Title
Massively Multiplexed DNA Sequencing with Ultrahigh-throughput Droplet Microfluidics

Author
Lan, Freeman

Publication Date
2016

Peer reviewed|Thesis/dissertation

Massively Multiplexed DNA Sequencing with Ultrahigh-throughput Droplet Microfluidics.

by

Freeman Lan

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

AND

UNIVERSITY OF CALIFORNIA, BERKELEY

[redacted]

Dedication

To my family who always supported me…

To my teachers and mentors who showed me the path…

“Always a student”

ACKNOWLEDGEMENTS

Elements of this dissertation has been published elsewhere or is in preparation for publication in a peer reviewed journal. Elements of Chapter 2 and Chapter 3 are published in *Nature Communications* under the title “Droplet barcoding for massively parallel single molecule sequencing”. Chapter 4 has been submitted for publication in a peer-reviewed journal.

The road to my dissertation began long before my time at UC Berkeley and UCSF. As a second year undergraduate at the University of Toronto, I was lucky enough to be taken under the wing of Professor Zhong-ping Feng in the Physiology department. She took a chance on me, as an engineering student, to learn the ways of a scientist. During my time as an undergraduate in her lab, she imparted in me the inklings of scientific thought as well as the skills of scientific reading and writing. I must also thank my other undergraduate mentors Professor John Glover and Professor Scott Gray-Owen, as well as the Biochemistry Department at the University of Toronto for the excellent scientific training that I received during my undergraduate education.

At UCSF and UC Berkeley, this dissertation would not have been possible without the help and support of many who have been part of my life during these 5 years. First and foremost, I would like to thank my dissertation advisor Professor Adam Abate for his enduring encouragement, support, and positivity. Adam, your quick thinking, limitless creativity, and can-do attitude is the greatest inspiration to me, and something I’ve strived to emulate over these years. I thank my thesis and qualification examination committee Professor Ryan Hernandez, Professor Raul Andino, Professor Laurent Coscoy, Professor Patricia Babbitt, Professor John Dueber, and Professor

Dorian Liepmann for their support, care, and advice. I would also like to thank Professor Joe DeRisi, Professor Kenneth Stedman, and Professor Katherine Pollard for taking the time to advise me despite not being a part of my committee. Professors, it is my utmost privilege to interact with you, and every interaction I have had with every one of you has given me inspiration to become a better scientist and person.

I thank the members of the Abate lab for their companionship during the last five years. I especially thank the first batch of graduate students Tuan Tran, Shea Thompson, John Haliburton, and Shaun Lim strived with me through the thick and thins of graduate school. Without you, this journey would be lonely. I thank the Postdocs of the Abate lab, past and present, especially Philip Romero, Maen Sarhan, Leqian Liu, and Sean Poust, for the excellent scientific discussions that contributed much to my learning.

A very special thanks to the 4th floor Genentech crew Ben Demaree, Angus Sidore, Aaron Yuan, Noorsher Ahmed, and Meiyi Tee for your amazing support in our projects together. Your talents and go-getter attitudes was what enabled us to tackle the toughest projects. You made the last year of my PhD an absolute pleasure.

Lastly, and most importantly, I thank my mother and father Jennifer and Chris for their unwavering and unconditional support of all my decisions. You are the giants' shoulders on which I stand, and with your support I will continue my journey of learning...

ABSTRACT

The ability to conduct and read-out assays in high-throughput is a powerful tool for studying complex biological systems. For example, fluorescence activated cell sorting, which enables high-throughput assays of cellular markers on single cells by fluorescent staining has been instrumental to understanding the immune system. Although fluorescence can be a powerful readout for high-throughput assays, it also has significant drawbacks. First, a fluorescent assay must be available for the target of interest, and second, only a few different assays can be used at once, owing to the limited spectrum available to fluorophores and detectors. In my dissertation, I develop a novel method of high-throughput assay readout using massively parallel DNA sequencing and droplet microfluidics. By conducting assays inside picoliter sized droplets generated using microfluidic channels, molecular biology assays that generate nucleic acids as a readout can be performed at kHz throughput. By uniquely barcoding the nucleic acids in each droplet, the results of each individual assay can be read in parallel using in a massively parallel sequencer. With this approach, I develop a method of deep sequencing single molecules and a method of sequencing single cell genomes at low-coverage, to generate highly accurate and haplotyped long DNA sequence reads and characterize diverse microbial populations in an unprecedented manner, respectively.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Background and Introduction

1.1 High-throughput assays in biological research

The process of evolution has gifted this world a fascinating array of complex biological phenomena. Evolution is random, therefore the biological phenomena it produces are unpredictable (Bonner, 1988). In the face unpredictability, logic and theory in absence of data can only take us so far; the scientific investigation of biology then, is first and foremost driven by experimental evidence. If obtaining experimental evidence is the main work of researchers of biology, then rapidly iterating through experimental assays (high-throughput assaying) would significantly improve his/her effectiveness. Indeed high-throughput assays is a cornerstone of biological research and has enabled important discoveries, such as the discovery of novel drug candidates(Feng et al., 2007), novel cell types in the human body(Mosmann and Sad, 1996), novel useful enzymes (Olsen et al., 2000), and molecular structures of proteins (Stevens, 2000).

The defining characteristic of high-throughput assays is the confinement of an assay in a way that a multitude of similar assays can be performed rapidly in a systematic manner. For example, in high-throughput drug screening (figure 1.1), cells are confined into wells on a plate where pipetting robots are used to add drug candidates to each well; each well is then assayed for response to these drugs (Burns et al., 2006). For another example, in functional metagenomics screening, potential enzyme coding DNA segments are randomly inserted into a large population of cells. Each cell is then screened for enzyme activity by searching for colorimetric response in colonies on a plate (figure 1.2) (Uchiyama and Miyazaki, 2009) or for fluorometric response of single

cells in a cell sorter (Yun and Ryu, 2005). Regardless of the exact method, all high-throughput assays involve confinement of an assay to a small volume and a method to rapidly interrogate the confined assays.

Figure 1.1 An example a high-throughput assay. Pipetting robots are used to reproducibly mix reagents in well plates.

Figure 1.2 An example a high-throughput assay. Each bacterial cell colony on a plate of thousands of colonies can be used as containers for an individual assays.

1.2 Brief introduction to droplet microfluidics

Microfluidics is a field of research characterized by controlling fluids at the microscale, where the fluid flow is usually in the low Reynald's number regime, hence all fluid flow is essentially laminar (Squires and Quake, 2005). Droplet microfluidics is a subfield of microfluidics that utilizes the generation of monodisperse droplets (figure 1.3) resulting from flowing of two immiscible phases at defined geometries (Thorsen et al., 2001). These droplets are stabilized by surfactants and can be used to conduct chemical and biological reactions in a tiny confined volume (Teh et al., 2008). Millions of droplets can be incubated in a test tube, acting as millions of individual compartments, or they can be manipulated using specially designed microfluidic devices. Microfluidic devices have been designed for adding reagents to (Abate et al., 2010a), splitting (Link et al., 2004), merging (Ahn et al., 2006), and sorting (Abate et al., 2010b) droplets. The combination of these functionalities enable many biochemical workflows to be conducted inside the droplets, make droplet microfluidics a potent platform for conducting high-throughput assays.

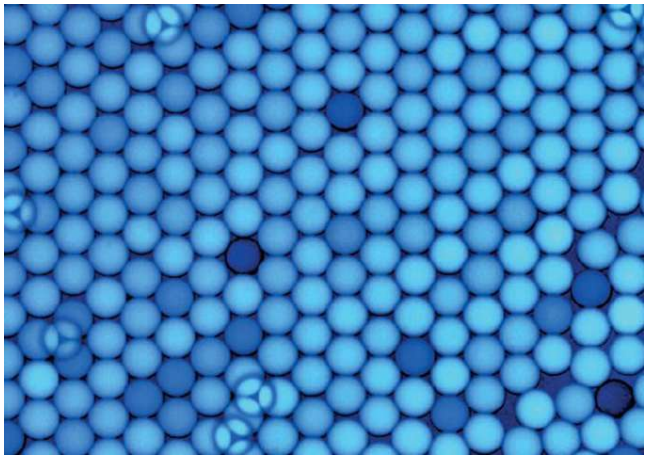

Figure 1.3. Monodisperse water-in-oil droplets generated using droplet microfluidics.

1.3 Droplet microfluidics as a platform for high-throughput biological assays

Droplet microfluidics, with its ability to generate millions of confined volumes for biochemical reactions lends itself naturally to conducting high-throughput assays in biology. Using just a droplet maker, droplets can be generated and used to screen large arrays of conditions for protein crystallography (Zheng et al., 2003). By adding a droplet sorter, which sort droplets based on fluorescence, single molecules, enzymes, and cells can be screened based on a fluorimetric assay (Agresti et al., 2010; Fallah-Araghi et al., 2012; Hindson et al., 2011). While these two methods constitute the major forms of high-throughput droplet assays demonstrated thus far, droplet microfluidics assays can also be combined with massively parallel DNA sequencing as a readout.

1.4 Brief introduction to massively parallel DNA sequencing

Massively parallel DNA sequencing, also called Next Generation Sequencing, is a recent development that has revolutionized biological research by enabling cost-effective sequencing of millions to billions of short DNA sequences in parallel (Metzker, 2010). To sequence DNA, it is first chopped up into small fragments <1kb long, then sequence specific adaptors are added to the ends of the fragments. These adaptor-added fragments are then loaded into a sequencer where all fragments are sequenced in parallel using a sequencing by synthesis chemistry(Shendure and Ji, 2008), resulting in millions to billions of short DNA sequences in a single run. Massively parallel sequencing has enabled rapid sequencing of whole genomes and metagenomes (Afshinnekoo et al., 2015; Smith et al., 2008). Furthermore, it is an excellent counting

tool for DNA and RNA, enabling whole transcriptome profiling (Mortazavi et al., 2008) and accurate enumeration of target sequences as the read out of an assay (Romero et al., 2015).

The key to the success of massively parallel sequencing is the ability to generate a massive amounts of data at once. The output capacity of a massively parallel sequencer is often so large that multiple samples and experiments can be sequenced in parallel in a single run. Here, the use of molecular indices or barcodes have been instrumental to the efficient use of sequencing capacity. Unique DNA sequence adaptors (barcodes) are added onto the DNA fragments to be sequenced so that the barcode sequence is read out along with the sequence of the fragment. Sequence fragments associated with the same barcode sequence are presumed to belong to the same sample (Smith et al., 2010). Hence, barcoding is the key to using massively parallel DNA sequencing as the read-out of high-throughput assays. For example, by barcoding the genomes of 100 cells for massively parallel sequencing, researchers were able to glimpse the clonal evolutionary signature of breast cancer tumor(Potter et al., 2013).

1.5 Massively parallel sequencing as read-out for droplet microfluidics assays

Although droplet microfluidics can rapidly generate millions of individual confined reactions, the practical throughput of droplet microfluidic assays is often much lower. This is because obtaining a read-out for each droplet is often more difficult than generating them in the first place. The advent of massively parallel sequencing opens

up new opportunities for high-throughput droplet microfluidics assays because using massively parallel sequencing, every droplet can be assayed in parallel, thus avoiding the bottleneck inherent in assaying each single droplet using conventional methods. However, to use massively parallel sequencing as a read out, the contents in each droplet must be uniquely barcoded prior to sequencing. However, the ability to barcode tens of thousands of samples for sequencing has not yet been demonstrated. In my dissertation, I develop a method to uniquely barcode hundreds of thousands of droplets for massively parallel sequencing, and I use this method to achieve ultrahigh-throughput sequencing of single molecules and single cells.

1.6 References

Abate, A.R., Hung, T., Mary, P., Agresti, J.J., and Weitz, D.A. (2010a). High-throughput injection with microfluidics using picoinjectors. Proc. Natl. Acad. Sci. *107*, 19163–19166.

Abate, A.R., Agresti, J.J., and Weitz, D.A. (2010b). Microfluidic sorting with high-speed single-layer membrane valves. Appl. Phys. Lett. *96*, 203509.

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S., et al. (2015). Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Syst. *1*, 72–87.

Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.-C., Marquez, M., Klibanov, A.M., Griffiths, A.D., and Weitz, D.A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proc. Natl. Acad. Sci. *107*, 4004–4009.

Ahn, K., Agresti, J., Chong, H., Marquez, M., and Weitz, D.A. (2006). Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. Appl. Phys. Lett. *88*, 264105.

Bonner, J.T. (1988). The evolution of complexity by means of natural selection (Princeton University Press).

Burns, A.R., Kwok, T.C.Y., Howard, A., Houston, E., Johanson, K., Chan, A., Cutler, S.R., McCourt, P., and Roy, P.J. (2006). High-throughput screening of small molecules

for bioactivity and target identification in Caenorhabditis elegans. Nat. Protoc. *1*, 1906–1914.

Fallah-Araghi, A., Baret, J.-C., Ryckelynck, M., and Griffiths, A.D. (2012). A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. Lab. Chip *12*, 882–891.

Feng, B.Y., Simeonov, A., Jadhav, A., Babaoglu, K., Inglese, J., Shoichet, B.K., and Austin, C.P. (2007). A high-throughput screen for aggregation-based inhibition in a large compound library. J. Med. Chem. *50*, 2385–2390.

Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., Legler, T.C., et al. (2011). High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Anal. Chem. *83*, 8604–8610.

Link, D.R., Anna, S.L., Weitz, D.A., and Stone, H.A. (2004). Geometrically Mediated Breakup of Drops in Microfluidic Devices. Phys. Rev. Lett. *92*, 054503.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Mosmann, T.R., and Sad, S. (1996). The expanding universe of T-cell subsets: Th1, Th2 and more. Immunol. Today *17*, 138–146.

Olsen, M., Iverson, B., and Georgiou, G. (2000). High-throughput screening of enzyme libraries. Curr. Opin. Biotechnol. *11*, 331–337.

Potter, N.E., Ermini, L., Papaemmanuil, E., Cazzaniga, G., Vijayaraghavan, G., Titley, I., Ford, A., Campbell, P., Kearney, L., and Greaves, M. (2013). Single-cell mutational profiling and clonal phylogeny in cancer. Genome Res. *23*, 2115–2125.

Romero, P.A., Tran, T.M., and Abate, A.R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. Proc. Natl. Acad. Sci. *112*, 7159–7164.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Smith, A.M., Heisler, L.E., St.Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G., Pourmand, N., et al. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. gkq368.

Smith, D.R., Quinlan, A.R., Peckham, H.E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W.F., Tusneem, N., Stromberg, M.P., et al. (2008). Rapid whole-genome

mutational profiling using next-generation sequencing technologies. Genome Res. *18*, 1638–1642.

Squires, T.M., and Quake, S.R. (2005). Microfluidics: Fluid physics at the nanoliter scale. Rev. Mod. Phys. *77*, 977–1026.

Stevens, R.C. (2000). High-throughput protein crystallization. Curr. Opin. Struct. Biol. *10*, 558–563.

Teh, S.-Y., Lin, R., Hung, L.-H., and Lee, A.P. (2008). Droplet microfluidics. Lab. Chip *8*, 198.

Thorsen, T., Roberts, R.W., Arnold, F.H., and Quake, S.R. (2001). Dynamic Pattern Formation in a Vesicle-Generating Microfluidic Device. Phys. Rev. Lett. *86*, 4163–4166.

Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. Curr. Opin. Biotechnol. *20*, 616–622.

Yun, J., and Ryu, S. (2005). Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. Microb. Cell Factories *4*, 8.

Zheng, B., Roach, L.S., and Ismagilov, R.F. (2003). Screening of protein crystallization conditions on a microfluidic chip using nanoliter-size droplets. J. Am. Chem. Soc. *125*, 11170–11171.

Chapter 2: Droplet barcode libraries for ultrahigh throughput barcoding

Molecular barcoding is the essential element of massively parallel sequencing that allows the sequencing of multiple samples in one run. However, to sequence tens of thousands of samples in parallel in a droplet microfluidic workflow requires at least tens of thousands of molecular barcodes that can be used to uniquely identify each sample. In this chapter, I develop a cost-effective and rapid method of generating millions of unique molecular barcodes capable of uniquely labelling millions of droplets in a droplet microfluidic workflow.

2.1. Motivation

Droplet microfluidics is a promising novel platform technology for conducting high-throughput assays. With its ability to generate millions of small compartments of biochemical reactions, it can be revolutionary for the throughput of high-throughput assays. Whereas conducting hundreds to thousands of assays would be considered a high-throughput by current standards, up to millions of assays at a time should be attainable using droplet microfluidics, in what I call “ultrahigh-throughput assays”. For example, using droplet microfluidics, drugs can be screened at more than 10,000 dosage conditions at a time, a two order of magnitude improvement over current standards (Miller et al., 2012). Despite its enormous potential, successful demonstrations of droplet microfluidic ultrahigh-throughput assays have been limited in scope. This is because while it is easy to generate millions of droplet assays, generating a readout from each droplet has been difficult. Generally, fluorometric readouts is the

method of choice to assay droplets, but fluorometric readouts are limited by the availability of a suitable fluorogenic assay and in cases requiring multiple fluorophores, the optical spectrum available to distinguish between multiple signals. A more powerful method of readout is massively parallel sequencing, because every droplet can be assayed in parallel and the number of assays is virtually unlimited. However, to sequence the contents of every droplet in parallel, its contents must be uniquely barcoded prior to sequencing. Therefore, a method of barcoding droplets at ultrahigh-throughput must be developed before massively parallel sequencing can be used as a readout for droplet microfluidic assays.

2.2 Barcoding in massively parallel sequencing

The use of molecular indices or barcodes is essential to sequencing multiple samples in one run, thereby fully utilizing the output capacity of a massively parallel sequencer. Barcodes are typically defined sequences of 8-10 bases which comprise part of the adaptors which are added to DNA fragments before sequencing. Commonly used methods to add barcoded adaptors are using PCR with primers containing 5' overhang (figure 2.1), or by ligation of double stranded barcoded adaptor sequences (figure 2.2). In both cases, high concentration and purity of barcoded adaptor DNA, which is obtained through chemical synthesis of the DNA oligomers, is required.

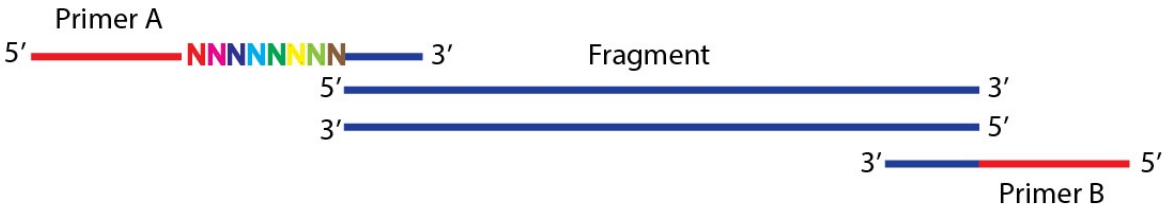

Figure 2.1 PCR method of attaching barcodes to fragments for sequencing. PCR primers containing barcode and adaptor sequence can be used to amplify the fragment in order to attach barcode sequences.

Figure 2.2 Ligation method of attaching barcodes to fragments for sequencing. Double stranded adaptor DNA containing barcode sequences can be ligated to fragments in order to attach barcodes.

2.3 Ultrahigh-throughput barcoding in droplets

In typical barcoding methodologies, chemical synthesis of barcoded adaptors is required for each unique barcode. This strategy is unpractical for barcoding at ultra-high throughput because the cost of chemically synthesizing millions of unique barcoded adaptor sequences individually is exceedingly high. An alternative approach is required to enable barcoding at ultrahigh-throughput. To this end, I develop a hybrid chemical synthesis and enzymatic amplification in droplets strategy that can generate millions of clonal pools of unique barcode sequences in what I call a droplet barcode library.

2.4 Generating a droplet barcode library

Barcode molecules consisting of 15 random N-mers flanked by constant sequences are chemically synthesized generating a single pool of single stranded DNA molecules with 4^{15} diversity. To generate a droplet barcode library, these molecules are individually encapsulated at a limiting dilution of ~1 in 10 droplets using a microfluidic droplet maker with PCR reagents for amplification (figure 2.3). Barcode molecules distribute among the droplets following a Poisson distribution: ~ 90% of droplets contain no barcodes, ~ 9% of droplets contain one unique barcode, and ~1% of droplets contain multiple barcodes. The droplets are thermally cycled, generating within each loaded droplet a clonal population of amplified product; these droplets can then be merged with target droplets to introduce unique barcodes. Using this approach, I generate ~10 million barcode droplets in < 1 hour for ~$10 of PCR reagent, which is sufficient to barcode ~1 million templates in the SMDB workflow.

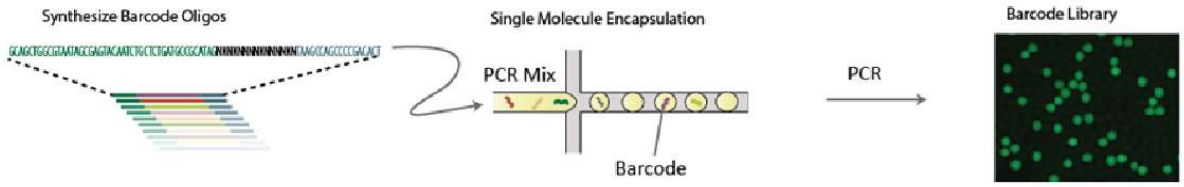

Figure 2.3 Schematic depiction of generating a digital droplet barcode library. Barcode droplets are stained with SYBR green for visualization.

2.5 Attaching barcodes to DNA fragments

To uniquely label DNA fragments, the barcode sequences must be attached to the fragments as a part of the sequencing adaptor. To accomplish this, I merge a droplet from the barcode droplet library with a droplet containing fragments to be barcoded using a droplet merger junction (Ahn et al., 2006). These fragments must contain universal sequence adaptors on either ends, which can be accomplished through adaptor ligation or transposase tagmentation (Picelli et al., 2014). Once barcodes and fragments are present in the same merged droplet, I attach barcodes to fragments using sequence overlap extension PCR, where sequence overlap regions between the barcodes and the constant sequence adaptor on the fragments are used to splice together the two molecules during PCR (Horton et al., 1989).

2.6 Sequence diversity of the droplet barcode library

Because barcode sequences are random, it is possible for two barcodes of the same sequence to be randomly sampled to label two different templates. The probability of this event happening is represented by equation 2.1 where n is the number of barcodes sampled from a sequence space of N possible sequences, and is diminishingly small for a sequence space that is sparsely sampled.

$$1 - \frac{N!}{(N^n)(N-n)!} \qquad \text{Equation 2.1}$$

This equation is computationally intractable for large N, therefore, to determine the likelihood of such an event in the droplet barcode library, I conduct in *in silico* simulations of the random barcode selection process (figure 2.4). I find that the likelihood of this undesirable event is extremely low for barcodes of sufficient length.

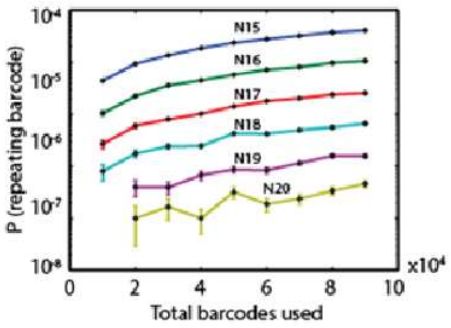

Figure 2.4 Probability of generating two barcode droplets with the same barcode derived using computer simulations.

2.7 Barcode mutations during PCR amplification

During PCR amplification and sequencing of the barcodes, errors and mutations generate a cloud of related sequences around the original barcode sequence. By sequencing the barcode library, I find that the original barcode sequences are on average three Hamming distances away from their nearest neighbor, while the sequences within the “cloud” of mutated barcodes around each original barcode are, on average, only 1 Hamming distance from their nearest neighbor (figure 2.5). However, the mutated barcodes typically comprise < 5% of all reads and do not represent a significant source of inefficiency. To address this issue, I cluster the mutated barcodes and their parent sequences into a single “barcode group” using a depth first search algorithm.

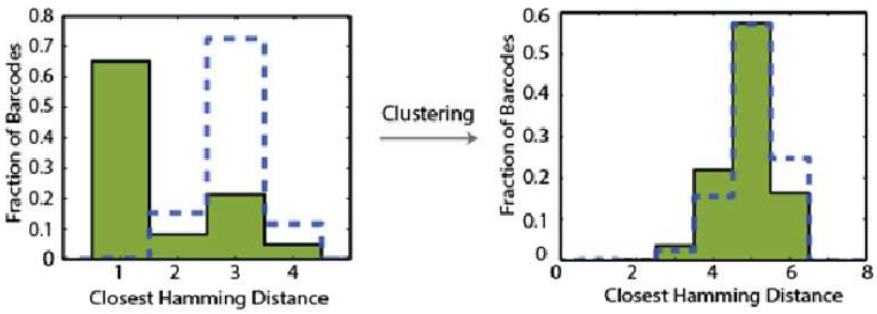

Figure 2.5 Hamming distances between 1000 randomly sampled barcodes and their closest Hamming neighbors before and after clustering mutated barcodes. Green line represents real data, blue dashed line represents simulations based on random sampling.

2.8 Algorithm to cluster all barcodes with their 1 Hamming distance neighbors

Given the assumption that all PCR mutations of an original barcode are connected to the original barcode by single base-pair mutations, the original barcode and all its mutant forms can be found by a depth first search algorithm as follows:

1. All unclustered barcodes are stored as a hashed set.
2. Take one barcode from the set, generate a set of all sequences that are 1 Hamming distance away. (There should be n x 3 sequences where n is the length of the barcode).
3. Search the set of unclustered barcodes for any of the set of sequences generated above.
4. Store the original barcode and 1 Hammings distance barcodes that are present in a new cluster.

5. For every barcode added to a new cluster, repeat steps 2 – 3 for that barcode and store the barcodes that are 1 hammings distance away into the new cluster.
6. Repeat from step 2 until the set of unclustered barcodes are empty.

This algorithm runs in linear time and is implemented in python using dictionaries as hashed sets. The rapid nature of dictionary lookup in Python results in fast runtimes.

2.9 Uneven amplification of barcodes during PCR

The stochasticity involved in amplification of single molecules results in some droplets in the barcode library amplifying faster than others. The consequence of uneven amplification is differences in barcoding efficiency, which translates to unevenness of representation of barcode groups in the resultant sequencing reads (figure 2.6).

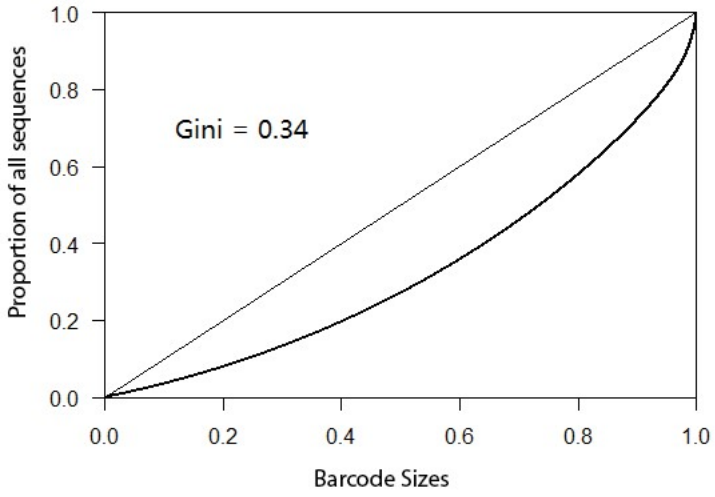

Figure 2.6 Lorenz curve of barcode representation in barcoded sequencing output.

To investigate the effects of PCR cycles and primer concentration on the unevenness of amplification, I sequence droplet barcode libraries generated using 20, 30, and 40 cycles of PCR and 400nM, 40nM and 4nM primers and plot Lorenz curves for each library (figure 2.7). Droplet barcode libraries generated with 40 PCR cycles are less biased compared to those generated with less cycles. This is likely because at 40 PCR cycles, the primers in most droplets have been used up and are thus unable to generate more PCR products, allowing droplets that are slower to amplify to catch up to the “PCR-saturated” droplets. Interestingly, reducing the primer concentrations, which in theory should result in “PCR-saturated” droplets with less PCR cycles, actually result in more biased libraries. A possible explanation is that at lower starting primer concentrations, the stochastic amplification of single molecules becomes less likely, thus resulting in an even wider gap in amplification rate between droplets.

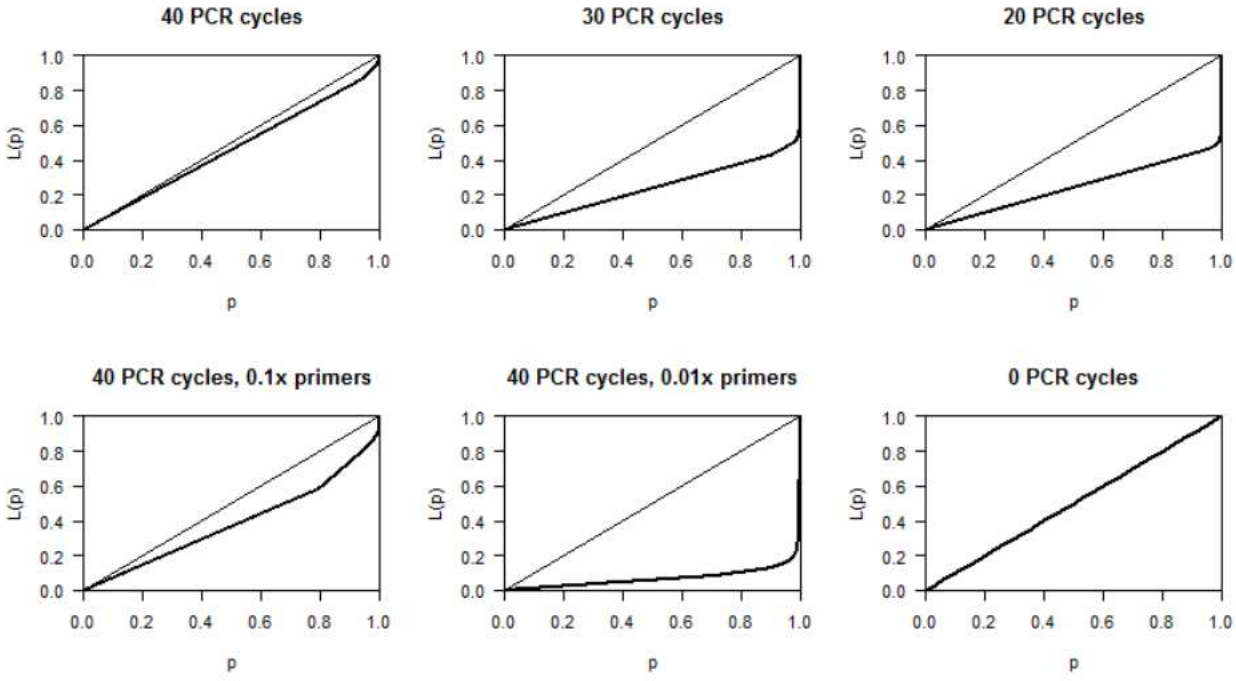

Figure 2.7 Lorenz curves of droplet barcodes generated under different PCR conditions.

2.10 Conclusions

The droplet barcode library can be cheaply generated requiring only PCR reagents, chemically synthesized oligonucleotides, and a microfluidic droplet maker. Using a random 15 nucleotide sequence as the barcode results in a vast sequence space where it is highly unlikely to resample the same barcode sequence twice in any given barcode library. Mutated barcodes can be re-associated with the original barcodes through a fast computational algorithm. Evenness of amplification can be ensured by "PCR-saturating" the droplets through many cycles of PCR. In the following chapters, I will use the droplet barcode library to develop useful ultrahigh-throughput droplet sequencing assays.

2.11 References

Ahn, K., Agresti, J., Chong, H., Marquez, M., and Weitz, D.A. (2006). Electrocoalescence of drops synchronized by size-dependent flow in microfluidic channels. Appl. Phys. Lett. *88*, 264105.

Horton, R.M., Hunt, H.D., Ho, S.N., Pullen, J.K., and Pease, L.R. (1989). Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. Gene *77*, 61–68.

Miller, O.J., Harrak, A.E., Mangeat, T., Baret, J.-C., Frenz, L., Debs, B.E., Mayot, E., Samuels, M.L., Rooney, E.K., Dieu, P., et al. (2012). High-resolution dose–response screening using droplet-based microfluidics. Proc. Natl. Acad. Sci. *109*, 378–383.

Picelli, S., Björklund, A.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. *24*, 2033–2040.

Chapter 3: Barcoding single molecules of DNA for long read sequencing

The ability to accurately sequence long DNA molecules is important across biology, but existing sequencers are limited in read length and accuracy. In this chapter, I describe a method, using ultrahigh-throughput droplet barcoding, to leverage short read sequencing to obtain long and accurate reads. Using droplet microfluidics, I isolate, amplify, fragment, and barcode single DNA molecules in aqueous picoliter droplets, allowing the full-length molecules to be sequenced with multi-fold coverage using short-read sequencing. I show that this approach can provide accurate sequences of up to 10 kilobases, allowing us to identify rare mutations below the detection limit of conventional sequencing and directly link them into haplotypes. This barcoding methodology can be a powerful tool in sequencing heterogeneous populations such as viruses.

3.1 Motivation

Massively parallel sequencing, also called Next Generation Sequencing (NGS), has tremendously impacted biomedical research due to its ability to acquire massive amounts of sequence data (Metzker, 2010; Shendure and Ji, 2008). Currently, the most widely adopted sequencing platform produces billions of short (<250bp) reads at a low cost of ~$50 per billion bases. However, short NGS reads pose challenges for many applications. For instance, piecing together short reads into long contiguous sequences can be challenging when assembling new genomes, particularly when repetitive sequences are present (Li et al., 2010; Nagarajan and Pop, 2013). When sequencing

metagenomes comprising thousands of species, it is often impossible to assemble the short reads into longer sequences that allow discovery of useful information, such as identification of the species to which a sequence belongs, or detection of gene clusters encoding useful molecules or phenotypes (Saeed et al., 2011; Wommack et al., 2008; Wooley and Ye, 2009). Furthermore, NGS is error-prone, generating an error in every thousand bases; this is often above the rate of biological variation and, consequently, prevents detection of true variants within the cloud of sequencing error (Bansal, 2010; Nielsen et al., 2011). The ability to obtain massive amounts of long and accurate reads would thus be a major step forward in our ability to characterize genomes accurately, and to study the impact of sequence variation in a variety of systems, such as in rapidly evolving virus populations (Acevedo et al., 2014), rare polymorphisms in human populations (Tennessen et al., 2012), and diverse and uncultivable species in microbial communities (Scholz et al., 2012).

3.2 Methods to obtain long and accurate reads

To obtain longer and more accurate reads, one approach is to directly improve the sequencing instrument (Chin et al., 2013; Laszlo et al., 2014). In addition to providing accurate reads, the instrument must be widely available, easy to use, and cost-competitive. Currently, no platform can match short-read NGS in these aspects and as such, short read sequencers dominate the market. Rather than inventing a new sequencing instrument, an alternative is to synthetically reconstruct long reads from short read data, leveraging the widespread popularity of short read NGS. An elegant approach is using unique molecular barcodes, which were first used to detect duplicated

NGS reads for error correction, and digital counting of molecules(Casbon et al., 2011; Kinde et al., 2011). To reconstruct long reads using molecular barcodes, long template molecules are broken into short fragments and labeled with “barcode” sequences identifying the template from which they originate(Amini et al., 2014; Hiatt et al., 2010; Kuleshov et al., 2014; Lundin et al., 2013). All short fragments can then be pooled and sequenced, and fragments of individual templates grouped by barcode. The reads in each group are then used to reconstruct synthetic long reads. Implementations of this approach rely on intramolecular reactions to attach barcodes to the fragments; however, this reaction becomes inefficient for templates above 3 kb. Alternatively, molecules can be physically isolated into wells, followed by fragmentation and barcoding. This approach can theoretically be extended to molecules of any length, but is limited in the number of templates that can be sequenced due to the limitations in throughput of liquid handling in well plates. Throughput can be increased by barcoding multiple templates in each well, but then single-molecule identity is lost (Amini et al., 2014; Kuleshov et al., 2014). To enable long and accurate DNA sequencing, an optimal approach would combine physical isolation of molecules with ultrahigh-throughput fluid handling.

To this end, I describe Single Molecule Droplet Barcoding (SMDB), an ultrahigh-throughput method to barcode long molecules for short read sequencing. Using droplet microfluidics, I isolate and barcode single molecules in aqueous droplets, which enables massively parallel sequencing of these single molecules at multi-fold coverage. To validate the method, I sequence a library of known DNA templates of 3-5 kb long and reconstruct long reads fully covering the templates. To demonstrate the ability to sequence large DNA molecules, I apply the method to the *E. coli* genome, obtaining

synthetic read-lengths up to 10 kb in length. Finally, to illustrate the power of the method for detecting variants below the detection limit of conventional sequencing, I apply it to a library of β-glucosidase genes mutated by PCR. While SMDB detects 457 SNPs in 81 haplotypes in the library, conventional short read sequencing detects only one SNP and cannot generate haplotypes. The ability to characterize variants and haplotypes below the inherent detection limit of the sequencer should be powerful for studying systems in which rare variants play an important role, such as in microbial community dynamics and viral quasispecies.

3.3 Development of single molecule droplet barcoding workflow (SMDB)

Single Molecule Droplet Barcoding (SMDB) is an ultrahigh-throughput method to barcode long molecules for short read sequencing. In SMDB, I use droplet microfluidics barcode fragments of single DNA molecules, performing all steps of template amplification, fragmentation, and barcoding in a microfluidic workflow (Fig. 3.1). DNA barcodes uniquely tag all reads derived from a template, which allows the reads to be unambiguously clustered to generate a long and accurate consensus sequence for the

template.

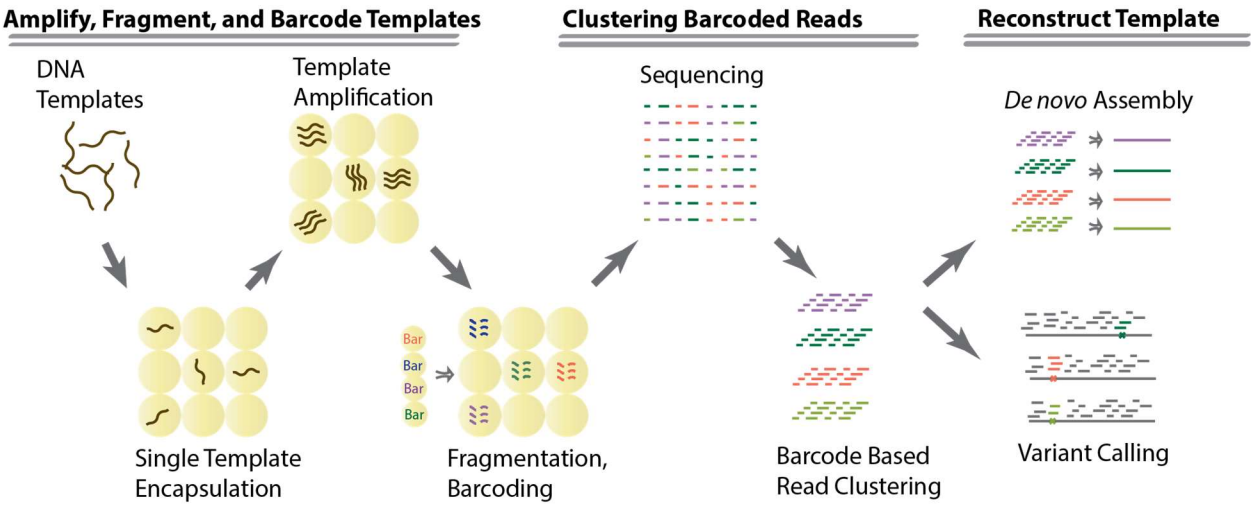

Figure 3.1 Overview of SMDB workflow.

3.3.1 Template encapsulation and amplification

The first step of SMDB is to isolate and amplify the template molecules, which is accomplished by introducing them into a microfluidic flow focus droplet generator that encapsulates them in ~50 μm diameter droplets of PCR reagent (Fig. 3.2). The template concentration is controlled so that ~1 in 10 droplets contains a single molecule, in accordance with Poisson statistics(Hindson et al., 2011). The droplets are collected into a PCR tube and thermal cycled for amplification, generating within each droplet a clonal population of the single molecules so that, once fragmented and barcoded, I can obtain multi-fold coverage of each template.

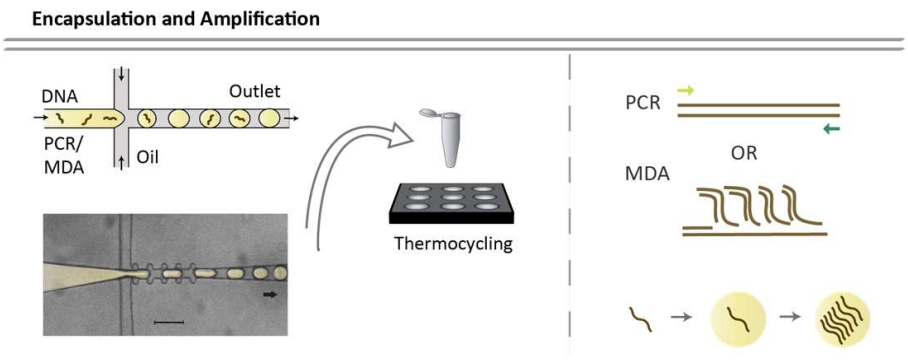

Figure 3.2 Template encapsulation and amplification for SMDB. A flow focus drop maker is used to encapsulate single templates into droplets. Inside droplets, PCR or MDA is used to clonally amplify the single template. Scale bars represent 100 μm.

3.3.2 Template Fragmentation

Following amplification, the templates must be fragmented to a length compatible with short-read sequencing. Importantly, fragmentation must be performed while maintaining compartmentalization, to prevent pieces of different templates from mixing before barcodes have been attached. To fragment in the droplets, I use a microfluidic device to add Tn5 transposase into each droplet, which randomly fragments and attaches short sequences to the amplified templates(Adey et al., 2010). Because transposases are single-turnover enzymes, an optimal stoichiometric ratio of transposase to templates must be maintained with a 10-fold dilution of the template droplet into the fragmentation droplet. To address this need, I develop a module combining droplet splitting and merging (Fig. 3.3). The incoming droplets pass through a junction sampling ~1/10th of their volume which is then merged with a new droplet ~ equal to the size of the original droplet. This device accomplishes the necessary tasks

of diluting the starting droplet and adding the new reagent, while maintaining the droplet size constant throughout the process. After the transposase is added, the droplets are collected into a syringe and incubated in a water bath at 55°C for the transposase reaction.

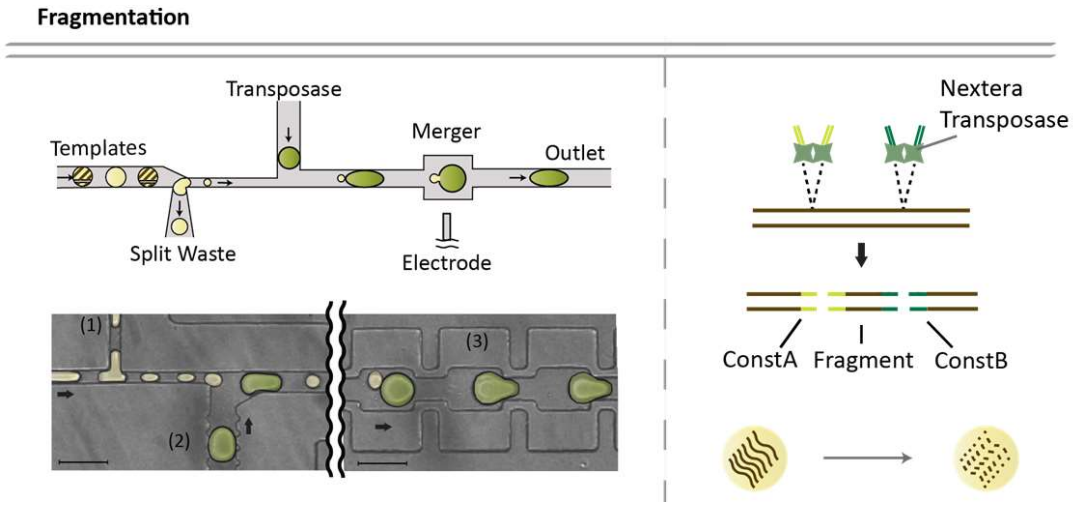

Figure 3.3 Template fragmentation for SMDB. The splitmerger device is used to add transposases into template drops while achieving a 10x dilution of the templates. The template droplets are injected on the left side, split at junction (1) so that 1/10th of the droplet continues on to pair with a reagent droplet generated on chip at (2) and the pair merges at the channel widening (3). The transposase reaction inside droplets fragments templates while adding adaptors to each fragment. Scale bars represent 100 μm.

3.3.3 Barcoding of fragmented templates

After the templates have been fragmented, the barcodes used to tag fragments belonging to the same template are attached by overlap-extension PCR in the droplets (Fig. 3.4). In this reaction, barcode sequences attach to the fragments through regions

of sequence homology on the adaptor sequences added by the transposase. This step thus requires merging three droplets: template, barcode, and PCR reagent. I design a triple merger device for merging three droplets at once. Improving on the designs of conventional mergers(Jin et al., 2010), I concatenate multiple merging junctions which act independently to achieve robust merging of all three droplets (Fig. 3.4). The volumes and reagent concentrations of the droplets are controlled to ensure correct stoichiometry for PCR barcoding. In addition, the channels enable one of each type of droplet to combine in the electro-coalescence junction. The resultant droplets are 90μm spherical diameter and can coalesce during thermal cycling. To make them more robust to coalescence, I split the merged droplets into four smaller droplets using a splitter(Abate and Weitz, 2011). The split droplets are collected into PCR tubes and thermally cycled to attach the barcodes.

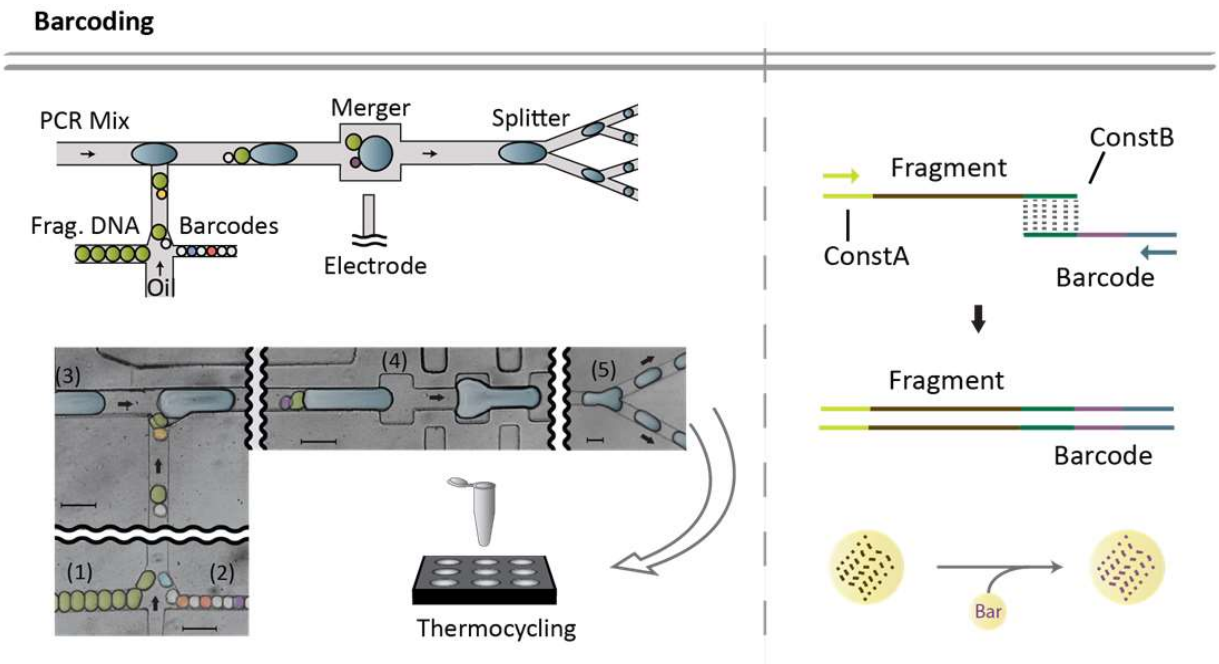

Figure 3.4 Attaching barcodes to fragments in SMDB. The microfluidic device used for attaching barcodes to DNA fragments. Template droplets (1) and barcode droplets (2) are injected into the device where they pair with each other and a large PCR reagent droplet generated on chip (3). The three droplets merge at the electrode (4) and are split into smaller droplets for thermal cycling (5). Barcodes are spliced onto fragments by overlap extension PCR. Scale bars represent 100 μm.

3.3.4 Massively parallel sequencing of barcoded templates

Approximately 10-50% of droplets coalesce after thermocycling, which is undesirable since it can lead to multiple templates or barcodes in a single droplet, and hence improper barcoding. I therefore remove these droplets using pinched-flow fractionation(Yamada et al., 2004). Alternatively, I find that manually removing large coalesced droplets using a micropipette can also work well. The remaining droplets are

chemically ruptured and the DNA contents are purified over a spin column, then size selected to remove free barcodes, resulting in a sequence-ready library. The library is then sequenced through a standard Miseq protocol, with the exception that a custom indexing primer for I7 is used to accommodate our own barcodes.

3.4 Validation of SMDB

The key property of SMDB is that *single* molecules are barcoded with unique barcodes. To examine the efficiency of single molecule sequencing with SMDB, I use SMDB to sequence a set of 8 known templates from 3-5 kb long. Because only one tenth of barcode droplets contain barcodes, I expect only one tenth of encapsulated templates to be barcoded. Starting with ~1M template droplets encapsulated at one in ten droplets containing templates, I expect a theoretical yield of ~10,000 barcoded templates. Practically, the yield of sequenced templates would be lower due to sample losses incurred during the start-up of microfluidic devices and during the removal of coalesced droplets. Sequencing the library, I obtain ~10 million reads using a MiSeq 2x250 run, yielding 3563 clusters which represents ~35% of theoretical yield.

3.4.1 Validation of Single molecule barcode groups

A key property of SMDB is its ability to barcode *single* molecules, which greatly simplifies bioinformatic analysis since all reads in a given cluster are known to originate from only one template. For perfect barcoding of single molecules, all reads in all clusters should map to only one template. Aligning reads from each cluster to the eight

reference sequences, I calculate for each barcode group the fraction of reads mapping to the dominant template, defined as the single (out of 8 possible) template to which the majority of reads in a cluster map (Fig. 3.5). I find that > 90% of clusters contain > 90% reads mapping to the dominant template. Nevertheless, I observe a low background of < 2% of reads mapping to the non-dominant template in less than half of the barcode groups, which I attribute to mis-tagging, a phenomenon often observed in barcoded sequence libraries prepared in well plates, and thought to originate from chimeric PCR products generated during library amplification and sequencing(Carlsen et al., 2012). Since many barcode groups contain some degree of non-dominant template reads, I define clusters containing > 90% dominant template as single template clusters. The overwhelming majority (~90%) of clusters are single-template clusters (Fig 3.6). Instances of multiple templates in the same barcode group are infrequent, and consistent with the rate of co-encapsulation expected by Poisson statistics. Multiple-encapsulations can be reduced by lowering template concentration, which reduces the instances of multiple templates in the same barcode groups at the expense of barcoding throughput.

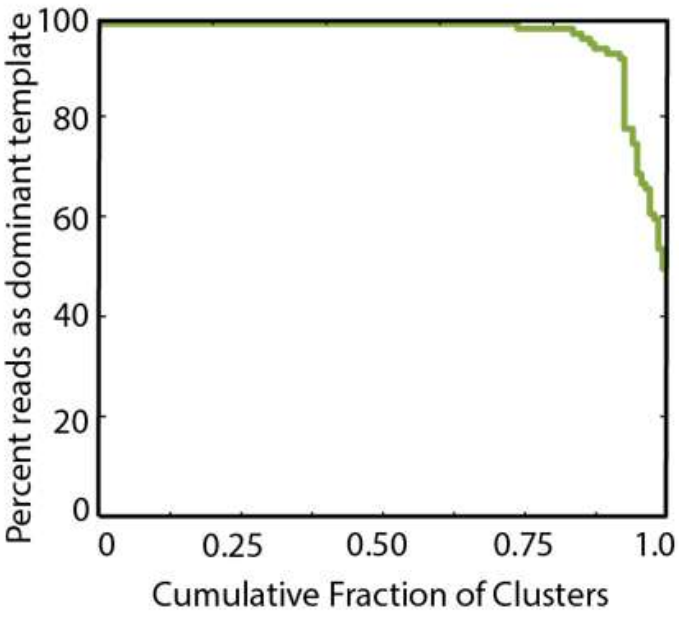

Figure 3.5 Purity of barcode groups. Cumulative histogram of the percent of reads mapping to dominant template (purity) of the barcode groups. The majority of barcode groups contain >90% reads mapping to the dominant template.

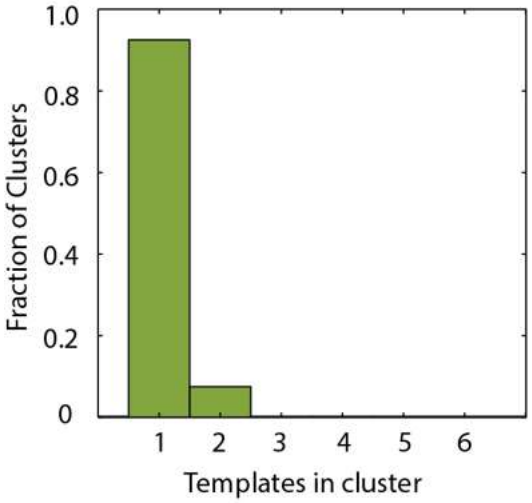

Figure 3.6 Number of templates detected inside barcode groups.

3.4.2 Coverage distribution of barcoded sequences

The ideal sequencing data provides full-length, high-accuracy coverage of all templates in the sample. However, bias in sequencing can yield excessive coverage in certain regions and insufficient coverage in others. To investigate whether our approach is susceptible to such bias, I plot the coverage distribution for each template (Fig. 3.7 shows two representative templates). I observe systematic coverage bias for all templates, much of which correlates with local GC content and, hence, is likely the result of the PCR amplification of the libraries for sequencing(Aird et al., 2011). I also observe decreased coverage at the ends of templates, a known bias of transposase fragmentation(Adey et al., 2010). Thus, the primary forms of bias in our data are the same as those observed in standard NGS, and result from the same sources.

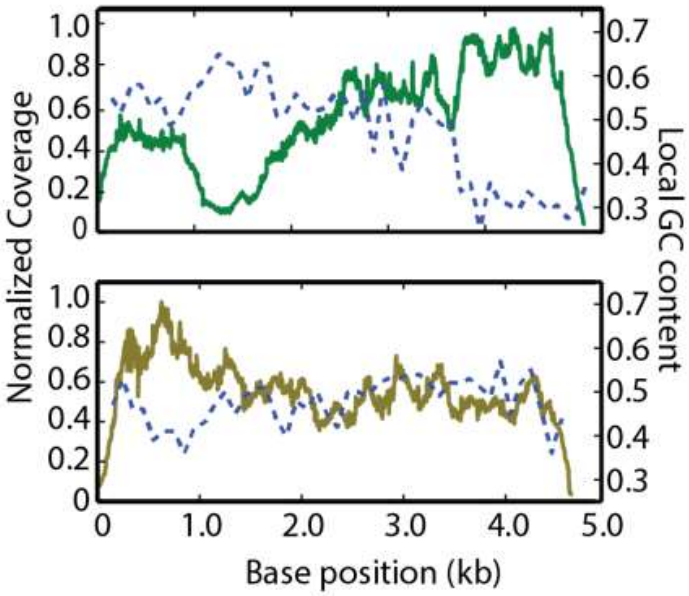

Figure 3.7 Sequencing coverage distribution over templates. Aggregate read coverage over the long molecule templates (solid lines) and corresponding local GC content (dashed lines).

3.5 Using single molecule barcoding to improve sequencing accuracy

3.5.1 Motivation

An important application of NGS is to detect rare single nucleotide polymorphisms (SNPs) in heterogeneous populations, such as viruses, cells, or human beings(Acevedo et al., 2014; Bansal, 2010; Meyerson et al., 2010; Out et al., 2009). Characterizing which SNPs are physically linked on the same template, called haplotyping, is important for understanding how multiple variants at distant loci can contribute to a given phenotype. However, performing these tasks with conventional NGS is often extremely challenging or impossible due to the inability of the short reads to span multiple SNPs. Moreover, standard NGS is error prone, generating one error in every ~1000 bases; this prevents confident detection of rare variants without accepting a large proportion of false positives(Bansal, 2010; Nielsen et al., 2011; Simen et al., 2009). To enhance sensitivity, known patterns of error production can be modeled and used to correct data, providing modest improvements(Bansal, 2010). Molecular techniques can greatly increase sensitivity to detect rare SNPs, but reduce read length even further(Lou et al., 2013).

3.5.2 High sensitivity SNP detection

SMDB is able to confidently detect rare SNPs because each molecule is sequenced to great depth, allowing reads to be “averaged together” to obtain an accurate consensus for every base. To demonstrate this, I generate a population of DNA templates via 35 cycle PCR of a bacterial plasmid extracted from a culture grown

from a single colony. In this population, every sequence shares significant homology, but rare variants exist. Variants like these can have important biological consequences, such as allowing HIV to evolve drug resistance, or the development of rare alleles that increase risk for disease in human populations(Simen et al., 2009; Tennessen et al., 2012). I sequence the population using SMDB on a MiSeq 2x150 run, obtaining 4.6 million reads in ~6,000 barcode groups. Because each barcode group represents fragments amplified from a single molecule, I expect a fraction of the fragments – and therefore reads – to contain amplification errors. In the worst case scenario where an error is made in the first round of amplification, I expect ~50% of the reads to be erroneous for any one position in the sequence. Since these cases are reported as di-allelic SNPs by the SNP-caller, I keep only the mono-allelic SNP calls to ensure the highest accuracy of our mutation calls. I identify 457 high confidence SNPs in ~10% of templates whereas ~90% of the templates contain no SNPs compared to the reference (Fig. 3.8). With the exception of SNP C1067G existing in ~5.5% of templates, all others are present in < 0.1% of the templates, far below the limit of detection for standard NGS. To compare our results to standard SNP calling methods which do not use barcode information, I call SNPs while disregarding the barcode grouping of reads and detect only the C1067G variant. Hence, SMBD amplifies the sensitivity of sequencing and allows capture of biological information invisible to standard methods. Unlike conventional NGS, the limit of detection of SMDB scales with the number of molecules sequenced and can be easily orders of magnitude more sensitive than conventional NGS (Fig. 3.9).

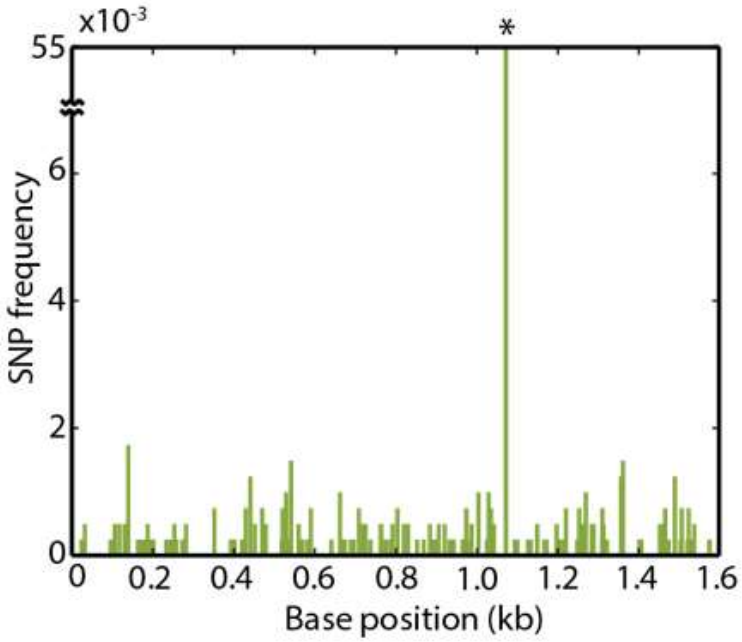

Figure 3.8 Frequency of detected SNPs by base position.

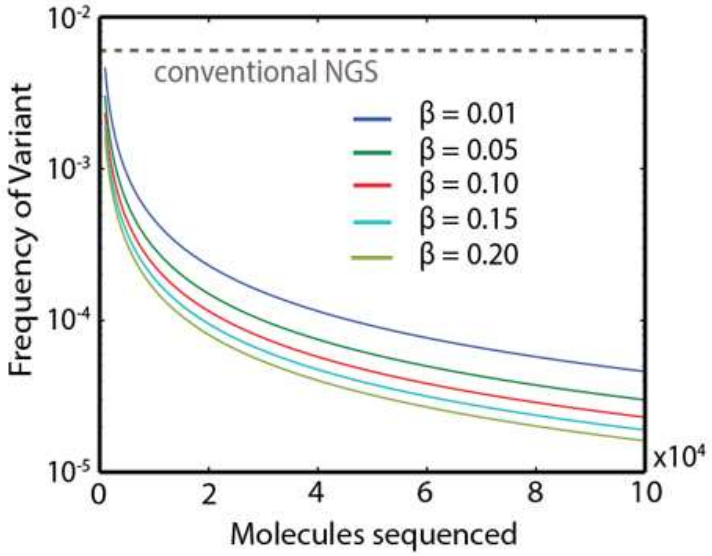

Figure 3.9 Theoretical detection limits (false negative errors) of SMDB based on number of molecules sequenced.

3.5.3 Haplotyping

In addition to detecting rare SNPs, SMDB naturally generates haplotypes, which are important for characterizing mutations that have synergistic effects and are broadly relevant from virus evolution to human genetics(Giallonardo et al., 2014; Sabeti et al., 2002). SMDB provides haplotyping information because SNPs that occur on the same template are grouped into the same barcode group, allowing haplotypes to be confidently identified for each template. To demonstrate SMDB haplotyping, I plot the haplotypes determined by SMDB in a phylogenetic tree, allowing us to determine the order of mutations that occurred during replication (Fig. 3.10)**.** The mutations in the population are generated by replication and, thus, in the absence of selection, ones that occur early in replication exist in a large subset of the progeny. The phylogenetic tree shows that C1067G was the first mutation that arose in the population, consistent with the fact that C1067G mutation is the most abundant SNP.

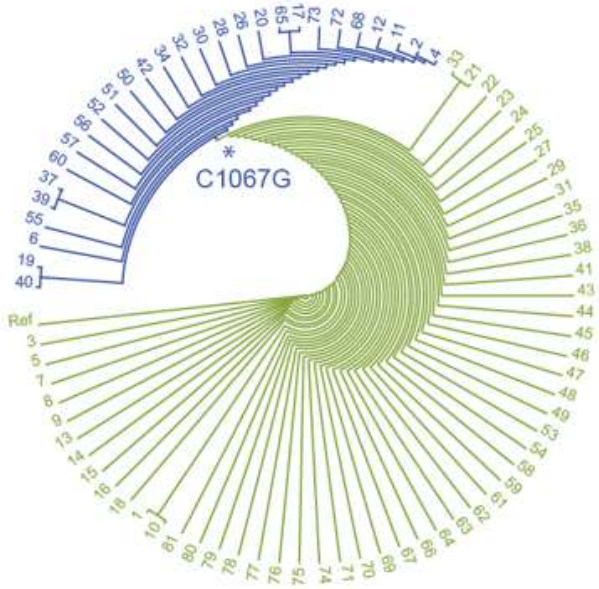

Figure 3.10 Phylogenetic lineage of haplotypes constructed from SMDB data.

3.6 Other uses for SMDB

I have applied SMDB to the barcoding of single DNA molecules from virus and microbial genomes, but the principle of encapsulating and barcoding nucleic acids in microfluidic droplets is broadly applicable. For example, droplet microfluidics has been used to encapsulate, lyse, and amplify single viruses and cells(Tao et al., 2015a, 2015b). As I show in the next chapter, the SMDB workflow could be combined with these methods to barcode the genomes of these organisms, to perform whole genome single virus or cell sequencing. This could make the barcoding workflow valuable for characterizing genetic reassortment in seasonal influenza. Indeed, while barcoding up to ~10,000 single entities is immediately practical with the methods I describe, if single cells rather than long templates were to be barcoded, the number of individual genomes that can be sequenced is limited by the sequencing throughput of NGS. Even with the massive capacity available with present-day instruments, it is not enough to fully leverage the throughput of our droplet method. However, as sequencing instruments continue to decrease in cost and increase in throughput, sequencing large barcoded populations of cells and viruses should become practical, impacting applications in which genetic diversity is important, such as in microbial communities.

3.7 References

Abate, A.R., and Weitz, D.A. (2011). Faster multiple emulsification with drop splitting. Lab. Chip *11*, 1911–1915.

Acevedo, A., Brodsky, L., and Andino, R. (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature *505*, 686–690.

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias

construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. *11*, R119.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J.O., Vijayan, K., et al. (2014). Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat. Genet. *46*, 1343–1349.

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics *26*, i318–i324.

Carlsen, T., Aas, A.B., Lindner, D., Vrålstad, T., Schumacher, T., and Kauserud, H. (2012). Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecol. *5*, 747–749.

Casbon, J.A., Osborne, R.J., Brenner, S., and Lichtenstein, C.P. (2011). A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. *39*, e81.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569.

Giallonardo, F.D., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz, S., Campbell, N.K., Joos, B., Lecca, M.R., et al. (2014). Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. *42*, e115–e115.

Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C., and Shendure, J. (2010). Parallel, tag-directed assembly of locally derived short sequence reads. Nat. Methods *7*, 119–122.

Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., Legler, T.C., et al. (2011). High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Anal. Chem. *83*, 8604–8610.

Jin, B.-J., Kim, Y.W., Lee, Y., and Yoo, J.Y. (2010). Droplet merging in a straight microchannel using droplet size or viscosity difference. J. Micromechanics Microengineering *20*, 035003.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., and Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. *108*, 9530–9535.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M. (2014). Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol. *32*, 261–266.

Laszlo, A.H., Derrington, I.M., Ross, B.C., Brinkerhoff, H., Adey, A., Nova, I.C., Craig, J.M., Langford, K.W., Samson, J.M., Daza, R., et al. (2014). Decoding long nanopore sequencing reads of natural DNA. Nat. Biotechnol. *32*, 829–833.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. *20*, 265–272.

Lou, D.I., Hussmann, J.A., McBee, R.M., Acevedo, A., Andino, R., Press, W.H., and Sawyer, S.L. (2013). High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc. Natl. Acad. Sci. U. S. A. *110*, 19872–19877.

Lundin, S., Gruselius, J., Nystedt, B., Lexow, P., Käller, M., and Lundeberg, J. (2013). Hierarchical molecular tagging to resolve long continuous sequences by massively parallel sequencing. Sci. Rep. *3*.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nat. Rev. Genet. *11*, 685–696.

Nagarajan, N., and Pop, M. (2013). Sequence assembly demystified. Nat. Rev. Genet. *14*, 157–167.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

Out, A.A., van Minderhout, I.J.H.M., Goeman, J.J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P.E.M., Tops, C.M.J., et al. (2009). Deep sequencing to reveal new variants in pooled DNA samples. Hum. Mutat. *30*, 1703–1712.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Saeed, I., Tang, S.-L., and Halgamuge, S.K. (2011). Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. Nucleic Acids Res. gkr1204.

Scholz, M.B., Lo, C.-C., and Chain, P.S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr. Opin. Biotechnol. *23*, 9–15.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Simen, B.B., Simons, J.F., Hullsiek, K.H., Novak, R.M., Macarthur, R.D., Baxter, J.D., Huang, C., Lubeski, C., Turenchalk, G.S., Braverman, M.S., et al. (2009). Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. J. Infect. Dis. *199*, 693–701.

Tao, Y., Rotem, A., Zhang, H., Chang, C.B., Basu, A., Kolawole, A.O., Koehler, S.A., Ren, Y., Lin, J.S., Pipas, J.M., et al. (2015a). Rapid, targeted and culture-free viral infectivity assay in drop-based microfluidics. Lab. Chip *15*, 3934–3940.

Tao, Y., Rotem, A., Zhang, H., Cockrell, S.K., Koehler, S., Chang, C.B., Ung, L.W., Cantalupo, P., Ren, Y., Lin, J.S., et al. (2015b). Artifact-free Quantification and Sequencing of Rare Recombinant Viruses Using Drop-based microfluidics. Chembiochem Eur. J. Chem. Biol.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science *337*, 64–69.

Wommack, K.E., Bhavsar, J., and Ravel, J. (2008). Metagenomics: Read Length Matters. Appl. Environ. Microbiol. *74*, 1453–1463.

Wooley, J.C., and Ye, Y. (2009). Metagenomics: Facts and Artifacts, and Computational Challenges*. J. Comput. Sci. Technol. *25*, 71–81.

Yamada, M., Nakashima, M., and Seki, M. (2004). Pinched Flow Fractionation: Continuous Size Separation of Particles Utilizing a Laminar Flow Profile in a Pinched Microchannel. Anal. Chem. *76*, 5465–5471.

Chapter 4. Single cell barcoding for ultra-high throughput single cell genome sequencing

Single cell genome sequencing is revolutionizing biology by enabling characterization of heterogeneous populations at single cell resolution; however, isolating and preparing single cell genomes for sequencing presents a bottleneck, increasing cost and limiting the number of cells captured. In this chapter, I describe SiC-seq, an ultrahigh-throughput method to sequence many single cell genomes. Using droplet microfluidics, I isolate, fragment, and barcode the genomes of >50,000 cells per run, allowing single cell information to be recovered by grouping reads by barcode. The stored genomes are amenable to computational sorting based on characteristic sequences, which I use to describe the distributions of antibiotic resistance genes, virulence factors, and phage sequences in microbial communities. The ability to routinely sequence large populations of single cells is a powerful and general tool for de-convoluting cellular heterogeneity across biology.

4.1 Motivation

Organisms are living expressions of their genomes and, hence, genome sequencing is a powerful way to study how they are structured, grow, and function. Organisms are also phenotypically diverse, and this diversity is mirrored by heterogeneity at the genomic level and plays important roles in populations as a whole, particularly among populations of single cells. For example, genomic heterogeneity fuels cancer's ability to evolve resistance to therapy and progress as a disease(Navin et

al., 2011; Potter et al., 2013), and is also found in the normally functioning cells of the brain(McConnell et al., 2013), skin(Abyzov et al., 2012), and immune system(Yancopoulos et al., 1986). Heterogeneity also plays important roles in microbial ecosystems, which constitute by far the most diverse organisms on the planet. As in other many-cell systems, heterogeneity impacts community dynamics, including a population's ability to colonize an environment or evolve against selective pressures(Hay et al., 2004). Because the heterogeneity exists among single cells, sequencing single cell genomes is critical to de-convoluting the complexity of the system and has already yielded breakthroughs in my understanding of microbial ecology(Kashtan et al., 2014; Marcy et al., 2007) and cancer(Navin et al., 2011; Ni et al., 2013).

A common challenge when applying single cell sequencing to heterogeneous systems is that they often contain massive numbers of cells: A centimeter-sized tumor can contain hundreds of millions of mutated cancer cells(Monte, 2009), while a milliliter of sea water can contain millions of microbes(Roxane Maranger and David Bird, 1995). Moreover, each cell has a tiny quantity of DNA, making it challenging to accurately amplify and sequence so many single cells. For example, cutting-edge methods based on optical tweezers(Zhang and Liu, 2008), flow cytometry(Rinke et al., 2014), and microfluidics(Gawad et al., 2014) can isolate and process hundreds of single cells for sequencing, but this constitutes a minute fraction of most communities. The sparseness of the sampling limits the questions that can be addressed, with the majority of findings relating to the most abundant subpopulations. For example, common environmental communities contain >1500 taxa, with rare taxons present at < 0.1%(Afshinnekoo et al.,

2015), most of which are missed by single cell sequencing; indeed, the difficulty of capturing these cells is the basis of “microbial dark matter” – the overwhelming abundance of species thought to exist, but that have never been characterized. Therefore, a method that could markedly increase the number of cells sequenced would impact a broad range of problems across biology in which heterogeneity is important.

Droplet microfluidics is a powerful technology for performing millions of independent picoliter reactions and has recently been used to profile the transcriptomes of single cells at high throughput (Klein et al., 2015; Macosko et al., 2015; Rotem et al., 2015). However, sequencing the *genomes* of single cells presents unique challenges, because genomic DNA must be purified from the cell body and processed through a series of enzymatic steps to prepare it for sequencing. Consequently, while droplet microfluidics provides the potential for ultrahigh-throughput single cell genome sequencing, as of yet, no approach for accomplishing this has been described.

In this chapter, I describe single cell genomic sequencing (SiC-seq), a droplet microfluidic method for sequencing > 50,000 single cell genomes per run, and ultimately scalable to millions. I validate the method by sequencing an artificial population of microbes containing known species at controlled proportions, obtaining ~0.1% average coverage per cell, uniform genomic sampling, and accurate species proportion estimates. Moreover, SiC-seq generates a novel metagenomic database in which reads are grouped by single cells. This database, in turn, enables a new kind of “*in silico* cytometry” similar to conventional flow cytometry, except that all sorting occurs computationally and sequence biomarkers need not be specified *a priori*. To demonstrate this, I apply SiC-seq and *in silico* cytometry to a sample of marine

microbes, and use it to measure how antibiotic resistance genes, virulence factors, and phage-associated sequences are distributed. The ability to repeatedly sort a sample of genomes without having to perform additional wet lab experiments allows rapid iteration through hypotheses and enhances what can be discovered by what is learned. It is valuable for generating correlation maps between characteristic sequences, to infer how different phenotypes are correlated within single cells, and how genetic elements spread through a community.

4.2 Developing the single cell barcoding (SiC-Seq) workflow

4.2.1 Overview and Challenges

Droplet microfluidics, with its ability to encapsulate and perform biological reactions on thousands of single cells per second, affords unparalleled potential for single molecule and single cell applications, including to uniformly amplify(Sidore et al., 2016), accurately quantitate(Hindson et al., 2011), and deeply sequence (Lan et al., 2016) single molecules, and to culture(Clausell-Tormos et al., 2008), screen (Romero et al., 2015), and sequence the transcriptomes(Macosko et al., 2015) of single cells. However, single cell *genome* sequencing presents the unique challenge that each cell's genome is protected behind a membranous barrier that must be removed before enzymatic processing is possible. The reagents required for efficient cell lysis, however, including detergents, proteases, and high pH buffers, are detrimental to sequencing preparation enzymes, requiring that these steps be performed separately. In SiC-seq, I address this challenge by encasing cells in hydrogel microspheres (microgels) that are permeable to molecules with hydraulic diameters smaller than the pore size, including

enzymes, detergents, and small molecules, but sterically trap macromolecular genomes. This allows us to use a series of “washes” on millions of encased cells, to perform the requisite steps of cell lysis and genome processing, while maintaining compartmentalization of each genome. Using a combination of microgel and microfluidic processing steps, I lyse the cells, fragment the genomes, and attach unique barcodes to all fragments, in a workflow that processes >50,000 cells in a few hours. The barcoded fragments for all cells can then be pooled and sequenced, and the reads grouped by barcode, providing a library of single cell genomes that can be subjected to additional downstream processing, including demographic characterization and *in silico* cytometry (Figure 4.1).

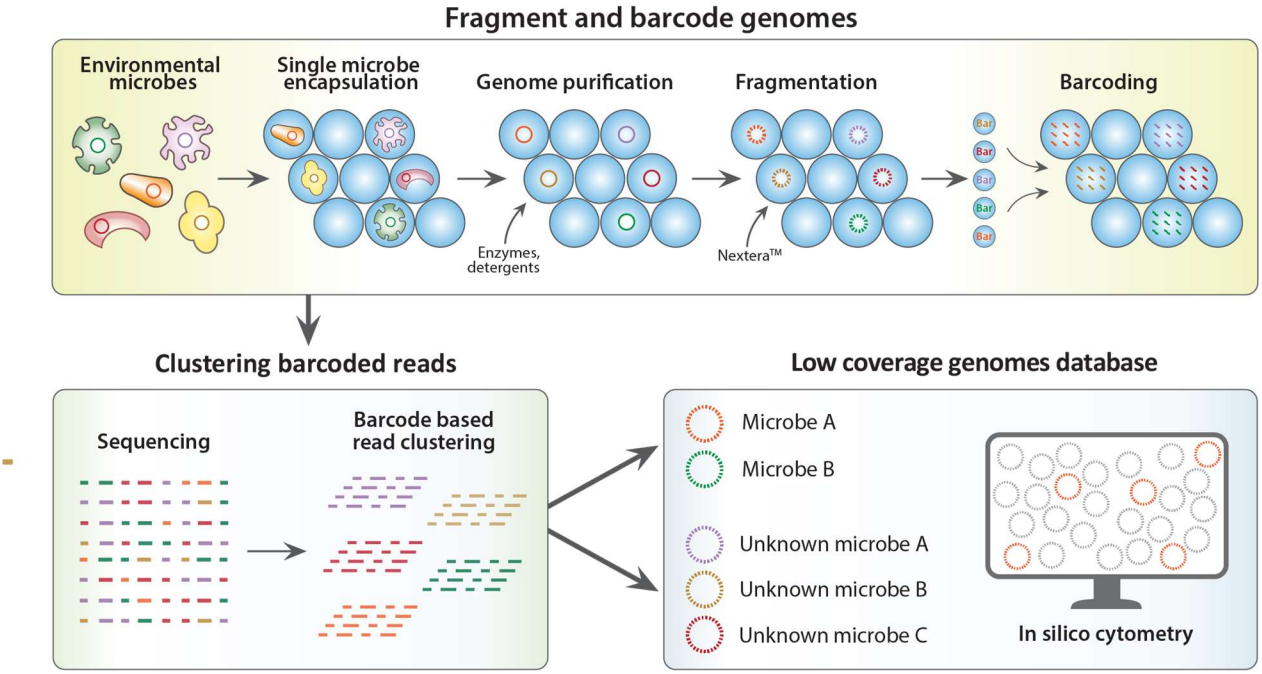

Figure 4.1 Overview of SiC-seq workflow

4.2.2 Single cell encapsulation

Before the single cell genomes can be barcoded, they must be physically isolated and purified from the cell body and fragmented. I isolate single cells in agarose microgels using a two-stream co-flow droplet maker, which merges a cell suspension stream with a molten agarose stream, forming a droplet consisting of an equal volume of both streams (Figure 4.2). The droplet maker runs at ~10 kHz, allowing us to generate 10 million 22 μm droplets in ~20 minutes, a total volume of aqueous emulsion fraction ~60 μL. Hence, droplet generation is fast and the total volume consumed small, allowing us to load cells at a rate of < 1% to minimize multi-cell encapsulation. The molten agarose droplets are collected into PCR tubes on ice, solidifying them. The solid microgels can then be transferred from oil to water, while maintaining encapsulation of the cells, which can then be subjected to lysis and genome purification.

4.2.3 Purification and fragmentation of genomes

To lyse the cells, I incubate the microgels overnight in a mixture of lysozyme, mutanolysin, and lysostaphin, digesting the protective microbial cell walls (Figure 4.2). I then incubate them in a mixture of sodium dodecyl sulfate and protease K for 30 minutes, solubilizing lipids and digesting proteins, preserving only high molecular weight genomic DNA, which I verify by staining with SYBR green dye. To fragment the genomes and attach the Illumina sequencing adaptors, I incubate the gels in the Nextera™ reaction for two hours, where transposases fragment the DNA while simultaneous add sequence adaptors to the fragments. Importantly, because the

transposases are dimeric, the fragmented genome remains intact as a macromolecular complex, so that it remains sterically encased within the hydrogel network.

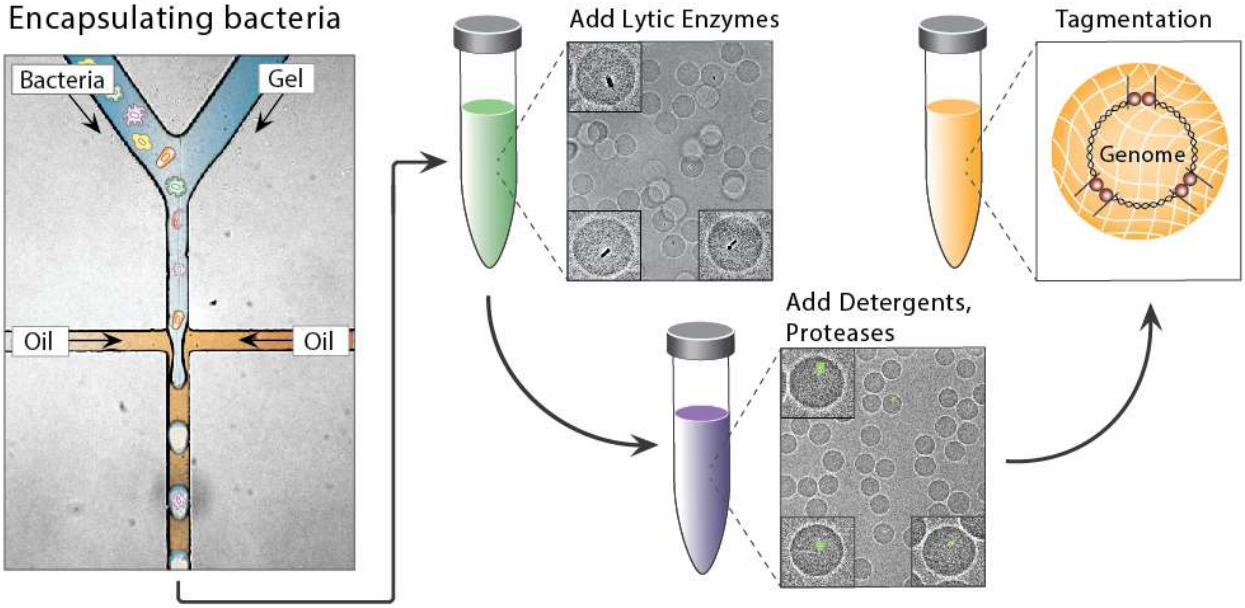

Figure 4.2 SiC-seq workflow, encapsulation of cells, purification and fragmentation of genomes in microgels.

4.2.4 Barcoding genomic fragments

After the genomes are purified and fragmented, they are barcoded for sequencing. I use a microfluidic device that encapsulates each microgel in PCR reagents, then merges it with a barcode droplet **(**Figure 4.3**)**. Monodisperse microgels have the unique and valuable property that, because they are compliant, they can flow at high volume fraction (> 0.65) through microfluidic channels without clogging, causing them to order and flow periodically into a droplet generator. By matching the droplet period with the microgel injection period, it is possible to achieve efficient loading of microgels in droplets. The droplets containing fragmented genome and barcode are

collected into a PCR tube and thermal cycled, splicing the barcode sequences onto the genomic fragments via complementarity through the adaptor sequence added by the transposase. During thermal cycling, some droplets coalesce, generating barcode groups corresponding to multiple cells. I remove these coalesced droplets using a micropipette at the end of thermal cycling, then the purified droplets are chemically ruptured using perfluoro-octanol, and their contents pooled and prepared for sequencing.

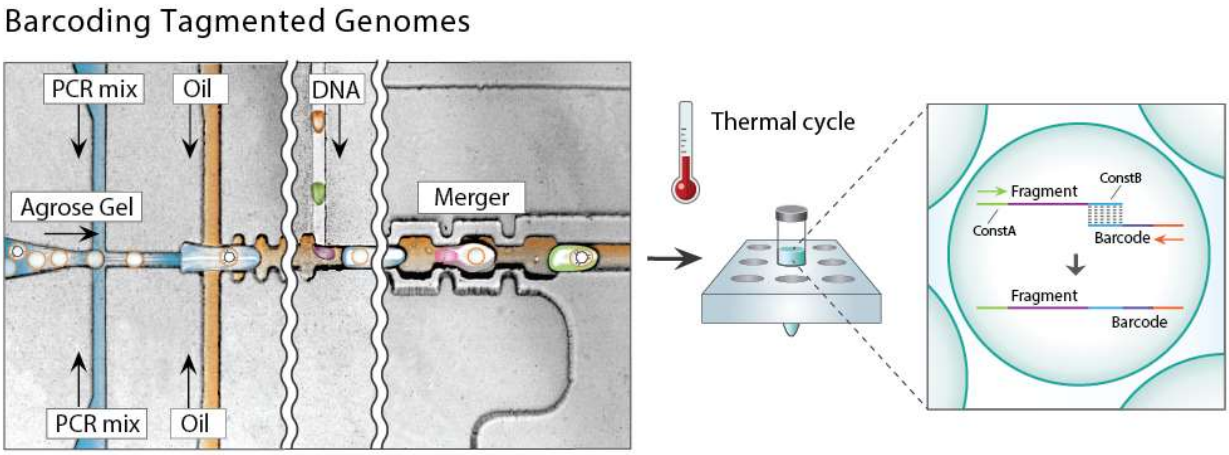

Figure 4.3 SiC-seq workflow, barcoding fragmented genomes in droplets.

4.2.5 Sequencing barcoded fragments

DNA is recovered from the pooled droplets using a spin column, then it is size selected using SPRI beads to remove the short unspliced barcodes from the long spliced barcoded genomic fragments. The resulting library is sequenced on an Illumina Miseq platform using a custom I7 read primer to accommodate for the spliced on

custom barcode. After sequencing, the reads are filtered by quality and grouped by barcode, providing single cell genomic sequence data.

4.3 Validating single cell barcoding workflow

The objective of SiC-seq is to provide single cell genomic sequence information bundled in barcode groups. To validate that SiC-seq generates single cell barcode groups, I apply it to an artificial community of mixed cultures of *Staphylococcus epidermidis, Saccharomyces cerevisiae, and Bacillus subtilis* at ratios of approximately 100:10:1, respectively. This mixture represents gram-positive bacteria and fungi, which are typically difficult to lyse to confirm that my lysis procedures are reasonably general. I prepare a single-cell library from this community using SiC-seq and sequence it on an Illumina MiSeq, yielding ~6 million paired-end reads of 75 bp after quality filtering. I group reads by barcode and discard clusters with < 50 reads, representing likely PCR-mutated barcode sequences, and yielding the final 51,500 barcode groups (Figure 4.4).

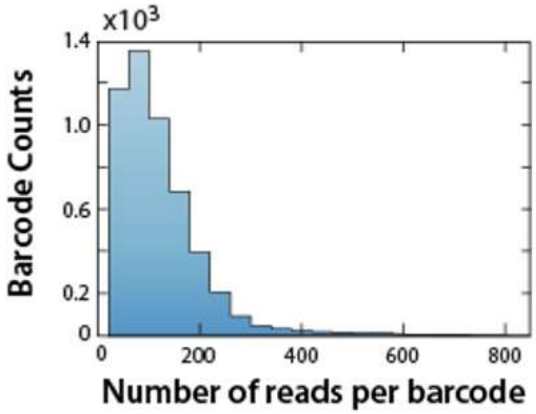

Figure 4.4 Distribution of number of reads in each barcode group.

4.3.1 Validation of one cell per barcode

To determine whether the barcode groups indeed correspond to *single* cells, I map all reads to the reference genomes of the three known species. If two microbes reside within the same cluster, reads will map to both genomes. I define a cluster purity score as the fraction mapping to the most mapped reference. The distribution of cluster purity scores is strongly skewed to high values, with an average of 94%, suggesting that most barcode groups represent single cells (Fig. 4.5).

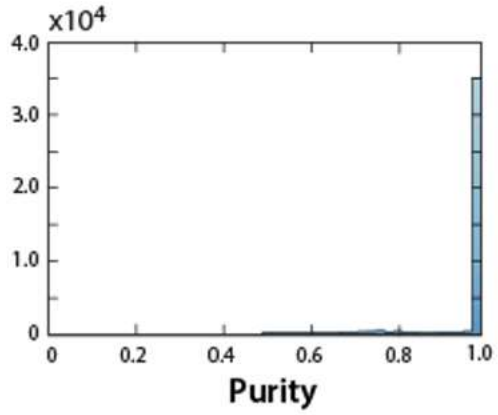

Figure 4.5 Distribution of purity of each barcode group, where purity is defined as the fraction of reads that map to the most dominant species in a barcode group.

4.3.2 Species abundance estimation using barcodes

To determine whether SiC-seq allows accurate quantification of species abundance within a mixed population, I compare abundance estimates measured in four ways: visual counting of cell cultures under a microscope when generating the mixed population, visual counting of cells after encapsulation into microgels, counting marker gene abundances in the reads based on Metaphlan assignment without using

barcodes, and by assigning a species identity to each barcode group based on the dominant reference mapped, and counting barcode groups. I find that all methods are in reasonable agreement for this sample (Figure 4.6). This demonstrates that SiC-seq enables accurate estimation of species abundance in a microbial population.

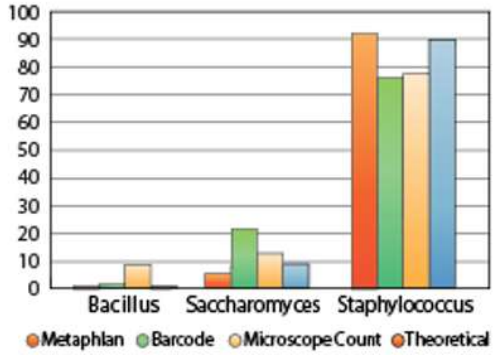

Figure 4.6 Estimation of relative abundance of species.

4.3.3 Coverage distribution of reads produced using SiC-seq

Sequencing the genome of a single cell typically incurs coverage distribution bias (Bourcy et al., 2014) due to uneven amplification of DNA starting from a single genome copy. To investigate coverage distribution bias in SiC-seq, I plot the normalized coverage distribution for all barcode groups detected for each microbe, an example shown in figure 4.7. With the exception of coverage gaps due to differences between the reference and actual genome sequenced, we observe no significant coverage bias. This indicates that the sampling of each genome within a barcode group is random, so that when all clusters are overlaid a uniform distribution is obtained. In addition, bias is minimal because each genome is amplified in a tiny volume of ~65 pL, which has been shown to curtail bias-inducing runaway of exponential amplification(Gole et al., 2013). Moreover, the sequencing library is composed of > 50,000 amplified cell genomes and,

as such, the amplification of each genome can be limited while still producing sufficient total DNA for sequencing.

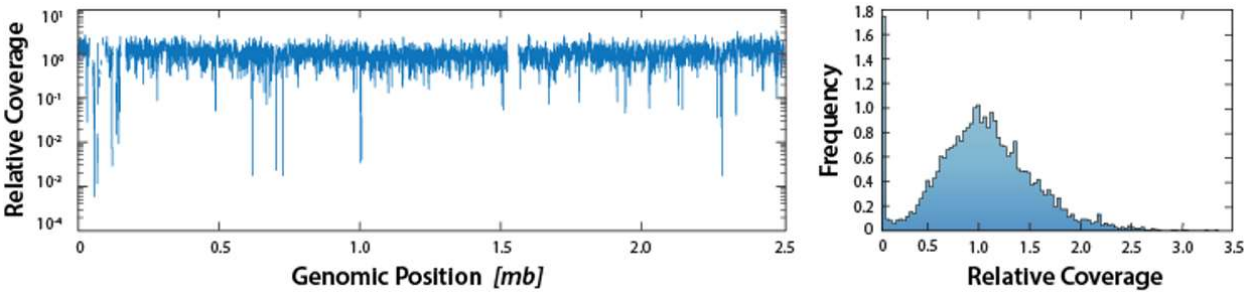

Figure 4.7 Coverage distribution over the *Staphylococcus* genome of all reads in *Staphylococcus* barcode groups and the same data plotted as a histogram of coverage distribution binned by relative coverage.

4.4 Analysis of SiC-seq data using *in silico cytometry*

The massive set of single cell genomes present in SiC-Reads provides new opportunities for discovering associations between sequences within single cells, in a process I dub *in silico* cytometry. SiC-Reads comprises a multidimensional collection of single cell genomes that can be sorted *in silico*, in analogy to what is commonly done with flow cytometry on single cells. However, while flow cytometry requires that a target biomarker be selected *a priori*, and is limited to the number of biomarkers that can be used, *in silico* cytometry can be performed as many times and with as many sequence biomarkers as desired, to sort and resort the database repeatedly to mine for connections between different genetic sequences and structures. Moreover, as new associations are learned, new sorting parameters can be defined, enabling new discoveries based on what is learned without having to repeat the experiment, ultimately limited only by the completeness and accuracy of the single cell database.

To demonstrate *in silico* cytometry, I apply SiC-seq to a microbial community recovered from coastal sea water of San Francisco. I obtain ~8 million reads of 150 bp length after quality filtering, with which I generate a SiC-Reads database. I assign barcode groups a taxonomic classification based on the reads they contain, following the rule that > 10% of reads must have a classification, and the group is classified according to the taxon with the most supporting reads. Most barcode groups are estimated to be high purity based on the classifiable sequences (~91%), in accordance with my control sample (~94%). Using this data, I demonstrate *in silico* cytometry by exploring the distribution of antibiotic resistance and virulence factors in the microbial community.

4.4.1 Using *in silico cytometry* to discover antibiotic resistance profile of a community

Antibiotic resistance (AR) is becoming increasingly common and represents a significant threat to global human health(Nathan and Cars, 2014). Because antibiotics are the primary tool for fighting bacterial infection, understanding how AR genes spread in the natural environment is essential. Microbes can gain AR through numerous mechanisms, including mutation, acquisition of resistance-conferring genes, or even deletion of genes. Complicating the study of these genes is that many are believed to serve non-killing purposes in the community, including as molecules serving as cues or coercions to alter the gene expressions of community members (Bernier and Surette, 2013). While AR genes have been identified in most environments by short-read sequencing, scant information on how they are distributed among taxa is available,

because obtaining this information usually requires testing or whole genome sequencing of single species; however, most species are uncultivable, precluding such analyses, and represent the “microbial dark matter”.

SiC-seq provides a unique opportunity to characterize the distribution of AR genes amongst all species in a sample, including uncultivable species. Species are classified based on reads in the barcode group, and then associated with AR genes also present. I search my database for known AR genes, finding 1,081 (0.012% of reads), representing 108 (0.30%) of barcode groups. The taxonomic distribution of AR genes has a clear structure (figure 4.8). The most abundant taxa associated with AR are not the most abundant taxa overall, suggesting that in this community certain taxa tend to associate more with AR genes. For example, Aminoglycoside resistance is primarily found in *Alteromonas spp.*, while Beta lactam resistance is common, and found in 4 out of 5 taxa. One explanation is that the broad-spectrum activity of Beta-lactams has encouraged their heavy use by humans and, correspondingly, has resulted in widespread resistance in the costal microbes of San Francisco. Aminoglycoside antibiotics, on the other hand, are less commonly used by humans and, thus, resistance against them is rare, with identified instances possibly representing natural *Alteromonas* biology. Similarly, *Enterobacter spp.* is associated primarily with Beta-lactam resistance, reflecting its heavy use by humans as the first line of therapy for *Enterobacter* infections.

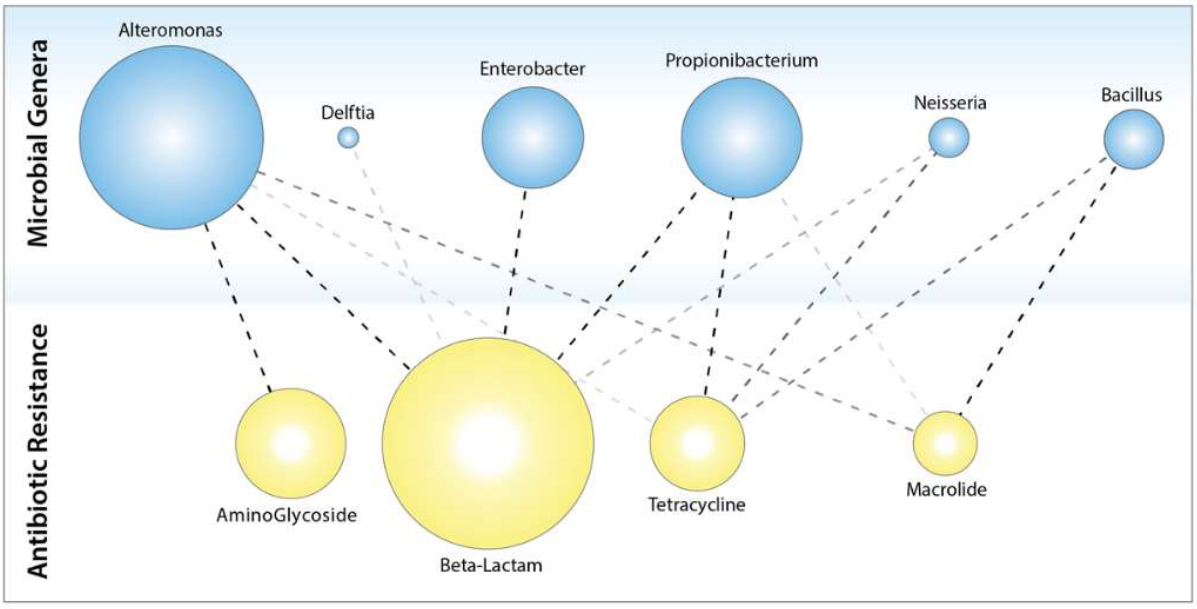

Figure 4.8 Distribution of antibiotic resistance genes discovered among taxa in San Francisco coast water community.

4.4.2 Using *in silico cytometry* to discover the virulence factor profile of a community

Virulence factors (VFs), like AR genes, are important in determining the threat that specific microbes in a community pose to human health. Many opportunistic pathogens reside in natural communities in the environment and cause outbreaks when transmitted to a suitable host (Yildiz, 2007). Therefore, monitoring and detecting potentially pathogenic microbes is important for public health. While metagenomics shotgun sequencing or DNA microarray methods can detect the abundance of VFs in a community, they cannot determine which microbes carry those VFs, or whether multiple VFs are present in the same microbes, both of which are important determinants for the

pathogenic potential of a species. Here, again, SiC-seq affords a unique opportunity to characterize VFs in a community and to associate them with specific host species. I search my coastal microbial community database for known virulence factor genes, yielding matches in 1,949 (0.022%) reads in 101 (0.28%) barcode groups consisting of 29 prevalent VFs distributed among 13 microbial genera. The abundances of taxa with VFs do not reflect that of the total population, suggesting that certain genera tend to carry more VFs than others. To quantify this, I calculate a VF ratio, the ratio between the abundance of barcode groups containing VFs and the total abundance of barcodes in the community for that species, and normalize the results so that the minimum ratio is 0 and the maximum 1 (figure 4.9). *Haemophilus* and *Escherichia* stand out amongst all species, both of which are known opportunistic human pathogens. Upon closer inspection, the main VFs detected in *Haemophilus* are lipo-oligosaccharides, which are the major constituents of *Haemophilus* outer membranes and an important determinant of host immune evasion (Preston et al., 1996). In *Escherichia,* the main VFs detected are the K1 capsule and Type III secretion system, both commonly present in virulent strains (Cross et al., 1984; Jarvis et al., 1995). Comparing my VF ratios with ones calculated for publically-available whole genomes, and down sampled to match my per-cell read depth, I find that the ratios are higher for the public genomes, indicating a bias towards pathogenic strains in currently-sequenced genomes.

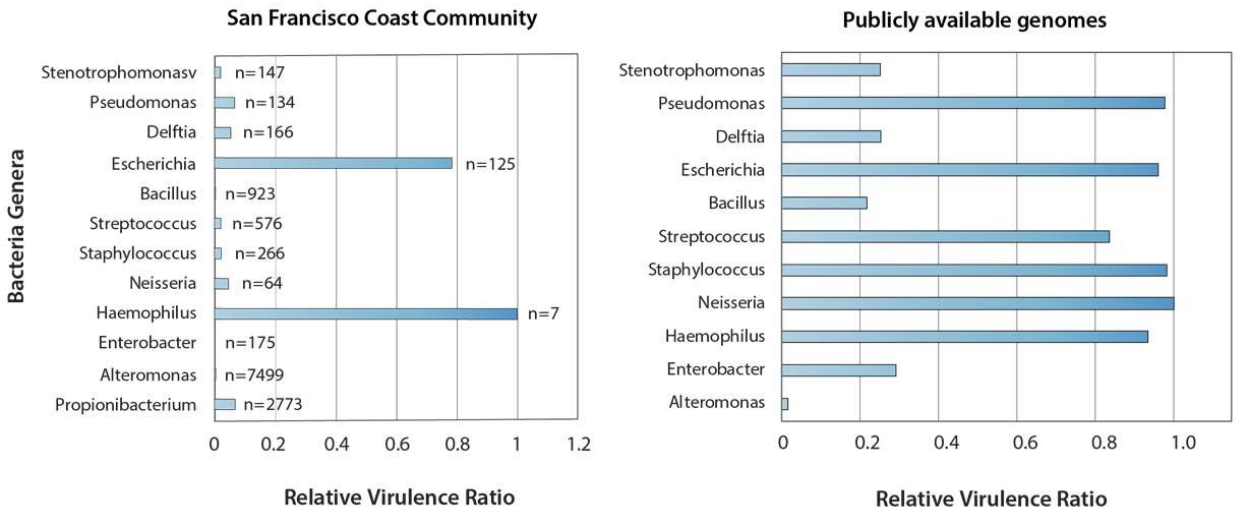

Figure 4.9 Normalized relative virulence ratios, calculated the proportion of barcodes detected with a virulence factor gene normalized to the genus with the highest proportion. The number of barcode groups of each genus used in calculations is indicated by n.

4.5 Other potential uses for SiC-seq

4.5.1 SiC-seq data as scaffold for genome assembly from shotgun metagenomics sequencing

The *de novo* assembly of genomes is a powerful tool for exploring uncultivable species and is a common goal in the field of metagenomics. Mate-paired sequencing can be used to bridge contigs in a library and, with sufficient coverage, even enables assembly of whole genome from shotgun metagenomics reads (Iverson et al., 2012). Though incredibly powerful, the method is limited by the construction of mate-paired libraries requiring micrograms of DNA that can be difficult to obtain from microbial

ecosystems. Furthermore, many mate-paired reads are required to assemble a whole genome, since each pair bridges only two contigs. SiC-seq should be a powerful alternative because it provides grouped reads from >50,000 cells, each of which bridges hundreds of contigs. Consequently, SiC-seq should allow generation of draft genomes from shotgun metagenomic data with far less DNA input requirement and sequencing effort, while also providing information about rare species present in the sample.

4.5.2 SiC-Seq of mammalian cell populations for characterizing genomic heterogeneity in cancer

While I focus on microbial communities, SiC-seq is applicable to populations of mammalian cells too, where it can have a more direct impact on human health. The grouped reads provided by SiC-seq should afford the information required to determine copy-number variations within the genome, which is relevant to cancer (Ni et al., 2013). The enormous size of mammalian genomes, however, limits the number of cells that can be sequenced for a desired target coverage. Nevertheless, as the cost of sequencing continues to decrease, more cells can be sequenced to greater depth, creating opportunities for characterizing mammalian tissues, cell-by-cell.

4.6 References

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438–442.

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., Maritz, J.M., Reeves, D., Gandara, J., Chhangawala, S., et al. (2015). Geospatial

Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. Cell Syst. *1*, 72–87.

Bernier, S.P., and Surette, M.G. (2013). Concentration-dependent activity of antibiotics in natural environments. Front. Microbiol. *4*.

Bourcy, C.F.A. de, Vlaminck, I.D., Kanbar, J.N., Wang, J., Gawad, C., and Quake, S.R. (2014). A Quantitative Comparison of Single-Cell Whole Genome Amplification Methods. PLOS ONE *9*, e105585.

Clausell-Tormos, J., Lieber, D., Baret, J.-C., El-Harrak, A., Miller, O.J., Frenz, L., Blouwolff, J., Humphry, K.J., Köster, S., Duan, H., et al. (2008). Droplet-Based Microfluidic Platforms for the Encapsulation and Screening of Mammalian Cells and Multicellular Organisms. Chem. Biol. *15*, 427–437.

Cross, A.S., Gemski, P., Sadoff, J.C., Ørskov, F., and Ørskov, I. (1984). The Importance of the K1 Capsule in Invasive Infections Caused by Escherichia coli. J. Infect. Dis. *149*, 184–193.

Gawad, C., Koh, W., and Quake, S.R. (2014). Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc. Natl. Acad. Sci. U. S. A. *111*, 17947–17952.

Gole, J., Gore, A., Richards, A., Chiu, Y.-J., Fung, H.-L., Bushman, D., Chiang, H.-I., Chun, J., Lo, Y.-H., and Zhang, K. (2013). Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. Nat. Biotechnol. *31*, 1126–1132.

Hay, M.E., Parker, J.D., Burkepile, D.E., Caudill, C.C., Wilson, A.E., Hallinan, Z.P., and Chequer, A.D. (2004). Mutualisms and Aquatic Community Structure: The Enemy of My Enemy Is My Friend. Annu. Rev. Ecol. Evol. Syst. *35*, 175–197.

Hindson, B.J., Ness, K.D., Masquelier, D.A., Belgrader, P., Heredia, N.J., Makarewicz, A.J., Bright, I.J., Lucero, M.Y., Hiddessen, A.L., Legler, T.C., et al. (2011). High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. Anal. Chem. *83*, 8604–8610.

Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., and Armbrust, E.V. (2012). Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. Science *335*, 587–590.

Jarvis, K.G., Girón, J.A., Jerse, A.E., McDaniel, T.K., Donnenberg, M.S., and Kaper, J.B. (1995). Enteropathogenic Escherichia coli contains a putative type III secretion system necessary for the export of proteins involved in attaching and effacing lesion formation. Proc. Natl. Acad. Sci. *92*, 7996–8000.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., Stocker, R., et al. (2014). Single-Cell

Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. Science *344*, 416–420.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell *161*, 1187–1201.

Lan, F., Haliburton, J.R., Yuan, A., and Abate, A.R. (2016). Droplet barcoding for massively parallel single-molecule deep sequencing. Nat. Commun. *7*, 11784.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Marcy, Y., Ouverney, C., Bik, E.M., Lösekann, T., Ivanova, N., Martin, H.G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D.A., et al. (2007). Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc. Natl. Acad. Sci. U. S. A. *104*, 11889–11894.

McConnell, M.J., Lindberg, M.R., Brennand, K.J., Piper, J.C., Voet, T., Cowing-Zitron, C., Shumilina, S., Lasken, R.S., Vermeesch, J.R., Hall, I.M., et al. (2013). Mosaic Copy Number Variation in Human Neurons. Science *342*, 632–637.

Monte, U.D. (2009). Does the cell number 109 still really fit one gram of tumor tissue? Cell Cycle *8*, 505–506.

Nathan, C., and Cars, O. (2014). Antibiotic Resistance — Problems, Progress, and Prospects. N. Engl. J. Med. *371*, 1761–1763.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Ni, X., Zhuo, M., Su, Z., Duan, J., Gao, Y., Wang, Z., Zong, C., Bai, H., Chapman, A.R., Zhao, J., et al. (2013). Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. Proc. Natl. Acad. Sci. U. S. A. *110*, 21083–21088.

Potter, N.E., Ermini, L., Papaemmanuil, E., Cazzaniga, G., Vijayaraghavan, G., Titley, I., Ford, A., Campbell, P., Kearney, L., and Greaves, M. (2013). Single-cell mutational profiling and clonal phylogeny in cancer. Genome Res. *23*, 2115–2125.

Preston, A., Mandrell, R.E., Gibson, B.W., and Apicella, M.A. (1996). The lipooligosaccharides of pathogenic gram-negative bacteria. Crit. Rev. Microbiol. *22*, 139–180.

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., and Woyke, T. (2014). Obtaining genomes from

uncultivated environmental microorganisms using FACS–based single-cell genomics. Nat. Protoc. *9*, 1038–1048.

Romero, P.A., Tran, T.M., and Abate, A.R. (2015). Dissecting enzyme function with microfluidic-based deep mutational scanning. Proc. Natl. Acad. Sci. *112*, 7159–7164.

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Schnall-Levin, M., Zhang, H., Basu, A., Bernstein, B.E., and Weitz, D.A. (2015). High-Throughput Single-Cell Labeling (Hi-SCL) for RNA-Seq Using Drop-Based Microfluidics. PLoS ONE *10*, e0116328.

Roxane Maranger, and David Bird (1995). viral abundance in aquatic systems: a comparison between marine and fresh waters. Mar Ecol Prog Ser *121*, 217–226.

Sidore, A.M., Lan, F., Lim, S.W., and Abate, A.R. (2016). Enhanced sequencing coverage with digital droplet multiple displacement amplification. Nucleic Acids Res. *44*, e66–e66.

Yancopoulos, G.D., Blackwell, T.K., Suh, H., Hood, L., and Alt, F.W. (1986). Introduced T cell receptor variable region gene segments recombine in pre-B cells: Evidence that B and T cells use a common recombinase. Cell *44*, 251–259.

Yildiz, F.H. (2007). Processes controlling the transmission of bacterial pathogens in the environment. Res. Microbiol. *158*, 195–202.

Zhang, H., and Liu, K.-K. (2008). Optical tweezers for single cells. J. R. Soc. Interface *5*, 671–690.

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 Nov 17, 2016