Title
Droplet Microfluidic Platforms for Single-Cell Sequencing

Author
Haliburton, John

Publication Date
2017

Peer reviewed|Thesis/dissertation

Droplet Microfluidic Platforms for Single-Cell Sequencing

by

John Robert Haliburton

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

DEDICATION AND ACKNOLWEDGEMENTS

For Genevieve, Johannah, and Robert--who got less of me for this experience but have always thought more of me

ABSTRACT

DNA sequencing has defined an era of biological discovery by unveiling the genetic basis of biological processes. The completion of the human genome broadened understanding of the multiple scales of biological complexity from large scale structures of human physiology down to the individual cells of which they are comprised. To utilize DNA sequencing at the single-cell level new technologies must be leveraged that allow for the isolation and manipulation of individual cells. Droplet microfluidics is a technology that is uniquely suited to the challenge of rapid and high-throughput isolation and manipulation of single-cells. This thesis presents droplet microfluidic platforms for single-cell sequencing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

The study of human health proceeds through the development of new technologies and the discovery of new insights into the complex structures of human physiology. The human body is comprised of several macro-scale organs, each with highly specific functions and interconnected roles in health and homeostasis. Organs in turn are comprised of several tissues with an equally high degree of specialization; and tissues themselves are comprised of single cells that are often spatially and functionally distinct. Herein lies the challenge in unraveling the complexity of human health; precise understanding of biological processes at the highest levels necessitates understanding at the level of individual cells.

Several technologies have been adapted for the study of single cells. Each technology has a place of historical significance and represents a need to increase efficiency on two axes: the number of cells that can be analyzed and the number of parameters or variables that can be studied in each cell. The study of single cells was initially pioneered through the invention and advancement of microscopy. In 1665 Robert Hooke observed, described, and coined the term “cell” through his study of plant tissue under a microscope [1], but it wasn’t until almost 200 years later, in 1839, that Theodor Schwann and Matthias Schleiden proposed the cell theory, which eventually yielded key design principles of biological organisms and established the cell as the fundamental building block of life[2, 3]. Through the use of selective chemical staining for cellular structures and content, microscopy represented a low parameter, low-throughput

technology for cellular study. Microscopy dominated the study of cells and molecular biology until the advent of flow cytometry in the 1960's. Flow cytometers made it possible to query tens of thousands of single cells for a small number of parameters like cell-size or the presence of cell surface markers, and the ability to sort single cells made it possible to study them in isolation. Cell sorting was particularly useful with the birth of recombinant DNA technologies and molecular assays like PCR, which paved the way for the first multi-parameter technologies like microarrays. Thus with cell-sorting and microarrays, the first low-throughput but high-parameter studies of single cells were performed to investigate both genomic sequence and gene expression. The invention of next-generation sequencing in the early 2000's opened the floodgates of the genomics-era and promised to enable multi-parameter studies like genome and transcriptome profiling of single cells at high throughput. The need to prepare samples for sequencing through multi-step enzymatic reactions necessitates the isolation of cells into suitable compartments, and well-plate methods were quickly established and eventually miniaturized into automated microfluidic devices that preclude the use of a flow-sorter and reduce reagent needs. However, to achieve high-throughput single-cell sequencing, a technology is needed that enables fast and scalable means of isolating and compartmentalizing cells in a way that multi-step enzymatic workflows can be used to prepare them for sequencing.

Droplet microfluidics is a technology that offers high-throughput compartmentalization through controlled flow of immiscible fluids like oil and water. Shortly after its reduction to practice in the early 2000's, researchers began to adapt it for high-throughput

biological studies like enzyme function and qPCR [4-8]. One key to its adaptation for single-cell sequencing was the ability to manipulate drops by addition and subtraction of reagents [9]. With these tools in hand, droplets provided an ideal solution for the isolation and preparation of single-cell sequencing libraries.

One key advance in the development of microfluidics, and subsequently droplet microfluidics, is the ability to rapidly prototype devices through the use of soft lithography and polymer molding. Soft lithography technology, developed in the semiconductor industry, allows for highly tunable and facile creation of microstructures on silicon or glass surfaces. When curable polymers are poured over the surfaces, the microstructures become channels in the polymer mold[10]. This process is fast, cheap, and does not require harsh chemical etchants like hydrofluoric acid used to etch glass channels.

Droplet microfluidics leverages the properties of fluid flow in microchannels to generate emulsions from immiscible fluids. Emulsion have been studied for quite some time, and various methods for controlled formation of emulsions exist, from simple systems of stirring reactors[11] to more controllable and precise methods of fluid flow in nested capillaries. The first demonstration of droplet formation in microchannels was accomplished by flowing hydrocarbon oils through a channel that intersected a channel of aqueous flow[7]. Droplets form at the junction of the immiscible flows as a result of dominating viscous shear forces at high capillary and low Reynolds numbers. Surfactants added into one or both of the immiscible phases stabilize the droplets and

prevent coalescence of droplets and interchange of materials confined within the droplets[12]. Two of the most important features of droplet microfluidics that make it appealing for biological research are the speed of droplet formation and the monodispersity of the droplets formed. Aqueous reagents confined in droplets cannot interact with reagents in other droplets. Therefore, compartmentalizing chemicals and biological reagents inside of droplets is the equivalent of partitioning reagents into well strips or plates in conventional high-throughput methods, but represents a significantly increased level of high-throughput capability over existing plate based methods.

In this dissertation I present my work towards developing droplet platforms for single-cell sequencing. I will discuss two novel and key elements of my research, which are DNA barcoding and interaction screening through the use of droplet microfluidics.

Single cell sequencing of B cell receptors with high throughput droplet barcoding

Abstract

Analysis of immune repertoires provides key insights into disease progression and treatment, and may lead to the discovery of novel therapies. Immune repertoire sequencing provides a detailed and accurate picture of immune structure, but requires the ability to sequence large number of individual B cells. Here, we demonstrate a scalable means for targeted antibody sequencing of single B cells. Leveraging a workflow we've recently developed allowing efficient RT-PCR on single cells in picoliter droplets, we barcode the mRNA of individual B cell antibody genes at high throughput. We validate the efficacy of the method and demonstrate how the data can be used to differentiate between sequence variants resulting from error and those representing true cell variation. Our ability to use picoliter droplets affords a ~100-fold higher throughput than competing methods and is scalable to the barcoding of millions of cells in under a day.

Introduction:

Organisms have the remarkable ability to encode diverse cellular phenotypes within a single genome, a trait that is essential to building the complex, functional tissues of which they are composed. The cellular phenotypes found in the immune system are a

particularly inspiring example of the way in which genomically encoded diversity is used to achieve an important functional goal. B cells and T cells produce a diverse repertoire of immunoglobulin surface receptors and soluble antibodies that target and bind foreign and host antigens in the first step of a complex pathway comprising the immune response.

During development, each B and T cell produces a distinct antibody encoded by two unique gene sequences within the cell's genome. Additionally, B cells can further diversify their antibody genes through somatic hypermutation. The totality of all antibodies within an individual is known as the antibody repertoire, and is a valuable source of information for disease diagnosis and identifying new biologic therapies [13] [14]. Consequently, there is immense interest in methods to accurately and comprehensively characterize antibody diversity within an individual repertoire [15].

Next generation sequencing (NGS) is a powerful method for characterizing diversity in biological systems like the antibody repertoire [16]. Key to its power is the ability to obtain high volumes of sequence data using a massively-parallel strategy. However, two challenges impede the use of NGS to accurately profile antibody repertoires. One is that NGS requires DNA or mRNA from many cells be pooled together, thus abolishing the pairing-information between immunoglobulin genes. The other is that natural sequence diversity occurs at similar or even higher rates than errors introduced during sequencing, making it difficult to distinguish true sequence variation from error [17] [18] [19]. To overcome these challenges, the heavy and light chain immunoglobulin genes can be linked prior to sequencing, to preserve pairing-information, or pairs can be

inferred by splitting cells into multiple groups and then utilizing statistical techniques. Another solution is to perform library preparation on single cells, attaching unique sequence "barcodes" to the antibody genes of each cell [20]. Several single-cell libraries can then be pooled and sequenced and each single cell's data extracted by grouping reads by barcode [21] [22] [23]. This approach has recently been applied to transciptionally profile single cells in high throughput [24] [25] [26] [27]. However, these barcoding methods have not been used to sequence antibody repertoires because the sequenced portion of the gene is not the diverse region specific to each antibody chain.

Droplet microfluidics is unparalleled in its ability to perform efficient and high throughput fluid handling. With microfluidic devices, droplets can be generated, incubated, injected with reagents, and sorted at kilohertz rates [28]. Because the droplets are not much larger than cells, cell material remains concentrated for efficient molecular biology [29]. The approach has been applied to diverse applications, from enzyme evolution and biophysical characterization, to novel cell sorting and antibody discovery [5] [30] [31]. To enable accurate and high throughput sequencing of antibody repertoires with single cell resolution, an ideal platform would combine high throughput microfluidic library preparation with massively parallel sequencing.

In this paper, we describe a method to efficiently barcode populations of B cells using droplet microfluidics. Individual B cells are encapsulated in droplets, lysed, and mRNA transcripts encoding the heavy and light chain genes labeled with unique cell barcodes. Leveraging a microfluidic workflow we've developed allowing protease digestion of cell

lysates, we are able to perform the barcoding reaction in picoliter droplets in which the cell lysate is normally inhibitory to RT-PCR. Since the throughput and cost of the method scales inversely with the droplet volume, this affords unparalleled scalability for sequencing large numbers of single cells: While recently described methods enable thousands of cells to be prepared in 1 hr, we can prepare 100,000 cells in the same time. By grouping the mRNA transcripts by cell barcode, we generate accurate consensus sequences of the antibody genes, which allows us to distinguish between artifactual variation resulting from sequencing error, and true biological variation that occurs at similar or higher rates. Even though the heavy and light chains are transcribed separately, our novel barcoding strategy allows us to identify pairs that comprise each antibody and is scalable to barcoding multiple other genes within the cell.

Methods:

Microfluidics

The microfluidic devices are fabricated using soft lithography[Basic Microfluidic and Soft Lithographic Techniques]. SU-8 photoresist (MicroChem Corp) is spun onto a 3”silicon wafer (University Wafer) to a desired thickness and baked at 135°C to remove solvent. A photo transparency mask (CAD/Art Services) containing the device features is placed on the wafer and exposed to UV light to crosslink the photoresist in the desired pattern. Following UV exposure, the wafer is post-baked at 135°C for 1 minute and placed into a developing bath of propylene glycol methyl ether acetate (PGMEA, Sigma) to dissolve uncrosslinked resist. To produce a device containing features with two different heights,

a first layer is spun, baked, and exposed as normal using a mask containing features of the first height; then, a second layer of SU-8 is spun on top of the first and the wafer baked for 10 min prior to exposure with a mask containing features for the second height. The second mask must be aligned to the features from the first layer to ensure correct fabrication of the channels; this is accomplished by hand alignment using a boom microscope and alignment marks designed into the masks. Following development with PGMEA, the masters are washed with isopropanol and post-baked at 135°C for 30 minutes. The masters are placed into plastic petri dishes and covered with degassed poly(dimethylsiloxane) (PDMS) prepared from 10:1 ratio of elastomer:crosslinker (Sylgard 184, Dow Corning). The dish is evacuated to remove entrapped air bubbles and baked at 65°C for at least 2 hours to crosslink the PDMS. The PDMS devices are cut with a scalpel and peeled away from the master. Holes for inlets and outlets are punched using a biopsy core (Harris Uni-Core), the devices are rinsed with isopropanol, and they are plasma-bonded to glass slides. The devices are flushed with Aquapel to render the channels hydrophobic and enable water-in-oil emulsification, and baked at 65°C for 20 min to remove excess Aquapel. To operate the microfluidic devices, Polyethylene (PE) tubing (Scientific Commodities) is used to connect device inlets to syringes containing reagents, and a custom Python script used to control syringe pumps and inject liquids into the device. Droplet merger is accomplished using salt water electrodes (Generating electric fields in PDMS microfluidic devices with salt water electrodes) energized by a 1500 V cold cathode fluorescent inverter (CCFL) powered by a Mastech supply. The oil used in all devices is Novec HFE 7500 (3M) containing 2% fluorosurfactant (RAN Biotechnologies) and

droplets are collected into tubes or syringes, as required by the protocol. Prior to subjecting droplets to thermal treatments, the HFE oil is removed from beneath the droplets with a pipette fitted with a gel loading tip and an equal volume of FC-40 oil (Sigma) containing 5% surfactant is added above the emulsion, allowing the buoyant droplets to cream into it.

Barcode Encapsulation and PCR

The PCR mix used to generate digital droplet barcodes consists of 50 fM barcode template and 400 nM each of forward and reverse amplification primers (Integrated DNA Technologies) dissolved in detergent free phusion polymerase buffer (New England BioLabs) augmented with 1 mM $MgCl_2$ and 8 units of Phusion HSII enzyme (Thermo Scientific), 2.5% PEG 6K (w/v) (Santa Cruz Biotech) and 2.5% Tween 20 (v/v) (Sigma). The barcode droplet maker is a flow focus device with channel height 30 mm and nozzle width 40 mm. The device is run with an aqueous flow rate of 600 mL/hr and oil flow rate of 1200 mL/hr. Droplets are collected into a PCR tube, the HFE oil swapped with FC-40, and thermal cycled at 98°C for a 3 min hot-start, followed by 30 cycles of 98°C for 20 sec, 58°C for 20 sec, and 72°C for 20 sec. After thermal cycling, the droplets are transferred into a 1 mL syringe containing HFE oil with 2% surfactant for reinjection into the barcode addition device.

Cell Encapsulation and Lysis

Cells are grown in suspension in RPMI media supplemented with 10% fetal bovine serum and pen/strep antibiotics (UCSF cell culture facility). Prior to encapsulation, cells

are removed from the incubator and counted. One million cells are washed twice with phosphate buffered saline (PBS) containing 0.1% Pluronic F-68 and re-suspended in 200 mL of cell re-suspension solution comprising 20 mM Tris-HCl pH 7.5, 100 mM KCl, 17% Optiprep (v/v), and 0.1% Pluronic F-68. In a separate tube, 200 mL of cell lysis solution is prepared, comprising 100 mM Tris-HCl pH 8.0, 5% PEG 6K (w/v), 5% Tween 20 (v/v), 4mM EDTA (Sigma), and 100 mg of Proteinase K (New England BioLabs). Cells are encapsulated with lysis buffer in a co-flow droplet maker of height 30 mm and nozzle width 40 mm [32]. The flow rate for the aqueous phases is 200 mL/hr and the oil 800 mL/hr. The droplets are collected into a 1 mL syringe, the HFE swapped with FC-40, and the syringe capped and loaded into a custom clamp, which is incubated upright in a custom incubation block (Fig S1) at 55°C for 30 min. The temperature of the block is ramped to 95°C and held for 10 min to inactivate the proteinase K; after the syringe has cooled to room temperature, the clamp and cap are removed and the oil swapped back to HFE with 2% surfactant for injection of the droplets into the barcode addition microfluidic device.

Single Cell Barcoding Device

200 mL of RT-PCR mix for linkage RT-PCR is prepared from the SuperScriptIII One Step High Fidelity RT-PCR kit (Life Technologies) by combining 100 mL of 2X Master mix with 400 nM each of barcode forward amplification primer and target RT primers, 200 nM of target forward amplification primers, 5 mL of enzyme mix containing reverse transcriptase and DNA polymerase, 2.5% PEG 6K (w/v), and 2.5% Tween 20 (v/v). The device for pairing barcode and cell lysate droplets is fabricated with two heights, a first

of 30 mm and a second of 80 mm. The moat channel for electric field shielding is loaded with 5M NaCl. The barcode and cell lysate droplets to be paired are introduced at 30 mL/hr and spaced by HFE with surfactant at 200 mL/hr. To generate large RT-PCR droplets, RT-PCR mix is injected at 400 mL/hr and HFE oil of 500 mL/hr, generating droplets ~100 mm in diameter. Contact with the electrode for merger is accomplished by clipping the positive output of the inverter to the needle of a syringe containing the salt water connected to the merger electrode with an alligator clip. The power supply output voltage is varied to adjust the electric field in the merger junction to optimally merge the droplets. The merged barcode, cell lysate, and RT-PCR droplets are collected into a 0.5 mL thin-walled PCR tube. Prior to PCR, the HFE oil is swapped with FC-40 and the droplets thermal cycled at 50°C for 30 min followed by a 94°C hot-start for 2 min and 30 cycles of 94°C for 20 sec, 58°C for 30 sec and 68°C for 1 min, and a final extension at 72°C for 5 min. After PCR, the droplets are chemically coalesced by adding an equal of volume of 1:1 of perfluorooctanol and HFE. The tube is briefly centrifuged to separate the oil and aqueous phases and the oil is removed from the bottom. The aqueous is purified with a Zymo Clean and Concentrator-5 column and eluted in 20 mL of water.

Library Preparation

DNA is quantitated by Qubit hsDNA and a Bioanalyzer chip to confirm the presence of linked products. The DNA is prepared for sequencing by isolating barcoded products and attaching sequencing adaptors through limited-cycle PCR. Magnetic streptavidin beads are washed three times with 2X BWT buffer (10 mm Tris-HCl pH 7.5, 1 mM EDTA, 1 M NaCl, 0.02% Tween 20) and re-suspended in 20 mL of 2X BWT buffer. 20uL

of DNA is mixed with streptavidin coated beads and allowed to bind at room temperature for 15 minutes. The beads are washed two times with 100 mL of 1X BWT and one time with 100 mL of TNT buffer (20 mM Tris-HCl pH 7.5, 50 mM NaCl, 0.02% Tween 20). The beads are re-suspended in 20 mL of 100 mM NaOH and incubated at room temperature for 5 min. They are then placed on the magnet and the supernatant containing single stranded barcoded DNA transferred to a clean tube and neutralized by addition of 20 mL of 100 mM HCl followed by 10 mL neutralization buffer (200 mM Tris-HCl pH 7.5, 0.05% Tween 20). Single stranded DNA is quantitated by the Qubit ssDNA kit. The library is amplified prior to sequencing in a 50 mL PCR reaction is prepared containing 25 mL of 2X KAPA HiFi master mix, 10 mL of barcoded ssDNA template, 200 nM of Illumina P5 primer, and 200 nM of custom P7 adaptor primer. The reaction is thermal cycled at 98°C for 3 min followed by 5 cycles of 98°C for 20 sec, 67°C for 20 sec, and 72°C for 20 sec. The PCR is purified with a Zymo Clean and Concentrator 5 column and quantitated by Qubit hsDNA and Bioanalzyer high sensitivity DNA Chip.

Sequencing and Bioinformatics

The library is sequenced on an Illumina MiSeq desktop sequencer using 2 x 250 bp paired end reads. Raw reads are processed from fastq files with a script that strips the barcode from Read 1 and places it in the read header for each read and its corresponding Read 2 pair, matched by read ID. The barcode sequence is then stripped from Read 1. The end product is a pair of barcode fastq files containing Read 1 and Read 2 sequences that have their corresponding barcodes in the read header. A consensus sequence of the B cell receptor is generated by randomly sampling 10,000

reads from the library and clustering by sequence identity. The consensus sequence is used to build a reference sequence and the barcoded fastq files are aligned against the reference using the bowtie aligner with a tolerance of three mismatches [33]. Bowtie output is formatted to report the barcode from the read header along with the mismatch variables: chain identity, mismatch position, and base composition, which are denoted as the sequence variant profile. The bowtie output is processed to combine information by barcode and outputs a database file containing all of the barcodes and their corresponding sequence variant profiles. This database file serves as the starting point for the analyses.

Results:

Overview of the method

Our strategy for enabling the high throughput sequencing of single B cells is to barcode antibody transcripts using droplet microfluidics. This is accomplished by adding unique barcodes sequences to each droplet containing a lysed cell. The barcode sequences are then attached to targeted mRNAs using overlap extension RT-PCR, as illustrated in Fig. 1. To barcode cellular antibody mRNAs with this approach, we thus require droplets containing unique barcode sequences. To produce these droplets, we use digital PCR. We encapsulate random template molecules in droplets at limiting dilution, controlling concentration so that most droplets are empty but some contain single molecules. We then amplify the templates, generating within each droplet a clonal population of the

original encapsulated sequence. Because the nucleic acids remain compartmentalized throughout the process, the amplified sequence within each droplet is unique, allowing it to be used to uniquely barcode the nucleic acids of a single cell. In parallel, we produce a second emulsion comprising droplets with single cells and a protease-based lysing reagent. The protease digests cellular proteins that inhibit RT-PCR, readying the cell lysate for the barcoding reaction. One barcode droplet is then merged with one lysate droplet, and RT-PCR reagent added, as shown in Fig. 1. The droplets are thermal cycled, amplifying the target mRNA transcripts and attaching the barcodes. The nucleic acids of all cells and droplets can then be extracted, pooled and sequenced, and the reads computationally grouped by barcode to aggregate single cell data.

The number of cells that can be barcoded with our approach is limited by the rate of droplet processing and the droplet volume. Using our novel protease digestion workflow [30] we are able to perform the barcoding reaction in ~750 pL droplets. Reducing

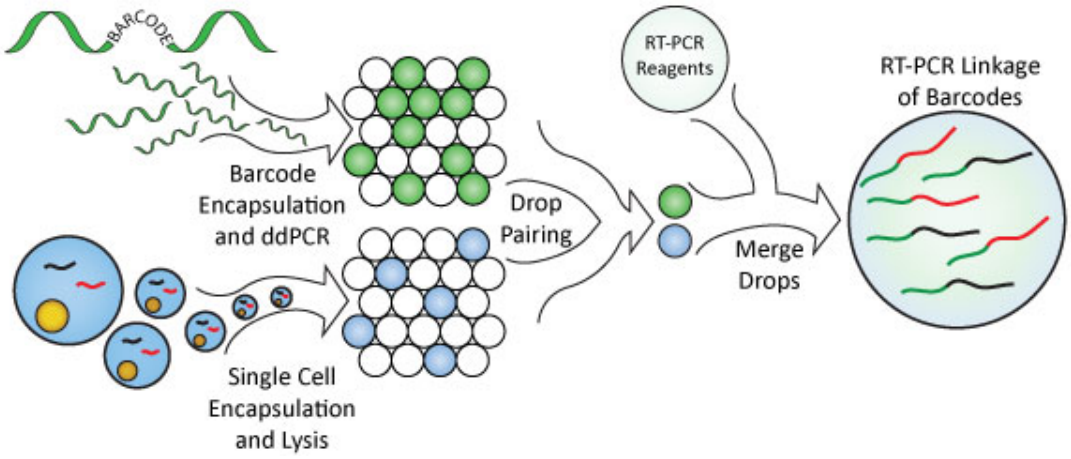


Figure 1: Droplet microfluidic workflow for single-cell barcoding

droplet volume has a double effect, increasing the rate of droplet processing and reducing the volume consumed per cell. For our flow rates and droplet sizes, we barcode ~100,000 cells per hour using 500 mL reagent, a hundred-fold increase in throughput and reduction in volume usage compared to competing approaches.

Microfluidic devices for high throughput single cell barcoding

The barcode droplets are generated using microfluidic flow focusing to emulsify a PCR solution containing oligos with randomized sequences. The droplet maker has a nozzle 40 mm wide and 30 mm tall, generating droplets ~32 mm in diameter, which are collected into PCR tubes for thermal cycling (Fig. 2a, upper-left); to-scale schematics of the devices are shown in Fig. 2b, and an image of the droplet maker generating barcode droplets in Fig. 2c, upper. The cell droplets are generated in a similar way, emulsifying a cell-laden suspension (Fig. 2a, lower left). The droplet maker has the same nozzle dimensions, but also contains a second inlet into which lysis buffer is introduced. The cell and lysis streams are injected at equal flow rates, causing them to merge before the droplet generation junction. Due to the low Reynolds number and high Péclet number the streams do not mix. Oil is introduced at the droplet maker, generating droplets comprising equal parts cell and lysis streams. Once encapsulated, the solutions diffusively mix in under a minute, exposing the cells to the lysis buffer. The droplets are collected into a syringe and incubated at 55°C for 30 min to allow proteinase K to digest inhibitory proteins, followed by 95°C for 10 min to inactivate the protease prior to addition of RT-PCR enzymes.

With the barcode and cell droplets prepared, the next step is to combine one of each droplet with a droplet containing RT-PCR reagent; this mixes the lysate of one cell with the amplified product of a single barcode sequence, and allows the barcodes to be added via RT-PCR. Pairing is accomplished using a third device (Fig. 2a, right). The barcode and cell emulsions are introduced via two inlets visible in the to-scale schematic (Fig. 2b, right). The inlets contain filters allowing correctly-sized droplets to pass without resistance, while large droplets that coalesce during thermal cycling or reinjection are captured; this reduces the frequency of mixed barcode and cell droplets, increasing data quality [A Sciambi ,unpublished work]. The filters empty into a cross junction where oil is introduced from a central channel (Fig. 2d). The flow rates are set

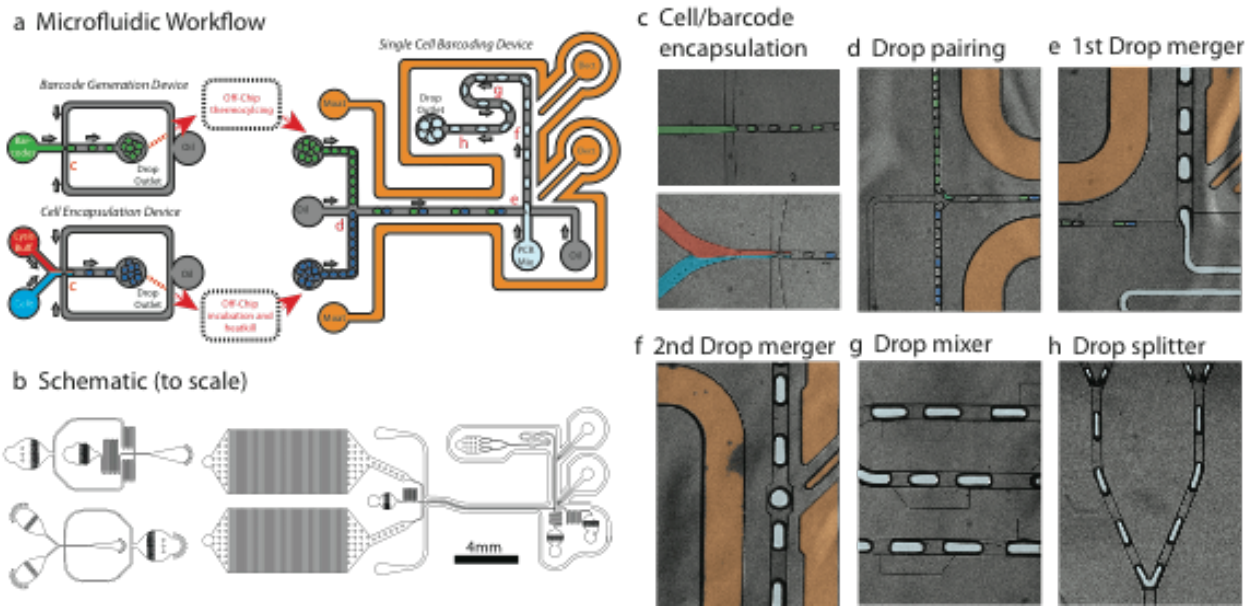


Figure 2: (a) Modular microfluidic workflow including barcode generation ddPCR, cell encapsulation and lysis in droplets, and a Single Cell Barcoding device to accomplish pairing of barcode and cell droplets and addition of RT-PCR reagents. (b) Schematic of devices used in the workflow. (c-h) Functional aspects of the workflow include the barcode and cell encapsulation devices for drop making (c), drop pairing in the Single-Cell Barcoding Device (d), droplet merger with RT-PCR reagents (e), a second drop merging chamber (f), a droplet mixer(g), and a droplet splitter (h).

to pair one barcode and one cell droplet (Fig. 2e) and the pairs are directed into a junction forming ~750 pL droplets of RT-PCR mix (Fig. 2e). A salt water electrode generates the electric field that merges the reinjected droplets with the forming RT-PCR droplet; a merger junction downstream merges any droplets that remain uncoalesced in the first junction (Fig. 2f).

The stability of a given emulsion depends on droplet size since each surfactant stabilizes droplets over a specific range. For our surfactant, droplets 750 pL exhibit poor stability during thermal cycling. To increase stability, a simple solution is to reduce droplet size by splitting the droplets as they exit the device [34] [35]; however, it is essential that the droplet contents be thoroughly mixed before splitting, or the split droplets will contain different concentrations of cell lysate and barcode. After the merger step, we thus flow the droplets through a mixing module consisting of switchbacks outfitted with “fan-blade” mixers. The fan blade mixers consist of horizontal expansions in the channel that are shorter than the channel; as a droplet passes a blade, the surrounding oil flows into the blade, but the droplet does not because, to do so, it would have to adopt an energetically-unfavorable squeezed shape. The rush of oil into and out of the blade generates a cross-flow that drags the interface of the droplet; when combined with the recirculating flow already present in the droplet [36], this efficiently mixes the droplet contents [37]. We’ve found these mixers to be more efficient than simple zigzag designs at the capillary number at which our device operates and for our fluids. Once mixed, the droplets pass into the splitting module, dividing first in half, and then into quarters, as shown in Fig. 2h. This yields droplets ~72 mm in diameter (188

pL), which are more stable during thermal cycling. After thermal cycling, the droplets are chemically coalesced and the PCR products purified as a mixture of barcoded and unbarcoded molecules, and unincorporated barcodes. The barcoded products are purified using a biotin-labeling strategy (Fig S2). Sequencing adaptors are added by limited-cycle PCR and the library sequenced on a MiSeq with paired end 200 bp reads.

Droplet barcoding reliably labels the mRNA of single cells

To illustrate that our barcoding methodology enables high-throughput single cell sequencing, we use it to barcode the heavy and light chain antibody genes of Ramos cells, a B cell lymphoma line. Ramos cells maintain a small amount of somatic hypermutation, so that even a flask-grown population expresses a diverse repertoire of antibodies [38]. To sequence this repertoire with single cell resolution, we process the cells through our barcoding workflow, targeting the heavy and light chains for sequencing. The linkage reaction yields barcoded heavy and light chain genes which we sequence using paired end reads. The number of bases sequenced in our paired end format captures almost the entirety of the heavy and light chain sequence. We used 200 base pair reads, sufficient to capture the barcode and most of the variable regions of the heavy and light chains, including the third complementarity region, which contains the majority of the variation in antibody sequences. Recent advances in sequencing chemistry provide sequences up to 300 base pairs and could easily provide full-length sequences from the same libraries.

Table 1: Sequencing Information

Total Reads	6,986,736
Barcode Groups	214,476
Reads aligned with ≤ 3 errors	3,259,790
Barcode Groups	152,293
Barcode Groups pass Coverage Filter (n≥5)	21,457
Sequence Variants in Coverage Filter	297,502
Sequence Variants pass SVR Filter (SVR≥0.85)	855

The barcodes are sampled from the barcode emulsion randomly; consequently, it is essential that the diversity of barcode sequences be significantly larger than the number of cells to be barcoded, or one barcode might be used to label multiple cells, resulting in the loss of single cell information. To ensure the needed diversity, we use 15 nucleotide random templates for the barcodes, yielding >1 billion unique permutations. With such a large barcode space, sequencing 1 million cells samples ~0.1% of the library. During sequencing, we observe a total of ~152,000 unique barcodes. To determine if this agrees with the anticipated diversity, we calculate a Hamming distance between sequenced barcodes and compare it to the theoretical distribution (Fig. 3b). The distribution agrees with the theoretical prediction indicating that, as anticipated, the barcode sequences are randomly sampled from the barcode emulsion. Moreover, the average Hamming distance is ~11, indicating that most barcodes are highly distinct from one another; this simplifies their association with barcode groups even if sequencing errors occur.

A potential concern of using PCR to amplify barcodes is that amplification bias may skew representation of the sequences. To investigate this, we analyze the base compositions of the barcodes, Fig. 3b, inset. The compositions are relatively uniform and match with the known biases of the oligonucleotide synthesis used to generate the templates. This low bias is likely due to the compartmentalized digital amplification of the barcodes, which allows each sequence to amplify to saturation without competition. To observe how reads are distributed across barcode groups, we plot the number of reads per barcode (Fig. 3c). Although 152,000 barcode groups are represented in the data, the majority contain few reads (blue line, Fig 3c). By plotting the cumulative distribution, we find that a vast majority of the data exists within the largest 20,000 barcode groups, and only a small fraction (~15%) in the upper 120,000 groups (red line, Fig 3c); these groups are likely the result of errors in the barcode sequences which can be adopted into the confident groups based on sequence similarity. These results demonstrate that droplet digital PCR is an effective means by which to generate barcodes for single cell sequencing.

Droplet barcoding allows measurement of true sequence variation within the backdrop of sequencing error

A challenge of characterizing the diversity of an antibody repertoire is that true sequence variation resulting from somatic hypermutation is often difficult to distinguish from artifactual variants generated by sequencing error. To correct for error, data can be filtered based on quality scores reported by the sequencing instrument or known patterns of error generation; however, even then, it is often not possible to completely

remove it from the data. A major advantage of our approach is that it allows us to unambiguously identify the true sequence variants from sequencing error without having to make assumptions about patterns of somatic hypermutation or error generation. Moreover, it allows correction of errors generated at any point in the process, including reverse transcription, amplification, or sequencing – something not possible unless multiple transcripts from a single cell are sequenced.

To obtain high confidence sequence data, we use two filters unique to our approach. The first filter removes sequences that originate from droplets containing a target transcript but not a cell. During encapsulation, cells are maintained in a syringe for several minutes, over which some may lyse and release their transcripts into solution. These transcripts, in turn, can be encapsulated in droplets, amplified, and barcoded. Because the reads originating from such “digital background” droplets do not represent a single cell, they must be discarded, which we accomplish by throwing out all barcode groups that do not contain reads from a heavy and light chain. Another important filter is barcode group coverage. As we increase the threshold for coverage, the number of barcode groups that pass this filter decreases inverse-exponentially (blue line Fig 3d), but the percent of the total reads passing decreases less rapidly (red line Fig 3d); this indicates that the majority of the data resides in high coverage barcode groups corresponding to true, single cell data. With deeper sequencing, more barcode groups pass this filter, yielding data on more single cells.

There are ~300,000 sequence variants within the reads that pass our coverage filter. Many are artifacts resulting from preparation and sequencing errors, and must be discarded to provide the true biological diversity of the antibody repertoire. To reveal which variants are artifacts of sequencing and which represent true, biological diversity, we assign a sequence variant ratio (SVR) to each variant. The SVR is the number of reads within a given barcode group that exactly match that sequence variant, divided by the number of all reads in the group for that chain (Fig. 3a, far right). If a given sequence variant represents the actual sequence of the gene, then a majority of reads will agree with that sequence, while reads that contain random errors will tend to mismatch at different bases. Hence, biological variants should have a high SVR, while sequencing errors should have a low SVR.

The confidence with which we can use the SVR to differentiate true from artifactual variants, depends on sequencing coverage: True mutations maintain high SVR as coverage is increased, whereas errors should decrease in SVR with coverage. A heat map of how the number of variants passing the coverage and SVR filters is shown in Fig. 3e. As anticipated, as coverage increases, the number of variants maintaining high SVR decreases, indicating that most do not represent true, biological variation. Taking a slice of the heat map at 0.85, we find a precipitous drop in the fraction of sequence variants that maintain high SVR as coverage increases, as shown in Fig. 3f. Above a coverage filter of 5 reads, the number of sequence variants passing the SVR filter stabilizes, indicating that setting a more stringent coverage filter does not exclude an appreciable amount of the data.

The SVR is a powerful means by which to differentiate artifactual variants resulting from sequencing error from true, biological variants, but is only possible by sequencing single cells: While artifactual variants do not represent true variations and should thus

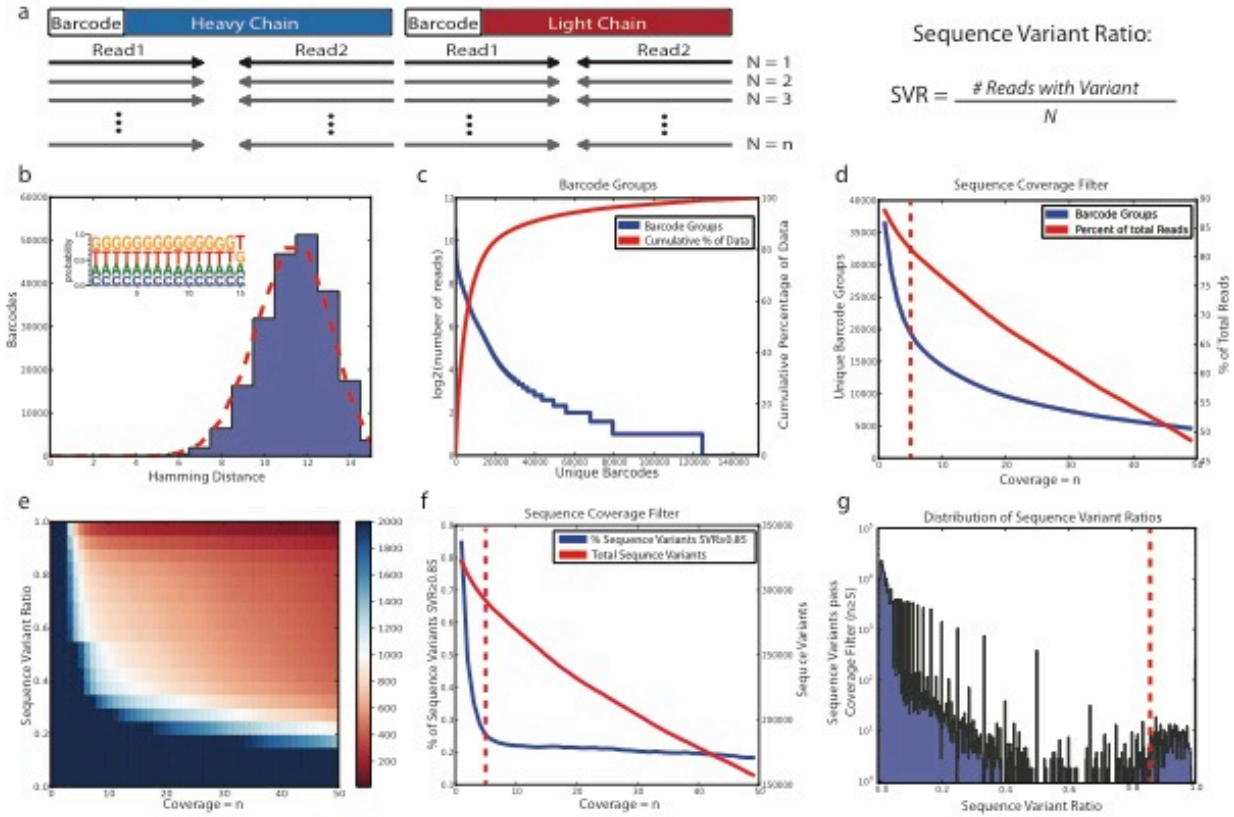


Figure 3: (a) Structure of the barcoded products showing read position and depth. Read depth within a barcode group is used to assign a sequence variant ratio (SVR) to all sequence variants. (b) Distribution of hamming distances between sequenced barcodes with the expected distribution shown in red. Inset: The distribution of bases across the 15 nucleotide barcode. (c) Number of reads per unique barcode in rank order (blue) and the cumulative percent of sequence data (red). (d) Coverage filter showing the number (blue) and percent of total reads (red) for barcode groups that pass the filter (dashed line is the coverage filter used, n≥ 5). (e) Heat map showing the total number of sequence variants that pass both coverage and SVR filters. (f) Total number of sequence variants (red) and number of sequence variants with SVR ≥ 0.85 (blue) for a given coverage filter (dashed line is the coverage filter used n≥ 5). (g) Distribution of SVR ratios for all variants that pass coverage filter n≥ 5 (dashed red line is the SVR filter used, SVR ≥ 0.85).

decrease in SVR with increased coverage, true variants will be reconfirmed with each new sequenced read, and should thus maintain high SVR. Consequently, the distribution of SVRs should self-segregate into two populations that become more distinct with increased coverage: low SVRs representing artifactual variants, and high SVRs representing true, biological variants. This provides a robust means by which to identify true biological variants within the sea of sequencing errors that outnumber them by hundreds of times. To illustrate this, we plot the histogram of SVRs for all 300,000 sequence variants that pass our coverage filter. Of this population, a scant 855 variants maintain an SVR of 0.85 at a coverage of 5 reads. The two populations correspond to the two peaks in the distribution, allowing unambiguous differentiation between artifactual and true variation.

Accurate heavy and light chain antibody sequences provide insight into B cell lineages

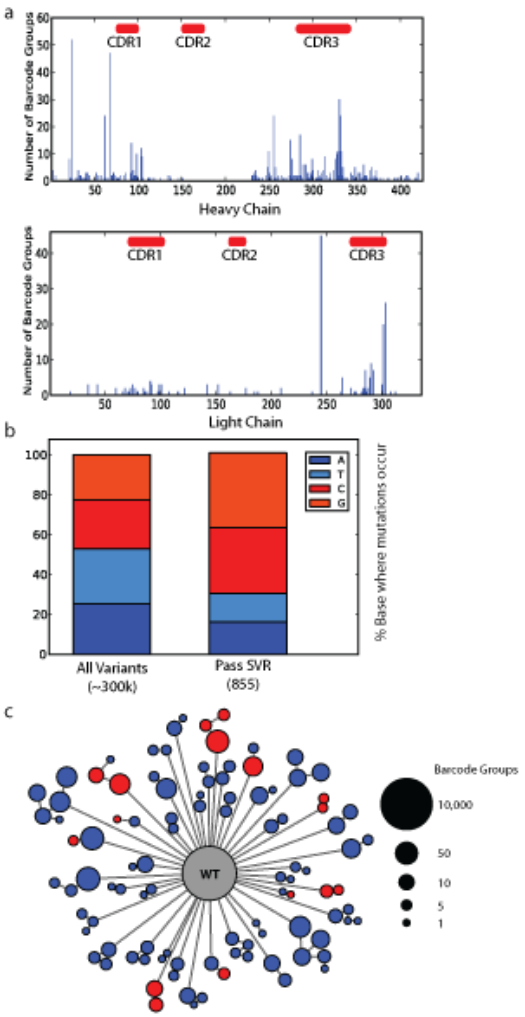


Figure 4 – (a) Distribution of mutations in barcode groups across the heavy and light chain sequences with the CDR regions marked in red at the top. (b) Nucleotide bias of mutations within the high confidence SVR filter. (c) Distributed network of mutations showing clone lineages with multiple mutations from the wild type sequence.

Somatic hypermutation (SHM) is an important mechanism by which B cells evolve and select for antibody binding. SHM centers on the complementary determining regions (CDRs) of the antibodies – the regions most important to antigen binding. In Ramos cells, SHM exhibits base composition bias, with mutations occurring most often at guanines and cytosines. A unique challenge of detecting variants resulting from SHM is that the frequency of SHM induced mutation is on the same order as artifactual sequencing error. To illustrate that our approach is uniquely suited to accurately characterize of SHM, we map the locations of the 855 high confidence variants passing our stringent coverage and SVR filters onto the heavy and light chain sequences (Fig. 4a). As anticipated based on the known biology of SHM, most mutations cluster in the CDR regions of the genes, especially the third CDR, the most variable portion of the antibody and important for binding [39]. Interestingly, while the base composition of the low confidence 300,000 variants

matches known error patterns for NGS, the high confidence variants demonstrates a strong guanine and cytosine bias, further supporting that these variants represent the true variants of our Ramos cells (Fig 4b) [40] [41]. Another unique advantage of our approach is that we can associate sequences for completely separate genes originating within the same single cell, such as the heavy and light chain sequences. Within our 20,000 barcode groups, we find mutations distributed across both chains (Fig. 4c). Viewing these mutations as a distributed network, we observe signatures of lineage expansion, whereby mutations occurring early are passed on to later generations. Because single mutations can significantly alter binding efficiency, the ability to track lineage expansion across heavy and light chains is important to understanding the evolution of epitope recognition by SHM.

Conclusions

High-throughput single cell sequencing is a powerful tool for characterizing complex biological systems comprehensively. As we have shown, it allows true biological variation to be profiled accurately, even when its frequency is of the same order as the error rate of sequencing. This is particularly important in antibody repertoires, in which somatic hypermutation can appear indistinguishable from sequencing errors. In addition to allowing the generation of accurate consensus sequences for single genes, our barcoding strategy also allows mutations occurring on distinct genes to be associated together, essential for capturing the combinatorial diversity of heavy and light chain antibody sequences. Moreover, whereas methods that fuse genes are limited to associating a small number of loci, our method of barcoding the loci with sequence

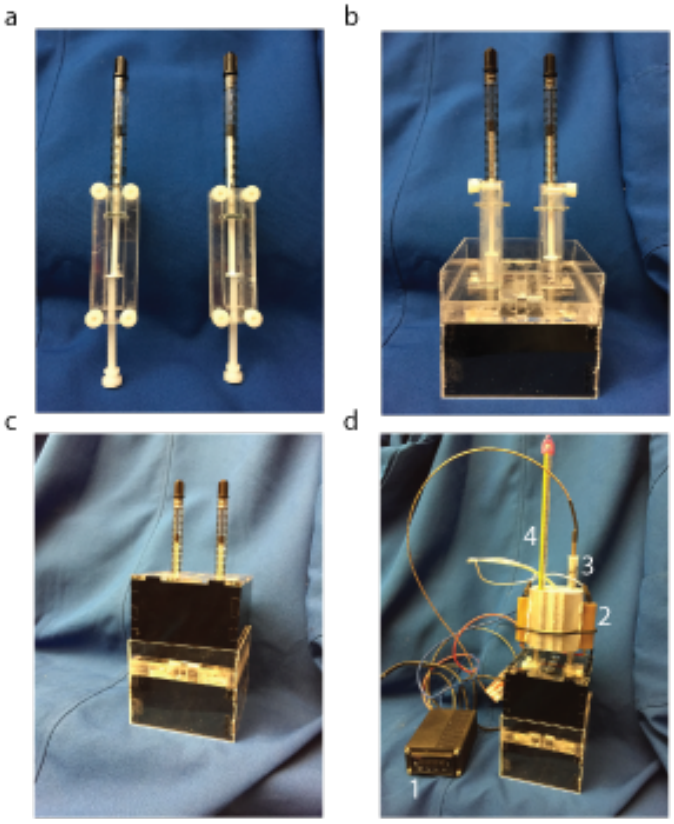


Figure S1 – An upright syringe incubation block. Syringes containing droplets are first secured into custom syringe clamps (a) and placed into a base plate that holds them upright (b). A lid is placed over the baseplate (c) and a milled out aluminum heat block is placed over the top of the syringes (d). The heat block is controlled by a PID controller (1) powering resistive heaters (2) with a thermocouple (3) for feedback control. A chamber for a thermometer (4) is also included so that the block temperature can be verified.

identifiers is extendable to many genes per cell, and even to whole transcriptomes or genomes. By leveraging new reagents comprising antibodies labeled with DNA identifiers, our approach should also allow highly multiplexed proteomic profiling of single cells, which can be performed simultaneously with mRNA profiling. The ability to track the flow of information through a single cell's genome, transcriptome, and proteome, for large populations of single cells promises to open a new frontier in systems biology and reveal insights into the molecular determinants of many diseases.

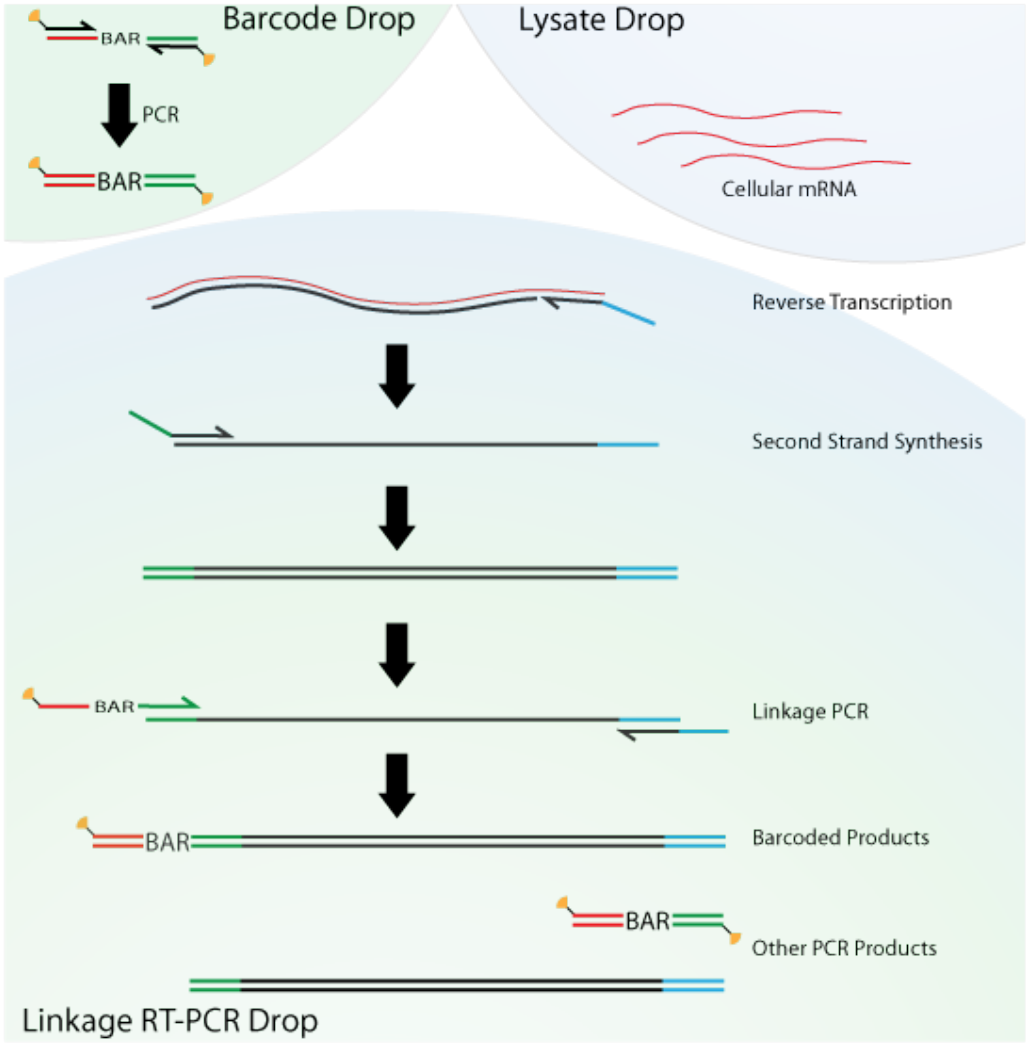


Figure S2 – Biotinylation scheme for purification of Linkage PCR products. Barcodes are produced by ddPCR with biotinylated forward and reverse primers (upper left), resulting in products that contain biotin on both DNA strands. When these barcodes are used for linkage PCR with cellular mRNA from a drop containing cell lysate (upper right) the result is a mixture of three products: single biotinylated and barcoded linkage PCR products, double biotinyalted free barcodes, and non-biotinylated RT-PCR products. When the mixture is purified with Streptavidin coated beads the non-biotinylated RT-PCR products will be washed out while the double biotinylated barcodes will bind the beads irreversibly. The single biotinylated Linkage PCR products can then be denatured off the beads and sequenced.

Genetic Interaction Mapping with Microfluidic-Based Single Cell Sequencing

Abstract

Genetic interaction mapping is useful for understanding the molecular basis of cellular decision making, but elucidating interactions genome-wide is challenging due to the massive number of gene combinations that must be tested. Here, we demonstrate a simple approach to thoroughly map genetic interactions in bacteria using microfluidic-based single cell sequencing. Using single cell PCR in droplets, we link distinct genetic information into single DNA sequences that can be decoded by next generation sequencing. Our approach is scalable and theoretically enables the pooling of entire interaction libraries to interrogate multiple pairwise genetic interactions in a single culture. The speed, ease, and low-cost of our approach makes genetic interaction mapping viable for routine characterization, allowing the interaction network to be used as a universal read out for a variety of biology experiments, and for the elucidation of interaction networks in non-model organisms.

Introduction

Cells rely on interactions between biomolecules to achieve complex and dynamic capabilities[42]. For example, cells use genetically encoded signaling proteins to interrogate environmental conditions necessary for adaptation and survival, such as by detecting competitors and responding by secreting an antibiotic. The complete set of biomolecular interactions that a cell uses is often depicted as a connected network known as a genetic interaction diagram[43-45]. With complete knowledge of the interaction network of a cell it is

possible, in theory, to predict how the cell will respond to any given stimulus. While achieving such predictive power in practice is not currently possible, even partial understanding of the interaction network is valuable and is a core concept in systems biology[46, 47]. For example, in the study of human health genetic networks are useful for understanding how pathways are dysregulated in disease or drug metabolism. Additionally there is interest in using genetic interactions to better understand novel and synthetic properties of microorganisms, such as the ability to digest environmental contaminants or produce biofuels from cellulosic biomass. Consequently, there is immense interest in novel methods to systematically map genetic interaction networks [48-52]. _ENREF_8

One way to infer the genetic interaction diagram of a cell is to apply genetic perturbations and observe the impact on a phenotype. By performing two such perturbations simultaneously, it is possible to infer an interaction between a pair of genes [53-55]. For example, if two genes do not interact, the removal of both genes should have a multiplicative effect on phenotype, whereas genes that do interact will produce more complex phenotypes that include suppression or synthetic lethality[56]. The utility and power of a genetic interaction network grows as an increasing number of pairwise interactions are characterized, and is greatest and most detailed by an exhaustive mapping of all possible pairwise interactions[57].

Model systems, like the budding yeast *Saccharomyces cerevisiae* and the bacterium *Escherichia coli*, were some of the first used for systematic genetic interaction mapping, due to the ease with which they can be manipulated[51, 55, 58-62]. This facilitated the development of the single- and double-knockout libraries needed for these studies[63, 64]. However, while generating massive libraries of double knockouts is technically feasible in these microorganisms, screening their phenotypes is far more difficult. For example, screening every pairwise genetic knockout in the *S. cerevisiae* genome, comprising ~6,000 genes, requires

screening of ~20 million strains. Even with recently-developed high-throughput colony methods, only thousands of combinations can be measured simultaneously, a minute subset of the space of possible combinations[59]. Consequently, to make best use of these screens, much care must be taken in selecting which genes to test as queries; this is not always possible and, even when it is, represents a biased means of mapping the interaction network, since it is only possible to detect interactions that are tested[65].

In this paper, we describe a method for comprehensively mapping genetic interactions. The key to the method is the use of microfluidics to isolate single cells in picoliter droplets at extremely high-throughput. Once confined in the droplets, single-cell linkage PCR physically links the genetic perturbations into a single DNA sequence for analysis by next-generation sequencing (NGS) [65, 66]. This, in essence, converts a library of living cells into a library of DNA molecules, wherein each molecule contains sufficient information to determine genotype of the cell from which it originated. Moreover, since the sequencing depth of a specific sequence is proportional to the relative abundance of the corresponding strain in the culture, the fitness of each strain can be estimated by comparing its membership in the population[67, 68]. This makes our approach supremely scalable: Whereas comprehensive screening of double knockout libraries of yeast or *E. coli* would require >10,000 high-density plates, our method can theoretically perform the same screen in a single culture. The speed and ease of our approach will enable the generation of genetic interaction networks in a variety of experimental conditions in diverse microorganisms. For example, rather than just screening a double mutant library in a single conditions (such as rich media), our approach can be adapted to screening multiple conditions, including temperature changes, altering the starting composition of the population, or including chemical perturbations. The availability of conditional genetic interaction networks will be useful for elucidating the cellular logic that underlies environmental sensing and adaptation, and may enable the identification of new drug targets.

Methods and Results:

Mapping genetic interactions requires comparing the phenotypes of single gene perturbations to the phenotypes of double gene perturbations. Making libraries of single genetic knockouts is straightforward, but producing libraries of double mutants is supremely challenging. A common way to produce this library is to cross libraries of single knockouts to generate strains containing defined double-knockout combinations. Alternatively, the single-knockout library can be complemented with a library of additional genes of complementary function (Fig 5a). Genetic interactions within the libraries are scored by measuring the fitness (or growth) of each double

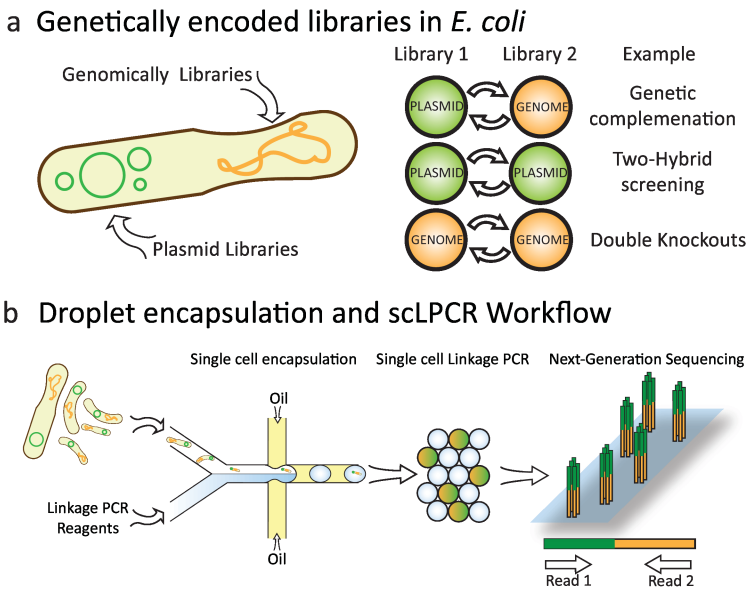


Fig 5: Screening Genetic Interaction Libraries by Single-cell Sequencing with Droplet Microfluidics
(a) Genetic libraries can be genomically encoded or introduced through episomal DNA like plasmids. Interaction libraries are created by combining two genetic libraries. Some of the most common types of interaction libraries are noted. (b)Libraries are screen by microfluidic encapsulation and single-cell linkage PCR (scLPCR) inside picoliter droplets. Confining cells inside of droplets allows PCR to link cellular DNA without crossover contamination of DNA from other cells. The PCR products are sequenced using paired-end chemistry on an Illumina platform to decode linkage products.

mutant strain. Moreover, the culture conditions can be varied, such as by depriving the cells of an important nutrient or adding a drug, to study how genetic interactions change under these conditions. This can be used, for instance, to elucidate the targets of a drug or to deduce key proteins important for signal processing.

A challenge in performing the mapping is tabulating all double knockouts with their fitness under the screening conditions. One way to accomplish this is to isolate each strain at a known location on a plate, and to measure colony growth at that spot. Since the knockout combination at each spot is known, it is straightforward to assign a fitness value to the perturbations. A limitation of this method, however, is that it is onerous to scale at the level needed to completely map genetic interactions in even the simplest cells, due to the need isolate each combination at a spot; this necessitates expensive robotics in addition to immense amounts of reagents and person-hours. Consequently, in most genetic interaction screens, only a small subset of possible interactions is tested. However, deciding which genes to test is not always straightforward and, even when it can be done, the screen will be biased, capable of discovering only interactions that are tested.

An alternative would be to combine all library members into a single, pooled culture, and to quantify population abundance afterwards without having to position each knockout combination on an array. While this is possible with single knockout libraries by “barcoding” strains prior to screening[69], it is not with double knockouts. To barcode strains, a unique identification sequence is associated with each knockout. To quantify population abundance, the barcodes can be amplified with PCR and counted by sequencing[67, 70]. While it is possible to barcode each perturbation separately in a double mutant, it is not currently possible to determine which *combination* of barcodes exists within each cell in a random, high-throughput manner. For example, if a population of double-knockout strains is subjected to PCR to amplify the barcodes,

the resultant amplicons for all cells would mix in solution, abolishing information about which pairs existed within the original cells. Retaining this information requires a means for associating together barcode pairs within single cells. Such a method would be very powerful because it would allow a large number of genetic interactions to be screened and retroactively scored in a single, pooled culture.

Our strategy to enable this optimally scalable approach to genetic interaction mapping is to use single cell droplet PCR to fuse barcode combinations into single molecules; these chimeric molecules can be sequenced in massively parallel fashion using NGS (Fig 5b). Moreover, since the sequencing depth of a particular barcode (or barcode pair) is proportional to its abundance in the culture, the fitness of each strain can be estimated by relative membership of its barcode in the sequence data. The key enabling feature of our approach is the ability to perform PCR on millions of single cells using microfluidics, an approach we term single cell linkage PCR (scL-PCR). The principle of scL-PCR is predicated on the ability to rapidly encapsulate single cells inside of microdroplets, where PCR can be used to link cellular DNA without contamination from the DNA of other cells (Fig 5b).

To investigate if scL-PCR faithfully enables the accurate identification of heterogeneous strain combinations in a mixed culture, we prepared two *E. coli* strains (Fig 6a, left). Strain ECK1365 contains a knockout at the *ynaA* locus with the 1365 barcode and strain ECK0679 contains a knockout at the *ybfH* locus with the 0679 barcode. The unique barcodes comprise known sequences of 20 bases embedded in a chloramphenicol selection marker. We perform linkage PCR using primers that will link the barcode sequences with each genetic locus. Performing linkage PCR in bulk, as expected, yields chimeric products comprising all four random combinations (two barcodes, two open reading frames); this is because bulk PCR allows the amplicons of both cells to mix in solution, generating chimeric products that consist of

sequences from both cells, and which do not represent the genotypes of either cell. By contrast, if the linkage PCR is performed on single cells the only fusions that are generated correspond to the true genotypes of the cells (Fig 6b).

Our method confines this single cell reaction in picoliter droplets using a microfluidic dropmaker (S3 Fig). Because these droplets can be generated at >1 kHz, our approach can process

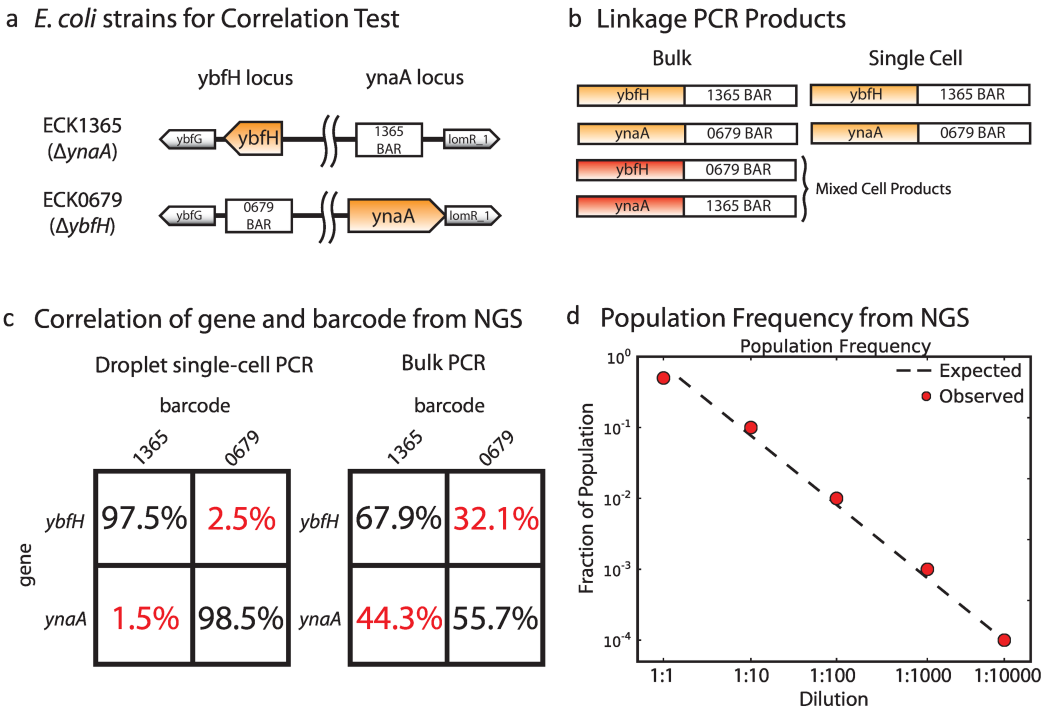


Fig 6: Droplet based single-cell sequencing preserves genomic structure and population membership
(a) KEIO collection strains of *E. coli* used to test linkage PCR: a barcode has been inserted into the genome at defined loci, creating gene knockouts of ynaA and ybfH in strains ECK1365 and ECK0679, respectively. (b) Linkage PCR to fuse the sequence from both genomic loci in the two strains yields a mixture of four products in bulk (left), two of which reflect the true genomic organization and two that reflect spurious mixed cell products. However, single-cell linkage PCR (scLPCR) only yields the two products that reflect true genomic organization. (c) Deep sequencing of products from bulk linkage PCR or scLPCR showing percent of reads that reflect true genomic organization (in black) or spurious mixed cell products (in red), indicating the scLPCR on a culture of mixed cell types recovers reports on the genomic variation within the population. (d) The fraction of the population determined by sequencing depth (red dots) when one KEIO strain is spiked into a culture of the other strain at defined dilutions shown on the x-axis. The expected results are shown as a dashed line.

millions of cells per hour; using higher throughput droplet generation techniques, throughputs of billions of cells are achievable. To demonstrate this, we grew the two E. coli KO strains described above separately and pooled them before encapsulation. The cells are individually encapsulated in droplets using microfluidic flow focusing at a concentration limiting dilution such that only 1 in 10 drops contains a cell. For comparison, we also aliquot a portion of the mixed cell population into a PCR tube and perform the LPCR in bulk. The products of the droplet and bulk reactions were prepared for NGS and sequenced using a paired end format, where the sequence from each read reports on a single genetic locus. The droplet workflow yields products accurately reflecting the genotypes of the original populations (Fig. 6c, *left*), whereas the bulk reaction shows the expected mixed products (Fig. 6c, right). This demonstrates that scLPCR in droplets preserves the genotypes of the strains.

In addition to determining the genotype of each double mutant strain as described above, genetic interaction mapping also requires that we assign a fitness value to each double mutant combination. This can be accomplished by counting the number of instances of each barcode fusion in the sequencing data. As an illustration, prior to encapsulation in droplets we mix the strains at different ratios from 1:1 to 1:10,000 cells. We find that sequencing depth accurately reflects membership library over the four order-of-magnitude range (Fig. 6d) that we tested. This demonstrates that read counting is an accurate means by which to quantify strain fitness.

Genetic interaction mapping can be accomplished by performing gene perturbation combinations that are genome-to-genome or genome-to-plasmid. Alternatively, they can also be performed via plasmid-to-plasmid interactions (for example, with two CRISPR-Cas9 constructs). To conceptually illustrate this, we created a library of 64 *E. coli* strains containing unique barcodes encoded on two separate plasmids (Fig. 7a). These plasmids were adapted from a two-hybrid strategy for detecting protein-protein interactions in bacteria. The 64 individual

strains were grown from frozen glycerol stocks and combined into a single, pooled population. The pool was subjected to the droplet workflow and the resulting scLPCR products were sequenced. As a control, we grew and performed the linkage PCR for the 64 strains individually. The percentage of reads that match the known barcode combinations are the same for the droplet scLPCR method and the individual mapping method, demonstrating that the droplet method performs optimally. In contrast, a bulk reaction control again yields mostly mixed products (Fig. 7b). Interestingly, only ~94% of reads for either droplet or isolated strain experiments match the known strain genotypes. We believe this to be due to spurious gene fusions (chimeras) generated during NGS library preparation, which requires a bulk PCR on the mixed products, which may lead to additional fusions. The frequency of these fusions may be reduced by optimizing sequencing preparation and by employing compartmentalized amplification methods, such as emulsion PCR. See the supplementary methods for a more in-depth discussion of noise and experiment design.

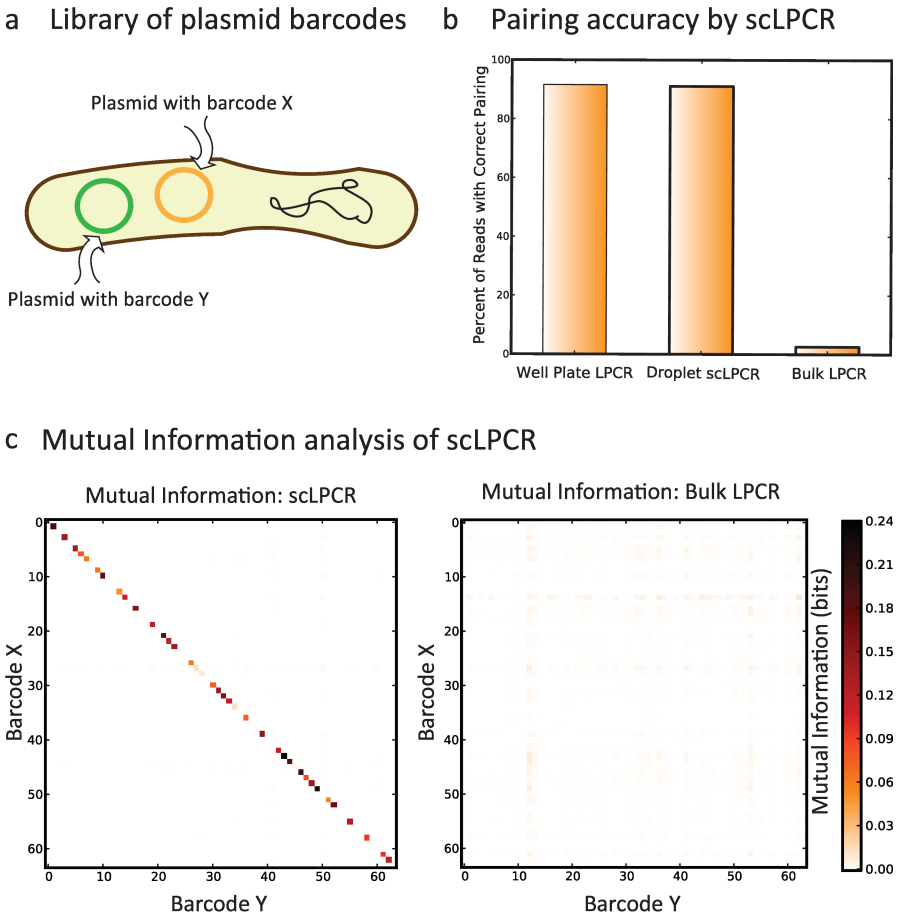


Fig 7: Screening complex libraries with droplet sc-Seq
(a) Library of *E. coli* containing 64 strains, each containing a pair of known barcodes (denoted X and Y) located on separate plasmids. (b) The accuracy of barcode X and Y pairing from NGS (as percent of sequencing reads that report a correct X/Y pair) is the same when using linkage PCR on isolated strains (well plate LPCR) or when using single cell linkage PCR in droplets (Droplet scLPCR), while linkage PCR from all strains in bulk (Bulk LPCR) yields mostly random X/Y pairs. (d) The amount of mutual information between specific X/Y barcode pairs in the NGS data shows strong correlations along the diagonal, representing true X/Y pairs. In the same data for libraries from linkage PCR in bulk there is no correlation between represented barcodes.

A valuable means by which to estimate the effectiveness of our method is to plot the mutual information (MI) between the known and measured pairwise gene interactions (Fig. 7c). Mutual information is a measure of the confidence with which the presence of one barcode can be associated with that of another. The barcode identities are plotted along the axes and ordered such that correct fusions fall along the diagonal, where the color of the bin is proportion to the MI between the barcodes. For the droplet scLPCR, there is substantial MI between the

barcodes on the diagonal, which represent the true sequences of strains in the library. In contrast, the bulk PCR shows little MI for all combinations, which indicates that pairing is random. Peculiarly, there are gaps where known barcode pairs should be present (Fig. 7c, *left*). This is likely due to the level of that strain in the population being too low to detect with the sequencing depth that we used. Likely, deeper sequencing would pull out these less-abundant strains.

The speed, ease, and low cost of scLPCR make it valuable for screening the *conditions* under which the cells are cultured, which is useful for investigating how genetic interactions mediate responses to environmental conditions. To illustrate how this can be used to answer a biological question, we generated a new genetic interaction library for amino acid auxotrophy. The library contains 6 strains of *E. coli* with single gene deletions, wherein a unique DNA barcode has been inserted into the genome of each strain at that locus. In five of the strains, the deleted gene is essential for amino acid biosynthesis, such that these strains are unable to grow in media not supplemented with the essential amino acid. We also construct four barcoded complementation plasmids that express one of the amino acid biosynthesis genes. If the strain with the deleted gene is complemented with a plasmid encoding that gene, it can synthesize the needed amino acid and, thus, should go in the deficient media. We transformed the set of six strains with the library of four complementation plasmids (24 total strains). The transformed library was recovered for a short time in rich media, washed 3 times in minimal media, and split into two new cultures. One culture was grown in rich media and the other was grown in minimal media. The cultures were grown for 16 generations with periodic dilution to keep them in exponential phase. The cultures were sampled periodically and analyzed by the droplet scLPCR workflow to measure culture membership. We kept the optical density of the cultures low to minimize crosstalk between cells in the culture and to ensure that other media nutrients do not become limiting.

Using scLPCR, we tracked the culture membership over the 16 generations (S4 Fig), finding that culture composition changes in rich and minimal media (Fig 8a). The proportional difference in composition between rich and minimal media at each time point reflects the biological impact of amino acid auxotrophy (Fig 8b). The *ynaA* knockout, which contains no amino acid auxotrophy, should grow equally well in rich or minimal media. As expected this strain is significantly enriched in the minimal media culture. Conversely, the *tyrA* knockout cannot grow in minimal media and cannot be complemented by any of the plasmids in our library; therefore this strain drops out of the culture grown in minimal media. In addition to tracking the membership of the culture by strain, we can track the membership of plasmids within each strain. We find that there is no enrichment for the knockouts of *hisB*, *leuB*, *metA*, and *proA* at the strain level (Fig 8b), but within each strain there is enrichment for cells harboring the needed complementation plasmid (Fig 8c) across the 16 generations of growth. Peculiarly, we also found that cells with the *metA* complementation plasmid persisted in the culture. This observation turns out to be consistent with recent findings suggesting that overexpression of the MetA protein can drive cells towards a persistor phenotype.[71]

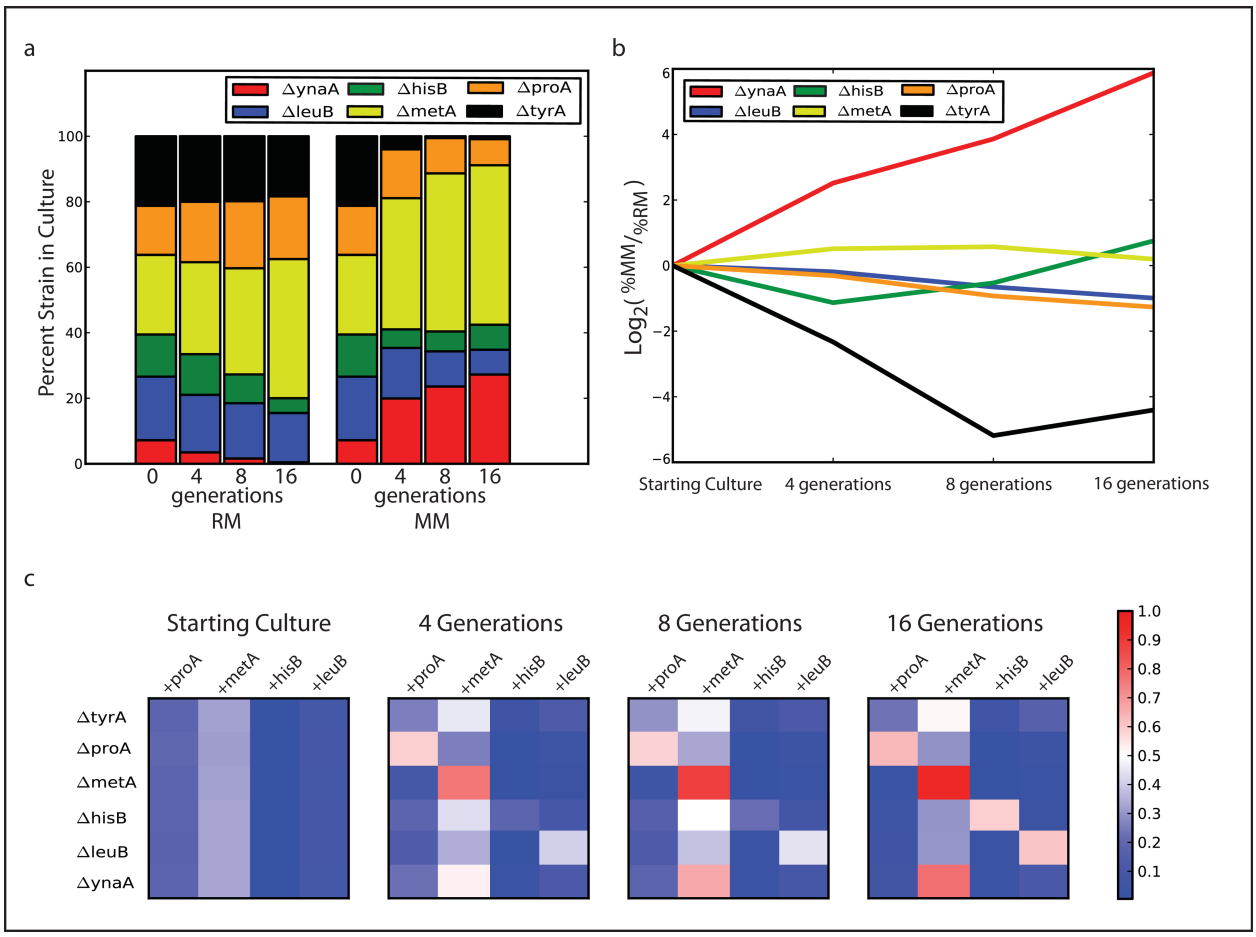


Fig 8: Screening a combinatorial library of amino acid auxotrophy with sc-Seq
(a) The membership (by strain) of heterogeneous cultures is tracked by droplet scLPCR for cultures grown in rich media (RM) or minimal media (MM) at 0,4,8, or 16 doublings after inoculation. (b) The fractional fold change in minimal media vs. rich media over 16 doublings shows that auxotrophic strains with no complement drop out of the population (ΔtyrA, black line) while prototrophic experience no selection and take over (ΔynaA, red line). (c) Droplet scLPCR shows the culture composition by strain and plasmid and unmasks the mechanism of complementation, whereby auxotrophic strains persist in the culture through selective outgrowth of only those strains that harbor the corresponding complementary gene (color corresponds to fraction of sequencing reads within each strain that are specific for the corresponding complement gene).

Conclusions

We have demonstrated a method to rapidly screen genetic interactions in a single culture. We produced genetic interaction libraries comprising two genetic perturbations and used single cell linkage PCR and NGS to reliably quantify the levels of every member in the library. This should

make our approach useful for non-model bacterial systems, wherein genomic modification (by transposons, CRISPR-Cas9, or targeted modification) is the only requirement. In addition, the massive scalability afforded by droplet microfluidics should enable higher order interactions, such as 3-gene interactions, to be tested.

A key advantage of our approach is the speed and ease with which libraries can be screened across multiple conditions. This allows our approach to be adapted to multiple library types, including genetic knockouts, the addition of biosynthetic pathways and non-native genes, and protein interactions like the classic two-hybrid screen. We envision that this method can be extended to eukaryotic systems for use in medical research and drug development. The elucidation of genetic interaction networks in model systems like *S. cerevisiae* and *E. coli* capitalized on decades of development in microbiology and precise molecular tools. Though library creation is time consuming, the pictures that emerge are rich in information and provide key insights into genomic design principles, and these libraries continue to be screened and mined for information. The arrival of new molecular tools like the Cas9 system allow these same concepts to be extended to new organisms with ease and it is expected that the library creation process will no longer be rate limiting. The use of droplet microfluidics to deconvolute complex cell libraries is a powerful tool that can be combined with next-generation methods of library creation to allow for truly rapid interaction profiling in a multitude of conditions, time points, and formats.

Supporting Information

Methods and Materials:

Fabrication of microfluidic dropmakers

The microfluidic devices are fabricated using soft lithography[Basic Microfluidic and Soft Lithographic Techniques]. SU-8 photoresist (MicroChem Corp) is spun onto a 3”silicon wafer (University Wafer) to a desired thickness and baked at 135°C to remove solvent. A photo transparency mask (CAD/Art Services) containing the device features is placed on the wafer and exposed to UV light to crosslink the photoresist. Following UV exposure, the wafer is post-baked at 135°C for 1 minute and placed into a developing bath of propylene glycol methyl ether acetate (PGMEA, Sigma). Following development with PGMEA, the masters are washed with isopropanol and post-baked at 135°C for 30 minutes. The masters are placed into plastic petri dishes and covered with degassed poly(dimethylsiloxane) (PDMS) prepared from 10:1 ratio of elastomer:crosslinker (Sylgard 184, Dow Corning). The dish is evacuated to remove entrapped air bubbles and baked at 65°C for at least 2 hours to crosslink the PDMS. The PDMS devices are cut with a scalpel and peeled away from the master. Holes for inlets and outlets are punched using a biopsy core (Harris Uni-Core), the devices are rinsed with isopropanol, and they are plasma-bonded to glass slides. The devices are flushed with Aquapel to render the channels hydrophobic and enable water-in-oil emulsification, and baked at 65°C for 20 min to remove excess Aquapel. To operate the microfluidic devices, Polyethylene (PE) tubing (Scientific Commodities) is used to connect device inlets to syringes containing reagents, and a custom Python script used to control syringe pumps and inject liquids into the device. The oil used is Novec HFE 7500 (3M) containing 2% fluorosurfactant (RAN Biotechnologies) and droplets are collected into PCR tubes. Prior to subjecting droplets to thermal treatments, the HFE oil is removed from beneath the droplets with a pipette fitted with a gel loading tip and an equal volume of FC-40 oil (Sigma) containing 5% surfactant is added above the emulsion.

Strains used in this study

Unless otherwise noted, all knockout strains in this study were taken from the ASKA knockout collection [72].

Culture conditions for two-strain experiment

Strains ECK1365 and ECK0679 were separately inoculated in to LB Broth containing 30ug/mL chloramphenicol. Strains were grown to saturation overnight and then used to inoculate fresh cultures at an O.D. ~ 0.005. Cultures were allowed to grow to mid-log (O.D. 0.2) and then pooled at various ratios (1:1, 1:10, etc). Pooled cultures were diluted to an O.D. of 0.005 in ddH2O (~1 cell per 200pL).

Diluted cells were encapsulated with PCR Mix (Phusion polymerase, detergent free buffer) in a co-flow microfluidic device. Devices are 30um in height and use a dropmaking nozzle that is 30um wide, resulting in drops that are ~35um in diameter. Flow rates for each aqueous inlet are 200ul/hr and flow rate for the oil is 800ul/hr.

For bulk experiments, diluted cells are combined with LPCR mix directly in a PCR tube.

Culture conditions for 64 strain experiment

Freezer stocks of 64 strains were inoculated into a deep-well 96-well plate containing 200uL of LB broth 30ug/mL chloramphenicol and 50ug/mL kanamycin. Strains were allowed to grow at 37^{o}C for 3 hours. For the well plate control, each well was sampled individually for LPCR. For the droplet experiments strain were combined and diluted to an O.D. of 0.005 in ddH2O (~1 cell per 200pL).

Diluted cells were encapsulated with PCR Mix (Phusion polymerase, detergent free buffer) in a co-flow microfluidic device. Devices are 30um in height and use a dropmaking nozzle that is 30um wide, resulting in drops that are ~35um in diameter. Flow rates for each aqueous inlet are 200ul/hr and flow rate for the oil is 800ul/hr.

For bulk experiments, diluted cells are combined with LPCR mix directly in a PCR tube.

Construction of barcoded complementation plasmids

Barcodes were introduced into plasmid pBbA2k-RFP(gift from Jay Keasling (Addgene plasmid # 35327)) by overlap PCR with primers the contained a 7bp barcode[73]. The plasmid contains a constitutive Tet promoter driving expression of RFP. Each amino acid biosynthesis gene was amplified from genomic DNA and cloned into the plasmid to replace the RFP gene. Cloning was done using Clontech's In-Fusion kit.

Culture conditions for complementation assay

Each ASKA knockout strain (6 total) was made competent and transformed with the set of 4 complementation plasmids. Cultures were pooled and recovered for 3 hours in rich media. Cultures were washed 3 times with minimal media before being inoculated into 50mL of either EZ-Rich Media (Teknova) or EZ-Min Media (EZ-Rich without Amino Acid supplement) at an initial O.D. of 0.02. Cultures were grown at 37^{o}C until the O.D. reached ~0.32 (4 generations), at which point culture were sampled and diluted back to O.D. 0.02.

Sequencing on the MiSeq NGS Platform:

The products of each LPCR reaction were subjected to an additional bulk PCR in order to add sequencing adapters. Products from this second PCR were column purified (Zymo Research) and sequenced on a MiSeq platform using a paired end format and 200bp reads. Reads were analyzed with custom scripts that extracted barcode or gene signals from each read.

Experimental design for measuring genetic interactions by deep sequencing

Traditional methods of measuring genetic interactions use growth on solid agar plate to calculate a fitness value (W) for a particular strain, defined as the area of the colony at some time (t_2) when imaged by a camera[57]. Precise control over the initial seeding density at the beginning of the experiment (t_1) and spatial separation of strains eliminate significant sources of noise.

In liquid cultures of mixed strains the fitness value of a strain is conceptually similar and defined as the fold expansion of each strain relative to the rest of population and is mathematically expressed for strain *i* as:

$$Wi = \frac{\ln\left(N_i(t_2)/d(N_i(t_1)\right)}{\ln\left((1-N_i(t_2)\right) * d(1-N_i(t_1)))} \qquad (1)$$

Where $N_i(t_1)$ and $N_i(t_2)$ are the frequency of strain *i* in the population at time points t_1 and t_2 and *d* is the ratio of the optical densities at timepoints t_1 and t_2 and represents the growth of the culture[74] Others have shown that deep sequencing can be used on barcoded strains to obtain values for $N_i(t_1)$ and $N_i(t_2)$ by using sequence depth as a proxy for $N_i(t_1)$ and $N_i(t_2)$. Our method extends this approach to strains with two, and possibly more, barcodes by using single-cell linkage PCR to associate multiple barcodes from cells prior to sequencing. Our results are similar to previous results, as we show that sequencing accurately reflects culture composition across multiple orders of magnitude. When using this method it is important to consider certain parameters and possible sources of experimental noise and how they could convolute results.

Important parameters:

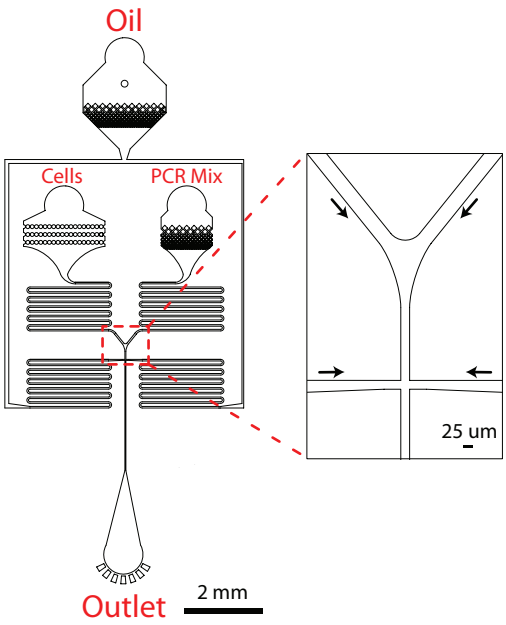
1. Library diversity and Sequencing depth: The confidence of fitness scores will grow as the culture is sequenced deeper. For larger libraries composed of several strains this will necessitate more sequencing. Importantly, because the method is a single-cell approach, it will also necessitate encapsulating more cells. Ideally the strain composition at the start of the experiment will be roughly equal, but for complex cultures that are made from hundreds of freezer stocks the distribution might be uneven and certain strains could drop out if not enough cells are screened and sequenced to capture them. As a general rule, at least 10 times as many cells as there are strains in the library should be screened, and at least 10 times as many reads as cells should be sequenced.
2. Controlling for dropouts: To control for the impact of dropouts, which could lead to false synthetic sick phenotypes, the culture should always be sequenced at the start of the experiment. From this sequencing data, all strains that are not detected with enough reads should be excluded from the experiment. Fit the read depth for all strains to a normal distribution and exclude those strains that don't meet a z-score threshold of at least 2 (~95% confidence).
3. Use of control strains: it can be helpful to include strains with known growth phenotypes, such as those that are known to be synthetic sick, synthetic lethal, or have no interaction. Using the strains will allow regions of the fitness spectrum to be assigned to these phenotype nomenclatures.

Possible sources of noise:

1. Multiple encapsulations: if the cells are seeded into the dropmaker at too high concentration then chimeric barcodes could be created. The impact of these products on the quality of the data should be minimal. In extreme cases there will be so many cells in each drop that the data will look like the mixed-culture control shown in Figure 3, where the frequency of every barcode pair is essentially the same, this level of contamination

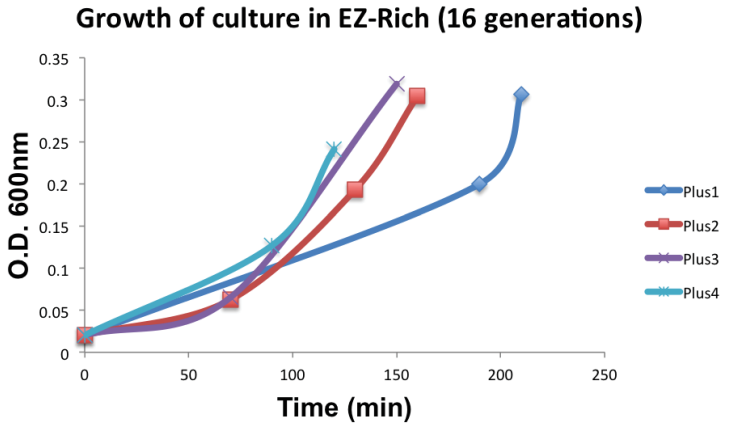
will be obvious to the user. Additionally, the use of control strains with known barcodes will enable the user to determine the amount of multiple encapsulations by observing the number of reads with spurious linkages between these barcodes and others.

2. PCR bias in drops: It is also possible that some strains will not amplify well in droplets. If that's the case these strains will fail to pass the dropout filter and will not be included for analysis
3. PCR bias in library preparation: It is known that bias is introduced into sequencing libraries through PCR. We observed a PCR cycle number dependent noise factor in our experiments and found that using a limited number of PCR cycles enabled us to produce libraries with high confidence. However, this source of bias could potentially create chimeric barcodes. Conditions should be optimized to use the bare minimum of PCR cycles needed to prepare sequencing libraries.
4. Culture density: Mixed cultures can influence each other through production of secondary metabolites, which could mask or exacerbate the effects of genetic interactions. It is crucial that the culture density be kept low enough to preclude the build-up of secondary metabolites. For that reason, the culture should be either diluted continuously or diluted when it reaches a certain O.D. threshold of early exponential phase.

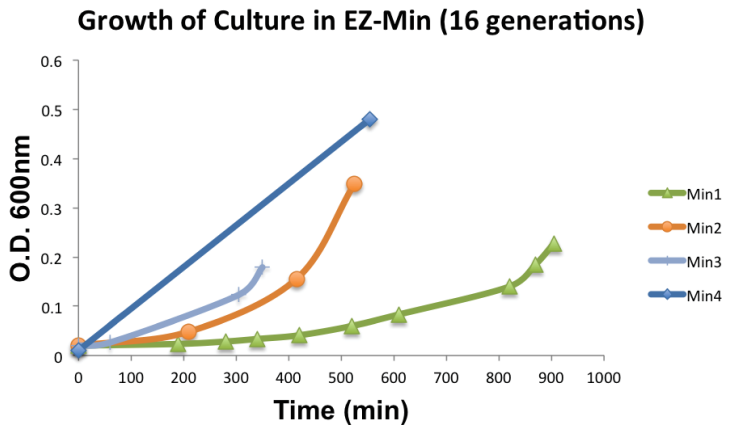


S3 Fig: Microfluidic device for sc-LPCR
The microfluidic droplet maker is a co-flow device consisting of a single outlet 3 inlets, one for oil and one each for cells and PCR mix. Aqueous mixes are flowed into a single channel that intersects a perpendicular channel of oil. Drops are made the junction and their size is function of the device geometry at the junction. This device has a width of 25 microns at the dropmaking junction and a height of 25um, which produces drops of approximately 30 microns in diameter.

a



b



S4 Fig: Growth of strains in auxotrophy experiment
(a) Library of complementation strains grown in EZ-Rich media for 16 generation. Each time the culture reaches O.D. ~0.32 (4 generations) it diluted back to O.D. 0.02. There is an initial lag of culture growth as the strains recover from transformation (Plus 1), but the culture quickly achieves uniform growth rate. (b) Library of complementation strains grown in EX-Min media for 16 generations. For this culture condition the lag phase is very long (Min 1), and each successive culture grows slightly faster.

Efficient Extraction of Oil from Droplet Microfluidic Emulsions

Droplet microfluidic techniques can perform large numbers of single molecule and cell reactions, but often require controlled, periodic flow to merge, split, and sort droplets. Here, we describe a simple method to convert aperiodic flows into periodic ones. Using an oil extraction module, we efficiently remove oil from emulsions to readjust droplet volume fraction, velocity, and packing, producing periodic flows. The extractor acts as a universal adaptor to connect microfluidic modules that do not operate under identical flow conditions, such as droplet generators, incubators, and merger devices.

Introduction

Microfluidics is a rapidly advancing field that is transforming multiple scientific disciplines by allowing precision control of fluids at picoliter scales[75-77]. Droplet microfluidics is a branch of this field in which a heterogeneous sample is partitioned into millions of distinct aqueous droplets in an immiscible carrier oil[78-80]. The ability to partition heterogeneous systems into subsamples is amazingly useful for applications across chemistry and biology. For example, when applied to molecules, it enables precision quantitation with digital ELISA[81, 82] and PCR[83, 84]. When applied to cells, it enables extremely high throughput single cell analysis, the evolution of enzymes to with unnatural properties, and the construction of pathways for biosynthesis of artificial molecules[4, 85, 86]. It can be used to characterize heterogeneous populations of cells and identify rare members, which is valuable in cancer, immunology, and infectious disease[87-90].

Most biological reactions require multiple steps of sample purification, incubation, and reagent addition, which are typically accomplished using microfluidic devices for droplet splitting, merging, and sorting[91-94]. Like any engineered system, microfluidic components have distinct regimes of optimal operation. Key factors that determine the

efficiency of these operations are the flow rates, oil volume fraction, and periodicity of droplets. For example, droplet formation typically requires a high fraction of oil, but incubation is most uniform when droplets are packed[95, 96]. Similarly, merger and picoinjection work best when droplets are periodic and can be synchronized, which requires close-packed emulsions[97]. Indeed, the packing of droplets and adjustment of the oil fraction is a common need when connecting microfluidic components together.

The simplest way to pack droplets is to collect the emulsion into an off-chip reservoir and allow them to “cream” due to their buoyancy. The packed droplets can then be reinjected into a second device to perform an additional operation, such as merging or sorting. While simple, off-chip collection has drawbacks. It is only applicable when the incubation between operations is long enough for emulsion transfer and requires a skilled user. Even then, it is error-prone, with droplets often coalescing due to dust, static charge, and flow through syringes, needles, and tubing. Indeed, even for skilled users, reinjection is unreproducible and the emulsions usually contain merged droplets, which can interfere with device operation and reduce data quality. A superior alternative would be to extract the oil on-chip to avoid off-chip handling. However, current methods are unable to extract the majority of oil from an emulsion and close-pack droplets; consequently, they are rarely used. A method to extract the majority of oil from an emulsion would make it easier to perform disparate microfluidic operations on a single chip.

In this paper, we describe a method to efficiently remove oil from an emulsion using an on-chip microfluidic extractor. This allows close-packing of initially dilute emulsion, making droplet flows periodic. We use the extractor to synchronize initially aperiodic droplet streams with periodic ones to perform pairwise merger. Our oil extractor is a universal adaptor for connecting microfluidic components that do not operate under identical flow and volume fraction conditions.

Materials and Methods

Device fabrication. The microfluidic device is fabricated using soft lithography[98] on a 3-inch silicon wafer (University Wafers). To facilitate accurate alignment of 5-μm-tall connecting channels to the rest of the layer's structures, the first mask only contains alignment marks. The multilayer master mold is fabricated using four photomasks as follows: (a) 25-μm-tall alignment marks are spin coated using SU-8 3025 photoresist (MicroChem), exposed and developed; (b) 5-μm-tall connecting channels (SU-8 3005) are spin coated, aligned and exposed; (c) 40-μm-tall drop making channels (SU-8 3025) are spin coated, aligned and exposed; (d) 90-μm-tall oil extracting channels and the remainder of the device including a large drop maker (SU-8 3025) are spin coated, aligned and exposed. The layers (b)–(d) are then developed together. Poly(dimethylsiloxane) (PDMS) (Momentive, RTV 615) is mixed at 10:1 ratio, degassed and poured onto the master in a petri dish. The PDMS is cured at 65°C for 2 hours and cut out using a scalpel. Inlet and outlet holes are punched with a 0.75-mm biopsy core (Harris, Uni-Core 0.75) to fit tightly polyethylene tubing (Scientific Commodities Inc, PE/2, ID 0.38 mm, OD 1.09 mm). The punched PDMS channel slab is bonded to a glass slide by activating with oxygen plasma for 60 s at 1 mbar in a plasma cleaner (Harrick Plasma, PDC-001) and baked at 65°C for 1 hour to complete bonding. The inner surface of the microchannels is treated with Aquapel to render it hydrophobic for stable droplet generation and flow.

Device operation. For the aqueous phase, PBS (pH 7.4) solution is loaded into plastic syringes (BD Luer-Lok syringe with 27G ½ needle) and connected to the inlets with PE/2 tubing. For the oil phase, hydrofluoroether (HFE; 3M Novec 7500) containing 2% (w/w) nonionic fluorosurfactant (RAN Biotechnologies, 008-Fluoro-Surfactant) is loaded into the same type of syringes. Syringe pumps (New Era Pump Systems, NE-501) are used to inject fluids at controlled flow rates. For the experiments shown in Fig. 2, flow rates are 100 μL/hr for aqueous phase and 400 μL/hr for oil. The oil extraction is controlled by setting the outlet tube (open to atmosphere) at a fixed height with respect to the microdevice. For the experiment in Fig. 4, flow rates are 80 and 250 μL/hr for aqueous and oil phases, respectively, for making the small droplets; 400 and 800 μL/hr for aqueous and oil phases, respectively, for making the large droplets; oil is extracted using a syringe pump at –220 μL/hr operating in withdrawing mode. Droplet formation is imaged on an inverted microscope using a fast-shutter camera (Unibrain, Fire-i 530b). Images are analyzed in LabVIEW and ImageJ using custom scripts to extract droplet

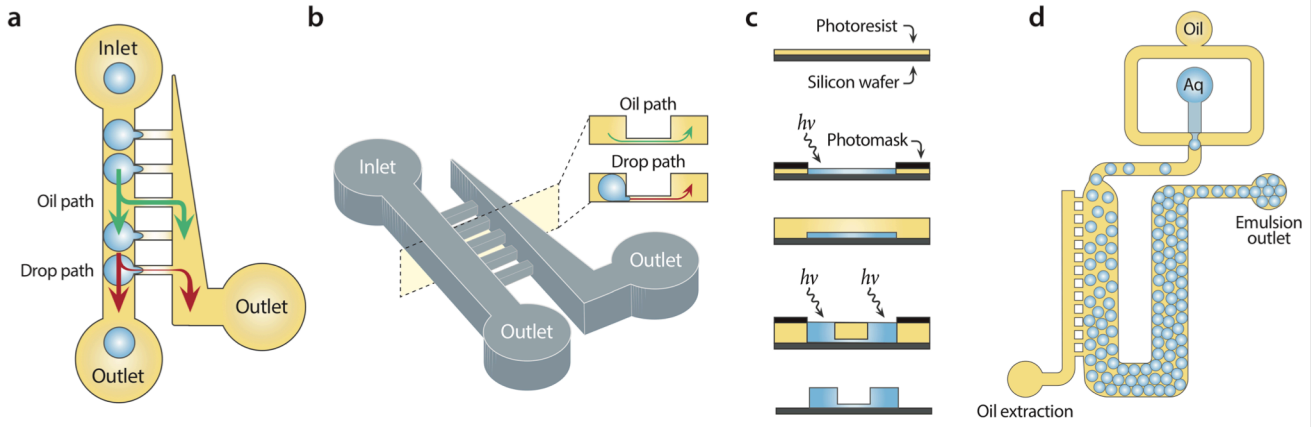


Fig. 9 Overview of oil extractor concept and design. The oil extractor consists of main and extraction channels connected by thin drainage channels; negative pressure is applied to the extractor outlet, drawing off oil but maintaining the droplets in the main channel due to their inability to deform through the connecting channels, (a). To extract a large fraction of oil while retaining droplets, the connecting channels are narrow and short, (b). The device thus requires two channel heights, which is produced using multi-layer fabrication, (c). By extracting the majority of oil, a dilute emulsion can be packed, (d).

positions and pairing ratios. For the merger experiments, high-speed imaging is used (Vision Research, Miro M110) to quantify the number of droplets merging.

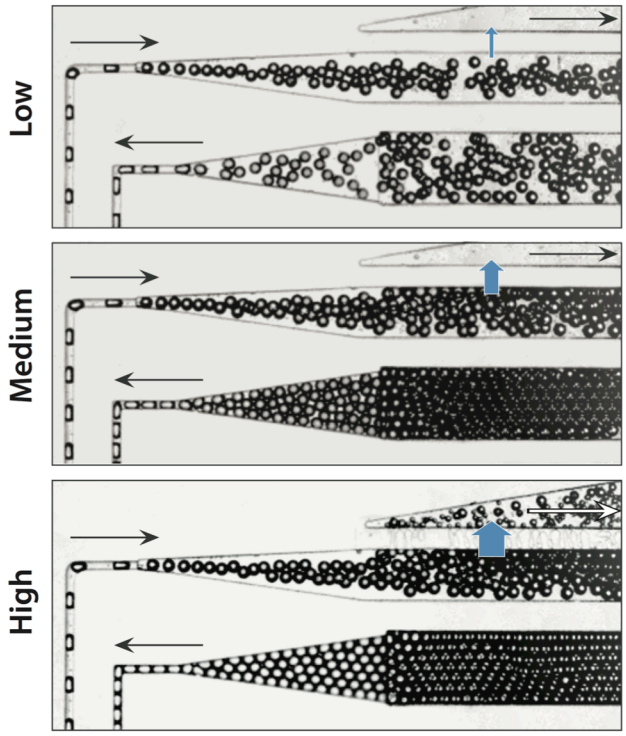


Fig. 10 The oil extractor can remove controlled volumes of oil from an emulsion. To control the amount of oil removed, a syringe pump withdrawals a controlled flow rate of oil from the extraction channel. For low draw rates, the droplets at the outlet are still unpacked, but for moderate and high draw rates, the droplets pack and order due to their monodispersity. High packing gives rise to plug flow, in which droplets travel through the delay line at equal speed.

Results

The concept of on-chip oil extraction is to remove a majority of oil from an emulsion while maintaining the droplets inside the channel. A straightforward way to do this is to draw off a controlled portion of oil from the emulsion using narrow channels perpendicular to the main channel. This is possible because for a large droplet to flow through a narrow channel, it must deform. However, deformation increases the Laplace pressure of the droplet, generating a force that opposes entrance into the narrow channel (Fig. 9a)[99]. This can be understood via the Laplace law,

$$\Delta P = \gamma(1/h + 1/w),$$

where ΔP is the pressure difference across the droplet interface, γ the interfacial tension, h the height, and w the width. Changing the width and height of a droplet by flowing it into a narrow channel thus increases the pressure in the droplet, allowing it to better resist entrance into the channel. The first oil extractors used channels with height equal

to the main channel but narrower width[96]. While these devices removed some oil, they could not remove the majority, because to do so requires extracting oil at higher flow rates, but this also extracts droplets. A simple solution would be to increase the Laplace stabilizing force using narrower extraction channels; however, this is difficult with described techniques due to the challenge of fabricating high aspect-ratio channels. Our solution is to reduce the heights and widths of the drainage channels (Fig. 9b), which allows a significant increase to the Laplace stabilizing force: While the minimum width of a channel is limited by the resolution of lithographic fabrication, height is controlled by spin coating (Fig. 9c), and can be made reliably below 5 μm; this provides >10X the Laplace stabilizing force and allows extraction of most of the oil from an emulsion (Fig. 9d).

A unique and valuable property of our oil extractor is that the amount of extraction is adjustable using a syringe pump to draw off oil to the desired fraction. To illustrate this, we form dilute emulsions and extract varying amounts of the oil (Fig. 10). At low extraction rates, little oil is removed and the droplets remain unpacked (Fig. 10, top). At moderate flow rates, a majority of oil is removed and droplets pack, (Fig 10, middle). At even higher flow rates, more oil is removed and droplets pack tightly; however, at these rates, pieces are also torn from the droplets (Fig. 10, bottom). This can be mitigated by fabricating even shorter extraction channels, although their hydrodynamic resistances must be carefully controlled to ensure the needed

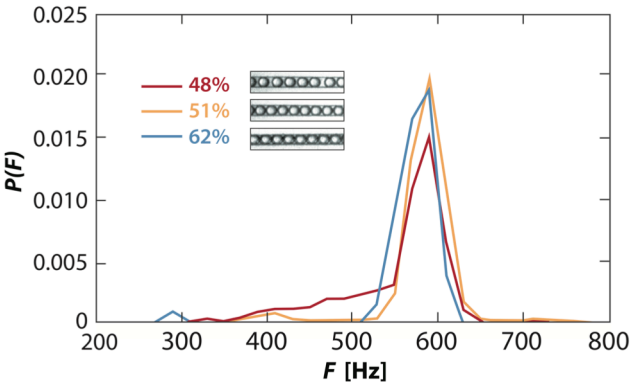


Fig. 11 Packed droplets flow periodically through channels. As the droplets become more packed, they order due to their monodispersity, yielding periodic flow and a narrow distribution of frequencies. The major frequency of ~570 Hz corresponds to two droplets touching while moving at the constant flow velocity.

extraction rate with the available pressure drop through the extractor. Interestingly, for high extractions, we find that droplets adjacent to the oil extractor tend to coalesce. This may be due to shear-induced coalescence and could be a major source of unintended merger during droplet reinjection from off-chip reservoirs that is difficult to see due to the inability to image within syringes, needles, and tubing.

Most droplet microfluidic devices are designed assuming periodic flow. This is essential for synchronizing streams for pairwise merger[92], or generating multiple emulsions with controlled numbers of cores and shells[100]. The ability to extract a large fraction of oil from an emulsion is valuable because it allows initially aperiodic streams to be made periodic. To illustrate this, we measure the periodicity of droplets flowing through our device for varying degrees of extraction (Fig.11). When we remove some oil (48% remaining aqeuous), we observe a broad distribution of droplet frequencies. Many droplets are emitted at 520-620 Hz, corresponding to two touching droplets moving at constant velocity, but also observe a sizable fraction of low frequency events, corresponding to droplets spaced by random volumes of oil; these droplets lead to aperiodicity in the flow. As we extract more oil, the drops pack (51%) and the tail nearly vanishes, indicating good periodicity. As we increase extraction further (62%), we maintain good periodicity and observe even fewer low-frequency events.

The ability to pack droplets by extracting oil allows us to transform an aperiodic flow into a periodic one. This is valuable when droplets must be synchronized on a microfluidic device. To illustrate this, we synchronize the flow and merger of two droplet streams, a first made upstream on the device at low volume fraction, packed by oil extraction, incubated for ~30 s, and paired with a second stream formed by another droplet maker. We adjust the frequency of the second droplet maker to achieve near-synchronization of the streams, and flow the pairs into a merger junction, where the droplets are coalesced via an electric field applied by salt-water electrodes (Fig.

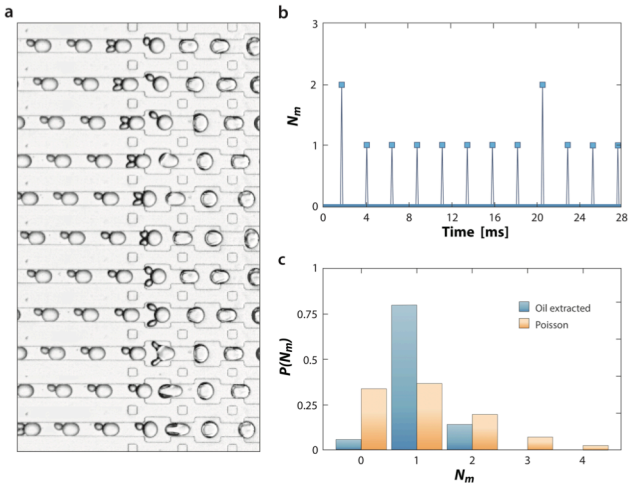


Fig. 12 Droplet periodicity allows precision synchronization of streams for efficient pairwise merger. Packed, smaller droplets are synchronized with generated, larger droplets by adjusting flow rates on a merger device, and the pairs merged via electrocoalescence with salt-water electrodes, (a). Synchronization requires that packed droplets be periodic and combined with the made droplets t equal frequency and phase, but small discrepancies can lead to "beat" patterns in which most events are pairwise mergers but some are three-way, (b). Nevertheless, by making the incubated droplets periodic, pairwise merger is achieved much more often than with random injection, given by a Poisson distribution, (c). N_m is the number of smaller droplets merged with the incoming, larger droplets; that is, N_m = 1 for pairwise merging. Blue bars show the distribution of pairing ratios with oil extraction obtained by analyzing 540 droplet merger events. Orange bars are the Poisson distribution with λ = 1.08.

12a)[94]. The droplets are periodic, although the streams are not perfectly synchronized and, in particular, the incubated droplets enter at a slightly faster rate than the made droplets, resulting in ~80% one-to-one fusions and ~14% two-to-one (Fig. 12b). Nevertheless, this is a major improvement over merger of unpacked droplets which enter at roughly random intervals and thus yield only ~37% one-to-one fusions, in accordance with Poisson statistics (Fig. 12c). This boost in pairwise merger is important because unmerged droplets waste regents and multiple mergers combine reactions, which can confound the results of an experiment. The ability to reliably synchronize droplet streams makes merging efficient and improves data quality.

Conclusions

We have presented a device to efficiently extract oil from an emulsion and pack droplets together. This allows oil volume fraction to be adjusted between steps in a workflow and aperiodic streams to be synchronized with other operations, such as merging, sorting, and double emulsion encapsulation. The ability to pack droplets yields plug-flow, in which all droplets move at identical speed, which is useful for incubating droplets for

controlled times, such as to allow a cell to secrete a molecule or an enzyme to catalyze a reaction. The oil extractor affords a universal adaptor for connecting microfluidic components that do not operate under identical conditions, and should thus enhance the reliability of multi-component devices. It should be valuable for applications requiring controlled delays, efficient mergers, or the generation of multiple emulsions with thin-shells[101].

Biblography

1. Hooke, R., *Micrographia.* 1665.
2. Schleiden, M.J., *Bertrage zur Phylogenesis.* Archiv fur Anatomie, Physiologie und wissenschaftliche Medizin, 1838. **Leipzig**: p. 137-176.
3. Schwann, T.H., *Microscopial researches into the accordance in the structure and growth of animals and plants. 1847.* Obes Res, 1993. **1**(5): p. 408-18.
4. Agresti, J.J., et al., *Ultrahigh-throughput screening in drop-based microfluidics for directed evolution.* Proceedings of the National Academy of Sciences, 2010. **107**(9): p. 4004-4009.
5. Romero, P.A., T.M. Tran, and A.R. Abate, *Dissecting enzyme function with microfluidic-based deep mutational scanning.* Proc Natl Acad Sci U S A, 2015. **112**(23): p. 7159-64.
6. Song, H. and R.F. Ismagilov, *Millisecond kinetics on a microfluidic chip using nanoliters of reagents.* J Am Chem Soc, 2003. **125**(47): p. 14613-9.
7. Thorsen, T., et al., *Dynamic pattern formation in a vesicle-generating microfluidic device.* Phys Rev Lett, 2001. **86**(18): p. 4163-6.
8. Hatch, A.C., et al., *1-Million droplet array with wide-field fluorescence imaging for digital PCR.* Lab Chip, 2011. **11**(22): p. 3838-45.
9. Eastburn, D.J., A. Sciambi, and A.R. Abate, *Picoinjection enables digital detection of RNA with droplet rt-PCR.* PLoS One, 2013. **8**(4): p. e62961.
10. Duffy, D.C., et al., *Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane).* Anal Chem, 1998. **70**(23): p. 4974-84.

11. P. B. Umbanhowar , V.P., and D. A. Weitz *Monodisperse Emulsion Generation via Drop Break Off in a Coflowing Stream.* Langmuir,, 2000. **16** (2): p. 347–351.

12. Holtze, C., et al., *Biocompatible surfactants for water-in-fluorocarbon emulsions.* Lab Chip, 2008. **8**(10): p. 1632-9.

13. Dimitrov, D.S., *Therapeutic antibodies, vaccines and antibodyomes.* MAbs, 2010. **2**(3): p. 347-56.

14. Haynes, B.F., et al., *B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study.* Nat Biotechnol, 2012. **30**(5): p. 423-33.

15. DeKosky, B.J., et al., *In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire.* Nat Med, 2015. **21**(1): p. 86-91.

16. Glanville, J., et al., *Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire.* Proc Natl Acad Sci U S A, 2009. **106**(48): p. 20216-21.

17. Georgiou, G., et al., *The promise and challenge of high-throughput sequencing of the antibody repertoire.* Nat Biotechnol, 2014. **32**(2): p. 158-68.

18. Weinstein, J.A., et al., *High-throughput sequencing of the zebrafish antibody repertoire.* Science, 2009. **324**(5928): p. 807-10.

19. Rubelt, F., et al., *Onset of immune senescence defined by unbiased pyrosequencing of human immunoglobulin mRNA repertoires.* PLoS One, 2012. **7**(11): p. e49774.

20. Shiroguchi, K., et al., *Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes.* Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1347-52.

21. Islam, S., et al., *Quantitative single-cell RNA-seq with unique molecular identifiers.* Nat Methods, 2014. **11**(2): p. 163-6.

22. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.

23. Shalek, A.K., et al., *Single-cell RNA-seq reveals dynamic paracrine control of cellular variation.* Nature, 2014. **510**(7505): p. 363-9.

24. Jaitin, D.A., et al., *Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types.* Science, 2014. **343**(6172): p. 776-9.

25. Fan, H.C., G.K. Fu, and S.P. Fodor, *Expression profiling. Combinatorial labeling of single cells for gene expression cytometry.* Science, 2015. **347**(6222): p. 1258367.

26. Macosko, E.Z., et al., *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 2015. **161**(5): p. 1202-14.

27. Klein, A.M., et al., *Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.* Cell, 2015. **161**(5): p. 1187-201.

28. Agresti, J.J., et al., *Ultrahigh-throughput screening in drop-based microfluidics for directed evolution.* Proc Natl Acad Sci U S A, 2010. **107**(9): p. 4004-9.

29. Eastburn, D.J., A. Sciambi, and A.R. Abate, *Identification and genetic analysis of cancer cells with PCR-activated cell sorting.* Nucleic Acids Res, 2014. **42**(16): p. e128.

30. Eastburn, D.J., A. Sciambi, and A.R. Abate, *Ultrahigh-throughput Mammalian single-cell reverse-transcriptase polymerase chain reaction in microfluidic drops.* Anal Chem, 2013. **85**(16): p. 8016-21.

31. Mazutis, L., et al., *Single-cell analysis and sorting using droplet-based microfluidics.* Nat Protoc, 2013. **8**(5): p. 870-91.

32. Tice, J.D., et al., *Formation of Droplets and Mixing in Multiphase Microfluidics at Low Values of the Reynolds and the Capillary Numbers* Langmuir, 2003. **19**(22): p. 9127-9133.

33. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

34. Christopher, G.F., et al., *Coalescence and splitting of confined droplets at microfluidic junctions.* Lab Chip, 2009. **9**(8): p. 1102-9.

35. Tan, Y., *Controlled fission of droplet emulsion in bifurcating microfluidic channels* TRANSDUCERS, Solid-State Sensors, Actuators and Microsystems, 12th International Conference on

, 2003. **1**(1): p. 28-31.

36. DeMello, A.J., *Control and detection of chemical reactions in microfluidic systems.* Nature, 2006. **442**(7101): p. 394-402.

37. Stroock, A.D., et al., *Chaotic mixer for microchannels.* Science, 2002. **295**(5555): p. 647-51.

38. Sale, J.E. and M.S. Neuberger, *TdT-accessible breaks are scattered over the immunoglobulin V domain in a constitutively hypermutating B cell line.* Immunity, 1998. **9**(6): p. 859-69.

39. Davies, D.R., E.A. Padlan, and S. Sheriff, *Antibody-antigen complexes.* Annu Rev Biochem, 1990. **59**: p. 439-73.

40. Xiao, Z., et al., *Known components of the immunoglobulin A:T mutational machinery are intact in Burkitt lymphoma cell lines with G:C bias.* Mol Immunol, 2007. **44**(10): p. 2659-66.

41. Harris, R.S., et al., *Epstein-Barr virus and the somatic hypermutation of immunoglobulin genes in Burkitt's lymphoma cells.* J Virol, 2001. **75**(21): p. 10488-92.

42. Regev, A. and E. Shapiro, *Cellular abstractions: Cells as computation.* Nature, 2002. **419**(6905): p. 343-343.

43. Appling, D.R., *Genetic approaches to the study of protein–protein interactions.* Methods, 1999. **19**(2): p. 338-349.

44. Bandyopadhyay, S., et al., *Functional maps of protein complexes from quantitative genetic interaction data.* PLoS Comput Biol, 2008. **4**(4): p. e1000065.

45. Baryshnikova, A., et al., *Genetic interaction networks: toward an understanding of heritability.* Annual review of genomics and human genetics, 2013. **14**: p. 111-133.

46. Benfey, P.N. and T. Mitchell-Olds, *From genotype to phenotype: systems biology meets natural variation.* Science, 2008. **320**(5875): p. 495-497.

47. Cusick, M.E., et al., *Interactome: gateway into systems biology.* Human molecular genetics, 2005. **14**(suppl 2): p. R171-R181.

48. Nichols, R.J., et al., *Phenotypic landscape of a bacterial cell.* Cell, 2011. **144**(1): p. 143-156.

49. Ashworth, A., C.J. Lord, and J.S. Reis-Filho, *Genetic interactions in cancer progression and treatment.* Cell, 2011. **145**(1): p. 30-38.

50. Babu, M., et al., *Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in Escherichia coli.* PLoS Genet, 2014. **10**(2): p. e1004120.

51. Peters, J.M., et al., *A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria.* Cell, 2016. **165**(6): p. 1493-506.

52. Li, P., et al., *An overview of SNP interactions in genome-wide association studies.* Brief Funct Genomics, 2015. **14**(2): p. 143-55.

53. Collins, S.R., et al., *Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map.* Nature, 2007. **446**(7137): p. 806-810.

54. Roguev, A., et al., *High-throughput genetic interaction mapping in the fission yeast Schizosaccharomyces pombe.* Nature methods, 2007. **4**(10): p. 861-866.

55. Tong, A.H., et al., *A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules.* Science, 2002. **295**(5553): p. 321-4.

56. Cheverud, J.M. and E.J. Routman, *Epistasis and its contribution to genetic variance components.* Genetics, 1995. **139**(3): p. 1455-1461.

57. Tong, A.H., et al., *Global mapping of the yeast genetic interaction network.* Science, 2004. **303**(5659): p. 808-13.

58. Collins, S.R., et al., *Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map.* Nature, 2007. **446**(7137): p. 806-10.

59. Costanzo, M., et al., *The genetic landscape of a cell.* Science, 2010. **327**(5964): p. 425-31.

60. Butland, G., et al., *eSGA: E. coli synthetic genetic array analysis.* Nat Methods, 2008. **5**(9): p. 789-95.

61. Kumar, A., et al., *Conditional Epistatic Interaction Maps Reveal Global Functional Rewiring of Genome Integrity Pathways in Escherichia coli.* Cell Rep, 2016. **14**(3): p. 648-61.

62. Typas, A., et al., *High-throughput, quantitative analyses of genetic interactions in E. coli.* Nat Methods, 2008. **5**(9): p. 781-7.

63. Baba, T., et al., *Construction of Escherichia coli K‐12 in‐frame, single‐gene knockout mutants: the Keio collection.* Molecular systems biology, 2006. **2**(1).

64. Giaever, G. and C. Nislow, *The yeast deletion collection: a decade of functional genomics.* Genetics, 2014. **197**(2): p. 451-465.

65. Segre, D., et al., *Modular epistasis in yeast metabolism.* Nature genetics, 2005. **37**(1): p. 77-83.

66. Zeitoun, R.I., et al., *Multiplexed tracking of combinatorial genomic mutations in engineered cell populations.* Nature biotechnology, 2015.

67. van Opijnen, T., K.L. Bodi, and A. Camilli, *Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.* Nature methods, 2009. **6**(10): p. 767-772.

68. van Opijnen, T. and A. Camilli, *Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms.* Nature Reviews Microbiology, 2013. **11**(7): p. 435-442.

69. Warner, J.R., et al., *Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides.* Nat Biotechnol, 2010. **28**(8): p. 856-62.

70. Pritchard, J.R., et al., *ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing.* 2014.

71. Mordukhova, E.A. and J.G. Pan, *Stabilization of homoserine-O-succinyltransferase (MetA) decreases the frequency of persisters in Escherichia coli under stressful conditions.* PLoS One, 2014. **9**(10): p. e110504.

72. Kitagawa, M., et al., *Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research.* DNA Res, 2005. **12**(5): p. 291-9.

73. Lee, T.S., et al., *BglBrick vectors and datasheets: A synthetic biology platform for gene expression.* J Biol Eng, 2011. **5**: p. 12.

74. van Opijnen, T., K.L. Bodi, and A. Camilli, *Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.* Nat Methods, 2009. **6**(10): p. 767-72.

75. Stone, H.A., A.D. Stroock, and A. Ajdari, *Engineering flows in small devices: microfluidics toward a lab-on-a-chip.* Annu. Rev. Fluid Mech., 2004. **36**: p. 381-411.

76. Squires, T.M. and S.R. Quake, *Microfluidics: Fluid physics at the nanoliter scale.* Reviews of modern physics, 2005. **77**(3): p. 977.

77. Whitesides, G.M., *The origins and the future of microfluidics.* Nature, 2006. **442**(7101): p. 368-373.

78. Teh, S.-Y., et al., *Droplet microfluidics.* Lab on a Chip, 2008. **8**(2): p. 198-220.

79. Guo, M.T., et al., *Droplet microfluidics for high-throughput biological assays.* Lab on a Chip, 2012. **12**(12): p. 2146-2155.

80. Tran, T., et al., *From tubes to drops: droplet-based microfluidics for ultrahigh-throughput biology.* Journal of Physics D: Applied Physics, 2013. **46**(11): p. 114004.

81. Kim, S.H., et al., *Large-scale femtoliter droplet array for digital counting of single biomolecules.* Lab on a Chip, 2012. **12**(23): p. 4986-4991.

82. Shim, J.-u., et al., *Ultrarapid generation of femtoliter microfluidic droplets for single-molecule-counting immunoassays.* Acs Nano, 2013. **7**(7): p. 5955-5964.

83. Hindson, B.J., et al., *High-throughput droplet digital PCR system for absolute quantitation of DNA copy number.* Analytical chemistry, 2011. **83**(22): p. 8604-8610.

84. Hindson, C.M., et al., *Absolute quantification by droplet digital PCR versus analog real-time PCR.* Nature methods, 2013. **10**(10): p. 1003-1005.

85. Kintses, B., et al., *Picoliter cell lysate assays in microfluidic droplet compartments for directed enzyme evolution.* Chemistry & biology, 2012. **19**(8): p. 1001-1009.

86. Romero, P.A., T.M. Tran, and A.R. Abate, *Dissecting enzyme function with microfluidic-based deep mutational scanning.* Proceedings of the National Academy of Sciences, 2015. **112**(23): p. 7159-7164.

87. Eastburn, D.J., A. Sciambi, and A.R. Abate, *Identification and genetic analysis of cancer cells with PCR-activated cell sorting.* Nucleic acids research, 2014: p. gku606.

88. Lim, S.W., T.M. Tran, and A.R. Abate, *PCR-activated cell sorting for cultivation-free enrichment and sequencing of rare microbes.* PloS one, 2015. **10**(1): p. e0113549.

89. Lim, S.W., et al., *PCR-activated cell sorting as a general, cultivation-free method for high-throughput identification and enrichment of virus hosts.* Journal of Virological Methods, 2016.

90. Lance, S.T., et al., *Peering below the diffraction limit: robust and specific sorting of viruses with flow cytometry.* Virology Journal, 2016. **13**(1): p. 201.

91. Link, D., et al., *Geometrically mediated breakup of drops in microfluidic devices.* Physical review letters, 2004. **92**(5): p. 054503.

92. Ahn, K., et al., *Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels.* Applied Physics Letters, 2006. **88**(26): p. 264105.

93. Baret, J.-C., et al., *Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity.* Lab on a Chip, 2009. **9**(13): p. 1850-1858.

94. Sciambi, A. and A.R. Abate, *Generating electric fields in PDMS microfluidic devices with salt water electrodes.* Lab on a Chip, 2014. **14**(15): p. 2605-2609.

95. Mary, P., et al., *Controlling droplet incubation using close-packed plug flow.* Biomicrofluidics, 2011. **5**(2): p. 024101.

96. Frenz, L., et al., *Reliable microfluidic on-chip incubation of droplets in delay-lines.* Lab on a Chip, 2009. **9**(10): p. 1344-1348.

97. Abate, A.R., et al., *High-throughput injection with microfluidics using picoinjectors.* Proceedings of the National Academy of Sciences, 2010. **107**(45): p. 19163-19166.

98. Xia, Y. and G.M. Whitesides, *Soft lithography.* Annual review of materials science, 1998. **28**(1): p. 153-184.

99. Dangla, R., S.C. Kayi, and C.N. Baroud, *Droplet microfluidics driven by gradients of confinement.* Proceedings of the National Academy of Sciences, 2013. **110**(3): p. 853-858.

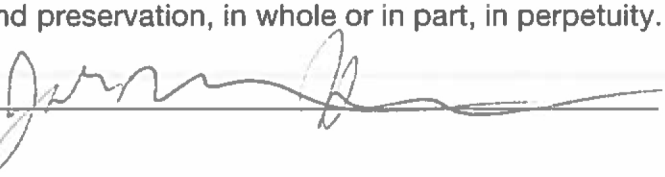
100. Abate, A. and D. Weitz, *High‐order multiple emulsions formed in poly (dimethylsiloxane) microfluidics.* Small, 2009. **5**(18): p. 2030-2032.

101. Kim, S.-H., et al., *Double-emulsion drops with ultra-thin shells for capsule templates.* Lab on a Chip, 2011. **11**(18): p. 3162-3166.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature  Date

03/28/17