear transformation so that the channel-wise intensity levels matched those of the reference data used for training. For this calibration, the iBBs of the dataset were once segmented and quantified by using the same image processing procedure as in the baseline approach (Figure S1, Supporting Information), and the intensity histograms of channels Cy5, dsRed, and DAPI were identified (Figure S6C, Supporting Information). Then, the highest peak values of the histograms and the background intensities were estimated and used for determining the linear transformation.

Sequencing Data Analysis: Read Processing: Reads were processed and aligned to the reference genome using UMItools,[24] CITE-seq Count,[25] and STAR aligner.[26] Alignment filtering and manipulations were performed with samtools version 1.1.[27] Further analysis tools were developed with Bash, R, and Python language.

Sequencing Data Analysis: Filtering and Normalizing scRNA Data: Unique molecular identifiers (UMIs) of single-cell RNA (scRNA) data were analyzed with a custom script developed using Scanpy version 1.8.1.[28] For each of the sequenced units (seq-units), we determined whether the unit contained human or mouse cells by thresholding UMIs mapped to the reference file GSM1629193_hg19_ERCC [GEO accession ID GSM1629193], which contained genes of both species. Concretely, the unit was regarded to contain human (mouse) cells if the number of human (mouse) UMIs was more than 1000 (500) and that of the mouse (human) UMIs was less than 100 (50). For the subsequent analysis, only the human cells were kept and left units with more than 50 expressed genes, then filtered genes that were expressed in two or more cells. Also, units with more than 20% mitochondrial gene expression were removed. The gene expression levels of each unit were normalized by scaling values to 10 000 in total, adding one, and taking the log values.

Sequencing Data Analysis: Identification of iBB Combinations from Sequencing Data: A method based on outlier detection was developed to identify the iBB combination of a seq-unit, i.e., a combination of the 25 iBB types in the unit, from sequencing data. UMI counts for each iBB barcode and unmapped were first transformed one by adding one and taking the log value as input values. Then, a background distribution of the input values for each unit was estimated by the elliptic envelope algorithm,[29] which robustly estimated the mean and variance of the signals with outliers. To determine the iBB combination of a seq-unit, the existence of each iBB type was called, when the UMI count for the iBB type was over 50, and the percentile of the input signal for the iBB type was above the 0.5% point of the right tail of the background distribution.

Sequencing Data Analysis: HEK293/K562 Classification with scRNA Data: For the experiments with sequenced units (seq-units) encapsulating HEK293 and K562 cells, a cell-type classification protocol based on scRNA data was developed. reference scRNA data of HEK293 cells and K562 cells were prepared, and a mixed dataset of HEK293/K562 cells. The UMI counts of the scRNA data were filtered and normalized as described in the previous section. Next, a differential gene set (DGS) from the reference dataset was identified, where the top 100 highly expressed genes for each of HEK293 and K562 cell types were identified as DGS based on Wilcoxon's rank-sum method (Data S1, Supporting Information). Principal component analysis (PCA) was applied to the normalized expression levels of DGS of the reference dataset.[30] The normalized expression levels of DGS were projected in the mixed dataset into the space of the top 50 principal components of the reference datasets. Then, the cell type probability of each unit as an initial cell-type score was assigned by a k-nearest neighbor classifier trained with the reference datasets, wherein k = 10 and the Euclidean distance on the projected space were used. The nearest neighbor graph of units in the mixed dataset was constructed based on the distance in the projected space. Then, units by the Leiden method[31] with resolution of 0.1 were clustered, in which the cell-type score of a cluster was defined as the average of the initial cell-type scores for units belonging to the cluster. Finally, the cell type of a unit in the mixed dataset was determined as that with the maximum score for the belonging cluster.

Linking Image And Sequenced Units: Image-unit and sequenced unit (seq-unit) encapsulating cells were linked based on matchings between the identified iBB combinations between those units (Figure 4B; Figure S5B, Supporting Information). For each seq-unit, the most probable image-unit was linked when such an image-unit was uniquely determined based on an estimation of the matching probability of iBB combinations between the seq-unit and the image-unit as described below.

Let x_i and N_{obs} denote an identified iBB combination of the i-th image-unit and the number of observed image-units in a study, respectively. For a unit with an iBB combination y, $P(y \rightarrow x)$ denotes as a probability that the corresponding iBB combination of image unit is missing, i.e., $x = \phi$ or observed as $x = x_i$. these quantities were estimated as follows:

$$P(y \rightarrow \phi) \propto \frac{1 - p_{obs}}{p_{obs}},\ P(y \rightarrow x_i) \propto \frac{P_{mod}(y \rightarrow x_i)}{P_{chance}(x_i)} \tag{3}$$

where p_{obs}, $P_{mod}(y \rightarrow x_i)$, and $P_{chance}(x_i)$ denote a percentage of observable image-units, a probability that an actual combination y is observed as a modified combination x_i, and a probability of finding a combination x_i by chance in a random pool of image-units, respectively. In this study, the following assumptions were used: i) $p_{obs} = 0.8$; ii) for the estimation of $P_{mod}(y \rightarrow x_i)$, only x_i within two edit steps from y were considered, in which each step modified an iBB type in the code to one of its neighboring types on the 5×5 matrix layout (Figure S8, Supporting Information) with a rate of $|y| \times 0.25\%$ or otherwise, unchanged, where $|y|$ denotes the number of distinct iBB types in y; and iii) the term $P_{chance}(x_i)$ was evaluated as follows:

$$P_{chance}(x_i) = 1 - [1 - P_0(x_i)]^{N_{obs}} \tag{4}$$

where $P_0(x_i) = 1 / {}_K C_{|x_i|}$ denote a probability of observing a pattern x_i when code patterns with the same number of distinct iBBs $|x_i|$ are uniformly sampled without replacement from the $K = 25$ iBB types. For the evaluation of P_{mod}, we only counted modified combinations of iBBs up to two edit steps from the original ones. The terms $P(y \rightarrow \phi)$ and $P(y \rightarrow x_i)$ $(i = 1, I, N_{obs})$ were normalized to 1. Finally, an image-unit was paired with a combination x_i to a seq-unit with a combination y only if the $P(y \rightarrow x_i)$ was the largest among all terms, and x_i was the closest pattern in terms of the edit distance from y; i.e., no other image-units were observed whose combinations had an equal or a closer edit distance from y than x_i.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

N.Y. and I.S. conducted a part of this work at Center for Advanced Intelligence Project, RIKEN. The authors thank Prof. Dr. Masashi Sugiyama for his support to this project. Computational resources were provided by the RIKEN AIP Deep Learning Environment (RAIDEN) supercomputer system. Funding: JST CREST (JPMJCR19H1, Japan); JSPS KAKENHI Grant Numbers (JP21H04636 and JP21H00416); Tateisi Science and Technology Foundation; The Noguchi Institute, the Naito Foundation; The Uehara Memorial Foundation; White Rock Foundation; The Canon Foundation; Senri Life Science Foundation; SECOM Science and Technology Foundation; The UTEc-UTokyo FSI Research Grant Program.

Conflict of Interest

Techniques and application related to this work has been submitted for patent application by FK, NY, SO.

Author Contributions

F.K. and T.M. contributed equally to this work. Conceptualization performed by F.K., T.M., H.A., N.Y., I.S., and S.O.; Methodologyperoformed by F.K., T.M., N.Y., I.S., S.O.; Investigation performed by F.K., T.M., Y.M.; Funding acquisition performed by S.O.; Writing – original draft FK, TM; F.K., T.M., H.A., N.Y., I.S., S.O. wrote, reviewed and edited the original manuscript

Data Availability Statement

The data that support the findings of this study are openly available in The BioImage Archive at https://www.ebi.ac.uk/bioimage-archive/, reference number S-BIAD705, ArrayExpress at https://www.ebi.ac.uk/biostudies/arrayexpress, reference number E-MTAB-12988, and Zenodo at https://zenodo.org, DOI 10.5281/zenodo.10202661.

Keywords

computational material design, droplet microfluidics, machine learning, multimodal barcoded microparticles, optical barcoding

Received: October 11, 2023
Published online:

[1] a) D. Dendukuri, D. C. Pregibon, J. Collins, T. A. Hatton, P. S. Doyle, *Nat. Mater.* **2006**, *5*, 365; b) J. Lee, P. W. Bisso, R. L. Srinivas, J. J. Kim, A. J. Swiston, P. S. Doyle, *Nat. Mater.* **2014**, *13*, 524; c) D. C. Pregibon, M. Toner, P. S. Doyle, *Science* **2007**, *315*, 1393.

[2] a) Y. Feng, A. K. White, J. B. Hein, E. A. Appel, P. M. Fordyce, *Microsyst. Nanoeng.* **2020**, *6*, 109; b) F. Hu, C. Zeng, R. Long, Y. Miao, L. Wei, Q. Xu, W. Min, *Nat. Methods* **2018**, *15*, 194; c) N. Martino, S. J. J. Kwok, A. C. Liapis, S. Forward, H. Jang, H.-M. Kim, S. J. Wu, J. Wu, P. H. Dannenberg, S.-J. Jang, Y.-H. Lee, S.-H. Yun, *Nat. Photonics* **2019**, *13*, 720.

[3] B. Harink, H. Nguyen, K. Thorn, P. Fordyce, *PLoS One* **2019**, *14*, e0203725.

[4] a) S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, A. Kreshuk, *Nat. Methods* **2019**, *16*, 1226; b) M. Y. Lee, J. S. Bedia, S. S. Bhate, G. L. Barlow, D. Phillips, W. J. Fantl, G. P. Nolan, C. M. Schürch, *BMC Bioinformatics* **2022**, *23*, 46; c) C. Mcquin, A. Goodman, V. Chernyshev, L. Kamentsky, B. A. Cimini, K. W. Karhohs, M. Doan, L. Ding, S. M. Rafelski, D. Thirstrup, W. Wiegraebe, S. Singh, T. Becker, J. C. Caicedo, A. E. Carpenter, *PLoS Biol.* **2018**, *16*, e2005970; d) C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, *Nat. Methods* **2021**, *18*, 100.

[5] S. Utech, R. Prodanovic, A. S. Mao, R. Ostafe, D. J. Mooney, D. A. Weitz, *Adv. Healthcare Mater.* **2015**, *4*, 1628.

[6] a) T. Moragues, D. Arguijo, T. Beneyton, C. Modavi, K. Simutis, A. R. Abate, J.-C. Baret, A. J. Demello, D. Densmore, A. D. Griffiths, *Nat. Rev. Methods Primers.* **2023**, *3*, 32; b) H. Yuan, Y. Chao, H. C. Shum, *Small* **2020**, *16*, e1904469.

[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, presented at 2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2017), Vancouver, British Columbia, Canada, September **2017**.

[8] S. Ji, W. Xu, M. Yang, K. Yu, *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221.

[9] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. Mccarroll, *Cell* **2015**, *161*, 1202.

[10] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, *Nat. Methods* **2017**, *14*, 865.

[11] L. McInnes, J. Healy, J. Melville, *arXiv:1802.03426* **2018**.

[12] a) S. Chelvanambi, J. M. Hester, S. Sharma, T. Lahm, A. L. Frump, *Am. J. Respir. Cell Mol. Biol.* **2020**, *62*, 112; b) T. N. Chen, A. Gupta, M. D. Zalavadia, A. Streets, *Lab Chip* **2020**, *20*, 3899; c) D. Feldman, A. Singh, J. L. Schmid-Burgk, R. J. Carlson, A. Mezger, A. J. Garrity, F. Zhang, P. C. Blainey, *Cell* **2019**, *179*, 787; d) M. S. Kowalczyk, I. Tirosh, D. Heckl, T. N. Rao, A. Dixit, B. J. Haas, R. K. Schneider, A. J. Wagers, B. L. Ebert, A. Regev, *Genome Res.* **2015**, *25*, 1860; e) K. Lane, D. Van Valen, M. M. Defelice, D. N. Macklin, T. Kudo, A. Jaimovich, A. Carr, T. Meyer, D. Pe'er, S. C. Boutet, M. W. Covert, *Cell Syst.* **2017**, *4*, 458; f) Z. Liu, J. Yuan, A. Lasorella, A. Iavarone, J. N. Bruce, P. Canoll, P. A. Sims, *Sci. Rep.* **2020**, *10*, 19482; g) S. Shah, E. Lubeck, W. Zhou, L. Cai, *Neuron* **2017**, *94*, 752; h) J. Yuan, J. Sheng, P. A. Sims, *Genome Biol.* **2018**, *19*, 227; i) J. Q. Zhang, C. A. Siltanen, L. Liu, K.-C. Chang, Z. J. Gartner, A. R. Abate, *Genome Biol.* **2020**, *21*, 49.

[13] a) C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, L. Cai, *Nature* **2019**, *568*, 235; b) C. Xia, J. Fan, G. Emanuel, J. Hao, X. Zhuang, *Proc. Natl. Acad. Sci. U S A* **2019**, *116*, 19490.

[14] a) P. Datlinger, A. F. Rendeiro, T. Boenke, M. Senekowitsch, T. Krausgruber, D. Barreca, C. Bock, *Nat. Methods* **2021**, *18*, 635; b) L. Mathur, B. Szalai, N. H. Du, R. Utharala, M. Ballinger, J. J. M. Landry, M. Ryckelynck, V. Benes, J. Saez-Rodriguez, C. A. Merten, *Nat. Commun.* **2022**, *13*, 4450.

[15] M. A. Wheeler, I. C. Clark, H.-G. Lee, Z. Li, M. Linnerbauer, J. M. Rone, M. Blain, C. F. Akl, G. Piester, F. Giovannoni, M. Charabati, J.-H. Lee, Y.-C. Kye, J. Choi, L. M. Sanmarco, L. Srun, E. N. Chung, L. E. Flausino, B. M. Andersen, V. Rothhammer, H. Yano, T. Illouz, S. E. J. Zandee, C. Daniel, D. Artis, M. Prinz, A. R. Abate, V. K. Kuchroo, J. P. Antel, A. Prat, et al., *Science* **2023**, *379*, 1023.

[16] R. Y. Cheng, J. de Rutte, A. R. Ott, L. Bosler, W. Kuo, J. Liang, B. E. Hall, D. J. Rawlings, D. Di Carlo, R. G. James, *Nat. Commun.* **2023**, *14*, 3567.

[17] E. Z. Macosko, M. Goldman, retrieved from dx.doi.org/10.17504/protocols.io.mkbc4sn (accessed: April 2019).

[18] M. Stoeckius, P. Smibert, https://ccp.bwh.harvard.edu/wp-content/uploads/2021/10/Cite-seq-and-Cell-hashing-protocol-Satija-Lab.pdf (accessed: April 2019).

[19] D. M. Blei, A. Kucukelbir, J. D. Mcauliffe, *J. Am. Stat. Assoc.* **2017**, *112*, 859.

[20] F. Chollet, **2015**. retrieved from https://github.com/fchollet/keras (accessed: Auguest 2021).

[21] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, et al., *arXiv:1603.04467* **2016**.

[22] D. P. Kingma, J. Ba, *3rd International Conference on Learning Representations*, ICLR, San Diego, CA, USA **2015**.

[23] X. Z. Kaiming He, S. Ren, J. Sun, Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, June-July **2016**.

[24] T. Smith, A. Heger, I. Sudbery, *Genome Res.* **2017**, *27*, 491.

[25] P. Roelli, retrieved from https://doi.org/10.5281/zenodo.2590196 (accessed: July 2020).

[26] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, *Bioinformatics* **2013**, *29*, 15.
[27] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, *Gigascience* **2021**, *10*, giab008.
[28] F. A. Wolf, P. Angerer, F. J. Theis, *Genome Biol.* **2018**, *19*, 15.
[29] P. J. Rousseeuw, K. V. Driessen, *Technometrics* **1999**, *41*, 212.
[30] C. Wu, I. Macleod, A. I. Su, *Nucleic Acids Res.* **2013**, *41*, D561.
[31] V. A. Traag, L. Waltman, N. J. Van Eck, *Sci. Rep.* **2019**, *9*, 5233.