Virtual microfluidics: a novel single-cell technology based on diffusion-restricted reaction that makes high-quality low-input genomic research accessible

by

Liyi Xu

B.S. with honors, University of California, Berkeley (2011)

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

Signature redacted

Author .....
Department of Biological Engineering
Sept 26, 2017

Signature redacted

Certified by...
Paul C. Blainey
Associate Professor of Biological Engineering
Signature redacted Thesis Supervisor

Accepted by .....
Mark Bathe
Associate Professor of Biological Engineering
Chairman, Graduate Program Committee

This doctoral thesis has been examined by a Committee of the Department of Biological Engineering as follows:

Signature redacted

Professor Angela Koehler .

Chairman, Thesis Committee
Karl Van Tassel (1925) Career Development Professor of Biological Engineering

Signature redacted

Professor Paul C. Blainey....

Thesis Supervisor
Associate Professor of Biological Engineering

Signature redacted

Professor Linda G. Griffith...

Member, Thesis Committee
S.E.T.I. Professor of Biological Engineering and Mechanical Engineering

Virtual microfluidics: a novel single-cell technology based on diffusion-restricted reaction that makes high-quality low-input genomic research accessible

by

Liyi Xu

Submitted to the Department of Biological Engineering
on Sept 26, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The extensive genomic diversity of complex systems, such as the human gut microbiome and the evolution of human cancer, has been revealed with advances in DNA sequencing. But we are still at an early stage in understanding this genomic diversity to expand our knowledge in biology and for biomedical applications. Taking the diverse human gut microbiome as an example, little is known about the rapid exchange of antibiotic resistance genes and virulence factors as part of the mobile gene flow between the microbes in the gut.

Understanding such heterogeneous systems often involves studying the nature and behavior of the individual cells that constitute the system and their interactions. However, it is technically challenging to probe the genomic material of cells, the smallest unit of life and amplify single genomes for sequencing. Current single-cell technologies require complex instrumentation and the data quality is often confounded by biased genome coverage and chimera artifacts. We address these challenges with a new single-cell technology paradigm to make high-quality low-input genomic research accessible to scientists.

We developed hydrogel-based *virtual microfluidics* as a simple and robust platform for the compartmentalization of nucleic acid amplification reactions. We applied whole genome amplification (WGA) to purified DNA molecules, cultured bacterial cells, human gut microbiome samples, and human cell lines in the *virtual microfluidics* system. We demonstrated whole-genome sequencing of single-cell WGA products with excellent coverage uniformity and markedly reduced chimerism compared with traditional methods. Additionally, we applied single-cell sequencing to identify horizontally transferred genes between the microbes in the gut and revealed human population activities' selective pressure in shaping the mobile gene pools.

Altogether, we expect *virtual microfluidics* will find application as a low-cost digital assay platform and as a high-throughput platform for single-cell sample preparation. This work offers a significant improvement in making high-quality low-input

genomic research accessible to scientists in microbiology and oncology.

Thesis Supervisor: Paul C. Blainey
Title: Associate Professor of Biological Engineering

Acknowledgments

This work would not have been possible without the help and support of many people.

First and foremost, I would like to thank my thesis advisor, Prof. Paul Blainey. Paul took a chance on me when I joined the lab as his first student and he has shown continuous support and trust in my ability to thrive professionally throughout the years. I am grateful to the current and former members of Blainey lab for their scientific support, constructive feedback and positive attitude in lab. Special thanks to Georgia Lagoudas, who shared her wisdom and grace with me over the past 5 years. Thanks to Jacob Borrajo, Anthony Kulesa, Navpreet Ranu, Dr. Soohong Kim, Atray Dixit, and David Feldman for many scientific and philosophical discussions. And thanks to Anthony Garrity and Francis McCarthy for encouraging me along the way.

I thank my thesis committee chair, Prof. Angela Koehler. Angela mentored me with patience as I developed my thesis, and has showed she cares a great deal in my success as a graduate student. I am grateful for her valuable time and advice in scientific delivery and professional development.

I am very grateful to Prof. Linda Griffith and members of Griffenburger lab who welcomed me into the lab and generously shared their knowledge and experience in gel chemistry and microscopy, which were fundamental in helping me get my projects started. I am especially grateful to have Hsinhwa Lee, Dr. Edgar Sanchez and Dr. Jorge Valdez, who have helped me think positively about my PhD.

I thank my collaborators, Prof. Ilana Brito (Cornell) and Prof. Eric Alm. Their dedication and perseverance in delivering the best scientific story helped me tremendously in getting the publication out.

I am very grateful to the Department of Biological Engineering for giving me the amazing opportunity to work on my thesis at MIT. I also thank the Broad Institute and the Lawrence Summers Fellowship for supporting me.

Finally, for helping me remain sane outside of the lab, I thank my friends and family for encouraging me throughout the years.

Contents

1 Introduction **15**

1.1 There is an increasing scientific and biomedical need for low-input nucleic acid analyses 15

1.1.1 Characterizing unculturable microorganisms requires single-cell sequencing 15

1.1.2 Pinpointing the human microbiome's therapeutic mechanism requires the help of single-cell sequencing technologies 18

1.1.3 Deciphering oncogenesis and tumor heterogeneity requires single-cell analysis 21

1.1.4 Clinical application of low-input DNA analysis requires a high-sensitivity single-molecule technology 24

1.2 State of the art of single-molecule and single-cell technologies 25

1.2.1 Single-molecule analysis of nucleic acids 25

1.2.2 Single-cell analysis for genomic studies 26

1.3 ***Virtual Microfluidics*** for digital quantification and single-cell sequencing 33

2 *Virtual Microfluidics*: a hydrogel-based system for simple and robust DNA digital quantification using *in situ* amplification **37**

2.1 Introduction 37

2.2 Results and Discussion 40

2.2.1 Digital PCR in-gel characterization 40

2.2.2 Digital MDA in-gel characterization 41

2.2.3 DNA amplification cluster analysis 43
2.2.4 Dynamic range of in-gel digital MDA 44
2.2.5 Analysis of reaction extent limitation and local competition among MDA clusters 44
2.3 Conclusion 46
2.4 Materials and Methods 46
2.4.1 PEG hydrogel cross-linking 46
2.4.2 In-gel digital PCR 47
2.4.3 In-gel digital MDA 47
2.4.4 In-gel real-time dMDA 48
2.4.5 Image acquisition and analysis 49

3 *Virtual Microfluidics* enables high-quality single-cell sequencing from a mixed population of cultured bacteria and the human gut microbiome 51

3.1 Introduction 51
3.2 Results and Discussion 52
3.2.1 In-gel single *E. coli* MDA 52
3.2.2 In-gel single-microbe MDA - cultured *E. coli* and *S. aureus* . . 52
3.2.3 In-gel single-microbe MDA - human gut microbiome samples . 57
3.2.4 Random Dispersion Model 59
3.3 Conclusion 60
3.4 Materials and Methods 61
3.4.1 In-gel single-microbe MDA - *E. coli* and *S. aureus* 61
3.4.2 Image acquisition and analysis 62
3.4.3 MDA product cluster retrieval 63
3.4.4 BLAST analysis and read assignment for *E. coli* and *S. aureus* 63
3.4.5 Secondary liquid MDA and PCR screening - cultured *E. coli* and *S. aureus* 65
3.4.6 WGS library construction and sequencing 65

3.4.7 NGS data analysis for *E. coli* and *S. aureus* 66
3.4.8 Custom reference generation by de novo assembly for *E. coli* and *S. aureus* 67
3.4.9 Chimera statistics for *E. coli* and *S. aureus* 69
3.4.10 In-gel single-microbe MDA - human gut microbiome samples . 71

4 The characterization of chimeric DNA rearrangements in single amplified human genomes across innovative single-cell technologies 75
4.1 Introduction 75
4.2 Results and Discussion 79
4.2.1 Chimera categories 81
4.2.2 Chimera rate analysis with respect to mapping quality filtering 84
4.2.3 Coverage uniformity and physical genome coverage performances 89
4.3 Conclusion 90
4.4 Materials and Methods 91
4.4.1 Experimental methods 91
4.4.2 Bioinformatic methods 92

5 Conclusion and future directions 95
5.1 Summary of advancements 95
5.1.1 Enabling equipment-independent high-throughput DNA target detection 95
5.1.2 Improving the whole genome sequencing data quality and success rate for characterizing uncultured microorganisms 96
5.1.3 Reducing structural variation artifacts for studying human cells using single-cell sequencing 97
5.2 Future directions 98
5.2.1 Future technical improvements of *virtual microfluidics* 98
5.2.2 The future of single-cell whole genome sequencing 100

A NCBI accession numbers 103

A.1 Bacterial single cells 103
A.2 Human single cells 103

List of Figures

1-1 An overview of single-cell whole genome sequencing 27

1-2 Conventional methods of digital quantification and single-cell sequencing vs. *virtual microfluidics* 34

1-3 A graphical abstract of the *virtual microfluidics* technique 35

2-1 The *virtual microfluidics* hydrogel structure. 38

2-2 DNA amplification clusters from PCR in hydrogel in capillary tubes. 40

2-3 Digital single-molecule MDA in hydrogel (Lambda phage DNA). . . . 41

2-4 Real time dMDA and MDA cluster size 42

2-5 Cluster size and location correlation analysis 45

2-6 DNA amplification clusters from PCR in frame-seal chambers. 47

2-7 Digital single-molecule PCR in hydrogel (Lambda phage DNA). . . . 48

2-8 Imaging analysis for a *E. coli* MDA cluster and a mammalian genome 49

3-1 Single-cell MDA on *Escherichia coli* 53

3-2 Single-cell whole genome sequencing from *E. coli* and *S. aureus* hydrogel WGA samples. 54

3-3 MDA chimera frequency with different insert sizes, *E. coli* and *S. aureus* data 57

3-4 Fiji microbiome project single-cell whole-genome sequencing. 58

3-5 NGS data analysis schematic for *E. coli* and *S. aureus* 64

3-6 Mapping single-cell genomes to references 67

4-1 The mechanism of MDA chimera 76

4-2 Comparison of single-cell technologies 79
4-3 Pair-ended sequencing for the chimera categorization. 81
4-4 Chimeric breakpoints per 10 kbp 82
4-5 The effect of mapping quality filtering on chimera detection. 84
4-6 The effect of pair-ended and single-ended mapping on chimera detection 85
4-7 Hierarchical clustering of chimera breakpoints per 10 kbp. 86
4-8 Chimera breakpoints shown in Circos plots (part 1) 87
4-9 Chimera breakpoints shown in Circos plots (part 2) 88
4-10 The coverage uniformity and the genome coverage performance 89
4-11 Bioinformatic workflow for chimera analysis 93

List of Tables

1.1 A comparison of single-cell isolation technologies 28

3.1 QPCR characterization of hydrogel punches 53
3.2 Sequence read classification of "other reads" 56
3.3 Overview of 117 FijiCOMP single-cell hydrogel samples 59
3.4 Microbe occurrence probability 60
3.5 PCR primer sequences 65
3.6 Mapping statistics, *E. coli* and *S. aureus* 68
3.7 *de novo* assembly statistics, *E. coli* and *S. aureus* 69
3.8 Downsampling on mapped reads from single-cell MDA samples 70
3.9 Metagenomic shotgun profiling weighted with single-cell samples . . . 72

4.1 Single-cell technology comparisons for chimera analysis 78
4.2 Data source for chimera analysis 80

A.1 SRA accession numbers for human single-cell chimera analysis 104

Chapter 1

Introduction

1.1 There is an increasing scientific and biomedical need for low-input nucleic acid analyses

Applications from microbial genome discovery to biomedicine[1, 2, 3, 4, 5, 6, 7, 8] are driving the broader application of high-throughput analyses of nucleic acids at the level of single molecules and single cells. In this introduction, I will describe the specific scientific and biomedical needs for single-molecule and single-cell analyses in the characterization of unculturable microorganisms, human microbiome research, cancer research, and clinical diagnosis. Following that, I will discuss the related technology landscape and ideal technology features for such analyses. Lastly, I will preview my technology, *virtual microfluidics*, which is the focus of this thesis. I will explain how this technology fits in the emerging single-cell field and discuss its potential to provide significant improvement and value in enabling easily accessible, high-quality, and low-input genomic research to a large scientific and biomedical field.

1.1.1 Characterizing unculturable microorganisms requires single-cell sequencing

Microbial communities (including bacteria and archaea) and their globally distributed networks are an essential part of human life on earth. They are highly diverse, with an

estimated 10^7 prokaryotic species [9]. These microbes play critical roles in ecosystems that break down environmental toxins, produce fermentated dairy products at an industrial scale, and associated with human health. For example, *D. radiodurans* was first discovered in soil as the most radiation-resistant organism known. It was then engineered with toluene dioxygenase to survive in highly radioactive waste sites and decompose hazardous chemicals such as toluene and chlorobenzene [10]. Also, it has been shown that long-term antibiotic treatment causes prolonged shifts in the gut microbial composition, which decreases amyloidβ plaque deposition and is associated with the Alzheimer's disease in humans [11]. Despite their importance, $90 \sim 99\%$ of microbes on earth have not been characterized because they are difficult to culture in the lab [12]. Studying the uncharacterized microorganisms can open the door to a huge reservoir of knowledge on microbial functions and gives us the ability to predict microbial responses to perturbations from human activities.

The highly diverse nature of environmental microbes, however, requires the characterization of the microbe not only collectively but also in isolation. Studying microbes in isolation is needed to decode their genome sequences, understand gene functions and microbes' relationship with different components of its community. Traditional methods of isolating a single species in culture are not only labor-intensive and slow, but they also require serial enrichment under culture conditions that are often suboptimal for microbes' growth [13]. Such a prolonged procedure increases the risks of genomic changes in the microbe of interest and may result in diversity loss due to competition or simple unculturability. Meanwhile, optimizing the culture conditions for a large number of diverse species of bacteria is difficult. Understanding the metabolic mechanism of an uncharacterized microbe could shed light on an axenic culture in the lab and in many cases, a symbiotic co-culture might be necessary [14].

To address these issues, the recent development of next-generation sequencing techniques has enabled us to discover a large number of microorganisms [15, 16, 17]. Specifically, these sequencing techniques have allowed scientists to identify the functional pathways and phylogenies of newly characterized microorganisms [7]. Furthermore, the evolutionary history and relationships among prokaryotic species are

developed based on marker genes, such as 16S sequences, gene panels and whole genome sequencing, instead of solely relying on qualitative morphology observations.

Recent studies have discovered new variations of genetic code encoding for amino acids. Across the diversity of microbes, the genetic code for amino acids has been shown to vary in size and codon assignment[18]. UGA is a common stop codon. A novel coding of glycine UGA was identified in newly discovered Gracilibacteria and SR1 [7, 19]. These discoveries broaden our definition of how to encode life's basic building blocks - amino acids. New phyla are also discovered from microbial samples. First discovered in human oral cavities [20] and soil [21], TM7 was one of the first elusive candidate phyla sequenced. The discovery shed light on potential virulence factors in TM7 that may contribute to oral disease [20]. Sequencing data from hospital sink biofilms was used to reconstruct partial genomes of a new candidate phylum TM6 and the periodontal pathogen *Porphyromonas gingivalis* [22, 23]. It also enabled the identification of key virulence factors and polysaccharides biosynthesis pathways that are proposed targets for new antibiotics.

Common sequence-based analysis methods, including metagenomic and single-gene studies on environmental microbial samples, can only provide a fragmented view of the common species present [13]. Metagenomic shotgun sequencing provides sequencing results in billions of short fragments (commonly 100 bp for short-read sequencing on Illumina machines). This method enables microbiologists to sample genes and detect the abundance of microbes in various environments. However, due to its fragmented nature, metagenomics alone is not suitable to assign genes to different phylogenetic groups. With extensive bioinformatic manipulations, it has been shown that metagenomic reads can be assigned into biologically informed bins based on the covariance of the k-mer read depth and thereby enables assembly of "individual genomes" [24]. Sequence reads alignment to reference genomes is another way to achieve binning. But the binned genomes are "composite genomes" that could be comprised of many different organisms. Even with long read sequencing technologies, such as PacBio and Nanopore, which allow sequencing reads of up to 400 kbp, there are still discontinuous assemblies with gaps that need to be bridged in order to

link genes and predicted pathways to a microorganism's phylogeny [25]. In addition, the current reference genome database is highly biased with culturable bacteria. Sequences from rare species can be easily discarded due to the difficulty in getting a mapping hit to the reference database.

Single-cell sequencing can phylogenetically connect unlinked metagenomic reads by cell of origin [7]. Single-cell assemblies have also revealed the microbial family Succinivibrionaceae is highly abundant in the gut microbiome but is difficult to detect by metagenomic analysis due to its poor representation in reference databases [26]. Directly, single-cell whole genome sequencing alone can yield a large number of *de novo* assembled genomes. But current single-cell technologies produce assemblies that are often confounded by biases and chimeras generated through the whole genome amplification process, resulting in mis-assemblies and low coverage completeness [27, 28]. A need exists for technologies that improve data quality from single-cell datasets in order to produce confident, finer-scale heterogeneity among microbial samples that are largely uncharacterized.

1.1.2 Pinpointing the human microbiome's therapeutic mechanism requires the help of single-cell sequencing technologies

The human body is also comprised of many uncharacterized microorganisms, which are of the same order as the number of human cells [29]. Among the many microbiome niches on and inside the human body, the gut microbiome has been shown to play a critical role in affecting human developmental variations [30] and modulating the host immune system [31]. There also has been increasing evidence indicating that bacterial microbiota plays a key role in carcinogenesis [32] and a wealth of studies in patients and mice have linked the microbiota to colorectal and lung carcinogenesis [33, 34]. In short, there are tremendous opportunities to leverage an understanding of the microbiome for diagnostic and therapeutic appictions in healthcare. To date, however, the majority of human microbiome research and clinical efforts have focused

on culturing consortia of microbes without knowing the individual microbe's contribution. It is difficult to predict such consortia's therapeutic efficacy or adverse effect if used to treat humans.

As one example, *Clostridium difficile* infection (CDI) caused around 500,000 incidences in the U.S. in 2011, with a mortality rate of 3% $\sim$ 4%. The cost of managing CDI was estimated to be at least $ 1 billion per year in the U.S. alone since 2010 [35, 36]. Remarkably, though, the basic pathophysiology of recurrent CDI is not completely understood. *C. difficile* is not pathogenic at low levels. Broad-spectrum antibiotics disrupt patients' gut microbial communities that normally keep *C. difficile* population in check. *C. difficile* spores can remain dormant during the antibiotic treatment and take over after treatment ends to proliferate, disrupt and cause inflammation in the gut by secreting toxins that damage the gut endothelial lining [37]. As a result, the current treatment uses Fecal Microbiota Transplantation (FMT) to reconstitute normal microbial homeostasis [38], but our biomedical understanding of FMT is limited. For instance, it is challenging to select a donor that guarantees a safe and efficacious FMT [39]. Fundamentally, scientists lack a complete understanding of FMT from a basic science perspective, and the key microbial populations that are responsible for beneficial outcomes and adverse effects remain unknown. A metagenomic sequencing study [40] has shown that the relative abundance of assembled composite genomes from the donor did not predict whether the microbes would colonize the FMT recipient or not. But the study showed a link between taxonomy and the colonization ability of a given assembled strain, while assembled strains from the same taxon have slightly different colonization properties. This highlights the importance of a better resolution in exploring the functional basis of FMT colonization and identifying precisely what microbes need to be transferred to maximally benefit a patient.

Single-cell whole genome sequencing is precisely the kind of technology that could help answer scientific questions in designing therapeutic tools using complex biological systems. Sequencing single cells from the targeted taxon could provide individually assembled genomes with strain-level resolution. These genomes are separate ecologi-

cal machines that contain genes related to sporulation that correlate with colonization efficiency [41]. Pinpointing the strains of microbes with high colonization efficiencies would likely guide the engineering of FMT with high manufacturing efficiency, high efficacy and low adverse effects. However, current single-cell sequencing technology is difficult to implement on a large number of gut microbes (with a density of 10^{11} cells/mL). It is also difficult to select the targeted taxon of interest from the gut microbiome consisting of microbes of different levels of unculturabilities. These challenges explain the current low adoption rate of single-cell technologies in the gut microbiome research. I have taken steps towards making such technology improvements in throughput, ease of implementation, optical screening and retrieval capability of microbes. This would enable a large fraction of the human gut microbiome to be analyzed with single-cell resolution to pinpoint its therapeutic mechanism.

Another class of problems that single-cell whole genome sequencing could address relates to the horizontal gene transfers (HGT) in the human gut microbiome. HGT is the acquisition of genetic materials (such as plasmids, transposons, prophages) from non-parental lineages. It allows rapid exchange of virulence factors [42], antibiotic resistance genes [43, 44], and xenobiotic metabolism genes [45] through the human gut microbiome. A better understanding of the distribution of antibiotic-resistance genes among different microbiomes could inform antibiotics overuse where the resistance of specific antibiotics is the highest [44]. Studying the mobile gene pool associated with HGT, however, is difficult with current short-read metagenomics sequencing technologies. Previous studies have been constrained to individual species [46] and limited mobile elements such as plasmids [47] and phages [48]. A more reliable method for cataloging mobile genes (to include transposons and prophages) depends on assembled genomes that are distantly related (more than 3% divergent) and share genes with exact sequence matches. This requires individual genomes with their phylogenies and genes coupled.

Single-cell technology could solve this problem by producing draft genomes that contain enough genome context to link genes to hosts and identify mobile genes with high confidence. Currently, research labs that are interested in applying single

microbe sequencing have been relying on services provided by large genomics centers, such as the Bigelow Single-cell Genomics Center and the Joint Genomic Institute. The success rate of obtaining pure single amplified genomes is less than 10% using FACS sorting and traditional whole genome amplification methods. In Chapter 3, we propose a new single-cell technology that can provide high-quality data with a much higher success rate (28%).

In summary, utilizing the microbiome for diagnostic and therapeutic tools holds enormous potential. High-resolution spatial, temporal, and functional analyses of the human intestinal microbiota are needed. In order to understand the microbiome's function and develop methods for intervention, it is necessary to implement single-cell whole genome sequencing technologies in academic and clinical settings. Later in this chapter, I will discuss the technology landscape and demand of single-cell sequencing. In Chapter 3, I will demonstrate *virtual microfluidics*' application on cultured microbes and human gut microbiome samples and show how it enables high-quality genomic analysis with ease of accessibility. At the end, I will also discuss the technology's potential to be implemented in the field for environmental microbiome studies.

1.1.3 Deciphering oncogenesis and tumor heterogeneity requires single-cell analysis

Understanding the nature of oncogenesis (how normal cells transform into cancer cells) and tumor genetic heterogeneity (different tumor cells show distinct phenotypic profiles) has been the focus of a continuous research effort for the past several decades [8, 49, 50, 51, 52, 53]. Most cancers carry 1,000-20,000 somatic point mutations and up to hundreds of insertions, deletions, and rearrangements [54, 55]. Cancer mutations' heterogeneous nature across a cell population is a factor of cancer treatment failure and disease recurrence, as the treatment for one tumor cell subpopulation may not work for another [56, 57]. By measuring the mutational heterogeneity of a tumor, researchers and clinicians hope to create targeted treatments and enable better clini-

cal outcome prediction. For example, among patients with non-small-cell lung cancer, elevated copy-number heterogeneity (such as in gene CDK4, FOXA1, and BCL11A) from subsections of a tumor was associated with shorter relapse-free survival (P=4.4×10^{-4}) [58]. This finding suggests that patients who have early-stage tumors with high levels of copy-number heterogeneity may represent a high-risk group who may benefit from close monitoring and early therapeutic intervention during follow-up. Tumor heterogeneity, as quantified by the clonal diversity measure from evolution and ecology, has been shown to predict progression to adenocarcinoma from a premalignant condition in Barrett's esophagus [59]. The Shannon diversity index is calculated as: $H = -\sum_i p_i \ln(p_i)$, where p_i is the frequency of clone i in the sample. A tumor with a high diversity index is expected to become resistant to chemotherapy as it harbors pre-existing resistance mutations [60].

In order to evaluate the clonal diversity, researchers have made qualitative observations on the chromosome aberrations directly in single cancer cells with karyotyping and fluorescence *in situ* hybridization (FISH). Recently, next-generation sequencing has enabled large-scale quantitative analysis [60], but the sensitivity of detection is limited to mutations that are present in about 20% of cells of a bulk sample [61]. In addition, in clinical samples, such as fine-needle aspirates and core biopsy samples, the number of cells is often limited. Single-cell sequencing is an effective solution to improve measurements of the extent of intratumor genomic heterogeneity even with low-input samples. Current single-cell technologies produce an uncharacterized amount of amplification artifacts that confound the profiling of genome-wide mutations (single nucleotide variations, copy number variations, and structural variations). The technology improvements I describe in this thesis will improve the ability to produce high-quality single-cell data with low-level of genome structural artifacts.

Single-cell measurements can also inform our understanding of oncogenesis. The process of mutational events in oncogenesis have two main explanations. The first explanation is that cells can acquire hypermutations (the "mutator hypothesis") [62]. It argues that normal mutation rates are insufficient to account for the multiple mutations observed in cancer cells [63]. Therefore, the hypermutation that increases

mutation rates (such as the p53 mutation that impairs the detection of and response to DNA damage) would account for a large number of mutations in human tumors. The other explanation – "driver mutation" – is that selection without increased mutation rates is sufficient, as the early driver mutations trigger clonal expansions and increase the pool of cells at risk for further driver mutations [64]. Driver mutation denotes mutations under positive selection within a population of cells. Passenger mutations are variants that have no phenotypic consequences.

In order to test the two (not necessarily mutually exclusive) hypotheses, quantifying the mutation rate in normal tissues and in different tumor types is needed. It has been shown by bulk sequencing that normal tissue has a frequency of spontaneous mutations less than 1×10^{-8} per base pair, while tumors from the same individual exhibit an average frequency of 210×10^{-8} per base pair [65]. However, this bulk-sequencing result does not differentiate between the two hypotheses; scientists cannot ascertain whether the high mutation frequency per basepair is due to a small collection of cells having high mutation rates with hypermutation or a majority of cells having driver mutations. Over the years, we have learned that the mutational process of cancer cells as population averages does not represent the mutational landscape because heterogeneous information is hidden in bulk samples [8, 66, 67]. By identifying mutations and mutation rates with single-cell resolution, the nature of hypermutations or driver mutations can be identified to better understand the oncogenesis process. Thus, there is a need for high-throughput, high-quality, low-cost single-cell sequencing methods to catalog and compare the mutation rate of a large number of normal and tumor cells for the study of oncogenesis.

Overall, there is a need for an easily accessible, high data quality single-cell technology to enable the robust measurement of tumor heterogeneity and to push the boundary of our knowledge in oncogenesis. With such a technology improvement, it will be possible to pinpoint cancer mutation mechanisms, to inform targeted cancer therapy for patients, and to bring accessible high-resolution genomic research to a wide scientific and biomedical audience. In Chapter 4, I will demonstrate *virtual microfluidics* on single human cells for whole genome sequencing and compare its

advantages in terms of ease of implementation, high data quality, and low artifact rate with several other recently developed single-cell technologies.

1.1.4 Clinical application of low-input DNA analysis requires a high-sensitivity single-molecule technology

Considerable effort has been spent to translate genomic data into personalized prognoses and treatments [68, 69, 70]. For example, patients with low genomic risks (using a 70-gene signature) during the early stage of breast cancer might not need to undergo chemotherapy. The 5-year survival rate was only 1.5% lower than the rate from the control group undergone chemotherapy [68]. This type of informed clinical decision-making has great potential to improve patients' quality of life and avoid unnecessary medical expenses. Repeated sampling of a tumor is ideally required to track patients' genetic profiles before and after therapies to optimally deliver targeted therapy [71]. A traditional biopsy is invasive and technically challenging depending on the site of sampling, while only sampling a single area of tumor underestimates the array of genetic aberrations in heterogeneous tumors [72]. Researchers have shown that the plasma provides a noninvasive source of tumor DNA for HER2 breast cancer [71, 73]. Other studies have demonstrated the monitoring of chronic myeloid leukemia with BCR-ABL1 fusion quantification in the plasma [74]. Common cancer mutations in KRAS and p53 have also been identified in plasma from patients with colorectal and pancreatic neoplasms [75, 76].

The biggest challenge facing liquid biopsy is the low occurrence of circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs) in the plasma (in the range of 1 cell in 20 mL). Current technologies limit the application to ctDNA in late-stage cancers and the detection rate of different gene targets in different cancer patients varies dramatically (33% $\sim$ 80%) [77, 78]. There is a need for technology improvement in order to achieve absolute quantification of single molecules and single cells from a dilute input. The technology should also enable target enrichment, optical screening of rare DNA fragments/cells, and eventually product retrieval for post-detection sequencing

validation. In the next section of the introduction, I will summarize the state-of-the-art technologies in DNA quantification by amplification, introduce the concept of a hydrogel-based quantification assay, and explain why it has a great potential for the clinical application of low-input DNA analysis. In Chapter 2, I will demonstrate and characterize the absolute quantification of nucleic acids by amplification.

1.2 State of the art of single-molecule and single-cell technologies

Previously, the need and significant impact of single-molecule and single-cell technologies in a wide range of scientific and biomedical contexts was discussed. In this section, I will discuss the current approaches to single-molecule and single-cell analysis with the goal of introducing the benefits and disadvantages of current methods. I will then close with a discussion on the gap between what current technologies offer and what hurdles need to be crossed to achieve a wide adoption of single-molecule and single-cell technologies.

1.2.1 Single-molecule analysis of nucleic acids

Accurate quantification of nucleic acids has been an integral part of biological science and has many applications in clinical research. A new class of 'digital', i.e. absolute quantification, single-molecule analyses require parallel clonal amplification of individual nucleic acid templates—typically a few fragments in milliliters of blood—to generate a sufficient number of genomic replicates for detection [73, 74, 79, 80, 81, 82, 83].

A wide variety of amplification-based approaches have been explored for the microfluidic compartmentalization of single molecules across a large number of small discrete reactors, such as high-density microfluidic arrays [84], engineered lab-on-chip systems [85, 86, 87, 88], and multi-phase micro-droplet systems [89, 90, 91, 92]. These systems provide platforms for performing assays such as digital PCR. Digital PCR (dPCR) has been demonstrated to have the least quantitative bias for measuring a

small fraction of DNA in liquid biopsy and has higher clinical sensitivity compared to traditional PCR-based assays, which rely on relative quantification based on "template standards" [80, 93, 94]. Digital PCR works by diluting the target molecules into many partitions in a microfluidic device or across many droplets, such that each partition has one molecule or less. After the dPCR reaction, one can easily read the number of fluorescent partitions, thus estimating the absolute number of molecules in a sample by counting [82].

However, dPCR requires tedious sample preparation and the dynamic range is restricted by the choice of commercial microfluidic chips or the number of droplets generated [91] as the number of partitions limits the range of molecule counting. Existing methods to conduct dPCR require complex instrumentation and microfabricated consumables that prevent broader deployment of digital assays. It is also difficult to retrieve amplification products from microfluidic chambers or droplets, which are required for follow-up sequencing and minimizing the false-positive rate. Other characteristics desired in single-molecule digital processing systems are resistance to extrinsic contamination, stability under temperature changes and good optical properties for digital readings. In this thesis, I will use a hydrogel-based method to isolate DNA fragments by selective diffusion restriction. It has clear advantages in terms of having a low equipment requirement, a high sample retrieval accessibility, and a low reaction volume with a high throughput.

1.2.2 Single-cell analysis for genomic studies

Single-cell sequence data quality is determined by the purity of cell isolation, DNA denaturation, whole genome amplification, library preparation quality (Fig. 1-1). In this thesis, I focus on the main factors and their effect on data quality: cell isolation (cross contamination, biases), lysis (efficiency, biases), and whole genome amplification (coverage uniformity, errors, artifacts).

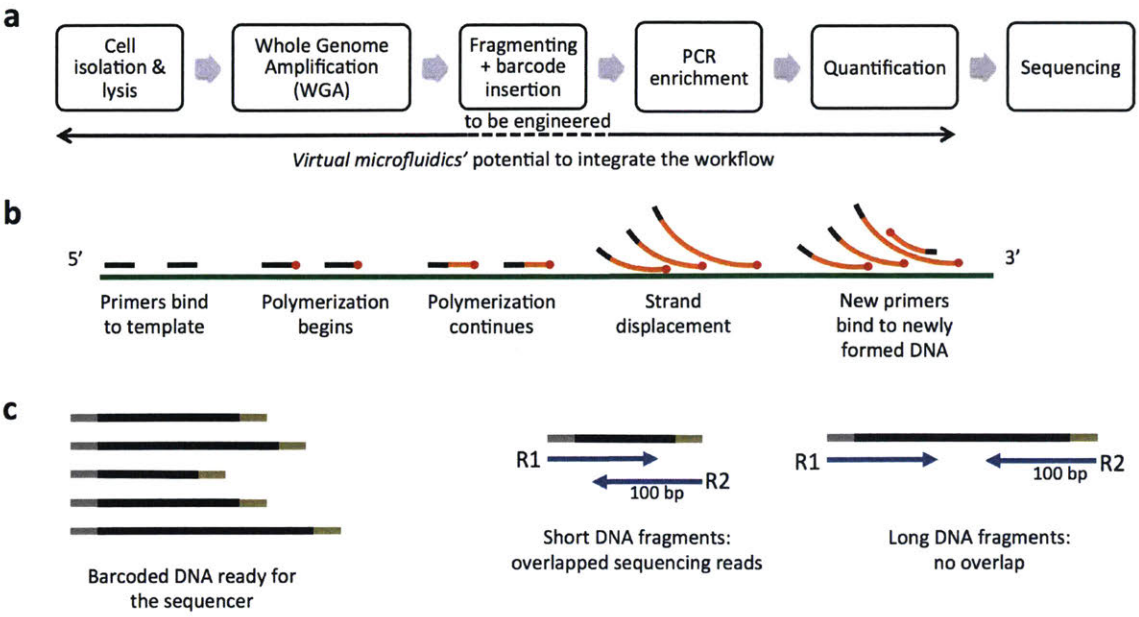

Figure 1-1: An overview of single-cell whole genome sequencing. (a) The single-cell analysis workflow. (b) Multiple displacement amplification (MDA) mechanism. (c) DNA are fragmented, barcoded and sequenced with pair-end sequencing.

Cell isolation and lysis

Like with single-molecule analysis, the first critical step in single-cell analysis is cell isolation. In order to separate cells of interest, relatively high-throughput technologies were adapted and developed. These technologies include fluorescent-activated cell sorting (FACS) into multi-well plates [95], high-density microfluidic arrays [84], engineered lab-on-chip systems [85, 86, 87, 88], and multi-phase micro-droplet systems [89, 90, 92, 96, 97] (see Table 1.1).

Currently, the most commonly used method for cell isolation is FACS. FACS automates the process of single-cell isolation of identifiable subpopulations. The cells being sorted need to have differentiating light-scattering characteristics, express certain fluorescent markers, or to be stained by fluorescent antibodies or DNA binding dyes. Studies have used FACS with DNA binding dyes to detect and localize genetic abnormalities such as whole chromosomal deletions and aneuploidy on single cells [98]. The human cell size (10 μm $\sim$ 20 μm in diameter) and surface markers make it a suitable sample for FACS. However, it is difficult to achieve a similar success

Single-cell Isolation Technology	Segregation Principle	Engineering requirement	Fixed Spatial Addressing	Reagent Addition	Product Recovery	Characteristic Reaction Volume	Max analytes per µL	Max analytes per mm2
SBS multi-well Plate	Macro-scale container array	Commodity plate & complex robotics	Yes	Liquid handler or manual pipetting		1 - 100 µL	1	0.10
Lab-on-a-chip microfluidics (e.g. Fluidigm)	Individually addressable microfluidic chambers	Specialized microdevice & controller		Automated by microdevice		0.1 - 500 nL	10,000	0.10
Open microfabricated array	Micro-fabricated container array			Liquid handler or manual pipetting (largest volume only)		0.05 pL* - 10 nL	20,000,000	20,000
Monodisperse microdroplets	Multi-phase system		No	Droplet merging	Droplet breakdown	5 - 1000 pL	200,000	10,000
Hydrogel	Selective diffusion restriction	No special equipment needed	Yes	Diffusion into or out of hydrogel	Physical punch or hydrogel breakdown	0.05 - 1 pL**	20,000,000	20,000

* Men *et al.*, Anal. Chem., 2012 used 3.3 micron diameter by 4.2 micron deep wells
** Defined by physical extent of product; reagents likely drawn from a larger volume

Table 1.1: A comparison of single-cell isolation technologies

rate when separating environmental microorganisms ($\sim$ 1 μm). Environmental microbial samples have a range of morphological shapes and this may result in a biased selection by light scattering, which contributes to a possible uneven representation of taxa obtained through single-cell analysis using FACS compared to bulk metagenomic sequencing. Sample preparations involving FACS, plate/well transfer, and dilution greatly increase the possibility of exogenous DNA contaminations from the lab environment [99]. In the case of single-cell whole genome amplification, one single molecule of exogenous DNA can be amplified with the random priming mechanism, pose challenges in data analysis and affect the quality of *de novo* assemblies [100].

To remedy the effect of sample handling, a wide variety of approaches have been explored for the microfluidic compartmentalization of single cells across a large number of small discrete reactors [87, 88]. However, similar to the current methods for single-molecule analysis, existing methods require complex instrumentation and are labor intensive if set up in-house. In addition to the characteristics desired as in single-molecule assays, single-cell studies often require the compartment access for the addition and product removal of reagents and samples. A recent development of a microfluidic system that incorporates DNA purifying capability for processing

microbial isolates may solve the reagent addition and retrieval problem but it is not easily accessible and scalable to single-cell resolution [101]. The operation of open microarray/microwells requires a specialized liquid handler such as Echo (Labcyte Inc.) and CellCelector^{TM} (Automated Lab Solutions). Emulsion droplet systems pose challenges in reagent addition and sample retrieval after the droplet formation. This approach is positioned well to produce a large number of picoliter partitions easily for digital counting in a single-molecule analysis that doesn't require follow up Sanger or whole genome sequencing (WGS). A recent development utilizing droplets as partitions for a single human genome WGA still requires FACS sorting or mouth pipetting for cell isolation [89].

The challenges in the single-microbe analysis are distinct from its mammalian counterpart. The difficulty in culturing prevents us from obtaining single colonies on an agar plate that provide enough starting material for sequencing (nanograms of DNA for Illumina benchtop procedures). Here is where single microbe Whole Genome Amplification (WGA) comes into play. WGA can amplify the femtograms of genomic DNA from a single microbe to nanograms (See next section on WGA for more detail). However, single microbial WGA faces numerous challenges, including low isolation efficiency and a high chance of contamination. In addition, microbial communities consist of organisms with diverse physiologies, meaning that a universal lysis strategy that works for all types is difficult. Most lysis methods are developed for bulk samples, which may not be suitable at the single-cell level if the lysis efficiency is low [13]. The undesired consequences of incomplete lysis might result in DNA locus damage and undetected microbial species.

Alkaline lysis is the most common method for single microbe lysis today and it was first described for single cells by Raghunathan *et al.* [102]. Other supplementary methods include heat lysis, repeated freeze-thaw [103, 104]. But the lysis success rates (obtaining pure single amplified genomes after WGA) vary widely and are often below 40% [105]. One particular type of microbes that does not lyse well in alkaline is the environmental extreme, such as *M. ruber* that was found in hot springs with alkaline pH [106]. On the other hand, Fleming *et al.* discussed that the best strategy might be

using a cocktail of enzymes that provide efficient but gentle lysis on the entire microbe types present [104]. But enzymatic cocktail unavoidably contains small fragments of DNA lodged in the enzymes themselves, which become contaminants through the WGA step. The enzymatic cocktail for one single cell in isolation is often excessive as not every enzyme present will be effective in lysing and the enzyme collections might not be compatible with downstream WGA method. And it is difficult to purify such a low quantity of unamplified genomic DNA. Consequently, an effective, adaptable and clean method to enable high-throughput single-microbe isolation and lysis is needed.

Cell isolation and lysis for mammalian cells is less technically challenging due to their relatively large cell size and the absence of the cell wall. The lipid bilayer cell membrane can be easily dissolved in a dilute solution of detergent, such as Triton-100X and NP-40. The enzymatic digestion and DNA deproteination methods have been widely implemented and optimized [66, 89, 107, 108, 109]. However, it was recently discovered that many single-cell studies could be confounded by the poor data quality caused by the DNA damage after cell lysis when biological "variations" are in fact extensive technical biases and errors [108, 110]. For sensitive applications such as measuring the single-nucleotide variations (SNVs) in cancer cells, treating the genomic DNA with uracil-DNA glycosylase to eliminate cytosine-deaminated uracil bases can reduce a significant amount of false-positive C-to-T SNVs [108, 111].

Whole Genome Amplification (WGA)

One of the most critical steps to analyze genomes of single cells is the whole genome amplification. Currently, the prevalent short-read library preparation and sequencing technologies require nanograms (10^{-9} g) of input DNA for on bench procedures, while the genome of a single cell ranges from femtograms (1.7×10^{-15} g, *Prochlorococcus* MED4) to picograms (6×10^{-12} g, human diploid). Thus, WGA is needed to replicate the genomic DNA from a single-cell to approximately $10^3 \sim 10^6$ fold. During the WGA process, artifacts and technical errors such as low physical genome coverage, non-uniform coverage due to GC% bias, false-positive (FP) errors and false-negative errors are often introduced. Achieving a high physical coverage of the genome with

a low error rate is crucial for calling mutations accurately at the same regions of multiple single cells. There is a need for WGA method improvement that provide high-quality single-cell WGA data in above-mentioned criteria.

Multiple Displacement Amplification (MDA) [109] is a well-characterized WGA method commonly used to enable single-cell genome sequencing [20, 89, 95, 102, 112, 113, 114]. MDA uses Φ29 polymerase with a strong strand displacement property and random exonuclease-resistant 6 bp primers to produce longer than 12 kbp amplification products at 30 °C (Fig. 1-1b). MDA provides greater genome coverage than the PCR-based methods such as degenerate oligonucleotide primed PCR (DOP-PCR) and lower error rate than *Taq* and *Bst* polymerases owing to the high fidelity of Φ29 polymerase [109]. However, the exponentially amplified genome through MDA has regions that are overrepresented and this bias positively correlates with the fold of amplification [87]. In addition, the random priming nature allows DNA amplification on any exogenous DNA contaminants, posing a threat in raising the false-positive rate in new genome discovery.

Recent technology innovations have been focused on improving MDA performance by varying methods of physical partitions. It has been shown that contaminating DNA was largely eliminated by moving MDA from microliter reaction volumes in tubes to a microfluidic format that used nanoliter volumes [87]. The MDA amplification gain is also limited by the nanoliter volume, thus improving its coverage uniformity. The trend of using sub-microliter partitions lead to the development of emulsion WGA (eWGA) [89] and Nanodrop MDA [107]. In eWGA, a single cell is FACS sorted and lysed, then randomly distributes in a large number (10^5) of picoliter droplets. Each droplet contains zero to a few fragments of DNA that go through MDA reaction. The results showed improved uniformity and high coverage at the amplification gain of 2×10^6 for human cells. This improvement is made possible by isolating different parts of the genome during MDA. Thus, this way, the over-amplified fragments would not compete globally with fragments in different droplets that have a late start in MDA. This results in a more uniform amplification depth that improves copy number variations calling accuracy. Nanodrop MDA utilizes commer-

cially available piezoelectric non-contact liquid dispenser to deposit nanoliter-ranged drops sequentially onto a planar substrate and each drop's single-cell occupancy depends on Poisson loading. MDA reagents are added to drops containing single cells in the same way and each 100 nL reaction is covered with mineral oil to prevent evaporation. Both eWGA and Nanodrop methods have shown improvements in terms of data quality in genome coverage percentage and coverage uniformity. But both methods still rely on FACS sorting or liquid dispensing and the rate of artifacts formed in MDA was not characterized. High level of chimeric DNA rearrangements (artifacts) can lead to inaccurate genome structural variation analysis but are rarely analyzed in single-cell technology development. We will discuss this subject in depth in Chapter 4 and considers how it can impact microbial *de novo* assembly, investigating mutational mosaicism in neurons, and single-cell cancer research.

In addition to improving the experimental setup for MDA, efforts have been made to develop new WGA chemistry that linearly amplifies the genome to reduce biased coverages the dependence on complex instruments [66, 108]. The most recent method Linear Amplification via Transposon Insertion (LIANTI) utilizes direct transposon insertion and *in vitro* transcription to linearly amplify RNA copies of the genome. This method eliminates the random non-specific priming used in traditional MDA method and has shown to improve coverage uniformity and reduce the false-positive rate in calling single nucleotide variations. The lower error rate is due to linear amplification's random error position on the same template. By sequencing single cells to a high depth (30×), amplification error would be corrected, which leads to fewer false positive calls. LIANTI has significant potential for wide adoption as the next generation of WGA method. According to the analysis in Chapter 4, a majority (70%) of LIANTI's sequencing reads contain barcode information and poor quality reads. A large portion of sequencing effort is wasted as a result ($2500 × 70% = $1750 wasted per cell in sequencing cost). More work needs to be done to integrate the transposon barcode insertion and its downstream library preparation.

After WGA, purified DNA will go through library preparation - fragmentation and barcode insertion. Accurately purified and quantified libraries, in terms of both

size ranges and quantities, will be loaded onto the sequencer (Fig. 1-1c).

Overall, the single-cell sequencing process from having cells in suspension to obtaining sequencing data is highly fragmented in terms of technology implementation. eWGA and LIANTI have to rely on traditional FACS sorting or mouth pipetting while Nanodrop requires sequential depositions for both cell isolation and reagents addition. Efforts have been made to integrate library preparation in microfluidic chips [101] and to package cell isolation and WGA in a commercially available droplet system (10X Genomics). Both examples are either cumbersome to implement or expensive to purchase. For 10X Genomics, for instance, it costs $50K per machine. Researchers working in microbial discovery and oncology studies are looking for ways to conduct single-cell sequencing with an economically reasonable and technically manageable method for a large number of single cells. Independent innovations addressing different stages of single-cell sequencing will push the field forward in small steps. Ideally, an integrated approach that provides an accessible platform to streamline the process while produces high-quality data is needed to realize single-cell sequencing's huge potential in scientific discovery and the biomedical field.

1.3 *Virtual Microfluidics* for digital quantification and single-cell sequencing

In the introduction, we reviewed the increasing scientific and biomedical need for single-molecule and single-cell analysis and the current technology landscape. Much improvement is needed to bridge the gap between the current technology state and the wide adoption of single-cell sequencing The necessary improvement areas I have discussed include high-throughput, the data quality, the ease of implementation, optical enrichment properties and the process integration. Thus, I developed the method *virtual microfluidics* to make an improvement in above-mentioned areas and to help push the single-molecule and single-cell analysis fields forward (Fig. 1-2).

Single-molecule and single-cell studies require individual molecules or cells sep-

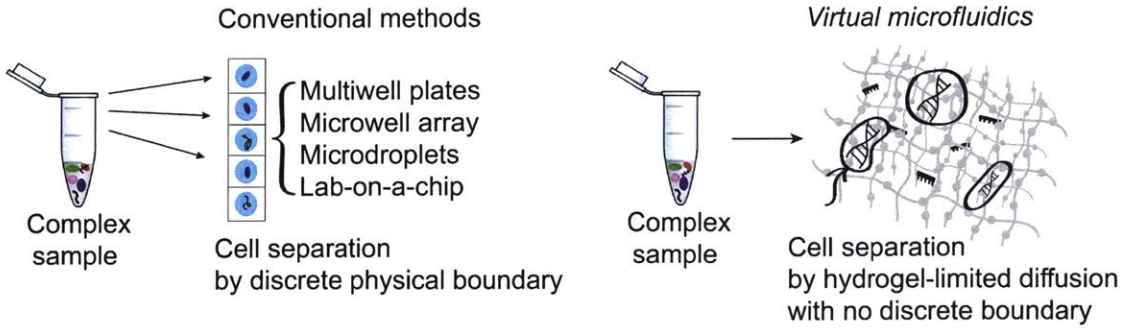

Figure 1-2: Conventional methods of digital quantification and single-cell sequencing vs. *virtual microfluidics*. Conventional methods require discrete physical boundaries. *Virtual microfluidics* relies on hydrogel-limited diffusion to compartmentalize templates and reaction products.

arated in partitions. But instead of creating physical partitions to isolate cells, we decided to build "invisible" dividers around them. Passive segregation of single molecules and the product of genome amplification would be able to facilitate the genomic analysis of thousands of nucleic acids in parallel. In order to achieve this, a porous structure that is able to fix many molecules and provide an aqueous environment for DNA amplification was chosen. Thus, the *virtual microfluidics* system was created to enable single-cell WGA *en masse* in polyethylene glycol (PEG) hydrogel. The hydrogel properties allow reagent exchange by diffusion, imaging accessibility, product retrieval easiness and improved WGA performances using MDA compared to instrumentation-based methods. *Virtual microfluidics* is also a versatile and compatible platform to integrate PCR, MDA and in-gel library preparation methods, which holds great promise in integrating single-cell analysis field and pushing the wider adoption of the technology (Fig. 1-3).

In Chapter 2, I will characterize *virtual microfluidics* in quantifying DNA targets for single-molecule analysis and demonstrated its improvement in the ease of implementation and its wide dynamic range in quantifying DNA targets. In Chapter 3, *virtual microfluidics* is applied to mixed cultures of bacteria and the human gut microbiome. It produces single amplified genomes with excellent coverage uniformity and markedly reduced chimerism compared with liquid MDA reactions. We demonstrate

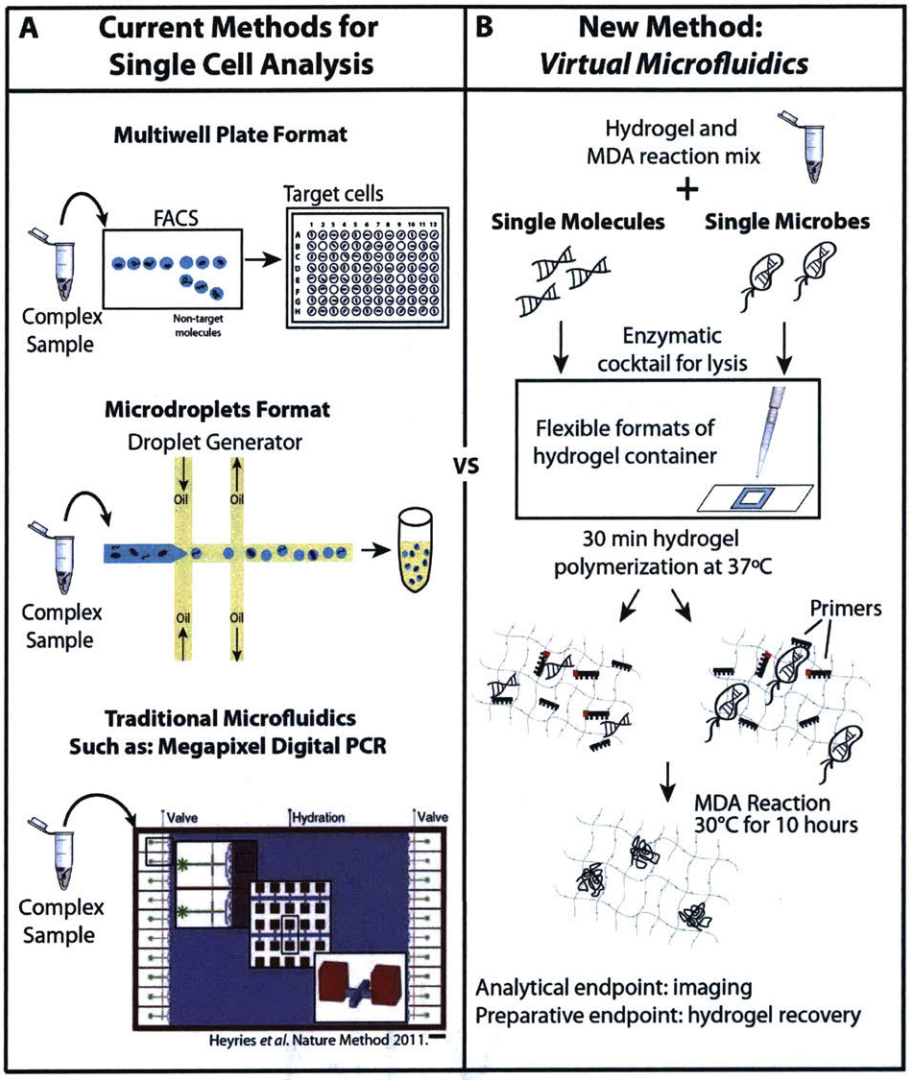

Figure 1-3: A graphical abstract of the *virtual microfluidics* technique.

single-cell sequencing on human gut microbiome samples and obtain 117 pure single draft genomes that enable the identification of more than 10,000 horizontally transferred genes that have unique population-specific and individual-specific features [44]. In Chapter 4, I will show how *virtual microfluidics* reduces the amount of chimera artifacts from MDA on single amplified human genomes compared to above-mentioned technologies.

The results described in Chapters 2 and 3 have been previously published as Xu *et al.* Nature Methods. 2016. The single-cell dataset generated from the human gut microbiome contributed to the publication of Brito *et al.* Nature. 2016.

Chapter 2

Virtual Microfluidics: a hydrogel-based system for simple and robust DNA digital quantification using *in situ* amplification

This thesis chapter is reproduced from a previously published paper, Xu *et al.*, Nature Methods, 2016 [26]. Experiments and data analysis were performed by Liyi Xu.

2.1 Introduction

The absolute quantification of DNA sequences and fragments in genomics [100] and prenatal diagnostics [81] requires assays that enable parallel clonal nucleic acid amplification. DNA quantification by amplification also is needed to overcome nonspecific background for the detection of rare sequence targets in microbial communities and blood-plasma-based diagnostics [4, 71, 115, 116].

The traditional method of single-molecule studies requires individual molecules separated in partitions. It is commonly done in engineered microfluidic systems and multi-phase micro-droplet systems, which prevent a broader deployment of digital assays in research labs and in the clinic. Instead of creating physical partitions to

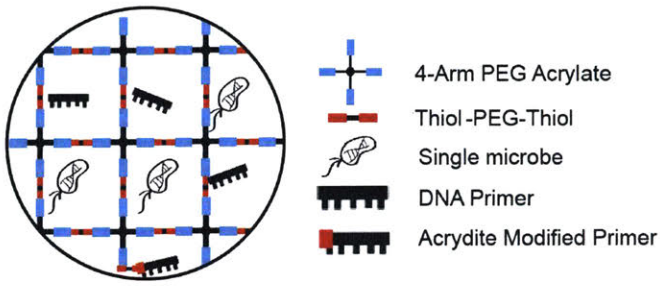

Figure 2-1: The *virtual microfluidics* hydrogel structure (not to scale).

isolate individual contents, we decided to build "invisible" dividers around them. Passive segregation of single molecules and the product of genome amplification will be able to facilitate genomic analysis of thousands of nucleic acids in parallel. In order to achieve this, a porous structure that is able to fix many molecules and provide a liquid environment for DNA amplification is needed.

Inspired by earlier work on culturing microbes in hydrogels [21], polymerase cloning in polyacrylamide gels using PCR [117] and in agarose gels using MDA in conjunction with flow cytometry [118], we developed and tested bulk polyethylene glycol (PEG) hydrogels as a general and facile platform for compartmentalizing single molecules and single cells without discrete partitions. This approach, which we call *virtual microfluidics* (Fig. 1-2 and Fig. 2-1), enables massively parallel single-molecule amplification in virtual sub-divisions without the need for engineered micro-devices, multi-phase liquid systems, or instrumentation for cell sorting or microfluidics control. We selected hydrolytically degradable PEG hydrogels that covalently crosslink under mild conditions [119]. The chemically selective crosslinking reaction used in our method does not damage templates or inhibit subsequent reactions and forms gels that are stable to high temperatures. The mesh size of the PEG gel allows diffusion of small molecules, oligonucleotides, and enzymes but immobilizes cells and high-molecular-weight nucleic acids [120]. If desired, PEG gels can be functionalized to selectively immobilize low molecular weight species by attachment to the gel matrix [121].

Hydrogels are formed by crosslinking polymer chains through physical, ionic or covalent interactions and are best known for their ability to absorb water [122], which

makes them an ideal candidate for solid-phase DNA amplification. Note that solid-phase here means in-gel, compared to liquid-phase reactions. Previously, the polyacrylamide hydrogel has been used as a scaffold for solid-phase DNA amplification [117]. However, polyacrylamide gel crosslinks with the free radicals initially produced by ammonium persulfate or by photochemical polymerization. The free radicals have been found to inhibit reverse-transcription PCR and inhibit DNA dyes such as SYBR Green and LC Green plus [123] for real-time monitoring, while photochemical polymerization might introduce DNA damage to the target of interest. Such DNA oxidative stress caused by the addition of free radicals is a pervasive cause of sequencing error and directly confound variant identification [110]. Luckily, a new generation of PEG-based multi-arm hydrogel has been developed that provide many advantages for our purpose [124]. Specifically, 4-arm PEG hydrogel with various end moieties has been applied for gene delivery and to bulid 3D scaffolds for tissue engineering. The hydrogels are formed via Michael addition chemistry by reacting a 4-arm acrylate terminated PEG with a thiol-functionalized PEG [125]. The use of Michael addition chemistry allows for *in situ* hydrogel formation under physiological conditions, which will cause minimal damage to the DNA and cells of interest in broad applications. In addition, while acrylamide powder is neurotoxic, 4-arm PEG components pose no such harm to researchers. In terms of mesh size, Raeber *et al.* and Kraehenbuehl *et al.* have shown that 4-arm PEG's pore size is between 25 nm and 100 nm depending on the weight percentage [119, 126].

Two types of DNA amplification—PCR and MDA are characterized for digital quantlification. PCR amplifies sequence specific region locally defined by forward and reverse primers, while MDA amplifies the template DNA globally through random primer binding. In this chapter, I descrbied the characterization on both amplification methods and focused on MDA method because of its wide application in enabling low-input genomics.

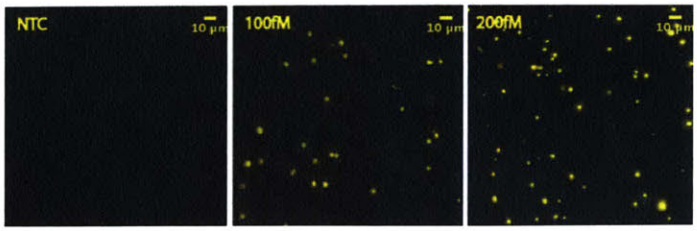

Figure 2-2: DNA amplification clusters from PCR in hydrogel in capillary tubes. Images are taking using a Nikon wide-field fluorescent microscope. The concentration of DNA template was labeled in units of femtomolar (fM). The capillary tube is 50 μm tall. Scale bars represent 10 μm.

2.2 Results and Discussion

2.2.1 Digital PCR in-gel characterization

In order to characterize *virtual microfluidics*, purified lambda phage DNA is used as the template to conduct DNA amplification in the PEG hydrogel. λDNA is a common, well-characterized substrate for restriction endonucleases and its sequence and properties have been well-understood [127, 128]. The 48502 bp length acts as a useful proxy for genomic DNA from bacteria or mammalian cells. In order to quantify the robustness of solid phase DNA amplification and estimate the dynamic range of the technology, a range of DNA template concentrations from serial dilutions of a stock were used. Reaction components consisting of 4-arm PEG-acrylate, dithiol-PEG, PCR reaction buffer, primers, dNTP, DNA polymerase and template are mixed thoroughly before loaded in a reaction chamber. Various experimental configurations such as PDMS channels, capillary glass tubes (Fig. 2-2), thin PDMS wells (10 μm to 50 μm), and frame-seal chambers (Biorad) have been tested, with the frame-seal chamber (Fig. 1-3 in Chapter 1 and Fig. 2-6 in methods) proved to be the most efficient and consistent loading method. Several types of DNA polymerase with different levels of fidelity, 3'-5' proofreading, and primer extension capacity, such as Jumpstart *Taq*, Vent (exo-), Vent, have been tested in hydrogel for amplification optimization. Fig. 2-7 indicates that accurate digital counting was achieved by PCR in gel.

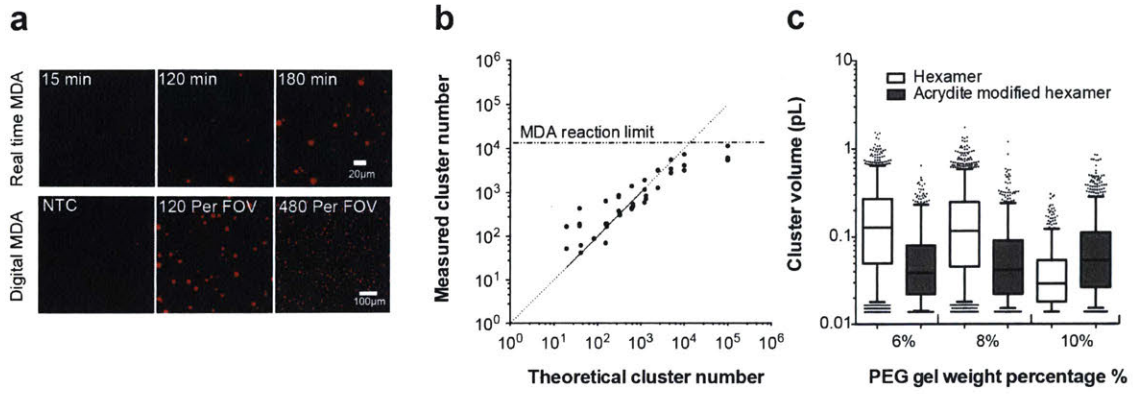

Figure 2-3: Digital single-molecule MDA in hydrogel (Lambda phage DNA).(a) Real-time and digital MDA in PEG hydrogel. Top, time-lapse epi-fluorescence images (SYTOX Orange DNA stain) illustrate MDA cluster growth from individual template molecules. Bottom, DNA cluster number increases with template concentration. FOV, 650 nm × 650 nm field of view. NTC, no DNA template control. (b) Calibration curve illustrates linear relationship between template concentration and cluster number (n = 2 or 3 FOV at each concentration). (c) MDA cluster size is correlated with gel weight percentage and affected by acrydite-modified hexamer anchorage. Data is shown as 5% – 95% box plots with scattered outliers and center line for median. (n = 1334, 1587, 684, 704, 869 and 1301).

2.2.2 Digital MDA in-gel characterization

In addition to PCR in hydrogel, Multiple Displacement Amplification (MDA) is conducted for whole genome amplification in hydrogel. MDA [109] is a popular amplification method for single-cell genome sequencing [20, 89, 95, 102, 112, 113, 114]. To evaluate the *virtual microfluidics* concept for WGA, we tested dMDA [92, 100] of purified, diluted Lambda phage DNA in the hydrogel format (Fig. 2-3a-c, Fig. 2-4 and Methods). Our estimate of 10 pg MDA product per cluster (Fig. 2-8) suggests that we approached endpoint product concentrations typical of conventional liquid MDA reactions (~800 ng/μL) [13]. We varied parameters to test how *in situ* single-molecule MDA reactions can be controlled (Fig. 2-3 and Fig. 2-4), observing that the smaller pore sizes in higher density gels limit the spread of DNA products.

A similar sample preparation and loading method are used for dMDA compared to dPCR in the hydrogel. The only difference for dMDA in the hydrogel is the UV decontamination step on all reaction components except the polymerase. A study by

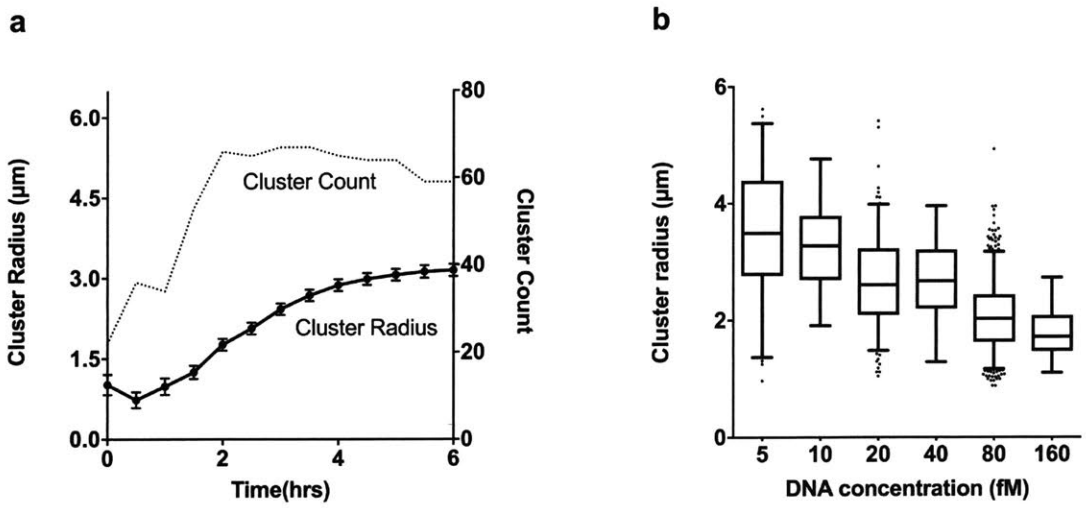

Figure 2-4: Real time dMDA and MDA cluster size. (a) Real time digital MDA for quantification of cluster growth (mean radius ± SEM) and count with time. The zero time point cluster count and radius data points reflect the properties of fluorescent contaminants. (b) MDA cluster size decreases with increasing DNA template concentration with the same gel condition. Data is shown as 5% – 95% box plot with outliers scattered and the centerline as the median. (n = 2 fields of view at each concentration, number of clusters for each field of view is: n = 43, 62, 89, 88, 191, 167, 321, 305, 478, 542, 711, 833).

Woyke *et al.* has shown that the calibrated UV decontamination step can effectively remove contaminant DNA without introducing significant coverage biases or variants [99].

With the success of dPCR and dMDA in the hydrogel, a robust but straightforward image analysis method is needed to obtain the absolute count of DNA molecules and the size of DNA amplification clusters. Currently, image analysis is conducted on the raw confocal stacks using the Fiji Image J. The Image J particle analyzer plug-in is widely used for counting particles and cells. Confocal stacks were taken with consistent laser power and integrated with a Z intensity gradient to minimize cluster fluorescent saturation. Each gel stack, roughly 300 μm thick, was first projected in Z direction with the maximum intensity for a faster 2D processing. Then, a thresholding algorithm is chosen manually based on the visual comparison between the original projection and the thresholded image. This image analysis method has been proved effective in measuring the number and the size of DNA amplification

clusters. However, a more consistent and robust image analysis routine might be needed for a higher-sensitivity application.

2.2.3 DNA amplification cluster analysis

Understanding the mechanism of cluster growth and effectively controlling the size of the end product are keys ensuring the broader applicability of this technology. Key parameters include initial DNA concentration, hydrogel weight percentage, primer modification chemistry and amplification time. One hypothesis regarding the size of clusters is that with a higher PEG hydrogel weight percentage, tighter clusters should be generated. Higher PEG weight percentage means more 4-arm PEG acrylate and dithiol-PEG in a fixed volume. Thus, the hydrogel network will provide a smaller mesh size for DNA template and prevent amplification product from moving further due to restricted diffusion. Another possibility is that a tighter mesh size might compromise the robustness of DNA amplification reaction due to the physical impedance.

Another way to control the DNA amplification cluster size could be to add a chemical moiety to the primers. Acrydite modification on DNA probes has been used frequently to attach primers to various hydrogel platforms [117]. In this case, acrydite reacts with the part of the thiol groups on dithiol-PEG and unreacted thiol groups will crosslink with 4-arm PEG acrylate. Varying the modified primers' concentration and its ratio to standard primer concentration gave us new insights on how to control cluster size (Fig. 2-3c).

Meanwhile, real-time monitoring on the DNA clusters' growing helps determine the optimal reaction time and cycle numbers for both MDA and PCR reactions (Fig. 2-4a). It will be also useful to help decide on a cut-off reading time to prevent false-positive readings including the small primer-dimer clusters and clusters from short DNA fragments mixed in a genomic DNA sample (Fig. 2-4a).

2.2.4 Dynamic range of in-gel digital MDA

When the input for digital MDA is very low (fewer than 100 per field of view), molecular counts are significantly inflated by contaminating fluorescence signals (contaminating DNA fragments or particles) that are not differentiated from true counts by our image analysis algorithm. At high target concentrations, the DNA clusters crowd one another, limiting the maximum useful concentration to 10,000 DNA molecules per field of view. For MDA, the smaller cluster sizes observed at higher template concentration (Fig. 2-4b) benefit assay dynamic range by improving cluster identifiability at the highest template concentrations. Based on our estimate of 10 pg DNA per cluster (Fig. 2-8), 10,000 - 100,000 clusters per field of view in our setup approximates typical maximum product concentrations of about 800 ng/μL achieved in conventional liquid MDA reactions. The assay dynamic range can be improved by manipulating the DNA cluster size, increasing the volume of gel imaged (e.g. by combining multiple fields of view), improving image processing methods, and further reducing the number of fluorescent contaminants.

2.2.5 Analysis of reaction extent limitation and local competition among MDA clusters

Based on Fig. 2-5, we concluded that a global auto-inhibition mechanism limits the growth of MDA clusters. We analyzed the variability in cluster number and DNA content around large and small reference clusters to test for local reagent competition among WGA reaction centers, finding little evidence for local competition. This observation is consistent with the high diffusion constants for enzymes, primers, and nucleotides measured in PEG hydrogels similar to ours [120, 129]. No specific limiting reagent was identified when reactants were supplemented individually (data not shown). The final reaction pH in our hydrogel reactions was measured to be 6.5 (initial pH = 7.5), which may limit cluster growth due to a global loss of polymerase activity at lower pH. Altogether, these data are consistent with density-dependent average size variation by global auto-inhibition, possibly by the pH drop. Variability

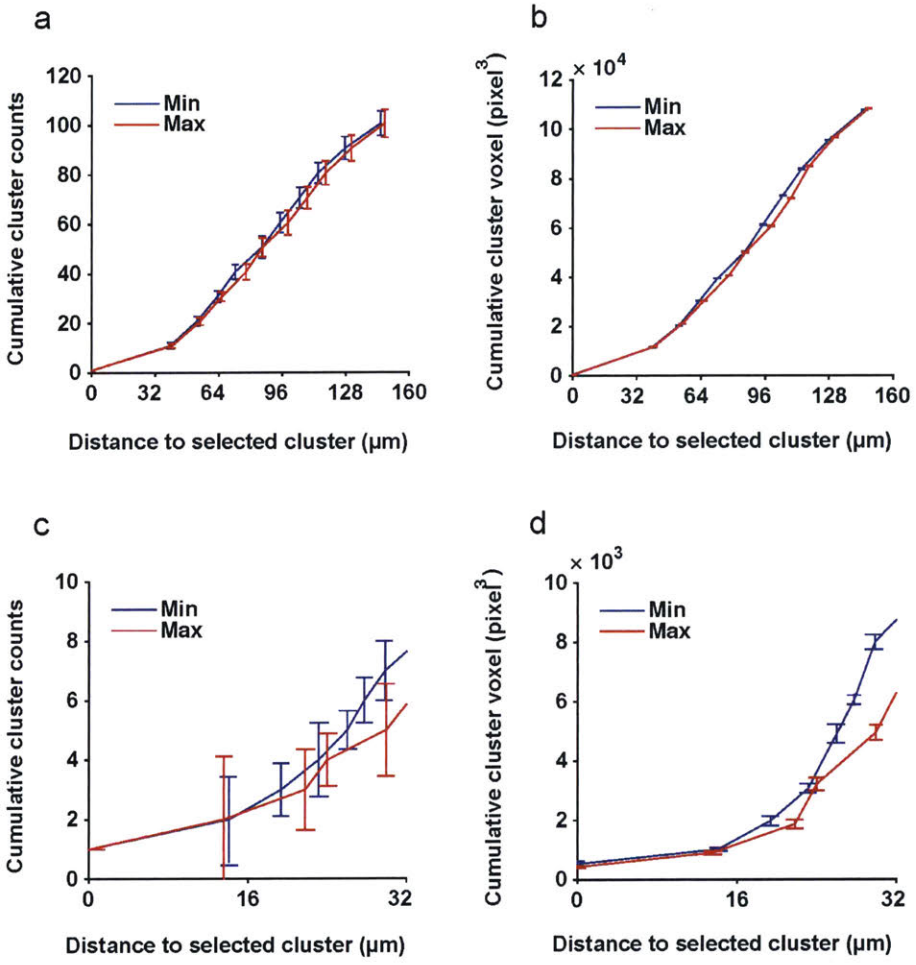

Figure 2-5: Cluster size and location correlation analysis. Five reference clusters with the largest radii and five reference clusters with the smallest radii were chosen among clusters in one field of view (100+ clusters) of one MDA hydrogel sample. a) By zooming out from selected clusters, the number of other clusters encountered at successively greater distances was plotted. b) The total volume of other clusters encountered was plotted. c) and d) are zoom-in views of a) and b) at 0 to 32 μm from reference clusters. We hypothesized that large clusters consume local resources and thus, would reduce the number and/or size of surrounding clusters. Although slight enhancements in the number and size of clusters surrounding the set of small reference clusters versus the set of large reference clusters exist, the effect is small. Error bars are SEM.

of cluster size in a single experiment may result from variable initial template conformation, the degree of template denaturation, or local inhomogeneities in the hydrogel structure.

2.3 Conclusion

Here we tested the performance of *virtual microfluidics* in-gel digital PCR and digital MDA amplification as an analytical method for molecular counting assays. *Virtual microfluidics* enables high-throughput digital assays and preparative whole-genome amplification without microfabricated consumables or expensive instrumentation. Up to 20,000,000 analytes per μL could be accommodated due to the nature of the diffusion-restricted reaction and the continuous virtual chambers. Throughput could be increased by using a thinner gel with more surface area. Its excellent optical accessibility allows potential fluorescent labeling of rare sequences, which is a key in identifying rare targets in liquid biopsy applications. We expect *virtual microfluidics* to find applications as low-cost, highly accessible digital assay platforms that offer superior sensitivity and dynamic range.

2.4 Materials and Methods

2.4.1 PEG hydrogel cross-linking

Hydrogel components, including 4-arm PEG acrylate (MW 10,000) and HS-PEG-SH (MW 3,400), were obtained from Laysan Bio. For every 25 μL of 10% (wt/v) cross-linked hydrogel, 1.6 mg of 4-arm PEG acrylate and 1.1 mg of HS-PEG-SH were dissolved in pH 7.4 PBS (Invitrogen). It was briefly vortexed and centrifuged to ensure mixing and it was allowed to sit on the bench for 10 min while the hydrogel components cross-linked through the reaction between the thiol and acrylate groups.

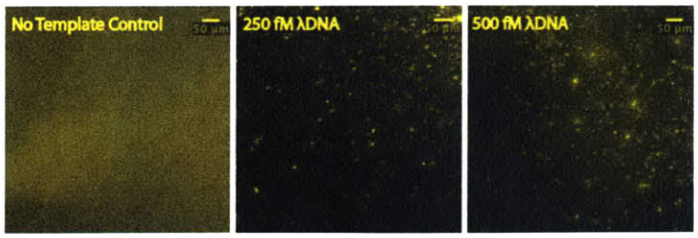

Figure 2-6: DNA amplification clusters from PCR in hydrogel in frame-seal chambers. Each image is a z-axis max projection from a confocal tiff stack taken by a Zeiss spinning disk confocal microscope. Scale bars represent 50 μm.

2.4.2 In-gel digital PCR

The primers (Table 3.5) used for PCR on purified λDNA (48 kbp, NEB) were ordered through IDT(Integrated DNA Technologies). A 25 μL hydrogel PCR reaction consisted of 2 U of VentR (exo-) polymerase (NEB), 1× ThermoPol Reaction Buffer (NEB), 0.4 mM dNTP (NEB), 1 μm Primers, 5% DMSO (Sigma), 0.5 mg/mL BSA (NEB), 1.6 mg 4-arm PEG acrylate in PBS, 1.1 mg HS-PEG-SH in PBS, and λDNA template (NEB) of various concentrations. The 25 μL above components were loaded in a 9 mm by 9 mm frame-seal chamber (Bio-rad). The following thermal protocol was ran on an MJ Research PTC–100 twin tower thermal cycler: 30 °C for 30 min (gel polymerization), 98 °C for 3 min; 98 °C for 30 sec, 57 °C for 30 sec, 72 °C for 1 min for 40 to 60 cycles; 72 °C for 5 min and hold at 4 °C. The gel was stained with 500 nM SYTOX Orange nucleic acid dye (Invitrogen) (Fig. 2-6 and 2-7).

2.4.3 In-gel digital MDA

A 25 μL hydrogel MDA reaction consisted of 0.5 μL of REPLI-g sc Polymerase (Qiagen), 1× Φ29 buffer (NEB), 50 μm random hexamers (IDT; including two phosphorothioate bonds at 3' terminus), 2.5% DMSO, 0.4 mM dNTP, 0.5 mg/mL BSA, 500 nM SYTOX Orange (Invitrogen) and denatured λDNA. λDNA was denatured using alkaline buffer "D1" (Qiagen) and neutralized using buffer "N1"(Qiagen) according to Qiagen REPLI-G sc kit protocol prior to hydrogel encapsulation. All MDA and gel components, except polymerase and SYTOX Orange dye, were UV treated for 30 min using the Stratalinker UV crosslinking instrument (Stratagene) to render con-

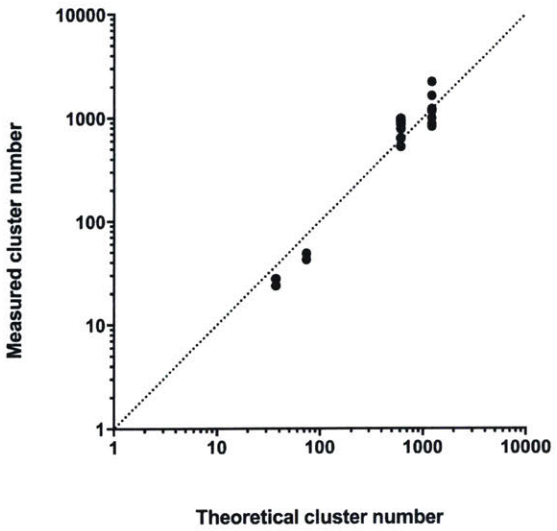

Figure 2-7: Digital single-molecule PCR in hydrogel (Lambda phage DNA). Measured cluster number per field of view versus calculated cluster number based on template concentration. Reaction conditions are described in methods (PCR). The dotted line indicates the ideal situation when measured cluster number equals to the theoretical cluster number.

taminating background DNA incompetent for MDA. The 25 μL reaction mixture was loaded in a 9 mm by 9 mm frame-seal chamber (Bio-rad, about 300 μm in height). The gel was sealed in the chamber with a plastic cover and maintained at 30 °C for 8 hours or longer in the MJ Research PTC–100 twin tower thermal cycler. After the reaction, we imaged the gel using Nikon ECLIPSE Ti inverted microscope or Nikon ultra-fast laser scanning confocal microscope (MIT Koch Institute Microscopy Core Facility) (Fig. 2-4a).

2.4.4 In-gel real-time dMDA

MDA hydrogel reactions were set up as described above and conducted at room temperature for 6 hours on a Nikon ECLIPSE Ti Epi-Fluorescence Microscope excited with a Lumencor Spectra X light engine (Lumencor) with fluorescent emissions filtered through a SpGold filter (Semrock) (Fig. 2-4b). MATLAB was used to capture time-lapse image stacks through a Nikon 20×/0.4 NA objective and Hamamatsu C11440 camera with 15 min intervals, 100 ms exposure time, and 10% Lumencor excitation

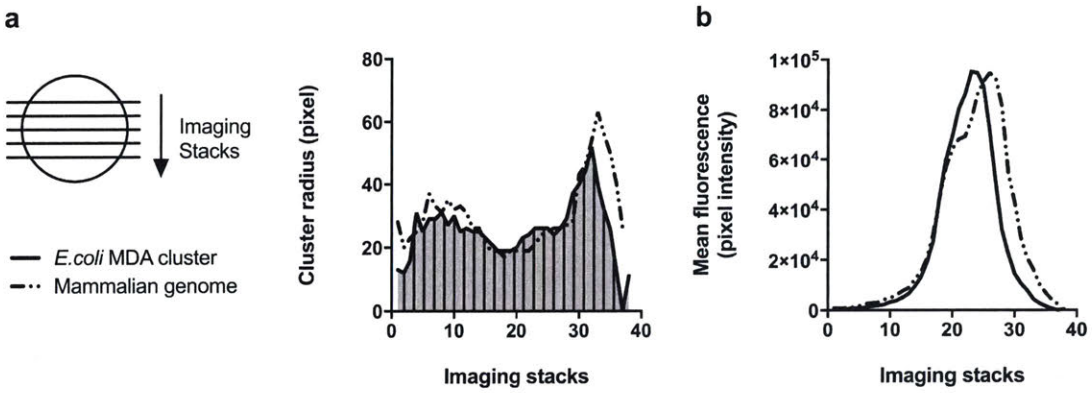

Figure 2-8: Imaging analysis for a *E. coli* MDA cluster and a mammalian genome. a) The *E. coli* MDA cluster and the mammalian genome were imaged with a Nikon confocal microscope. The cluster radius of each were ploted. b) The mean fluorescence of each stack was plotted for the *E. coli* MDA cluster and the mammalian genome.

power. All samples were stained with 500 nM SYTOX Orange. Each *E. coli.* MDA cluster or mammalian cell image stack was cropped and processed as described below.

2.4.5 Image acquisition and analysis

Z-stack images were taken by Nikon ultra-fast laser scanning confocal microscope with pinhole = 1.2, HV = 112, offset = 0, laser wavelength = 561 nm, laser power = 1.3 to 1.5, using a 20× objective on Galvano mode. Acquisition speed was 1 frame/sec and z step size was 0.95 μm. On the inverted microscope, z-stack images were taken with the exposure time 100 ms, Lumencor excitation power 10%, binning size 2 and z step size 10 μm. Both z stacks were first processed into max intensity projections in FIJI. Max projection tiff files were then loaded into MATLAB. The background was obtained by applying a Gaussian filter of hsize 200 and sigma 50. All max projections were background-subtracted and thresholded at 2 $\sim$ 2.5× standard deviations above the mean intensity. Cluster count, cluster area (radius), and cluster mean intensity were obtained with the bwconncomp and regionprops functions (Fig. 2-8).

Chapter 3

Virtual Microfluidics enables high-quality single-cell sequencing from a mixed population of cultured bacteria and the human gut microbiome

This thesis chapter is reproduced from a previously published paper, Xu *et al.*, Nature Methods, 2016 [26]. Experiments and data analysis on the cultured bacteria were performed by Liyi Xu. Ilana Brito and Liyi Xu conducted the data analysis of the gut microbiome data.

3.1 Introduction

In the burgeoning field of single cell analysis [1], high-throughput and high-fidelity whole-genome [89, 95, 102] and whole-transcriptome amplification (WGA and WTA) reactions are needed to produce sufficient material for sequence library construction to support the discovery and validation of new genomes [2, 112, 114], as well as the analysis of genomic and functional heterogeneity [3, 89, 112].

A variety of approaches have been explored for compartmentalization across a large number of discrete reactors, including SBS plates [95], high-density microfluidic arrays [84], engineered lab-on-chip systems [85, 86, 87, 88], and multi-phase micro-droplet systems [89, 90, 91, 92]. However, they require complex instrumentation and microfabricated consumables that hinder broad deployment. An ideal platform should resist external contaminants and cross-compartment mixing, exhibit high throughput in small reaction volumes, be stable under temperature change, allow optical access, and allow facile addition and removal of reagents and samples. Finally, it should generate high-quality amplified products and minimize biases and artifacts, such as chimeric fragments commonly formed in PCR, WGA and WTA, that can severely impact single-cell sequencing results.

Building on my work in Chapter 2 on characterizing *virtual microfluidics* for DNA digital quantification, I further developed the technology for single-cell sequencing.

3.2 Results and Discussion

3.2.1 In-gel single *E. coli* MDA

We applied digital MDA at the single-cell level using the *virtual microfluidics* system. Individual log-phase *Escherichia coli* could be identified in the hydrogel by fluorescence microscopy (Fig. 3-1). We lysed the embedded cells by heat treatment and carried out MDA on the denatured genomic DNA, observing the appearance of MDA clusters at the reaction endpoint.

3.2.2 In-gel single-microbe MDA - cultured *E. coli* and *S. aureus*

Next, we tested the potential of *virtual microfluidics* to support single-cell shotgun genome sequencing (Fig. 3-2). We mixed log-phase *Escherichia coli* (BL21) and *Staphylococcus aureus subsp. aureus* (RN6390/8325) strains at about 200,000 cells/ mL and embedded the cells in a 300 micron thick PEG hydrogel. We used a mixed-

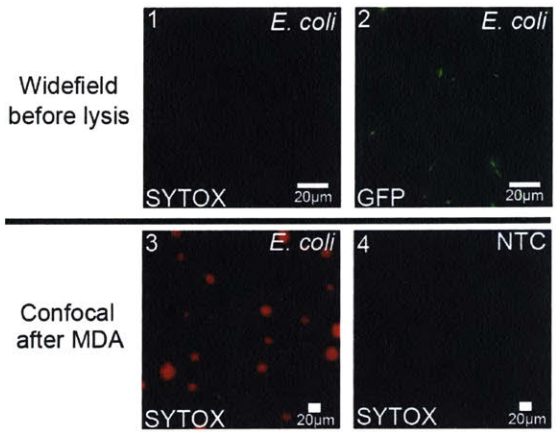

Figure 3-1: Single-cell MDA on *Escherichia coli*. Hydrogel-encapsulated *E. coli* express GFP and exclude SYTOX Orange before lysis. SYTOX Orange staining reveals product clusters after MDA. Top images are from the same field of view. NTC, MDA control lacking *E. coli*.

Total hydrogel punches	80
E. coli positive punches	7
S. aureus positive punches	36
Double positive	7
Double negative	30

Table 3.1: QPCR characterization of hydrogel punches

input approach to ensure sensitive identification of any cross-contamination among single-cell samples and any contamination of single cell samples from other sources (including *E. coli* DNA contamination). The embedded cells were lysed by enzymatic and heat treatment, and MDA reagents were introduced by diffusion into the gel. Eighty sub-samples from the gel (of 60 nL each) were recovered manually in a grid pattern as indicated in Fig. 3-2a Each punch sample was re-amplified to 10^9 - 10^{10} overall fold-amplification in a second-round 20 μL liquid MDA reaction. Real-time PCR (QPCR) assays for *E. coli* and *S. aureus* genome sequences were applied to diluted aliquots from each sample (Table 3.1 and 3.5). The QPCR results were well-approximated by a random cell dispersion model.

We sequenced Illumina short-insert libraries produced from randomly selected punch samples and positive-control gDNA samples (MiSeq v2 500 cycles). Quality-filtered reads were then mapped to *de novo* assemblies of the positive-control gDNA

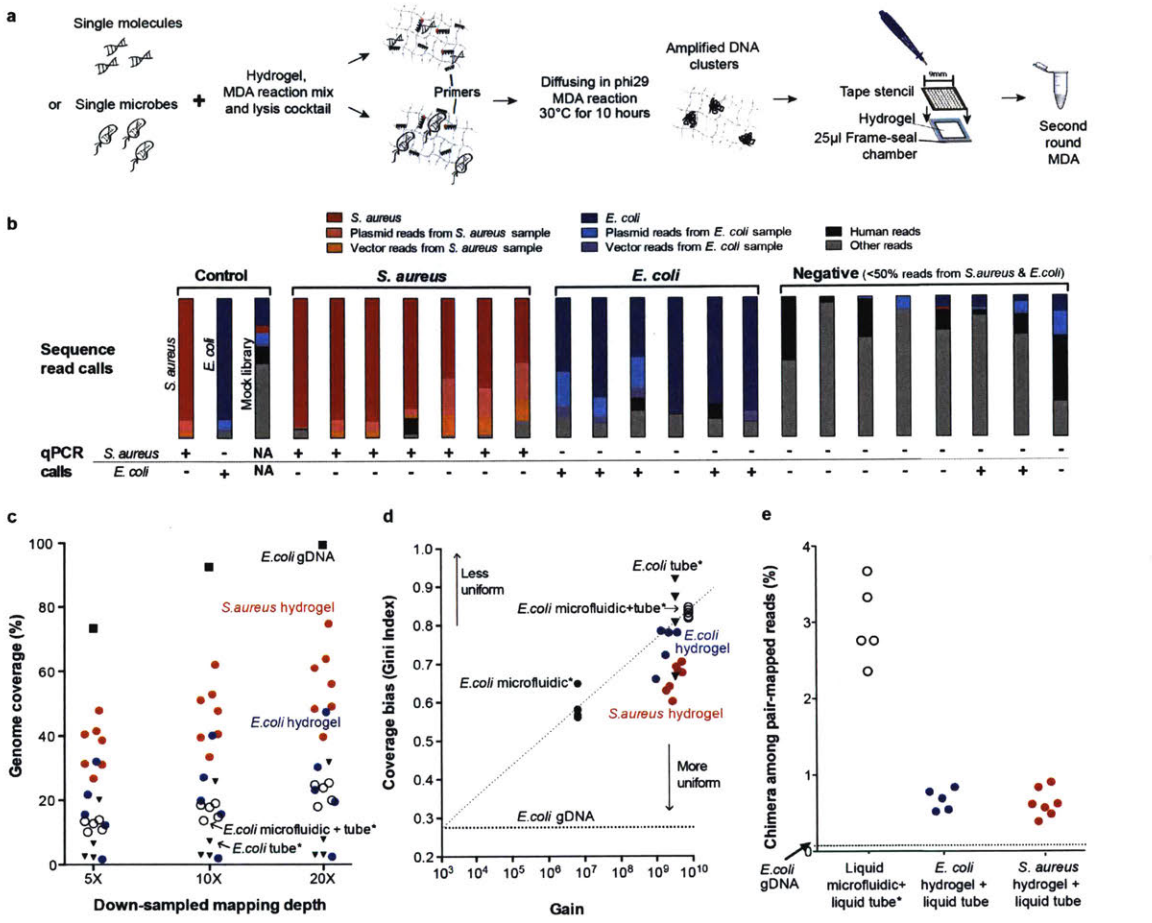

Figure 3-2: Single-cell whole genome sequencing from *E. coli* and *S. aureus* hydrogel WGA samples. (a) *Virtual microfluidics* WGA workflow. (b) Sequence read classification using BLAST against the corresponding databases. The samples were ordered based on the ratio of *S. aureus* to *E. coli* reads from shotgun sequencing and the fraction of reads not from *S. aureus* or *E. coli*. Two negative samples are classified as false positive PCR calls. Positive *S. aureus* and *E. coli* samples with matching PCR calls are included in downstream analyses. (c) Genome coverage in *S. aureus* and *E. coli* hydrogel punch samples compared with published single-cell *E. coli* data produced using conventional liquid MDA reactions*. One *E. coli* outlier library showed extremely poor genome coverage. This library had low complexity (37% duplicate reads), which points to poor library quality rather than MDA as the cause for low genome coverage. All samples were randomly down-sampled based on mapped reads and bootstrapped 10 times; error in all cases was smaller than the symbols plotted. (d) Coverage distribution bias. Gini Index (derived from Lorenz curve) reports the genome coverage bias of single-cell *E. coli* and *S. aureus* punches compared to the same published liquid-MDA single-cell *E. coli* data as a function of amplification gain. (e) Chimera frequency in the virtual microfluidics samples is significantly reduced versus published *E. coli* data produced using standard liquid MDA reactions. Indicates liquid MDA data from de Bourcy *et al.* 2014.

datasets and sequence databases (Table 3.6, Table 3.7). The positive punch samples showed strong enrichment (Fig. 3-2b) for reads mapping to the expected reference genome (Fig. 3-6) while the negative punch samples showed enrichment for human reads, reads with poor mapping quality, and *E. coli* (possibly contaminants from the reagents and/or laboratory environment), and were similar to the results from a mock library (Table 3.2). The lack of *E. coli* and *S. aureus* cross-contamination in the positive punch samples indicates that *virtual microfluidics* can resolve single-cell amplification products.

At 20× mean coverage (Table 3.8 in methods), approximately 30% of the *E. coli* genome and about 60% of the *S. aureus* genome were covered in each single-cell sample (Fig. 3-2c). The coverage values for *E. coli* are in-line with typical single microbe genome sequencing at similar sequencing effort [87, 99]. The superior coverage performance in *S. aureus* may be attributable to the lower GC content of *S. aureus* (33%) compared with *E. coli* (51%), better accessibility (deproteination) of the genome after lysis, and/or higher average genome equivalents per cell in *S. aureus* resulting from cell cycle dynamics.

To rigorously evaluate sequence coverage distribution, we calculated the Gini Index (a measure of inequity ranging from 0 to 1) for each of our single-cell datasets and previously published single-cell *E. coli* liquid MDA datasets for which raw read data were available and fold-amplification was known (Fig. 3-2d). The coverage uniformity in our single-cell punch samples compares favorably with published single-cell datasets at similar amplification gain.

We then analyzed the occurrence of chimeric reads, which are known to occur with high frequency in MDA by a cross-priming mechanism [130]. Chimeric reads directly confound *de novo* assembly, analysis of rearrangements, and mapped read counting. Our single-cell datasets contained about 0.5% chimeric reads, approximately five-fold lower than previously published short-read datasets produced using liquid single-cell MDA samples (Fig. 3-2e and Table 3.6). The occurrence of chimeric reads spanning more than 10 kb of the template is even lower (about 0.1%, Fig. 3-3), raising the possibility of extracting long-range information from single-cell MDA samples using

	"Other reads"	% Identified	Cloning/expression vector %	*Other E. coli* %	*Other S. aureus* %	Synthetic construct %	*Propoioniba cterium acens* %	Other major categories		Percent listed
E. coli genomic DNA	42941	64.27%	21.30%	6.07%	0.01%	72.33%	0.12%			99.82%
S.aureus genomic DNA	4720	38.24%	1.16%	0.11%	61.83%	3.66%	10.80%	Staphylococcus phage: 349	(19.34%)	96.90%
S1	11887	4.85%	1.22%	0.17%	20.31%	4.34%	19.62%	Assorted bacteria and fungus: 248	(43.06%)	88.72%
S2	1306	16.62%	3.23%		11.52%	1.84%	77.42%			94.01%
S3	34963	0.24%	9.64%		28.92%	15.66%	12.05%	Assorted bacteria and fungus: 28	(33.73%)	100.00%
S4	1800	25.78%	0.65%		61.64%	2.80%	16.16%			81.25%
S5	4571	27.28%	0.88%	0.24%	62.63%		16.84%	Staphylococcus phage: 214	(17.16%)	97.75%
S6	4899	46.87%			18.60%		57.23%	Assorted bacteria and fungus: 506	(22.04%)	97.87%
S7	14074	14.53%	0.29%		77.65%	1.32%		Staphylococcus phage: 361	(17.65%)	96.92%
E1*	352603	11.23%	1.15%	0.01%		3.99%		Assorted bacteria and fungus: 35284	(89.12%)	94.26%
E2	32744	76.79%	25.00%	2.57%		71.89%	0.14%			99.60%
E3	20547	59.69%	27.47%	0.21%		71.67%	0.44%			99.79%
E4	47915	77.10%	13.09%	0.03%		46.17%	0.92%	Malassezia globosa CBS: 13804	(37.36%)	97.57%
E5*	430977	12.60%	5.95%	0.33%		21.91%	15.95%	Assorted bacteria and fungus: 26635	(49.07%)	93.20%
E6	46757	1.64%	1.56%	0.65%	7.16%	5.34%	39.71%	Listeria seeligeri serovar 1/2b str : 245	(31.90%)	86.33%
E7	55490	77.20%	20.42%	0.01%		79.24%				99.67%
NTC1	365936	0.94%	1.51%		3.22%		12.73%	Assorted bacteria and fungus: 2809	(81.47%)	98.93%
NTC2	303923	86.55%			0.08%		2.99%	Human: 247884	(94.24%)	97.31%
NTC3	251727	89.49%					0.26%	Assorted bacteria and fungus: 218140	(96.84%)	97.09%
NTC4	243833	96.26%					99.96%			99.96%
NTC5	213052	90.75%					2.20%	Human: 187178	(96.81%)	99.02%
NTC6	45986	95.44%				0.04%	0.03%	Human: 43806	(99.81%)	99.87%
NTC7	74719	0.54%			37.87%		45.30%			83.17%
Mock	7204	11.30%	8.60%			29.85%	14.74%	Assorted bacteria and fungus: 373	(45.82%)	99.02%

* False-positive *E. coli* single-cell samples
** Values less than 0.01% were omitted for clarity

Table 3.2: Sequence read classification of "other reads"

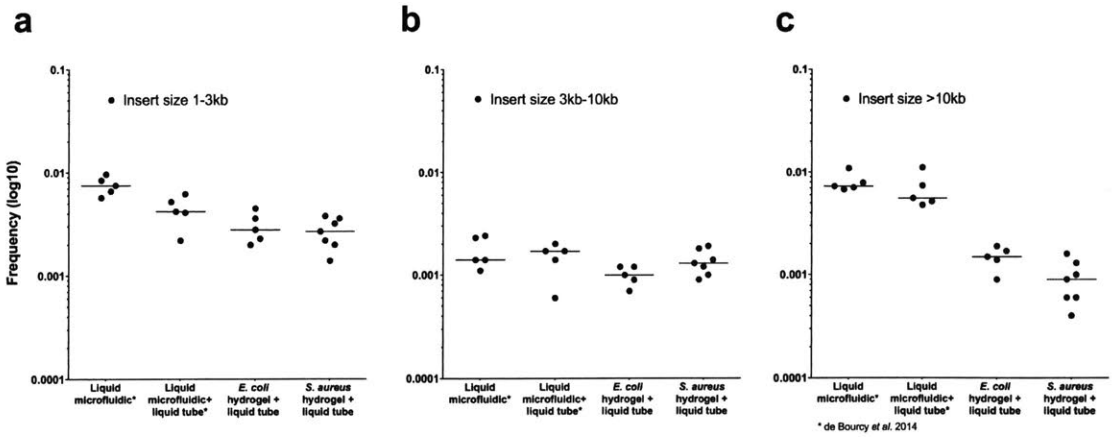

Figure 3-3: MDA chimera frequency with different insert sizes, *E. coli* and *S. aureus* data. a) 1 - 3 kb. b) 3 kb - 10 kb. c) Larger than 10 kb. Centerlines represent the mean value.

long-read sequencing. This dramatic reduction in the occurrence of chimeric reads can be understood by restricted diffusion of the MDA intermediates that prevents cross-priming by isolating each portion of the product mixture. It may be the case that substantially all of the chimeras we observed in the punch samples were generated during the liquid-phase secondary amplification reactions. Based on these results, it is likely beneficial to run MDA in PEG hydrogels for all applications at all scales.

3.2.3 In-gel single-microbe MDA - human gut microbiome samples

Next, we tested the potential of *virtual microfluidics* for single-cell genome sequencing using samples from the Fiji Community Microbiome project (FijiCOMP). The FijiCOMP samples contain a vast uncharacterized diversity of microbial species that differ from those found in the microbiome of Western subjects. The procedure for processing these human stool samples was similar to those for lab-cultured *E. coli* and *S. aureus*, with modifications for initial sample processing and lysis (methods).

We processed a total of 421 hydrogel punch samples and compared the distribution of organisms detected in our hydrogel samples with the distribution observed from shotgun metagenomic profiling, which showed that the same top microbial families

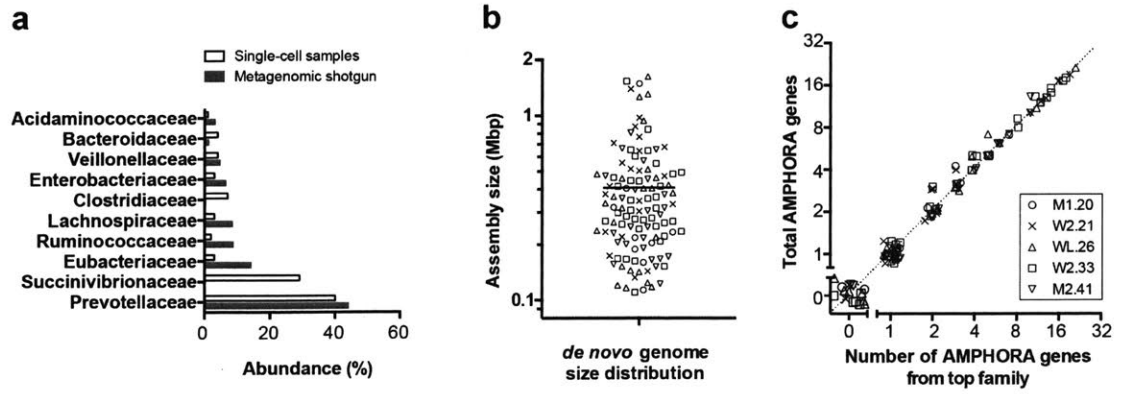

Figure 3-4: Fiji microbiome project (FijiCOMP) single-cell whole-genome sequencing. Here are the results for 117 single-cell data sets from five donor individuals. (a) Distribution of top ten microbial families from single-cell assemblies and metagenome shotgun sequencing. Samples were weighted according to the number of single cells analyzed (Table 3.9). (b) *De novo* assemblies from single-cell sequencing data ranged from 100 kbp to 2 Mbp. The line indicates the mean assembly size. (c) The total number of AMPHORA genes is nearly equal (dotted line) to the number of AMPHORA genes from the top phylogenetic family for each sample, supporting the assertion that each data set arises from an individual bacterial cell. A Gaussian-distributed random jitter ($\mu = 0$, $\sigma^2 = 0.1$) was applied to enhance visualization.

were observed using both approaches (Fig. 3-4a and Table 3.9 in methods). Interestingly, the second most abundant microbial family found in the single-cell dataset, the Succinivibrionaceae, was not initially detected but was later confirmed in the shotgun metagenomic data due to its rare representation in the established database using standard methods for the taxonomic assignment such as MetaPhlAn [131]. This discrepancy highlights the importance of unbiased approaches like single-cell analysis for organisms that are less well represented in reference databases.

We carried out *de novo* assembly of the single-cell datasets and assigned taxonomy to ribosomal gene sequences and 31 "single copy" bacterial marker genes at the family level [132]. This analysis enabled us to make crude assessments of sample purity in the FijiCOMP single-cell datasets (for which we lack strain-specific bona fide reference sequence). Of the 293 assemblies (up to 12 Mbp), we classified 117 as single amplified genomes with assembly size greater than 100 kb and strong enrichment of sequences from a single taxonomy (see Fig. 3-4 and Table 3.3) for the fate of all samples. The purity of single amplified genomes are evaluated with the identity

Sample categorizations	Number of punches
Low read counts; no assembly	16
Laboratory contamination (*E. coli*, *P. aeruginosa*)	29
Human cell sequences amplified	3
No phylogenetic markers	80
Enrichment of multiple taxonomies from assemblies	108
Assembly $<$ 100 kb	68
Single-cell assemblies	117
Total sequenced	421

Table 3.3: Overview of 117 FijiCOMP single-cell hydrogel samples

of AMPHORA (AutoMated PhylogenOmic infeRence Analysis) marker genes. AMPHORA genes are a collection of protein-coding marker genes that are single-copy in the genome, universally distributed, and are relatively recalcitrant to horizontal gene transfers [132]. We identify the number of AMPHORA marker genes and their phylogenines in each single amplified genome. If the total number of AMPHORA genes is nearly equal to the number of AMPHORA genes from the top phylogenetic family, it indicates the dataset arises from an individual bacterial cell (Fig. 3-4c). Overall, the data quality observed from these human microbiome bacteria was consistent with the results of our studies with lab-cultured Gram-negative and Gram-positive samples and demonstrates the applicability of the hydrogel method to real-world samples, including lysis and amplification of a variety of naturally occurring microbes.

3.2.4 Random Dispersion Model

Based on qPCR analysis of the 80 punches, the expected number of punches that have both *E. coli* and *S. aureus* is:

$$P_{E.coli} = \frac{14}{80} = 0.175; \quad P_{S.aureus} = \frac{43}{80} = 0.538$$
$$P_{negative} = \frac{30}{80} = 0.375; \quad < N_{both E.coli and S.aureus} >= \frac{14}{80} \times \frac{43}{80} \times 80 = 7.52$$

This result is in line with our qPCR result of seven double positive punches (Table 3.1), indicating the likelihood that the distributions of *E. coli* and *S. aureus* across

the punch samples are independent as we expected. Furthermore, if we assume a random (Poisson) distribution of microbes in the hydrogel:

$$P_{E.coli}\ (0, \lambda_E) = e^{-\lambda_E} = \frac{66}{80} = 0.825,\ \lambda_E = 0.192$$

$$P_{E.coli}\ (1, \lambda_E) = \lambda_E\ e^{-\lambda_E} = 0.158$$

$$P_{E.coli}\ (2, \lambda_E) = \frac{\lambda_E^2\ e^{-\lambda_E}}{2} = 0.01$$

$$P_{E.coli}\ (Single\ cell) = \frac{0.158}{1 - 0.825} = 90.3\%$$

$$P_{S.aureus}\ (0, \lambda_S) = e^{-\lambda_S} = \frac{37}{80} = 0.462,\ \lambda_S = 0.772$$

$$P_{S.aureus}\ (1, \lambda_S) = \lambda_S\ e^{-\lambda_S} = 0.356$$

$$P_{S.aureus}\ (2, \lambda_S) = \frac{\lambda_S^2\ e^{-\lambda_S}}{2} = 0.138$$

$$P_{S.aureus}\ (Single\ cell) = \frac{0.356}{1 - 0.462} = 66.2\%$$

The low probability value for the occurrence of single *S. aureus* is calculated based on the high number of hydrogel punches that were identified as *S. aureus* by qPCR. To bring down the value, a more dilute sample of *S. aureus* should be used (Table 3.4).

	$P_{E.coli}(0) = 0.825$	$P_{E.coli}(1) = 0.158$	$P_{E.coli}(2) = 0.01$
$P_{S.aureus}(0) = 0.463$	0.38	0.073	0.0046
$P_{S.aureus}(1) = 0.356$	0.29	0.056	0.0036
$P_{S.aureus}(2) = 0.138$	0.114	0.022	0.0014

Table 3.4: Microbe occurrence probability

3.3 Conclusion

Virtual microfluidics enables high throughput whole genome amplification and serial reagent exchange in an easy-to-use, benchtop format that requires no special equipment or environmental control. Here we show preparative amplification and recovery of single bacterial genomes for *ex-situ* analysis of lab-cultured control cells and the human gut microbiome by next-generation sequencing (NGS).

Virtual microfluidics establishes a new paradigm in single-molecule and single-cell analysis with dramatically different characteristics than established microfluidic approaches. Besides reducing the production of chimeras in MDA, the unique physical characteristics of the engineered hydrogel environment may provide a means for enhancing coverage extent and uniformity from WGA and WTA samples through the self-limiting reactivity within each virtual compartment, similar to a recently reported emulsion approach [89]. In addition, the straightforward addition and removal of reagents to/from product clusters *en masse* and excellent optical access ideally suit the *virtual microfluidics* system for rare-cell assays incorporating *in situ* labeling of cells or product clusters. We expect that *virtual microfluidics* will find application as a high-throughput platform for single-cell sample preparation.

3.4 Materials and Methods

3.4.1 In-gel single-microbe MDA - *E. coli* and *S. aureus*

Antibiotic resistant *Staphylococcus aureus subsp. aureus* (GFP) NCTC 8325 and *Escherichia coli* (RFP) BL21 strains were obtained as cryogenic stocks. For each culture, the frozen stock was inoculated in 5 mL LB broth and cultured at 37 °C overnight. 10 μL of 25 mg/mL Chloramphenicol was added to *S. aureus* culture and 5 μL of 50 mg/mL Ampicillin was added to the *E. coli* culture. 50 μL and 20 μL of each overnight culture were added to fresh 5 mL LB broth with the respective antibiotic concentration. After two hours incubation (to achieve exponential growth phase), 1 mL of each culture (O.D. 600 nm = 0.2) was centrifuged for 2 min at >10 krpm and the pellet was washed with 500 μL PBST (1% Tween–20) twice. The equal ratio mixture of microbes were diluted to 206,000 cells/mL and 1 μL of each was encapsulated in the same hydrogel sample to produce an average of less than 1 microbe per 500 μm diameter view. In addition to the hydrogel MDA reaction mix described above, lysozyme (Sigma, final concentration 2.5 mg/mL) and lysostaphin (Sigma, final concentration 0.1 mg/mL) were added to the mix. The hydrogel was

left at RT to let crosslink for 20 minutes and cross-linked hydrogels were incubated at 37 °C for 1 hour for microbe lysis and heated to 95 °C for 5 min to denature genomic DNA before rapid quenching on ice. 1 μL of REPLI-g sc Polymerase (Qiagen) diluted in 2 μL water was then added on top of the hydrogel and allowed to diffuse into the gel. Next, the gel chamber is resealed and MDA was conducted for 10 hours. After the MDA reaction, the sample was heated to 65 °C for 5 min to deactivate phi29 polymerase.

3.4.2 Image acquisition and analysis

Z stack images were taken by a Nikon ultra-fast laser scanning confocal microscope with the pinhole = 1.2, HV = 112, offset = 0, laser wavelength = 561 nm, laser power = 1.3 to 1.5, using a 20× objective on Galvano mode. The acquisition speed was 1 frame/sec and z step size was 0.95 μm. On the inverted microscope, z stack images were taken with the exposure time 100 ms, Lumencor excitation power 10%, binning size 2 and z step size 10 μm. Both z Stacks were first processed into max intensity projections in FIJI. Max projection tif files were then loaded into MATLAB. Background was obtained by applying a Gaussian filter of hsize 200 and sigma 50. All max projections were background-subtracted and thresholded at 2 $\sim$ 2.5× standard deviations above the mean intensity. Cluster count, cluster area (radius), and cluster mean intensity were obtained with the bwconncomp and regionprops functions.

For whole gel (25 μL, 9mm by 9mm) microbe density approximation, I imaged the gel with a 4× objective in a 5 × 5 grid with a 31% overlap. The 25 images were stitched using the FIJI stitching function. Fluorescent DNA clusters were counted and only gels with the appropriate clusters' range and dispersion (60 $\sim$ 80 per gel) were selected for hydrogel cluster retrieval. Images of the sampled locations were acquired but not used to guide sampling, sample preparation, or data analysis in this case.

3.4.3 MDA product cluster retrieval

In order to identify and retrieve a regular array of punches (not guided by cluster image data), we produced a tape stencil to guide the punch tool (Adhesive Applications High Tack Silicone Film Tape). We laser cut the double-sided tape with a 9×9 array of 500 μm diameter circles that has a center-to-center distance of 947 μm (Full Spectrum Laser LLC MLE–40). The tape stencil was applied on top of the frame seal plastic cover. The gel was peeled off the glass slide by allowing it to adhere to the plastic cover. The gel is then punched with a 1 mm diameter steel punch (Militek) and the micro-samples collected in a 96 well LoBind twin.tec plate (Eppendorf). The steel punch was cleaned with bleach and 70% Ethanol after each use.

3.4.4 BLAST analysis and read assignment for *E. coli* and *S. aureus*

To characterize all samples after quality trimming, each sample (R1 from each read pair) was blasted (task megablast) with the parameters listed in Fig. 3-5. The BLAST database for *E. coli* consists of three *E. coli* genomes (strain BL21, MG1655 and W3110). The *S. aureus* database consists of the genomes of strain 8325, TW20 and USA300. Univec, Plasmid and Human genome (GRCH38) databases were downloaded from NCBI. All databases were produced using makeblastdb and blastdb_alias tool. Each read was mapped to all five databases (*E. coli*, *S. aureus*, Univec, Plasmid, and Human db) and the results were ranked based on bit score, e-value and then percent identity. We assigned each read to one of the source databases based on the top hit. Using the filter_fasta.py tool in QIIME [133], we selected reads that did not map to any of the five databases for further analysis. We ran BLAST against the nt database to characterize these reads (Table. 3.2).

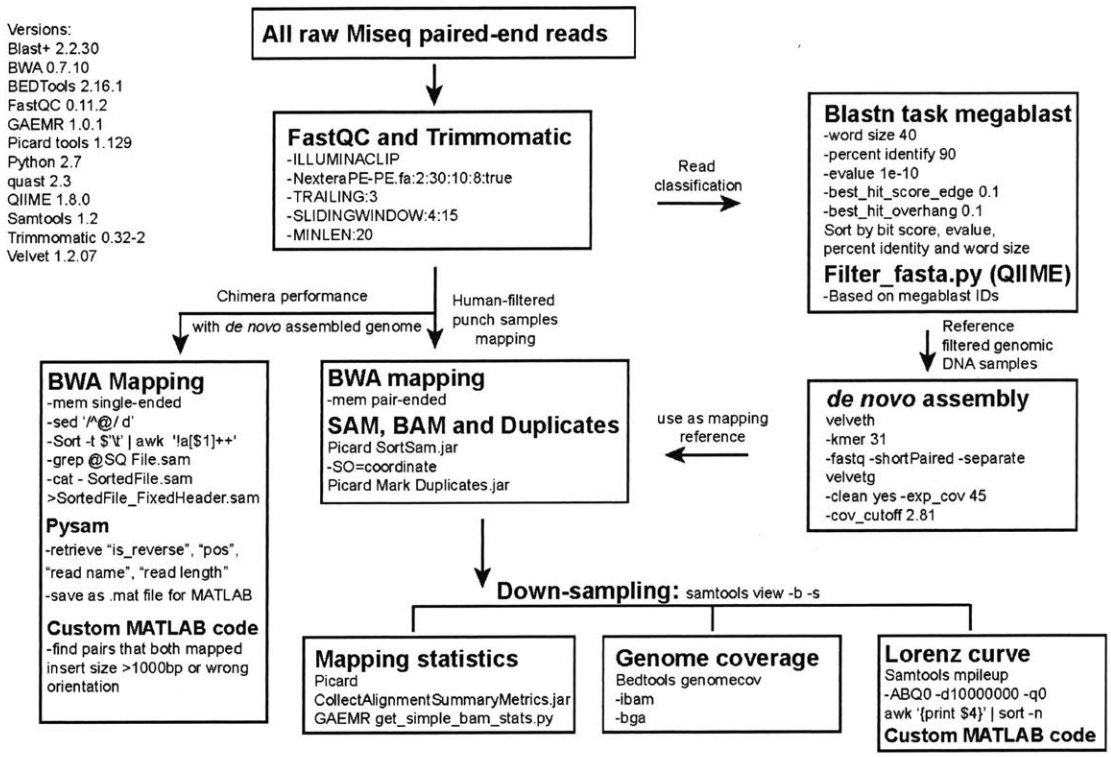

Figure 3-5: NGS data analysis schematic for *E. coli* and *S. aureus*. The analysis workflow is shown with a combination of bioinformatic tools, python scripts and MATLAB scripts.

Complete Primer list	
MDA hexamer	5'-NNNN*N*N-3'
S. aureus G1 F	TGC ACA TTT AAA CCC AGC GG
S. aureus G1 R	ATC GCA TGT GCA ATT CTC GG
S. aureus arc F	TTG ATT CAC CAG CGC GTA TTG TC
S. aureus arc R	AGG TAT CTG CTT CAA TCA GCG
E. coli G2 F	CAA CCA AAT TAT TGC CGC GC
E. coli G2 R	GCC ACG GTA ATT ACT GTC GC
E. coli uspA F	CCG ATA CGC TGC CAA TCA GT
E. coli uspA R	ACG CAG ACC GTA GGC CAG AT
λDNA 780bp F	CGG CAA ACG GGA ATG AAA CGC C
λDNA 780bp R	TGC GGC AAA GAC AGC AAC GG

* Represents phosphorothioated DNA bases
All sequences are listed from 5' to 3'

Table 3.5: PCR primer sequences

3.4.5 Secondary liquid MDA and PCR screening - cultured *E. coli* and *S. aureus*

The retrieved hydrogel punch was dissolved and denatured in 1 μL of 1 M KOH with 0.1 mM EDTA and 0.1 M DTT at 72 °C for 10 min before neutralization in 1 μL stop solution (Qiagen REPLI-g single cell kit. Approximately 0.06 μL of hydrogel and 10 pg of DNA was captured for a cluster. The neutralized product was added to 12.5 μL REPLI-g sc reaction mix with 1 μL of phi29 polymerase. The secondary MDA reaction was incubated for 10 hours before polymerase deactivation at 65 °C for 5 min. The DNA products from MDA were cleaned by the SPRI procedure in 1.8:1 beads to DNA volume (Beckman Coulter). Each sample was analyzed for the presence of *S. aureus* and *E. coli* marker loci by four sets of primers (Table 3.5) in standard qPCR reactions with Jumpstart Taq 2× ready mix (Sigma Aldrich), 1× Evagreen (Biotium), 1× ROX (Invitrogen) and 1 μM primers in Stratagene M3005. Both melting curve analysis and agarose gel electrophoresis (not shown) were used to support the QPCR results.

3.4.6 WGS library construction and sequencing

We quantified the purified MDA products using the Qubit/Quant-IT HS assay (Thermo Fisher Scientific) and normalized samples to 5 ng/μL. All SPRI procedures were con-

ducted on the Bravo robotic system (Agilent Technologies). Purified DNA (5 ng) was then added to 1 μL of 5× tagmentation DNA buffer, 2 μL H_2O and 1 μL Nextera Tagmentation DNA enzyme (Illumina). The mixture was first incubated at 58 °C for 10 min. With the addition of 0.5 μL of 1% SDS, it was then incubated at 68 °C for 10 min, 4 °C for 3 min and 25 °C for 3 min to stop the tagmentation reaction. Another SPRI clean-up was carried out, followed by PCR library barcoding using Index primer N7 and S5 (Illumina) with the thermal protocol: 72 °C for 3 min, 98 °C for 30 sec, 12 cycles of 98 °C for 10 sec, 60 °C for 30 sec, 72 °C for 30 sec and a 5 min final extension at 72 °C. Samples were barcoded uniquely in the PCR step using standardized custom sample barcodes (Broad Institute Genomics platform). The PCR products were purified with SPRI twice with 1:1 beads to DNA volume and library quantification was carried out with the Quant-It assay (Thermo Fisher Scientific) and the KAPA library quantification kit (KAPA Biosystems). Library normalization and pooling were conducted on the Janus Mini Varispan workstation (PerkinElmer). For *E. coli* and *S. aureus* samples, an average of 0.7 million paired-end reads were allocated for each sample in a MiSeq 500 cycle v2 run (Illumina). For stool samples, about 1 M reads (> 50×) were allocated to each sample on HiSeq 2500 2×101/125 runs (Illumina).

3.4.7 NGS data analysis for *E. coli* and *S. aureus*

Data quality was first visualized using FastQC (Babraham Bioinformatics). All data were trimmed using trimmomatic [134] and human reads were filtered out with BLAST and QIIME. Each pair of trimmed and filtered reads was piped into BWA and mapped to the custom reference sequences (Fig. 3-6). SAMtools view was used to produce BAM files, and Picard tools (Broad Institute) deployed to mark duplicate reads. The data analysis workflow is illustrated in Fig. 3-5. The samples included two positive-control purified genomic DNA samples, seven hydrogel MDA punches identified as *E. coli* only by qPCR, seven punch samples identified as *S. aureus* only by qPCR, and seven punch samples identified by qPCR as double negative. Mapping statistics were obtained using the GAEMR (Broad Institute)

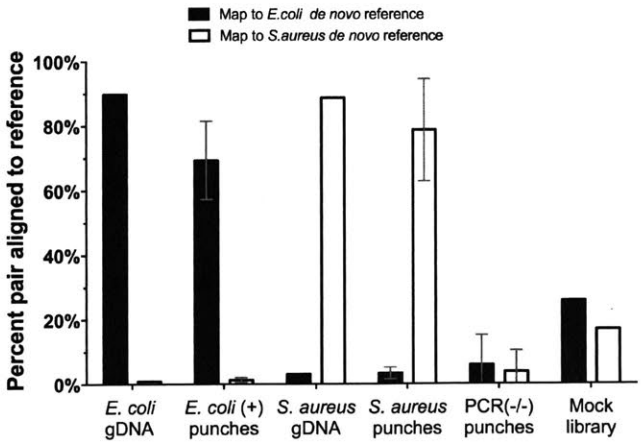

Figure 3-6: Mapping single-cell genomes to references. Single-cell samples were mapped to *E. coli* and *S. aureus* reference sequences with mean percent pair aligned and standard deviation shown. (n = 5 for *E. coli* punches, n = 7 for *S. aureus*, n = 7 for negative punches).

get_simple_bam_stats.py tool (Table 3.6). Genome coverage was obtained using Bedtools genomecov. Lorenz curves were obtained by first processing BAM files (duplicates marked) using samtools mpileup and then ranking the ascending coverage per base pair. Single-cell *E. coli* MDA data from de Bourcy *et al.* 2014 were downloaded from NCBI Sequence Read Archive (SRA) and analyzed by the same procedures.

Genome Coverage Completeness Estimation: note that some studies in the field report data from quality-filtered ('cherry-picked') cells, which dramatically improves quality statistics such as average coverage. In this study, we report data on complete sets of single-cell MDA reactions.

3.4.8 Custom reference generation by de novo assembly for *E. coli* and *S. aureus*

The gDNA *E. coli* (BL21) and *S. aureus* (8325) positive-control data were assembled and curated to create custom reference genome sequences. Raw sequencing files were quality trimmed using Trimmomatic (Fig. 3-5). We blasted the trimmed files against respective reference databases and filtered described previously with the parameters listed in in Fig. 3-5. Hit reads were filtered out of the sequencing files using the QIIME filter_fasta.py tool. The filtered and trimmed files were assembled into unordered

	# Raw reads	Median insert length (*E. coli*)	Median insert length (*S. aureus*)	Mapping to *E. coli* (%)	Mapping to *S. aureus* (%)	Coverage of *E. coli* genome	Coverage of *S. aureus* genome	Percent chimera on *E. coli* ref (%)	Percent chimera on *S. aureus* (%)
E. coli genomic DNA	1,888,337	120		89.86%	0.83%	99.99%		0.08%	
S. aureus genomic DNA	987,244		144	3.04%	88.70%		99.99%		0.12%
S1	873,658		150	1.81%	90.89%		64.35%		0.62%
S2	361,857		178	2.07%	89.02%		70.44%		0.84%
S3	594,951		153	5.17%	53.30%		53.85%		0.39%
S4	654,669		160	2.17%	90.16%		67.87%		0.57%
S5	470,077		155	5.32%	69.80%		66.95%		0.61%
S6	803,684		148	5.27%	64.47%		86.46%		0.49%
S7	450,240		166	1.39%	93.04%		47.34%		0.90%
*E1**	*822,345*	137		*9.81%*	*0.59%*	*7.14%*		0.68%	
E2	473,648	134		61.12%	0.64%	23.85%		0.69%	
E3	425,227	135		77.35%	2.38%	47.59%		0.55%	
E4	527,854	142		55.20%	1.01%	30.52%		0.78%	
*E5**	*1,200,678*	121		*9.76%*	*3.36%*	*14.10%*		0.56%	
E6	745,092	124		85.31%	1.56%	16.77%		0.84%	
E7	1,053,059	124		68.43%	0.38%	23.79%		0.52%	
NTC1	972,597			10.89%	16.96%				
NTC2	671,370			0.51%	1.59%				
NTC3	715,428			0.23%	0.20%				
NTC4	508,975			0.12%	0.07%				
NTC5	784,944			0.99%	3.27%				
NTC6	579,921			84.03%	0.06%				
NTC7	605,265			22.64%	0.13%				
Mock	27,429			25.71%	16.81%				

* False-positive *E. coli* single-cell samples

Table 3.6: Mapping statistics, *E. coli* and *S. aureus*

Assembly statistics	*E. coli* gDNA	*S. aureus* gDNA	*E. coli* gDNA (de Bourcy *et al.*)
# Contigs ≥ 0bp	126	113	115
# Contigs ≥ 1kbp	101	86	91
Total length ≥ 0bp	4,406,278	2,678,216	4,432,657
Total length ≥ 1kbp	4,396,297	2,666,092	4,422,013
Largest contig	295,162	148,028	326,226
Coverage	97%	95%	95%
GC %	50.80	32.69	50.75
N50	75,214	48,471	85,192

Table 3.7: *de novo* assembly statistics, *E. coli* and *S. aureus*

contigs with velvet. We mapped (BWA) unordered contigs to their closest NCBI reference genome (NC_012971 and NC_007795.1 respectively). The resulting SAM files were ranked on mapped length in descending order. Using a custom MATLAB function, we created a reference genome backbone consisting of only '-' with the same size as the reference genome and wrote sequences on it with only the top SAM mapping sequence for each contig. We conducted the same assembly process for the genomic DNA (*E. coli* DH10B) data from de Bourcy *et al.* 2014 using reference genome NC_010473.1. Assembly statistics are listed in Table 3.7. Custom MATLAB function, python codes and shell scripts are included in the supplementary software zip file in Xu *et al.* 2016.

Random subsampling of mapped reads for *E. coli* and *S. aureus* Duplicates-marked BAM files were down-sampled using samtools and bootstrapped with random number seed 0 to 9 for each depth. See Table 3.8 for more information.

3.4.9 Chimera statistics for *E. coli* and *S. aureus*

To make the chimera statistics comparable, we used *de novo* assembled genome sequences (described below) from bulk genomic DNA samples as the reference. We mapped read 1 and read 2 from each sample single-ended using BWA. We sorted the SAM file by read index. We used a custom python code to import pysam in order to pair up the 'read index', 'mapping position', 'is-reverse', and 'read length' information into a .mat file. With a customized MATLAB script, we calculated the insert size

Filename	Mapped reads	Fraction: 20X	10X	5X	Source
SRR1614004	1,953,388	0.181	0.091	0.045	de Bourcy tube
SRR1614005	124,353	--	--	0.712	de Bourcy tube
SRR1614006	1,552,882	0.228	0.114	0.057	de Bourcy tube
SRR1614007	222,175	--	0.798	0.399	de Bourcy tube
SRR1614011	2,191,740	0.162	0.081	0.040	de Bourcy MF
SRR1614012	2,034,992	0.174	0.087	0.044	de Bourcy MF
SRR1614013	5,476,888	0.065	0.032	0.016	de Bourcy MF
SRR1614014	2,631,391	0.135	0.067	0.034	de Bourcy MF
SRR1614015	11,448,119	0.031	0.015	0.008	de Bourcy MF
SRR1614016	3,833,180	0.092	0.046	0.023	de Bourcy MF+T
SRR1614017	2,720,938	0.13	0.065	0.033	de Bourcy MF+T
SRR1614018	2,377,409	0.149	0.075	0.037	de Bourcy MF+T
SRR1614019	2,217,047	0.16	0.08	0.04	de Bourcy MF+T
SRR1614020	5,331,247	0.066	0.033	0.017	de Bourcy MF+T
LX1	1,690,339	0.208	0.104	0.052	*E.coli* gDNA (No MDA)
LX11	283,308	1*	0.62	0.31	*E.coli* hydrogel
LX12	324,651	1*	0.54	0.27	*E.coli* hydrogel
LX21	259,540	1*	0.68	0.34	*E.coli* hydrogel
LX23	571,435	0.62	0.31	0.15	*E.coli* hydrogel
LX24	708,610	0.5	0.25	0.12	*E.coli* hydrogel
LX3	728,290	0.29	0.15	0.07	*S.aureus* hydrogel
LX4	335,567	0.64	0.32	0.16	*S.aureus* hydrogel
LX5	321,849	0.66	0.33	0.17	*S.aureus* hydrogel
LX6	609,813	0.35	0.18	0.09	*S.aureus* hydrogel
LX7	335,862	0.64	0.32	0.16	*S.aureus* hydrogel
LX8	528,641	0.4	0.2	0.1	*S.aureus* hydrogel
LX9	435,904	0.49	0.25	0.12	*S.aureus* hydrogel

* 20X coverage was not obtained for these samples: all available data were used

Table 3.8: Downsampling on mapped reads from single-cell MDA samples

for each read pair and checked their relative orientation. We filtered out pairs that were mapped one-ended. The chimera percentage was calculated as the (number of properly orientated read pairs with insert size more than 1000 bp + number of read pairs of wrong orientation)/Total number of read pairs.

3.4.10 In-gel single-microbe MDA - human gut microbiome samples

We received ethics approvals for human subjects research from the Columbia University IRB, Massachusetts Institute IRB, Broad Institute IRB, and two research ethics committees in Fiji: HRERC at CMNHS, FNU and FNHRERC at MoHFiji Ministry of Health. The Fiji Community Microbiome Project (FijiCOMP) study participants from 5 agrarian villages within the Fiji Islands. This project provided stool samples stored in 20% glycerol within 30 minutes of voiding and were frozen at –80 °C. Five participants were analyzed for single-cell studies: M1.20, W2.21, WL.26, W2.33, M2.41 (Table 3.9). 10 μL of thawed cells were resuspended in 500 μL PBST (0.1%). Samples were sonicated for 20 seconds and filtered through 35 μm Nylon mesh and 5 μm membrane (Pall Corp.) to collect filtrate with a 500 μL PBST wash. Samples were further diluted 1 to 500 $\sim$ 1 to 2000 fold in PBST to reach the final concentration of $\sim$ 30 cells/μL. The diluted cell samples (2 μl) then underwent alkaline lysis (1.5 μL D2 buffer) for 15 minutes at room temperature, after which the solution was neutralized (1.5 μL Stop solution). Hydrogel monomer mix (1.3 mg 4-Arm PEG Acrylate and 0.9 mg SH-PEG-SH) and MDA master mix were gently pipetted down the wall of each sample tube. MDA master mix includes 1× phi 29 buffer (NEB), 50 μM random hexamers with two phosphorothioate bonds at 3' terminus, 2.5% DMSO, 0.4 mM dNTP, 0.5 mg/mL BSA, 500 nM SYTOX Orange (Invitrogen) and 1 μL REPLI-g SC Polymerase (Qiagen). Only gentle tapping was used to ensure reagent mixing, in order not to disrupt the lysed microbes and denatured genomes. 25 μL of each microbial suspension was added into a frame-seal chamber, the sealed chamber was incubated at 30 °C for 12 hours, followed by 65 °C for 5 mins.

Sample ID	M1.20	W2.21	WL.26	W2.33	M2.41	Weighted average
Prevotellaceae	27.0	46.0	61.9	28.9	52.1	44.0
Succinivibrionaceae	0.0	0.0	0.0	0.0	0.0	0.0
Clostridiaceae	0.6	0.0	0.9	0.0	0.0	0.2
Bacteroidaceae	0.0	0.1	1.3	0.0	4.1	1.2
Veillonellaceae	0.4	3.4	2.9	4.4	9.0	4.6
Firmicute	0.0	0.0	0.0	0.0	0.0	0.0
Enterobacteriaceae	0.1	33.3	1.2	0.3	0.3	6.4
Lachnospiraceae	8.1	4.2	4.7	16.5	4.1	8.4
Eubacteriaceae	19.3	3.4	7.1	29.5	6.6	14.2
Ruminococcaceae	16.7	5.6	2.8	13.5	7.3	8.7
Megasphaera	0.0	0.0	0.0	0.0	0.0	0.0
Acetobacteraceae	0.0	0.0	0.0	0.0	0.0	0.0
Acidaminococcaceae	8.3	0.6	11.0	0.6	0.0	3.2
Clostridiales	0.0	0.0	0.0	0.0	0.0	0.0
Erysipelotrichaceae	16.0	2.0	0.8	3.4	1.3	3.0
Total (%)	96.5	98.5	94.7	97.0	84.7	94.0
Single cell count	8	21	25	37	26	All =117
Single cell percentage	7%	18%	21%	32%	22%	All= 100%

Table 3.9: Metagenomic shotgun profiling weighted with single-cell samples

Secondary In-gel MDA - human gut microbiome samples Hydrogel punches (approximately 0.24 μL of hydrogel and 10 pg of DNA if a cluster was captured) were dissolved and denatured in 1 μL of 400 mM KOH with 0.1 mM EDTA and 0.1 M DTT at 72 °C for 10 min before neutralization in 1 μL stop solution (Qiagen REPLI-g single cell kit). The neutralized product was added to 8 μL hydrogel and MDA master mix to reach a final volume of 10 μL for second round MDA reaction in hydrogel. The MDA reaction was incubated for 10 hours at 30 °C before polymerase deactivation at 65 °C for 5 min. The 10 μL gel was dissolved with 10 μL 400 mM KOH for 5 mins at 72 °C, and then neutralized with 6.6 μL 2.5% acetic acid.

Pre-processing and assembly of single-cell genomes from stool First, we removed the adapter sequences from single-cell libraries using TRIMMOMATIC [134] (TRAILING:3 MINLEN:40). To ensure that human DNA was not captured in our single-cell libraries, we screened single-cell amplicons against the human genome (GRCh38 reference) using BMTagger [135] (default). We screened our amplicons

against *E. coli* references (BL21 and DH10B) using BMTagger. Overall, the level of contamination was small (around 0.01%). We also screened against *Pseudomonas* (PAO1) and *Staphylococcus* (NCTC 8325) genomes, which were sequenced alongside our libraries, to ensure no chimeric reads formed during sample preparation with contaminating sequences from other cultures in our lab that confounded our analyses. Finally, single genome amplicons were quality filtered (Phred score > 3), and filtered for reads that were less than 45 bp. Amplicons were then assembled using SPAdes (v3.6.0) (–careful) [136]. We retained genomes where at least 100 kb could be assembled.

Assessing the fidelity of single-cell genomes from stool To further vet the quality and purity of our assemblies, we used BLAST to assign taxonomies to a set of 31 predetermined core genes that are both phylogenetically conserved and single copy in almost all genomes [132]. Although we could not identify the full set of 31 core genes in any of the assemblies, we were able to easily distinguish cases where two or more cells were sequenced together from those in which there was a single cell. Additional validation of the single-cell assemblies included quantifying the levels of contamination using CheckM [137] and examining the number and taxonomy identified using RNAmmer [138]. CheckM accesses the quality of a genome using a broader set of marker genes specific to its inferred lineage within a reference genome tree and provides estimates of genome completeness and contamination percentages. RNAmmer uses hidden Markov models trained from ribosomal RNA databases to predict the rRNA species. The extent and contiguity of our assemblies was documented by reporting assembled genome size, N50, the number of contigs, CheckM completeness percentage, CheckM contamination percentage and notes on RNAmmer classification in an Excel file in Xu *et al.* 2016.

Notably, some microbes can be difficult to isolate from human stool samples due to the cells' tendency to break or aggregate. Some of the punch samples with low numbers of AMPHORA genes could be the result of broken cells containing reduced genomic representation or free genomic DNA fragments, while samples with evidence

for multiple taxonomies could have resulted from cell aggregates. Stool samples are also fairly complex and contain a lot of particulate matter that complicates sample processing. In principle, genes from samples with sequences of variable taxonomy could arise for several reasons: the products of multiple cells being collected in a single punch, downstream contamination in the second round MDA or library construction steps, informatic demultiplexing, or from taxonomic mis-classification of hard-to-assign sequences.

Analysis of metagenomic shotgun reads from stool FijiCOMP metagenomic samples, each containing roughly 50 million paired-reads, were profiled using MetaPhlAn [131]. Metagenomic samples were also aligned to the SILVA rRNA database (v.115) to determine the presence of organisms from the Succinivibrionaceae family. Based on alignments to the SILVA rRNA database, we find that organisms within the Succinivibrionaceae family are in fact highly abundant in the FijiCOMP metagenomic data, with average FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values around 26,000.

Chapter 4

The characterization of chimeric DNA rearrangements in single amplified human genomes across innovative single-cell technologies

4.1 Introduction

Transposable elements (TEs, 'jumping genes') are discrete pieces of DNA that can move within the genome of a single cell and between the genomes of different cells. Nearly 45% of the human genome is derived from TEs [139, 140]. Studies have shown that TEs can cause mosaic copy-number variations (CNVs) and structural variations (SVs) on genes such as PIK3CA, AKT3 and mTOR during prenatal brain development. These mutations could result in brain malformation and neurological defects, including epilepsy, intellectual disability and hemimegalencephaly [141, 142]. Thus, it is important to identify such mutation mechanisms and the affected diverse cell types that disrupt the function of the cortical circuits. In order to achieve this, targeted qPCR assay can be used to detect an increase in the copy number of TE. However, a critical limitation is that the genomic location of the new insertion cannot

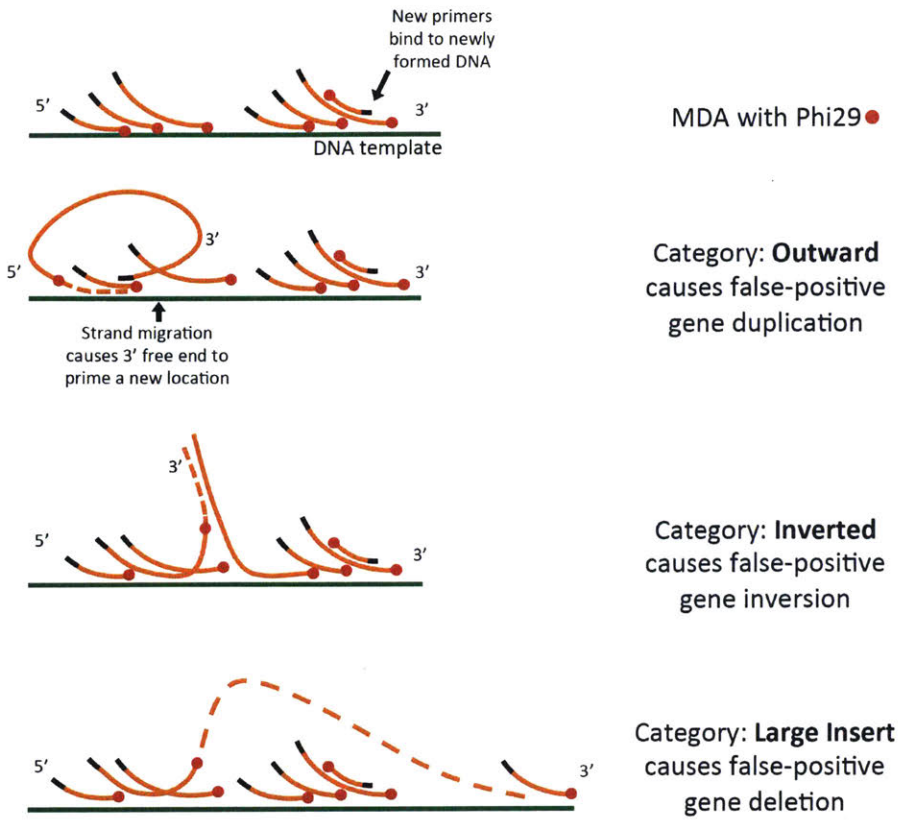

Figure 4-1: The mechanism of MDA chimera. The strand displacement action of MDA is shown on top. The strand migration dynamic causes the 3' end to freely bind to another part of the genome, resulting in outward, inverted and large-insert chimera. Cross-chromosome chimera is not shown but the mechanism is similar.

be identified [143].

Single-cell targeted sequencing and whole genome sequencing approaches have enabled the location identification of novel brain-specific TE insertions. But the results are often confounded by the inherent false discovery rate of current technology [143]. Single-cell technology has also been widely applied in several other scientific and biomedical field. For example, the screening of embryos using single-cell technologies on polar bodies and blastomeres has shown improved *in vitro* fertilization success by eliminating embryos with a high frequency (30 × above baseline) of chromosomal rearrangements, which often lead to miscarriages [144]. This technique also lowers the rate of Mendelian disease through genome-wide single nucleotide variations (SNVs) screening on the embryos [145, 146].

The data quality of single amplified and sequenced cells is a key to providing

accurate and trustworthy information for above-mentioned applications. During the genome amplification and library preparation steps, artifacts are often generated that confound the detection of genomic signatures and complicate data analysis. These artifacts are often due to the implementation of the multiple displacement amplification (MDA) reaction and PCR-based sequencing library preparation [130]. One type of artifact generated by MDA–chimeric DNA rearrangements (Fig. 4-1), caused by the highly branched DNA secondary structures with free 3' ends of single-strand DNA during the reaction [130], has raised flags in several single-cell studies. It has been shown that the preimplantation genetic diagnosis guided by single-cell genomics can be affected by false-positive SNVs and structural variations from MDA chimera [28, 147]. Researchers using single neurons to study developmental disorders discovered the complication of the chimera artifacts in identifying novel retrotransposon L1 as the existence of the chimera caused false-positive identifications of structural variations [5, 148]. This is also a problem for environmental microbes that lack a closely related reference genome. But in this chapter, we focus on human cells.

Experimentally, several studies have made attempts to reduce such chimera artifacts. It has been shown that nanoliter microfluidic device might generate less MDA chimera than microliter samples [20]. However, nanoliter to picoliter microfluidic devices often require clean-room fabrication and supporting pneumatic instruments, which makes it hard to implement for labs with limited resource and funding. Zhang *et al.* used a combination of Φ29 polymerase debranching, S1 nuclease digestion and DNA polymerase I nick translation to reduce the chimeric rate in the library preparation step for single microbes [95], but such methods do not directly affect the chimera generated during MDA reaction. A recent study by Picher *et al.* used a modified Φ29 enzyme for whole genome amplification but didn't show evidence of reduced chimera compared to using unmodified Φ29 [149]. A novel single-cell technology that is highly accessible and provides high-quality data, especially on reducing the MDA chimera artifact, is greatly needed.

Bioinformatically, the majority of single-cell studies have focused on characterizing coverage uniformity, CNVs, SNVs, purity and throughput metrics, but have

Technology Comparisons	Description	Engineering requirement	Throughput Bottleneck	Reagent addition	Product recovery
Virtual Microfluidics	MDA in hydrogel matrix	No special equipment needed	Possion loading	Diffusion into or out of hydrogel	Physical punch or hydrogel breakdown
Emulsion WGA (Fu *et al.* 2015)	A single cell divided in 10^5 picoliter aqueous droplets with MDA	Droplet generator	Single-cell isolation	N.A.	Droplet breakdown
Nanoliter droplet (Leung *et al.* 2016)	Nanoliter droplets dispensed on a planar substrate	Commercial liquid dispensing system	Possion loading	Sequential one by one addition	Sequential one by one retrieval
LIANTI (Chen *et al.* 2017)	Linear amplification via transposon insertion	Mouth pipetted or FACS sorted	Single-cell isolation	N.A.	N.A.

Table 4.1: Single-cell technology comparisons for chimera analysis

often neglected to quantify the amount of artifacts generated from single-cell whole genome amplification process and sequencing library preparation steps that affect assay performance. There exist established bioinformatic tools designed to filter chimera artifacts from 16S PCR reactions (comparing the phylogenies of fragments) [150, 151] and single-cell RNA-seq experiments (with the matching of unique molecular identifiers and cell barcodes) [152]. However, existing tools are not designed for chimera characterization and filtering in single MDA-amplified human genomes. A detailed bioinformatic characterization of MDA chimeras is needed to evaluate single-cell technologies and datasets that have been developed and produced.

In this study, we present the application of *virtual microfluidics* on single human cells to demonstrate our technology's capabilities in producing high-quality single amplified human genomes with minimum equipment requirement. We benchmarked its performance with recent innovations of single-cell technologies (Table 4.1) based on exponential amplification method—MDA (eWGA, Nanodrop) and quasi-linear amplification method—LIANTI and MALBAC [66, 89, 107, 108]. *Virtual microfluidics* has been shown as a hydrogel-based single-cell isolation and amplification technology that is highly accessible and can provide high-quality single-cell whole genome amplification (WGA) product on cultured bacteria and human gut microbiome samples [26]. We have previously demonstrated its advantages on small genomes (bacteria) in terms of multi-fold chimera reduction and coverage uniformity improvement. Our hypothesis is that the single amplified human genomes in *virtual microfluidics* will

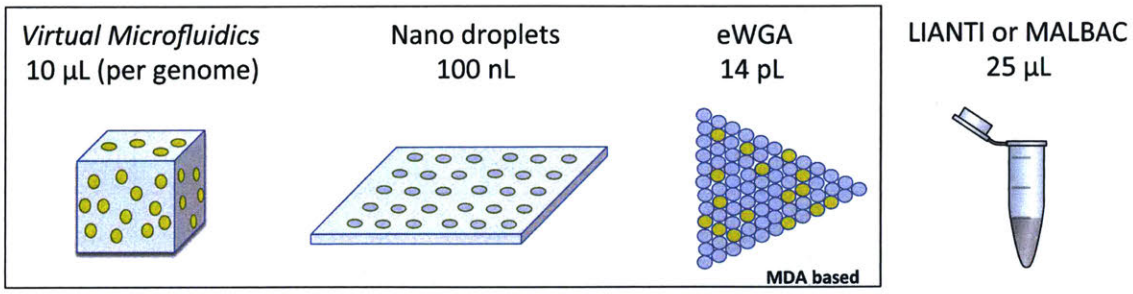

Figure 4-2: Comparison of single-cell technologies. The volume labeled is for each single-cell genome amplification reaction.

have a similar level of high-quality data to the previous study, with reduced chimeric DNA rearrangements and improved coverage uniformity.

As we mentioned in Chapter 3, the restricted diffusion of the MDA intermediates prevents cross-priming by isolating each portion of the product mixture. It limits cross-chromosome artifacts and chimera reads of large insert size (the distance between the forward read and reverse read mapped on the genome, Δ mapped coordinates + the length of reverse read). The hydrogel's micro-environment also physically limits the secondary DNA structures during MDA, thus reducing the chimera breakpoints out of total reads that are pair-mapped. To support the technology demonstration, we developed an algorithm to categorize the signatures of MDA chimeras across multiple single-cell platforms. This characterization will serve as a guidance for chimera analysis to be a non-negligible part of the analysis suite for evaluating future single-cell technologies.

4.2 Results and Discussion

We conducted modified *virtual microfluidics* single-cell sequencing on 8 RPE–1 cells in a HiSeq 2500 lane with 2×125 and obtained roughly 1× mapping depth to reference genome GRCh37-lite (methods). We characterized the single RPE cell dataset while benchmarking with unsorted/(unclear whether cherry-picked) single-cell datasets from eWGA (5 cells), MALBAC (2), tube MDA (2), Nanodrop (9) and LIANTI (3) (Table 4.2, Table A.1 and Fig. 4-2). All samples were trimmed and analyzed according

Data source	Number of single cells analyzed out of total available (0.1~10X depth)	Sequencing methods	Library Prep method	Amplification fold	Median insert size (bp)	Cell line
Virtual Microfluidics	8 MDA single cells	PE, 2×100bp HiSeq	PCR enriched (PCR free for bulk genomic DNA)	0.05e4~3e4	110	hTERT RPE-1 ATCC® CRL-4000
Emulsion WGA (Fu *et al.* 2015)	5 out of 10 eWGA 2 MDA (tube) 2 MALBAC (tube)	PE, 2×100bp HiSeq/MiSeq	PCR enriched	~2e6	160	HUVEC
Nanoliter droplet (Leung *et al.* 2016)	8 out of 15 MDA high depth 9 out of 95 low depth	PE, 2×125bp HiSeq	PCR enriched	~1e4	190	184- hTERT
LIANTI (Chen *et al.* 2017)	3 LIANTI cells	PE, 2×125bp HiSeq	PCR enriched	3300	300	BJ ATCC® CRL-2522

Table 4.2: Data source for chimera analysis

to the analysis workflow (Fig. 4-3). All mapped, de-duplicated, repeat-masked and sorted BAM files are down-sampled to 430,000 reads/sample ($\sim 0.01\times$, including both forward and reverse reads) for chimera analysis (methods) .

The benchmarking datasets were chosen to represent different methods of cell isolation (hydrogel-based *virtual microfluidics*, emulsion droplets, liquid dispensing), WGA chemistry (MDA, MALBAC, LIANTI) and library preparation methods (Nextera PCR-enriched and PCR-free ligation-based). According to Evrony *et al.* 2015, most chimera originated from library preparation after MDA reaction. However, our previous study showed a multifold chimera rate reduction from MDA process in cultured bacteria compared to standard tube reactions using the same library preparation procedure (Nextera). Including above-mentioned datasets will help parse out the chimera characteristics and sources from single amplified human genomes.

It has been shown from a couple hundred sequencing reads of *E. coli* that over 85% of MDA chimeras are inverted read pairs, which means the forward and reverse read pairs don't have the correct orientation (inward facing) [130]. This type of sequencing result can be easily filtered out in a single-chromosome organism with the well-established reference genome, such as culturable bacteria. However, it becomes bioinformatically difficult when the reference genome is enriched with repeat islands in mammalian cells and when the draft genome is often inaccurate for unculturable environmental microbes. Making improvements in single-cell technologies is a task

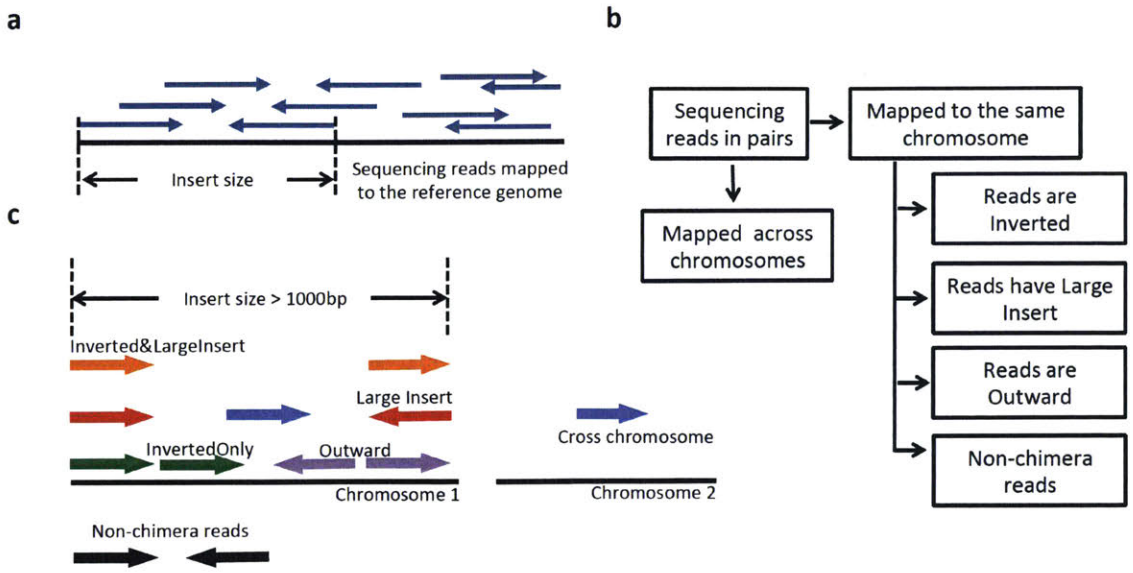

Figure 4-3: Pair-ended sequencing provides read orientations for the chimera categorization. (a) The insert size represents the length of DNA fragment. (b) The decision workflow to categorize different chimera reads. (c) Chimera and non-chimera reads are illustrated.

that needs to be done well both experimentally and bioinformatically.

4.2.1 Chimera categories

A correct read pair should map to different strands (+/-, or sense and antisense) within the insert size range controlled by sequence library size selection (200 bp ~ 800 bp). We defined five categories of chimera reads in this study (Fig. 4-3bc). Inverted means forward and reverse reads mapped to the same strand of DNA template (Fig. 4-3c). Within the inverted reads categories, the pairs of reads can be further categorized into inverted&LargeInsert (>1000 bp insert size) and InvertedOnly (<1000 bp). For reads with the correct orientation, the pairs of reads can be categorized into LargeInsert Only chimera (>1000 bp) and cross-chromosome chimera. The color code in Fig. 4-3c corresponds to the same categorizations in Fig. 4-4 chimera quantification. The five categories are mutually exclusive and collectively exhaustive for all chimeras that can be parsed out.

The median insert size distribution of sequencing reads across different datasets

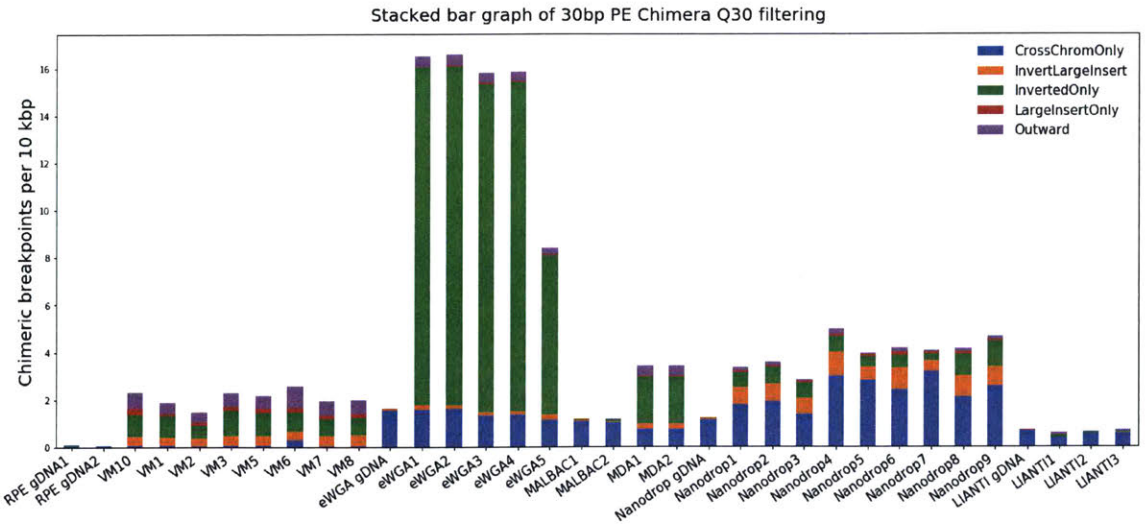

Figure 4-4: Chimeric breakpoints per 10 kbp across all samples are shown as 5 categories of chimera: cross-chromosome, inverted, large insert size (>1000 bp), outward, and inverted large insert.

varies from 100 bp $\sim$ 300 bp. In order to eliminate the factor of inconsistent insert size on chimera detection and possible read overlaps, we normalized the chimera breakpoints over total basepairs mapped and trimmed all sequencing reads to 30 bp [153]. Previous studies have revealed the nature of MDA chimera in terms of overlapping number of basepairs in the chimera junction [154, 153]. Here we focus on the categorization of MDA chimera in different single-cell whole genome amplification technologies on human cells and their genome-wide signatures. Single-ended and pair-ended mapping of the same sequencing dataset were implemented side by side to compare the read paring's impact on chimera detection. Here we define the chimera breakpoints as the total number of Read 1 and Read 2 that categorize as chimera reads (as there are two parts of the genome joined together, which represents two breakpoints).

$$\text{Chimera breakpoints per 10 kbp} = \frac{\text{Chimera reads } (R1 + R2) \times 10^4}{\text{Total number of bp mapped}}$$

$$= \frac{\text{Chimera reads } (R1 + R2) \times 10^4}{(\text{Total pairs of reads that are mapped}) \times \text{Average insert size}}$$

Fig. 4-4 shows the number of chimeric breakpoints per 10 kbp mapped in a stacked bar graph. Genomic DNA (gDNA) samples without WGA serve as the baseline of comparisons. *Virtual microfluidics* samples exhibit low chimera breakpoint frequency while confirming the 85% inverted composition in MDA chimera. RPE gDNA was processed with PCR-free library preparation and is used as the negative control of chimera detection. The residual chimera detected from the RPE gDNA represents the inherent genome structural variations of the cell line. This also serves as the comparison for the chimera introduced by PCR library preparation process. The eWGA, Nanodrop, and LIANTI gDNA all went through PCR enrichment. But it is unknown whether these 3 cell lines used (PCR-free datasets unavailable) have inherently higher structural variations than RPE cell line does.

Interestingly, the eWGA single cells show more than 90% enrichment in "Inverted Only" chimera reads and a $3 \sim 7$ fold chimera frequency increase compared to the gDNA baseline. This increase can be explained by the isolation of individual fragments in picoliter liquid droplets. The confinement of a single DNA template in picoliter volume resulted in mostly inverted artifacts from MDA while few template was available for cross-chromosome priming to happen. The nanodrop dataset has a different pattern for chimera signature, showing more than 50% chimeras that span across different chromosomes and 25% chimeras with large insert size. Both eWGA and nanodrop methods have a higher frequency of chimera reads occurrence compared to microliter-ranged in-tube MDA. Our explanation is that a smaller reaction volume does not affect the secondary structure of a single-strand DNA. But a higher density of DNA increases the chance of cross priming. Due to the smaller reaction volume that increases the DNA density in the reaction, more chimeras are produced in eWGA and Nanodrop.

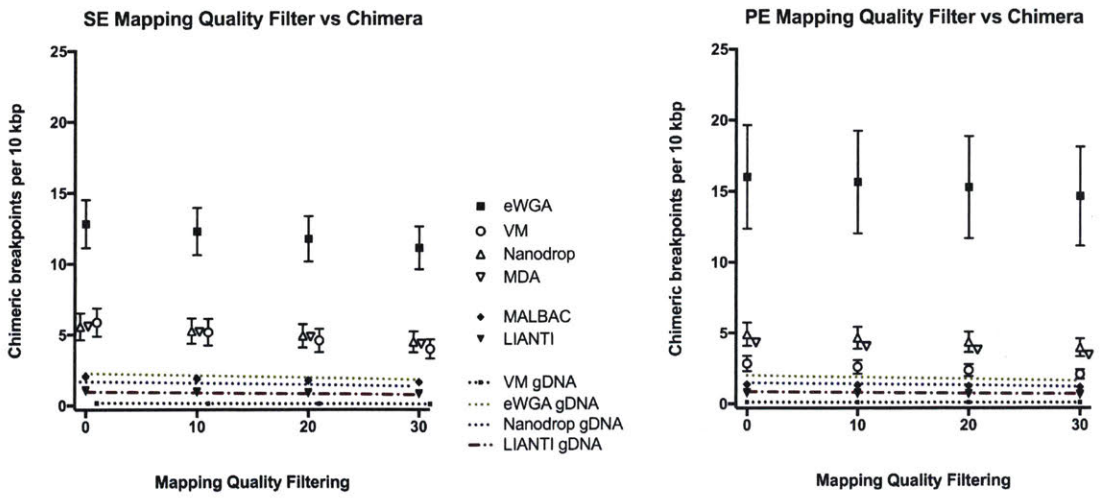

Figure 4-5: The effect of mapping quality filtering on chimera detection.

4.2.2 Chimera rate analysis with respect to mapping quality filtering

We implemented mapping quality filters (Q = 30, 20, 10) and found a slight decrease of chimera rate when increasing the mapping quality threshold as expected by the more stringent quality filtering (Fig. 4-5).

Most interestingly, we compared the effect of pair-ended and single-ended mapping on chimera detection (Fig. 4-6). For all datasets except eWGA, single-ended mapping analysis overestimates chimera rate compared to pair-ended mapping with the same downsampled reads. With pair-ended mapping, read pairs mapped within the range of insert size with the correct orientation is chosen as the primary mapping results. With the single-ended mapping, reads that mapped to multiple locations equally well were randomly chosen for the final output, thus, causing an overestimation of chimeric read out of total mapped reads. For eWGA, there is a 50% increase of chimera frequency by pair-ended mapping compared to single-ended mapping. This increase can be explained by the especially high content of inverted chimera that can be easily detected in the pair-end mode. In the single-end mode, a potential inverted chimera might be able to map to a reverse-complementary location with the same mapping quality, and thus, the single-end mode underestimate the chimera rate. We

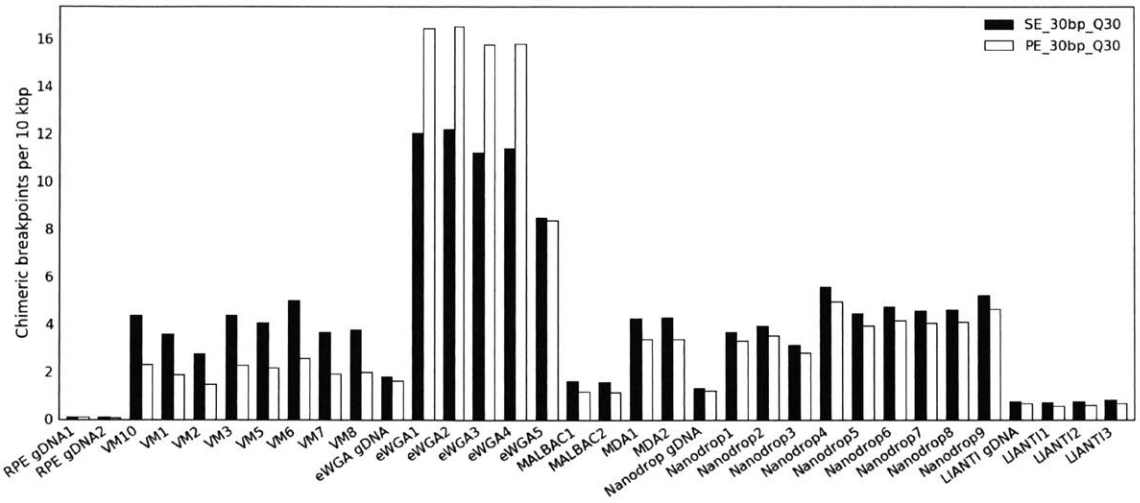

Figure 4-6: The effect of pair-ended (PE) and single-ended (SE) mapping on chimera detection.

are uncertain about why VM samples are affected the most by PE vs SE mapping ($\sim$ 50% change). Possible explanations include the differences between cell lines, library preparation procedures and the nature of DNA amplification in hydrogel. Further experiments will be needed to validate the cause.

Furthermore, by hierarchically clustering the datasets based on their chimera signatures (Fig. 4-7), we see the close similarities between *virtual microfluidics* samples and the LIANTI single cells in terms of chimera rate, while most of the nanodrop datasets are closely clustered with traditional MDA reactions. MALBAC samples are clustered with gDNA controls with PCR-enriched chimera baseline. LIANTI samples are closely clustered with PCR-free gDNA controls that represent chimera-free negative control, which indicates the *in vitro* transcription amplification could generate minimal chimera.

In order to visualize the location of chimera discovered, I plotted all chimera pairs throughout the genome in Circos plots (Fig. 4-8 and 4-9) [155]. All chimera pairs shown were based on the normalization of 430K mapped sequencing reads for each sample. Each Circos plot shows the chimera locations across the entire set of chromosomes. The VM gDNA represents the RPE bulk genomic DNA sample without any PCR enrichment, indicating that the unamplified genomic DNA contains a baseline of

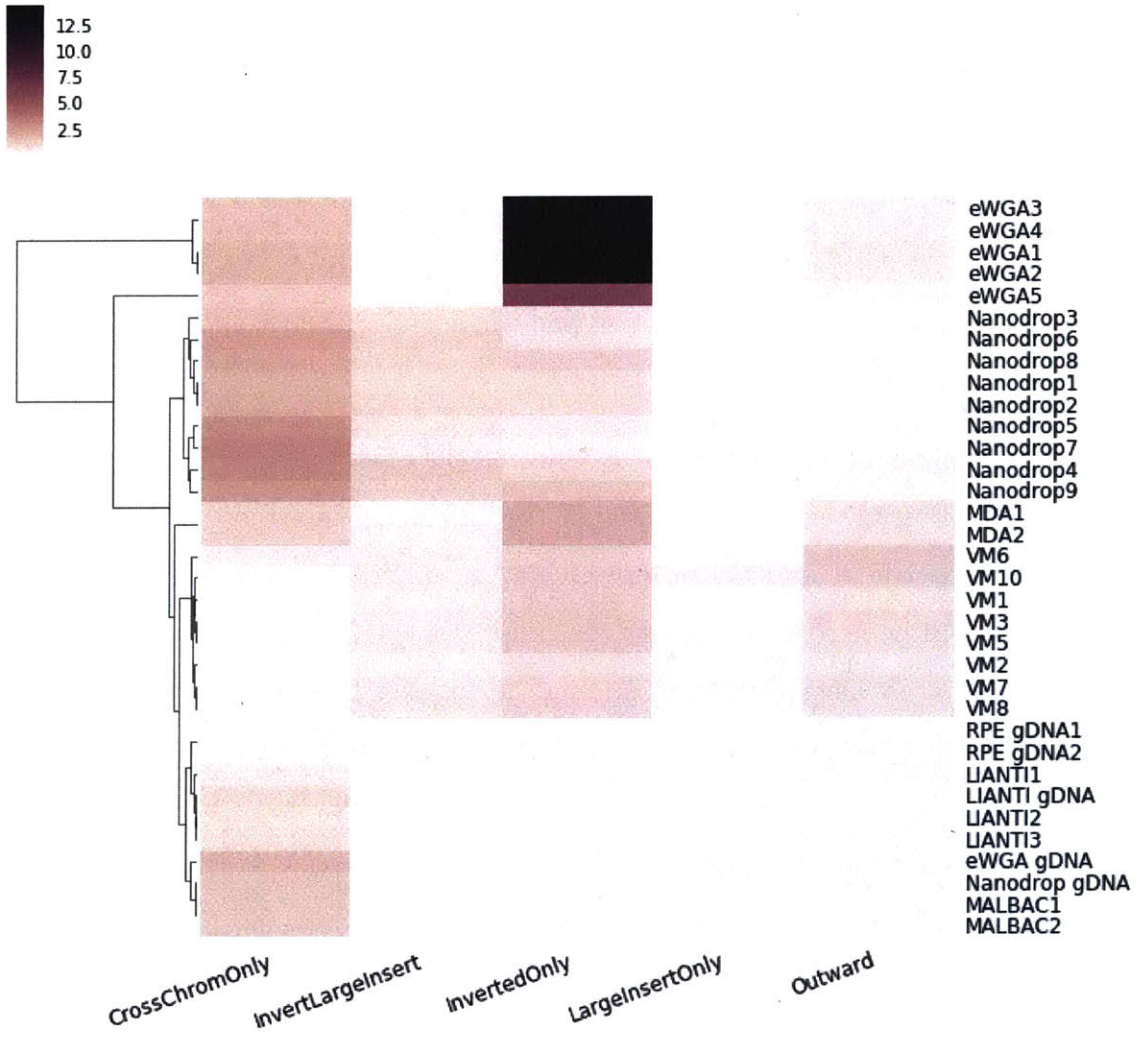

Figure 4-7: Hierarchical clustering of chimera breakpoints per 10 kbp for each single cells and genomic DNA controls. *Virtual Microfluidics* samples are closely clustered with genomic DNA controls, indicating low levels of chimera generated by hydrogel MDA

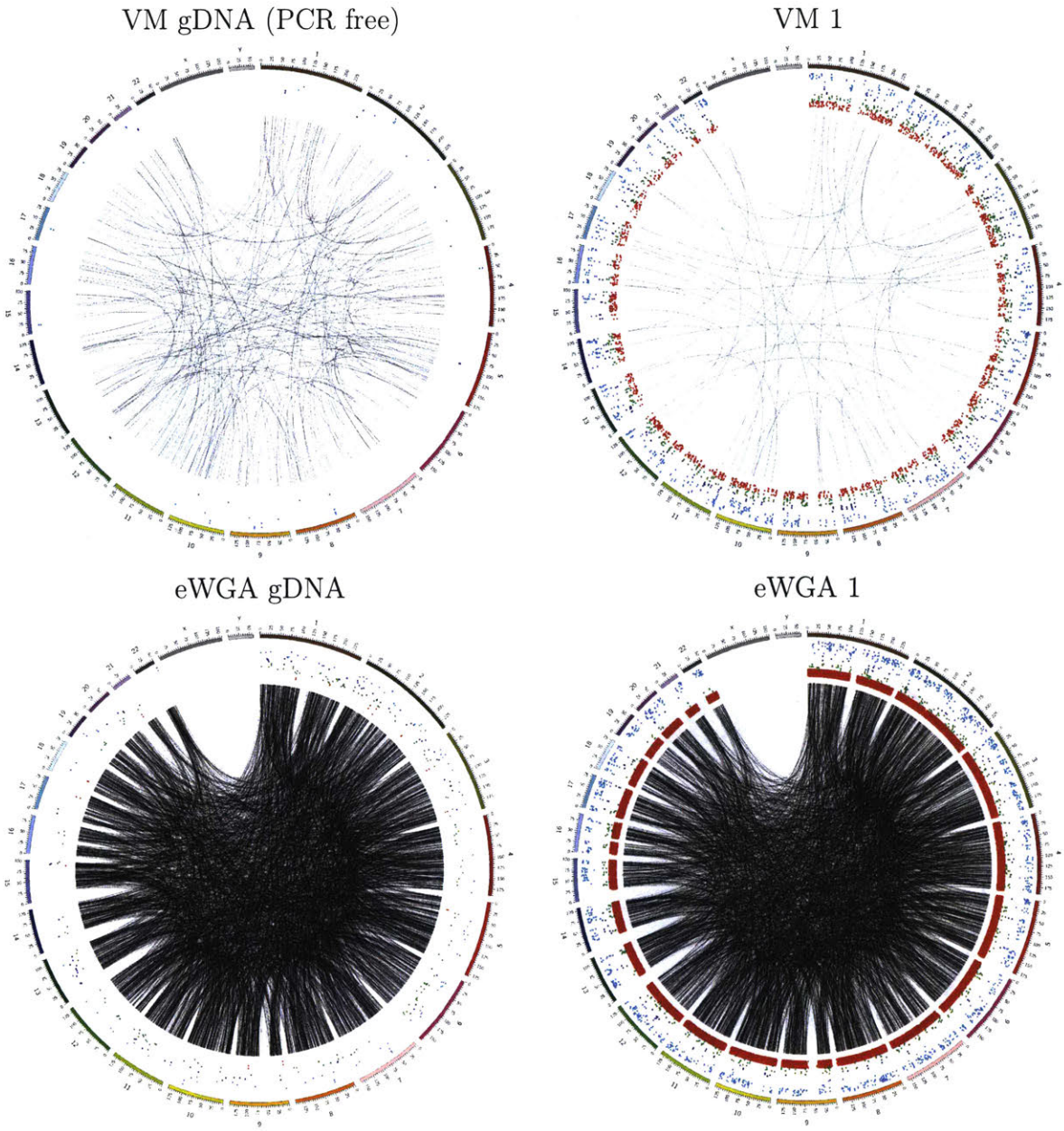

Figure 4-8: Chimera breakpoints shown in Circos plots for bulk genomic DNA and single-cell samples (part 1). Cross-chromosome chimera pairs are connected, shown as black lines in the center. Inverted chimera breakpoints are represented as red dots. Inverted & large-insert chimera breakpoints are green dots. Large-insert chimera breakpoints are purple dots. Outward chimera breakpoints are blue dots.

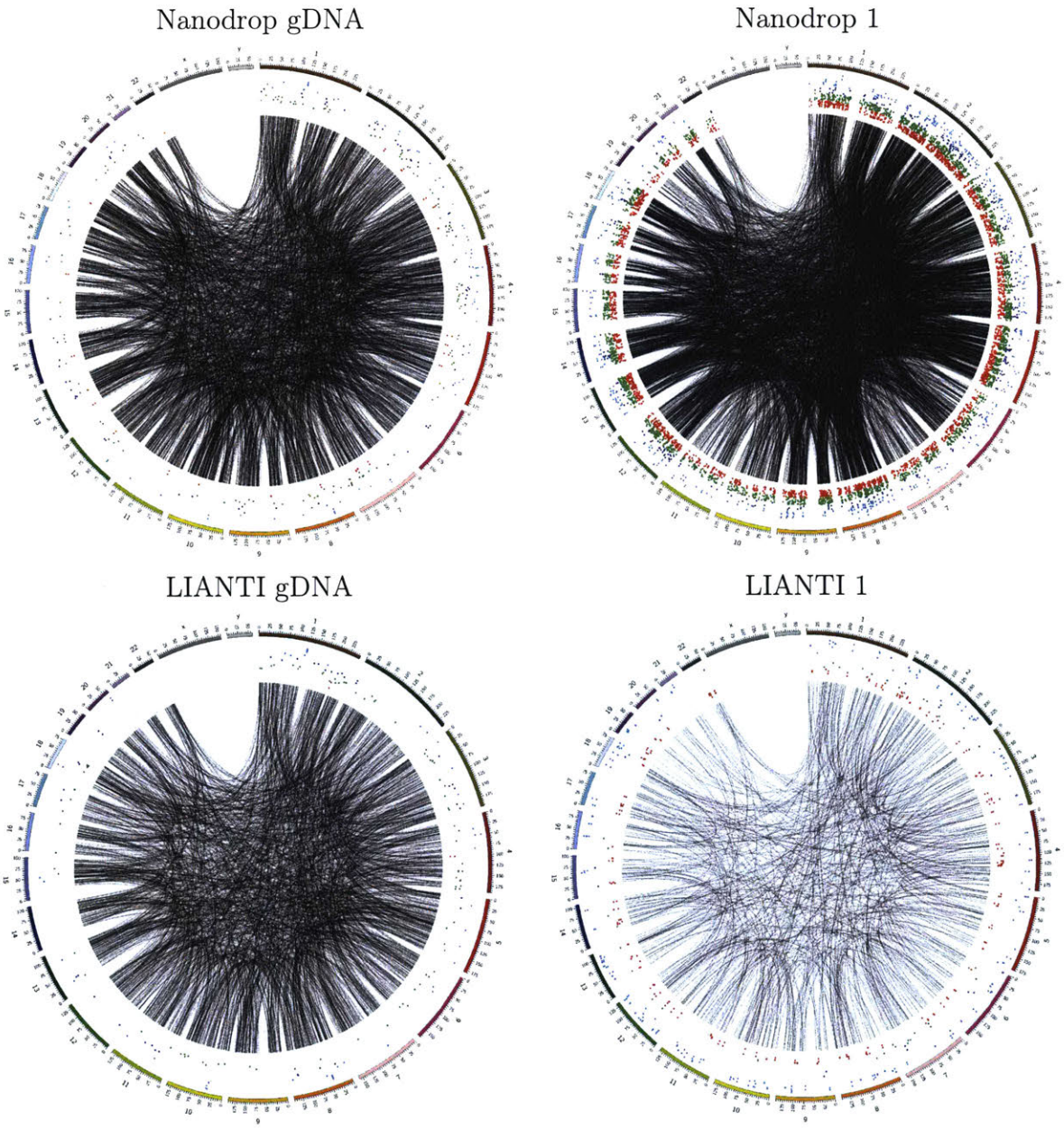

Figure 4-9: Chimera breakpoints shown in Circos plots for bulk genomic DNA and single-cell samples (part 2). Cross-chromosome chimera pairs are connected, shown as black lines in the center. Inverted chimera breakpoints are represented as red dots. Inverted & large-insert chimera breakpoints are green dots. Large-insert chimera breakpoints are purple dots. Outward chimera breakpoints are blue dots.

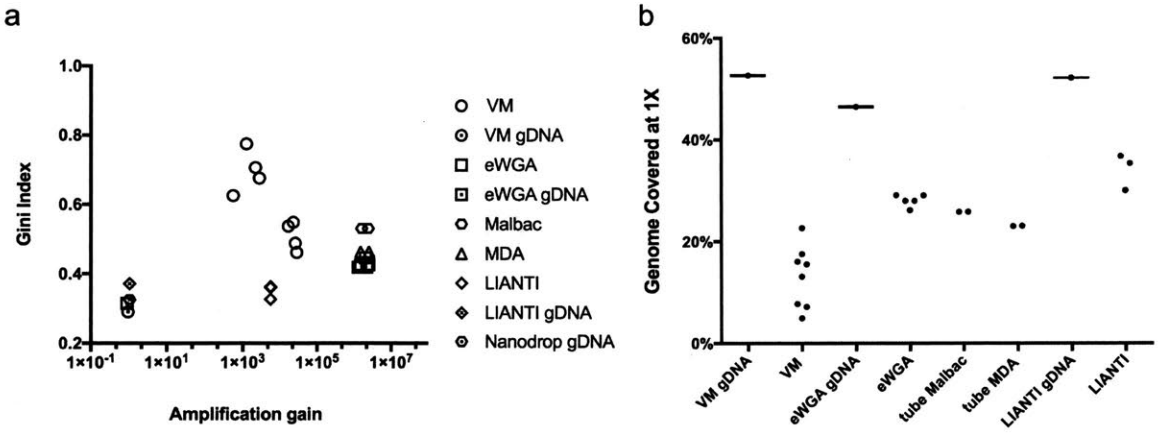

Figure 4-10: The coverage uniformity and the genome coverage performance. (a) The coverage uniformity is shown as Gini index vs amplification gain. Gini index of 1 means the maximum bias. (b) Genome coverage percentage is shown for all datasets.

structural variations with respect to the reference genome. The difference in frequencies of cross-chromosome chimera between the VM gDNA and other gDNA samples probably originate from PCR enrichment and are cell-line specific. This difference shows the difficulty in benchmarking chimera performances across different studies and the importance of standardizing single-cell model systems for future technology development and characterization.

4.2.3 Coverage uniformity and physical genome coverage performances

In order to evaluate the coverage uniformity and the physical genome coverage performance of single amplified human genomes, *virtual microfluidic* VM (8 cells), eWGA (5), MALBAC (2), MDA (2), LIANTI (3) were first downsampled to 1× mapped depth (about 30 million reads). Fig. 4-10a quantifies the coverage bias vs amplification gain. The coverage bias is quantified using the area under Lorenz curve and is represented as the Gini index. Including the amplification gain is important for quantifying coverage biases, as the literature has shown MDA over-amplification results in highly biased genomes [87]. Fig. 4-10b shows the genome coverage percentage across all samples at 1× mapping depth. *Virtual microfluidic* samples show

a range of performance in both coverage uniformity and physical genome coverage percentage. This is most likely due to the uneven amplification gain obtained for 8 different samples (from 28093 to 565 folds). Future experiments with amplification gain control of above 25000 fold should be able to produce much improved overall performance in the coverage uniformity and the genome coverage percentage. Overall, the LIANTI method shows superior performances in terms of coverage uniformity, physical genome coverage and chimera reduction. This new scheme of whole genome amplification might overtake MDA's dominant place in future single cell genomic applications.

The performances of single-cell analysis are often cherry-picked and selectively reported. The Nanodrop method's high sequencing-depth data were cherry-picked based on the quality of its low-depth dataset, thus we excluded it from the coverage and uniformity comparison. It is highly likely that eWGA and LIANTI single cells are cherry-picked as the best subsets but it is not yet confirmed. The effect of cherry-picking, known as the fallacy of incomplete evidence, gives a false impression on the overall quality of single cell sequencing technologies and inflates performance measurements such as coverage uniformity and genome coverage. In contrast, the 8 VM cells were the entire dataset that went through MDA, library preparation and sequencing (without cherry picking). I believe there is a great potential in improving data qualities and *virtual microfluidics* measurements represent the foremost of single-cell technology platform to this date.

4.3 Conclusion

In conclusion, *virtual microfluidics* enables high-quality single-cell genome sequencing with 1 (compared with Nanodrop) $\sim$ 8 fold (compared with eWGA) chimera rate reduction in MDA reaction while only requiring basic bench tools. It eliminates the need of creating ultra-small discrete chambers for sub-microliter MDA reactions. This chapter also showcase the importance of quantifying chimeric DNA rearrangements from single-cell genomic amplification and library preparation processes. Such study

is important in providing a baseline analysis of the chimera signatures and can be used for predicting the amount of false positive DNA rearrangements that are of interests to prenatal and cancer diagnosis. *Virtual microfluidics* also has a potential as a flexible platform for combining new WGA chemistry such as LIANTI and library preparation methods involving *in situ* tagmentation that will push the throughput and data quality from single-cell WGA to a new level.

4.4 Materials and Methods

4.4.1 Experimental methods

The hTERT RPE–1 (ATCC) cell line stably expressing GFP-H2B were cultured in 10% final concentration of fetal bovine serum (FBS) and 0.01 mg/ml hygromycin in ATCC-formulated DMEM:F12 Medium (Catalog No. 30–2006). When the culture was at $> 80\%$ confluence, it was serum-starved for 12 hrs overnight for cell cycle synchronization. A blank Costar 384-well plate with glass bottom was imaged for GFP fluorescence and under the white light before cell deposition. Cells were trypsinized, counted and diluted to 1 cell/μl, and 1 μl was added to each well of the 384 plate. The plate was spin down briefly and imaged for GFP fluorescence and under white light to confirm single-cell occupancy in each well. To the wells with single cells, 4 μl of lysis buffer (30 mM Tris-HCl, 10 mM NaCl, 5 mM EDTA, 0.5% Triton X–100 and 1 mg/ml proteinase K) with hexamer of final concentration 50 μM was added and heated at 50 °C for 3 hrs and at 70 °C for 30 mins to denature proteinase. Then the plate was heated at 98 °C for 4 mins and at 95 °C for 2 mins to ensure proper fragmentation based on eWGA paper. Finally, DNA denaturation happened at 95 °C for 5 mins, and the plate was cold quenched on ice for 20 mins.

After cold quenching, PEG hydrogel reaction mix was added to the well. Gels were formed in 20 mins at room temperature and maintain at 30 °C for 12 hrs for MDA reaction and 65 °C for 5 mins to deactivate Φ29. Reaction wells were imaged with SYTOX orange DNA intercalating dye. To retrieve DNA for library preparation, 6.6

μl of 400 mM KOH was added to incubate for 10 mins at 72 °C and 3 μl 3.75% acetic acid neutralization buffer was added. The neutralized gel-DNA mix was SPRI cleaned with 1X:1X volume ratio and library prepared with standard Nextera procedures with 12 cycles of PCR. The 8 cell libraries were loaded on HiSeq 2500 in the rapid run mode.

The experimental difference from *virtual microfluidics* on the microbial sample is that we diluted single cells and Poisson loaded them into 384 wells. A single genome was fragmented, evenly distributed in the PEG hydrogel and went through digital MDA. Only one round of MDA was conducted.

4.4.2 Bioinformatic methods

Raw sequencing fastq.gz files were quality and adapter trimmed using Trimmomatic. For fast processing of chimera analysis, fastq files were downsampled to 600,000 reads using Seqtk (https://github.com/lh3/seqtk with seed 100). Fastq files were mapped with BWA under default mode both pair-ended and single-ended. BAM files were sorted by mapping coordinates. Mark PCR and optical duplicates, and mask repeat region with the file downloaded from UCSC Genome Browser (assembly:GRCh37/ hg19, group: Repeats, track: RepeatMasker, output:BED). The mapping statistics were retrieved from BAM files using gaemr get_simple_bam_stats.py, and all BAM files were downsampled based the BAM stats resulting to 430,000 reads each sample —both forward and reverse reads. Single-ended mapped BAM files were sorted by query names and merged (Fig. 4-11).

Genome coverage was obtained using Bedtools genomecov. Lorenz curves were obtained by first processing BAM files (duplicates marked) using SAMtools mpileup and then ranking the ascending coverage per base pair (see Fig. 3-5 in Chapter 3 for detail).

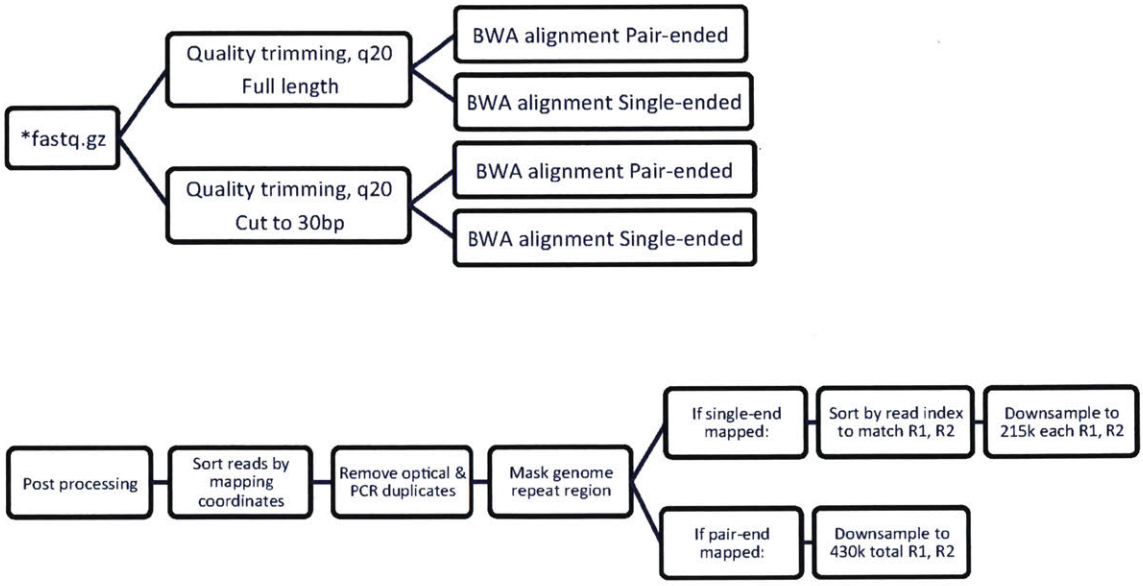

Figure 4-11: Bioinformatic workflow for chimera analysis.

Chapter 5

Conclusion and future directions

In this thesis, a novel single-cell whole genome sequencing technology termed *virtual microfluidics* was developed and applied to enable the study of genomic heterogeneity in complex biological systems. Our technology establishes a new paradigm in single-molecule and single-cell analysis with dramatically different characteristics than established microfluidic approaches. Applications of the technology on purified DNA, cultured bacteria, human gut microbiome samples, and human cell lines demonstrated the robustness of the system. Below, the key findings of this thesis are summarized and possible future research directions are proposed.

5.1 Summary of advancements

Virtual microfluidics enables high-throughput nucleic acid digital quantification and whole genome amplification in an easy-to-use, benchtop format that requires no special equipment or environmental control.

5.1.1 Enabling equipment-independent high-throughput DNA target detection

In Chapter 2, we demonstrated *virtual microfluidics* as a robust nucleic acid quantification platform. The dynamic range of the measurement from a traditional method

(such as dPCR in droplets or microfluidics devices) for DNA quantification is restricted by the number of partitions, usually up to 10^5. Due to the nature of *virtual microfluidics*' diffusion-restricted reaction and the continuous virtual chambers, up to 20,000,000 analytes per μL could be accommodated in our system. Specifically, we tested the performance of in-gel digital PCR and digital MDA as an analytical method for molecular detection and counting. We demonstrated high-throughput digital assays and preparative whole-genome amplification without microfabricated consumables or expensive instrumentation. As few as one DNA target can be detected microscopically with a high signal-to-noise ratio by DNA amplification. The in-gel amplification environment also seals potential infectious targets to minimize the handling of biohazardous materials for infectious disease diagnosis. We expect that *virtual microfluidics* will find application as a low-cost digital assay for detecting DNA biomarkers in the clinic.

5.1.2 Improving the whole genome sequencing data quality and success rate for characterizing uncultured microorganisms

In Chapter 3, we characterized whole genome amplification and recovery of single bacterial genomes for lab-cultured control cells and the human gut microbiome using next-generation sequencing (NGS). Compared to traditional methods of whole genome sequencing on single microbes (in tube and microfluidic device), we improved the uniformity of the whole genome amplification by 25% $\sim$ 33% and reduced the rate of the chimeric artifact by a factor of six. The success rate of *virtual microfluidics* single-cell sequencing is about 28%, which is limited by the Poisson distribution. In contrast, typical success rates (the percentage of amplified genomes that pass purity and genome-size threshold) of single-cell sequencing services provided by large-scale genomic centers based on the first-hand experience from our collaborators is about 10%. Such genomic centers routinely conduct single-cell sequencing in the clean room and utilize FACS for cell isolation, while our approach has a minimal engineer-

ing requirement. We demonstrated single-cell sequencing on human gut microbiome samples and obtained 117 pure single draft genomes. Working with collaborators, we were able to utilize the draft genomes to identify more than 10,000 horizontally transferred genes with unique population-specific and individual-specific features [44]. We expect that *virtual microfluidics* will find application as a high-throughput platform for single-cell sample preparation to study a diverse collection of uncharacterized microbes and environmental microbiome samples.

5.1.3 Reducing structural variation artifacts for studying human cells using single-cell sequencing

In Chapter 4, we demonstrated *virtual microfluidics* for high-quality single-cell genome sequencing on human cell lines with a 3 $\sim$ 6 fold chimera artifact reduction compared to several single-cell technologies. The chimera reduction feature of the *virtual microfluidics* reduces the false-positive rate of genome structural variation detection in studying tumor clonal heterogeneity. Bioinformatically, we characterized chimeric DNA rearrangements in several recently developed single-cell technologies. The unique chimera signatures across different platforms drew attention to the importance of characterizing chimera artifacts in newly developed single-cell technologies. The hydrogel environment also eliminates the need of creating ultra-small discrete chambers for sub-microliter MDA reactions for comparable high-quality data. Furthermore, all of the preparative steps of single-cell whole genome sequencing are accessible with basic lab equipment. We expect *virtual microfluidics* to be implemented widely by researchers who are interested in studying tumor heterogeneity with single-cell resolution.

To summarize, this thesis work centers on the development and demonstration of *virtual microfluidics*, a novel technique for high-quality low-input genomic research. This technique makes single-cell genomics more accessible to a wide range of scientific and biomedical researchers.

5.2 Future directions

5.2.1 Future technical improvements of *virtual microfluidics*

Additional technical improvements are needed in order to realize the full potential of *virtual microfluidics*. To begin, the throughput of *virtual microfluidics* can be improved further. Currently, the primary throughput limitation in the initial demonstration on microbial samples is the volume sub-sampled (60 nL) when product clusters are retrieved, which limits the number of sub-samples that can be retrieved from a single hydrogel. A number of approaches are worth exploring: using a thinner gel with more surface area and/or reducing the punch size from the 500 μm and 1 mm diameters we employed here could improve throughput. A second possibility could be to use imaging data to guide product retrieval and increase the fraction of retrieved samples containing a single-cell WGA reaction product. The thin hydrogel format affords excellent physical access for imaging, equipment, and reagents, which enables an assortment of sub-sampling approaches including punch/pickers, localized hydrogel dissolution, and localized affinity tagging or barcoding. Finally, barcoding approaches could conceivably enable retrieval of all amplified products *en masse* while allowing *in silico* demultiplexing to sort sequence reads according to the cell of origin [156].

Suitability of in-gel amplification format for product cluster labeling

Virtual microfluidics's excellent optical accessibility allows potential fluorescent labeling of rare sequences, which is essential to identify rare targets in microbial dark matter discovery and liquid biopsy applications. For these applications, the demand for single-cell assay throughput is not driven by the need to amass a large number of single-cell datasets, but rather to access cells that are rare in the population. The hydrogel format is ideally suited for this case as the WGA reaction endpoint is an opportune moment to genotype product clusters using hybridization probes in order to identify cells of interest for retrieval and sequencing analysis [157, 158]. In the post-reaction hydrogel, genomic sequences have been amplified and are not protected

by a cell envelope. In addition, the thin gel slab format facilitates the application of reagents for rapid template denaturation, labeling, and de-staining. Once labeled, the desired targets can be selectively retrieved for further analysis by image-guided selection. Sequence-specific labeling might also reduce the number of false-positive background spots that challenged the intercalating dye-based approach we used in this study and/or to lend molecular specificity to quantification assays. In fact, sequential FISH could be used to probe for large sets of functional genes within the gel itself, enabling the application of complex selection criteria [159].

Potential for amplification bias reduction

Up to 10 pg of DNA is produced by MDA from each template in the hydrogel format using our protocol. Although we re-amplified punch samples in the microbial study to microgram quantities, 10 pg is, in principle, enough product (of an order 1000 bacterial genome equivalents) to support deep sequencing directly. Given that we obtain good coverage distribution with our high-yield re-amplification protocol for bacteria, it may be possible for coverage distribution improvement by direct library construction from the 10 pg hydrogel product. Recent advances in ultra-efficient library construction have demonstrated library construction from sub-nanogram input levels [160].

Although a number of modified protocols have been proposed to improve coverage distribution in MDA, none has yet been widely adopted, with major single-cell genomics centers continuing to use Φ29 DNA polymerase reaction conditions very similar to those originally developed 30 years ago [161, 118]. In contrast, limiting fold-amplification reduces coverage bias, since the ratio of maximum possible fold amplification to minimum possible fold amplification is necessarily reduced when the average degree of amplification is reduced. When combined with the cost savings of micro-scaled reactions and increasingly efficient sequence library construction procedures, such an approach shows the future trend of single-cell WGA [87].

Today, investigators limit amplification-fold by reducing reaction volumes [86, 88] or by limiting reaction time [162]. Although it is currently unknown which approach is more fruitful in bias reduction, both approaches have drawbacks. The hydrogel reac-

tion format offers unique advantages in limited-extent WGA, as the product clusters from each template molecule only reach a few microns in size, even under dilute template conditions. This suggests that one can achieve uniform (limited) reaction extent across single-cell WGA reactions, even when the reactions occur asynchronously. The hydrogel format also enables maintenance of optimal amplification conditions for each template throughout the reaction time course if desired by reagent supplementation, possibly reducing sequence content and template fragment length biases. In order to further measure the amplification bias in the hydrogel, a random barcode library can be introduced into the hydrogel for amplification and sequencing analysis.

5.2.2 The future of single-cell whole genome sequencing

The single-cell sequencing process from having cells in suspension to obtaining sequencing data is highly fragmented in terms of technology implementation. On the other hand, due to the diverse applications of single-molecule and single-cell analysis, it is difficult to find a one-tool-fits-all solution. The key is to have a platform technology that allows modular changes of different processes involved, in order to strike a balance among the requirements of throughput, hands-on time, cost, and quality for various applications. This reality also explains the slow uptake of single-cell technology in various research and clinical settings. To address these issues, *virtual microfluidics* represents a platform technology that can be implemented with a diverse collection of methods in cell lysis, whole genome amplification, and barcoding strategies. With further technical optimizations, this technique could play a central role in integrating the single-cell sequencing field.

Beyond the traditional method of whole genome amplification on isolated single cells, I envision that it will be ideal to barcode a large number of single cells with minimal amplification and obtain accurate long-read sequencing results. Such a technology combination will revolutionize the landscape of single-cell sequencing. Because it minimizes the biases and artifacts from extensive amplification that often distort the output data from the original genomic sequences. Long-read sequencing (currently 10 kbp - 400 kbp) has the advantage of providing the genuine read

of a long genome sequence that could be highly repeated and of high GC%, which are challenging for the current short-read technologies (50 bp - 500 bp). At the same time, it also requires the development of more accurate sequencing technologies. Currently PacBio (long-read) has an error rate of $\sim$ 10%, compared to the 0.1 $\sim$ 1 % of the short-read Illumina sequencing. A recent development has demonstrated direct library preparation on single cells in nano-droplets for whole genome sequencing [163]. Its performance is below that of state of the art single-cell technologies, in terms of genome physical coverage and coverage uniformity. However, the concept of direct library preparation on a large number of samples (hundreds) without the expensive and often biased WGA process is attractive to researchers to obtain the most representative genomic information.

Beyond single-cell sequencing, there still exist challenges in implementing genomic testing in the clinic as the standard of care. First of all, the reimbursement and clinical adoption reply on the innovation of the sequencing cost reduction. Secondly, it is difficult for patients, doctors and researchers who lack the genomic related training to interpret the data. This is especially critical when doctors and patients have to make decisions with the consideration of the inherent false-positive and false-negative rate of the genomic data. Furthermore, a large percentage (40% - 60%) of patients often don't have actionable mutations according to recent cancer sequencing projects (GenomeWeb Mar 01, 2017). The uptake of standard genomic technologies will likely develop hand-in-hand with genomic-guided and targeted drug discovery. With the growing of related products such as 23andMe, and genomic service companies such as Foundation Medicine, Grail, and Color Genomics, the sharing of genomic data to promote research and diagnosis will spread rapidly. In my opinion, the future of single-cell sequencing field relies on the uptake of standard genomic technologies, in addition to more single-cell analysis innovations, in order to realize its full potential.

Appendix A

NCBI accession numbers

A.1 Bacterial single cells

Raw sequencing data on *E. coli* and *S. aureus* are accessible at the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA279815 with BioSample accession numbers SAMN03451478-SAMN03451501. FijiCOMP metagenomic reads can be found under BioProject accession number PRJNA217052 with the accession numbers: SRX345831, SRX344363, SRX344765, SRX343094, SRX344442, SRX346405, SRX343839, SRX343780, SRX345901, SRX344600, SRX343866, SRX343411, SRX344189, SRX344380, SRX346966, SRX345329, SRX343800, and SRX344616. FijiCOMP *virtual microfluidics* 117 single cells are accessible with the BioSample accession numbers SAMN04461233-SAMN04461349.

A.2 Human single cells

Raw sequencing data on RPE-1 bulk genomic DNA and single cells using *virtual microfluidics* are accessible under BioProject accession number PRJNA408301 with BioSample accession numbers SAMN07682898 and SAMN07682891.

Data source	SRA accession
Virtual Microfluidics	gDNA: SRR6075104 Single cells: SRR6075105 SRR6075106 SRR6075107 SRR6075108 SRR6075109 SRR6075110 SRR6075111 SRR6075112
Emulsion WGA (Fu *et al.* 2015)	gDNA: SRR1777284 Single cells: -MDA: SRR1777287 SRR1777288 SRR1777290 SRR1777291 SRR1777294 -MALBAC: SRR1777304 SRR1777305 -MDA in tube: SRR1777307 SRR1777308

Data source	SRA accession
Nanoliter droplet (Leung *et al.* 2016)	gDNA: SRR3749177 Single cells: -High depth: SRR3749178 SRR3749179 SRR3749180 SRR3749181 SRR3749182 SRR3749183 SRR3749184 SRR3749186 -Low depth: SRR3749174 SRR3749218 SRR3749230 SRR3749245 SRR3749252 SRR3749263 SRR3749274 SRR3749285 SRR3749296
LIANTI (Chen et al. 2017)	gDNA: SRR5365378 Single cells: SRR5365376 SRR5365375 SRR5365374

Table A.1: SRA accession numbers for human single-cell chimera analysis

Bibliography

[1] Paul C Blainey and Stephen R Quake. Dissecting genomic diversity, one cell at a time. *Nature Methods*, 11(1):19–21, January 2014.

[2] Ian P G Marshall, Paul C Blainey, Alfred M Spormann, and Stephen R Quake. A Single-cell genome for Thiovulum sp. *Applied and environmental microbiology*, 78(24):8555–8563, December 2012.

[3] Jianbin Wang, H CHRISTINA Fan, Barry Behr, and Stephen R Quake. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, 150(2):402–412, July 2012.

[4] Jim F Huggett, Simon Cowen, and Carole A Foy. Considerations for digital PCR as an accurate molecular diagnostic tool. *Clinical chemistry*, 61(1):79–88, January 2015.

[5] Gilad D Evrony, Eunjung Lee, Peter J Park, and Christopher A Walsh. Resolving rates of mutation in the brain using single-neuron genomics. *eLife*, 5, February 2016.

[6] Gilad D Evrony, Xuyu Cai, Eunjung Lee, L Benjamin Hills, Princess C Elhosary, Hillel S Lehmann, J J Parker, Kutay D Atabay, Edward C Gilmore, Annapurna Poduri, Peter J Park, and Christopher A Walsh. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, 151(3):483–496, October 2012.

[7] Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N Ivanova, Iain J Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K Swan, Esther A Gies, Jeremy A Dodsworth, Brian P Hedlund, George Tsiamis, Stefan M Sievert, Wen-Tso Liu, Jonathan A Eisen, Steven J Hallam, Nikos C Kyrpides, Ramunas Stepanauskas, Edward M Rubin, Philip Hugenholtz, and Tanja Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, July 2013.

[8] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, March 2011.

[9] Thomas P Curtis, William T Sloan, and Jack W Scannell. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10494–10499, August 2002.

[10] Cleston C Lange, Lawrence P Wackett, Kenneth W Minton, and Michael J Daly. Engineering a recombinant Deinococcus radiodurans for organopollutant degradation in radioactive mixed waste environments. *Nature Biotechnology*, 16(10):929–933, October 1998.

[11] Myles R Minter, Can Zhang, Vanessa Leone, Daina L Ringus, Xiaoqiong Zhang, Paul Oyler-Castrillo, Mark W Musch, Fan Liao, Joseph F Ward, David M Holtzman, Eugene B Chang, Rudolph E Tanzi, and Sangram S Sisodia. Antibiotic-induced perturbations in gut microbial diversity influences neuro-inflammation and amyloidosis in a murine model of Alzheimer's disease. *Scientific reports*, 6:30028, July 2016.

[12] Suzan Yilmaz and Anup K Singh. Single cell genome sequencing. *Current Opinion in Biotechnology*, 23(3):437–443, December 2011.

[13] Paul C Blainey. The future is now: single-cell genomics of bacteria and archaea. *FEMS microbiology reviews*, 37(3):407–427, May 2013.

[14] Hui Wang, Jürgen Tomasch, Michael Jarek, and Irene Wagner-Döbler. A dual-species co-cultivation system to study the interactions between Roseobacters and dinoflagellates. *Frontiers in Microbiology*, 5, 2014.

[15] Christian Rinke, Janey Lee, Nandita Nath, Danielle Goudeau, Brian Thompson, Nicole Poulton, Elizabeth Dmitrieff, Rex Malmstrom, Ramunas Stepanauskas, and Tanja Woyke. Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics. *Nature Protocols*, 9(5):1038–1048, April 2014.

[16] Ramunas Stepanauskas and Michael E Sieracki. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):9052–9057, May 2007.

[17] Paul C Blainey, Annika C Mosier, Anastasia Potanina, Christopher A Francis, and Stephen R Quake. Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS ONE*, 6(2):e16626, 2011.

[18] Laure Prat, Ilka U Heinemann, Hans R Aerni, Jesse Rinehart, Patrick O'Donoghue, and Dieter Söll. Carbon source-dependent expansion of the genetic code in bacteria. *Proceedings of the National Academy of Sciences*, 109(51):21070–21075, December 2012.

[19] Joshua D Campbell, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H Berger, Chandra Sekhar Pedamallu, Sachet A Shukla, Guangwu Guo, Angela N Brooks, Bradley A Murray, Marcin Imielinski, Xin Hu, Shiyun Ling, Rehan Akbani, Mara Rosenberg, Carrie Cibulskis, Aruna Ramachandran, Eric A Collisson, David J Kwiatkowski, Michael S Lawrence, John N Weinstein, Roel G W Verhaak, Catherine J Wu, Peter S Hammerman, Andrew D Cherniack, Gad Getz, Cancer Genome Atlas Research Network, Maxim N Artyomov, Robert Schreiber, Ramaswamy Govindan, and Matthew Meyerson. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, 48(6):607–616, June 2016.

[20] Yann Marcy, Cleber Ouverney, Elisabeth M Bik, Tina Lösekann, Natalia Ivanova, Hector Garcia Martin, Ernest Szeto, Darren Platt, Philip Hugenholtz, David A Relman, and Stephen R Quake. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of America*, 104(29):11889–11894, July 2007.

[21] Mircea Podar, Martin Keller, and Philip Hugenholtz. Single Cell Whole Genome Amplification of Uncultivated Organisms. In *Uncultivated Microorganisms*, pages 241–256. Springer Berlin Heidelberg, Berlin, Heidelberg, February 2009.

[22] Jeffrey S McLean, Mary-Jane Lombardo, Jonathan H Badger, Anna Edlund, Mark Novotny, Joyclyn Yee-Greenbaum, Nikolay Vyahhi, Adam P Hall, Youngik Yang, Christopher L Dupont, Michael G Ziegler, Hamidreza Chitsaz, Andrew E Allen, Shibu Yooseph, Glenn Tesler, Pavel A Pevzner, Robert M Friedman, Kenneth H Nealson, J Craig Venter, and Roger S Lasken. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proceedings of the National Academy of Sciences*, 110(26):E2390–9, June 2013.

[23] Jeffrey S McLean, Mary-Jane Lombardo, Michael G Ziegler, Mark Novotny, Joyclyn Yee-Greenbaum, Jonathan H Badger, Glenn Tesler, Sergey Nurk, Valery Lesin, Daniel Brami, Adam P Hall, Anna Edlund, Lisa Z Allen, Scott Durkin, Sharon Reed, Francesca Torriani, Kenneth H Nealson, Pavel A Pevzner, Robert Friedman, J Craig Venter, and Roger S Lasken. Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. *Genome research*, 23(5):867–877, May 2013.

[24] Brian Cleary, Ilana L Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J Alm. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature*, 2015.

[25] Scott Clingenpeel, Alicia Clum, Patrick Schwientek, Christian Rinke, and Tanja Woyke. Reconstructing each cell's genome within complex microbial communities—dream or reality? *Frontiers in Microbiology*, 5, 2015.

[26] Lily Xu, Ilana L Brito, Eric J Alm, and Paul C Blainey. Virtual microfluidics for digital quantification and single-cell sequencing. *Nature Methods*, 2016.

[27] Sébastien Rodrigue, Rex R Malmstrom, Aaron M Berlin, Bruce W Birren, Matthew R Henn, and Sallie W Chisholm. PLOS ONE: Whole Genome Amplification and De novo Assembly of Single Bacterial Cells. *PLoS ONE*, 4(9):e6864, 2009.

[28] Niels Van der Aa, Masoud Zamani Esteki, Joris R Vermeesch, and Thierry Voet. Preimplantation genetic diagnosis guided by single-cell genomics. *Genome Medicine*, 5(8):71, 2013.

[29] Ron Sender, Shai Fuchs, and Ron Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS biology*, 14(8):e1002533, August 2016.

[30] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, Andrew C Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J Gregory Caporaso, Catherine A Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I Gordon. Human gut microbiome viewed across age and geography. *Nature*, 72(7402):1027–227, May 2012.

[31] Naama Geva-Zatorsky, Esen Sefik, Lindsay Kua, Lesley Pasman, Tze Guan Tan, Adriana Ortiz-Lopez, Tsering Bakto Yanortsang, Liang Yang, Ray Jupp, Diane Mathis, Christophe Benoist, and Dennis L Kasper. Mining the Human Gut Microbiota for Immunomodulatory Organisms. *Cell*, 168(5):928–943.e11, February 2017.

[32] Claudia S Plottel and Martin J Blaser. Microbiome and Malignancy. *Cell Host & Microbe*, 10(4):324–335, October 2011.

[33] Shin Yoshimoto, Tze Mun Loo, Koji Atarashi, Hiroaki Kanda, Seidai Sato, Seiichi Oyadomari, Yoichiro Iwakura, Kenshiro Oshima, Hidetoshi Morita, Masahira Hattori, Masahisa Hattori, Kenya Honda, Yuichi Ishikawa, Eiji Hara, and Naoko Ohtani. Obesity-induced gut microbial metabolite promotes liver cancer through senescence secretome. *Nature*, 499(7456):97–101, July 2013.

[34] Joshua M Uronis, Marcus Mühlbauer, Hans H Herfarth, Tara C Rubinas, Gieira S Jones, and Christian Jobin. Modulation of the intestinal microbiota alters colitis-associated colorectal cancer susceptibility. *PLoS ONE*, 4(6):e6026, June 2009.

[35] Shashank Ghantoji, K Sail, David R Lairson, H L DuPont, and Kevin W Garey. Economic healthcare costs of Clostridium difficile infection: a systematic review. *The Journal of hospital infection*, 74(4):309–318, April 2010.

[36] Fernanda C Lessa, Yi Mu, Wendy M Bamberg, Zintars G Beldavs, Ghinwa K Dumyati, John R Dunn, Monica M Farley, Stacy M Holzbauer, James I Meek, Erin C Phipps, Lucy E Wilson, Lisa G Winston, Jessica A Cohen, Brandi M Limbago, Scott K Fridkin, Dale N Gerding, and L Clifford McDonald. Burden of Clostridium difficileInfection in the United States. *The New England journal of medicine*, 372(9):825–834, February 2015.

[37] Johan S Bakken, Thomas Borody, Lawrence J Brandt, Joel V Brill, Daniel C Demarco, Marc Alaric Franzos, Colleen Kelly, Alexander Khoruts, Thomas Louie, Lawrence P Martinelli, Thomas A Moore, George Russell, Christina Surawicz, and Fecal Microbiota Transplantation Workgroup. Treating Clostridium difficile infection with fecal microbiota transplantation. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 9(12):1044–1049, December 2011.

[38] Zain Kassam, Christine H Lee, Yuhong Yuan, and Richard H Hunt. Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis. *The American journal of gastroenterology*, 108(4):500–508, April 2013.

[39] Mark Ratner. Seres's pioneering microbiome drug fails mid-stage trial. *Nature Biotechnology*, 34(10):1004–1005, October 2016.

[40] Sonny T M Lee, Stacy A Kahn, Tom O Delmont, Alon Shaiber, Özcan C Esen, Nathaniel A Hubert, Hilary G Morrison, Dionysios A Antonopoulos, David T Rubin, and A Murat Eren. Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome*, 5(1):50, May 2017.

[41] Stephen Nayfach, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S Pollard. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome research*, 26(11):1612–1625, November 2016.

[42] John Chen and Richard P Novick. Phage-Mediated Intergeneric Transfer of Toxin Genes. *Science*, 323(5910):139–141, January 2009.

[43] Chris S Smillie, Mark B Smith, Jonathan Friedman, Otto X Cordero, Lawrence A David, and Eric J Alm. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241–244, October 2011.

[44] I L Brito, S Yilmaz, K Huang, L. Xu, S D Jupiter, A P Jenkins, W Naisilisili, M Tamminen, C S Smillie, J R Wortman, B W Birren, R J Xavier, P C Blainey, A K Singh, D Gevers, and E J Alm. Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439, July 2016.

[45] Jan-Hendrik Hehemann, Gaëlle Correc, Tristan Barbeyron, William Helbert, Mirjam Czjzek, and Gurvan Michel. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*, 2010.

[46] Michael J Coyne, Naamah Levy Zitomersky, Abigail Manson McGuire, Ashlee M Earl, and Laurie E Comstock. Evidence of extensive DNA transfer between bacteroidales species within the human gut. *mBio*, 5(3):e01305–14, June 2014.

[47] Brian V Jones, Funing Sun, and Julian R Marchesi. Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. *BMC Genomics*, 11:46, January 2010.

[48] Mya Breitbart, Ian Hewson, Ben Felts, Joseph M Mahaffy, James Nulton, Peter Salamon, and Forest Rohwer. Metagenomic analyses of an uncultured viral community from human feces. *Journal of bacteriology*, 185(20):6220–6223, October 2003.

[49] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, October 1976.

[50] Yoshihide Tsujimoto, Lawrence R Finger, Jorge Yunis, Peter C Nowell, and Carlo M Croce. Cloning of the chromosome breakpoint of neoplastic B cells with the t(14;18) chromosome translocation. *Science*, 226(4678):1097–1099, November 1984.

[51] Gloria H Heppner. Tumor heterogeneity. *Cancer research*, 44(6):2259–2265, June 1984.

[52] Jonathan Sebat, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Pär Lundin, Susanne Månér, Hillary Massa, Megan Walker, Maoyen Chi, Nicholas Navin, Robert Lucito, John Healy, James Hicks, Kenny Ye, Andrew Reiner, T Conrad Gilliam, Barbara Trask, Nick Patterson, Anders Zetterberg, and Michael Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, July 2004.

[53] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, May 2012.

[54] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, March 2013.

[55] Iñigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489, September 2015.

[56] Assieh Saadatpour, Shujing Lai, Guoji Guo, and Guo-Cheng Yuan. Single-Cell Analysis in Cancer Genomics. *Trends in genetics : TIG*, 31(10):576–586, October 2015.

[57] Philippe L Bedard, Aaron R Hansen, Mark J Ratain, and Lillian L Siu. Tumour heterogeneity in the clinic. *Nature*, 2013.

[58] Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas B K Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, Max Salm, Stuart Horswell, Mickael Escudero, Nik Matthews, Andrew Rowan, Tim Chambers, David A Moore, Samra Turajlic, Hang Xu, Siow-Ming Lee, Martin D Forster, Tanya Ahmad, Crispin T Hiley, Christopher Abbosh, Mary Falzon, Elaine Borg, Teresa Marafioti, David Lawrence, Martin Hayward, Shyam Kolvekar, Nikolaos Panagiotopoulos, Sam M Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Rajesh Shah, Leena Joseph, Anne M Quinn, Phil A Crosbie, Babu Naidu, Gary Middleton, Gerald Langman, Simon Trotter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Lesley Gomersall, Dean A Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-Sellers, Vineet Prakash, Jason F Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Stefan Dentro, Philippe Taniere, Brendan O'Sullivan, Helen L Lowe, John A Hartley, Natasha Iles, Harriet Bell, Yenting Ngai, Jacqui A Shaw, Javier Herrero, Zoltan Szallasi, Roland F Schwarz, Aengus Stewart, Sergio A Quezada, John Le Quesne, Peter Van Loo, Caroline Dive, Allan Hackshaw, and Charles Swanton. Tracking the Evolution of Non–Small-Cell Lung Cancer. *The New England journal of medicine*, page NEJMoa1616288, April 2017.

[59] Carlo C Maley, Patricia C Galipeau, Jennifer C Finley, V Jon Wongsurawat, Xiaohong Li, Carissa A Sanchez, Thomas G Paulson, Patricia L Blount, Rosa-Ana Risques, Peter S Rabinovitch, and Brian J Reid. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics*, 38(4):468–473, April 2006.

[60] Nicholas E Navin. Cancer genomics: one cell at a time. *Genome biology*, 15(8):452, August 2014.

[61] Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 2016.

[62] Lawrence A Loeb, Keith R Loeb, and Jon P Anderson. Multiple mutations and cancer. *Proceedings of the National Academy of Sciences*, 100(3):776–781, January 2003.

[63] Lawrence A Loeb. Mutator phenotype may be required for multistage carcinogenesis. *Cancer research*, 51(12):3075–3079, June 1991.

[64] I P Tomlinson, M R Novelli, and W F Bodmer. The mutation rate and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25):14800–14803, December 1996.

[65] Jason H Bielas, Keith R Loeb, Brian P Rubin, Lawrence D True, and Lawrence A Loeb. Human cancers express a mutator phenotype. *Proceedings of the National Academy of Sciences of the United States of America*, 103(48):18238–18242, November 2006.

[66] Chenghang Zong, Sijia Lu, Alec R Chapman, and X Sunney Xie. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, 338(6114):1622–1626, December 2012.

[67] Gene-Wei Li and X Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 2011.

[68] Fatima Cardoso, Laura J van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, Sylvain Causeret, Mauro DeLorenzi, Annuska M Glas, Vassilis Golfinopoulos, Theodora Goulioti, Susan Knox, Erika Matos, Bart Meulemans, Peter A Neijenhuis, Ulrike Nitz, Rodolfo Passalacqua, Peter Ravdin, Isabel T Rubio, Mahasti Saghatchian, Tineke J Smilde, Christos Sotiriou, Lisette Stork, Carolyn Straehle, Geraldine Thomas, Alastair M Thompson, Jacobus M van der Hoeven, Peter Vuylsteke, René Bernards, Konstantinos Tryfonidis, Emiel Rutgers, and Martine Piccart. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *The New England journal of medicine*, 375(8):717–729, August 2016.

[69] Elli Papaemmanuil, Moritz Gerstung, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Nicola D Roberts, Nicola E Potter, Michael Heuser, Felicitas Thol, Niccolo Bolli, Gunes Gundem, Peter Van Loo, Iñigo Martincorena, Peter Ganly, Laura Mudie, Stuart McLaren, Sarah O'Meara, Keiran Raine, David R Jones, Jon W Teague, Adam P Butler, Mel F Greaves, Arnold Ganser, Konstanze Döhner, Richard F Schlenk, Hartmut Döhner, and Peter J Campbell. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England journal of medicine*, 374(23):2209–2221, June 2016.

[70] Santhosh Girirajan, Jill A Rosenfeld, Bradley P Coe, Sumit Parikh, Neil Friedman, Amy Goldstein, Robyn A Filipink, Juliann S McConnell, Brad Angle, Wendy S Meschino, Marjan M Nezarati, Alexander Asamoah, Kelly E Jackson, Gordon C Gowans, Judith A Martin, Erin P Carmany, David W Stockton, Rhonda E Schnur, Lynette S Penney, Donna M Martin, Salmo Raskin, Kathleen Leppig, Heidi Thiese, Rosemarie Smith, Erika Aberg, Dmitriy M Niyazov, Luis F Escobar, Dima El-Khechen, Kisha D Johnson, Robert R Lebel, Kiana Siefkas, Susie Ball, Natasha Shur, Marianne McGuire, Campbell K Brasington, J Edward Spence, Laura S Martin, Carol Clericuzio, Blake C Ballif,

Lisa G Shaffer, and Evan E Eichler. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *The New England journal of medicine*, 367(14):1321–1331, October 2012.

[71] Heidrun Gevensleben, Isaac Garcia-Murillas, Monika K Graeser, Gaia Schiavon, Peter Osin, Marina Parton, Ian E Smith, Alan Ashworth, and Nicholas C Turner. Noninvasive detection of HER2 amplification with plasma DNA digital PCR. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(12):3276–3284, June 2013.

[72] Li Ding, Matthew J Ellis, Shunqiang Li, David E Larson, Ken Chen, John W Wallis, Christopher C Harris, Michael D McLellan, Robert S Fulton, Lucinda L Fulton, Rachel M Abbott, Jeremy Hoog, David J Dooling, Daniel C Koboldt, Heather Schmidt, Joelle Kalicki, Qunyuan Zhang, Lei Chen, Ling Lin, Michael C Wendl, Joshua F McMichael, Vincent J Magrini, Lisa Cook, Sean D McGrath, Tammi L Vickery, Elizabeth Appelbaum, Katherine Deschryver, Sherri Davies, Therese Guintoli, Li Lin, Robert Crowder, Yu Tao, Jacqueline E Snider, Scott M Smith, Adam F Dukes, Gabriel E Sanderson, Craig S Pohl, Kim D Delehaunty, Catrina C Fronick, Kimberley A Pape, Jerry S Reed, Jody S Robinson, Jennifer S Hodges, William Schierding, Nathan D Dees, Dong Shen, Devin P Locke, Madeline E Wiechert, James M Eldred, Josh B Peck, Benjamin J Oberkfell, Justin T Lolofie, Feiyu Du, Amy E Hawkins, Michelle D O'Laughlin, Kelly E Bernard, Mark Cunningham, Glendoria Elliott, Mark D Mason, Dominic M Thompson, Jennifer L Ivanovich, Paul J Goodfellow, Charles M Perou, George M Weinstock, Rebecca Aft, Mark Watson, Timothy J Ley, Richard K Wilson, and Elaine R Mardis. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999–1005, April 2010.

[73] Rory L Cochran, Karen Cravero, David Chu, Bracha Erlanger, Patricia Valda Toro, Julia A Beaver, Daniel J Zabransky, Hong Yuen Wong, Justin Cidado, Sarah Croessmann, Heather A Parsons, Minsoo Kim, Sarah J Wheelan, Pedram Argani, and Ben Ho Park. Analysis of BRCA2 loss of heterozygosity in tumor tissue using droplet digital polymerase chain reaction. *Human pathology*, 45(7):1546–1550, July 2014.

[74] Lawrence J Jennings, David George, Juliann Czech, Min Yu, and Loren Joseph. Detection and quantification of BCR-ABL1 fusion transcripts by droplet digital PCR. *The Journal of molecular diagnostics : JMD*, 16(2):174–179, March 2014.

[75] Philippe Anker, Francois Lefort, Valeri Vasioukhin, Jacqueline Lyautey, Christine Lederrey, Xu Qi Chen, Maurice Stroun, Huge E Mulcahy, and Michael J G Farthing. K-ras mutations are found in DNA extracted from the plasma of patients with colorectal cancer. *Gastroenterology*, 112(4):1114–1120, April 1997.

[76] Huge E Mulcahy, Jacqueline Lyautey, Christine Lederrey, Xu qi Chen, Philippe Anker, Elspeth M Alstead, Anne Ballinger, Michael J G Farthing, and Maurice Stroun. A prospective study of K-ras mutations in the plasma of pancreatic cancer patients. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 4(2):271–275, February 1998.

[77] Manel Esteller, Montserrat Sanchez-Cespedes, Rafael Rosell, David Sidransky, Stephen B Baylin, and James G Herman. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. *Cancer research*, 59(1):67–70, January 1999.

[78] Mingwei Ma, Hongcheng Zhu, Chi Zhang, Xinchen Sun, Xianshu Gao, and Gang Chen. "Liquid biopsy"-ctDNA detection with great potential and challenges. *Annals of translational medicine*, 3(16):235, September 2015.

[79] Jun Zou, Brian Duffy, Michael Slade, Andrew Lee Young, Nancy Steward, Ramsey Hachem, and T Mohanakumar. Rapid detection of donor cell free DNA in lung transplant recipients with rejections using donor-recipient HLA mismatch. *Human immunology*, 78(4):342–349, April 2017.

[80] Fiona M F Lun, Rossa W K Chiu, K C Allen Chan, Tak Yeung Leung, Tze Kin Lau, and Y M Dennis Lo. Microfluidics Digital PCR Reveals a Higher than Expected Fraction of Fetal DNA in Maternal Plasma. *Clinical chemistry*, 54(10):1664–1672, August 2008.

[81] Y M Dennis Lo, Fiona M F Lun, K C Allen Chan, Nancy B Y Tsui, Ka C Chong, Tze K Lau, Tak Y Leung, Benny C Y Zee, Charles R Cantor, and Rossa W K Chiu. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proceedings of the National Academy of Sciences of the United States of America*, 104(32):13116–13121, August 2007.

[82] Bert Vogelstain and Kenneth W Kinzler. Digital PCR. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16):9236–9241, 1999.

[83] P J Sykes, S H Neoh, M J Brisco, E Hughes, J Condon, and A A Morley. Quantitation of targets for PCR by use of limiting dilution. *Biotechniques*, 13(3):444–449, September 1992.

[84] Kerry Routenberg Love, Sangram Bagh, Jonghoon Choi, and J Christopher Love. Microtools for single-cell analysis in biopharmaceutical development and manufacturing. *Trends in Biotechnology*, 31(5):280–286, May 2013.

[85] Todd Thorsen, Sebastian J Maerkl, and Stephen R Quake. Microfluidic large-scale integration. *Science*, 298(5593):580–584, October 2002.

[86] Zachary C Landry, Stephen J Giovanonni, Stephen R Quake, and Paul C Blainey. Optofluidic cell selection from complex microbial communities for single-genome analysis. *Methods in enzymology*, 531:61–90, 2013.

[87] Charles F A de Bourcy, Iwijn De Vlaminck, Jad N Kanbar, Jianbin Wang, Charles Gawad, and Stephen R Quake. A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE*, 9(8):e105585, 2014.

[88] Yann Marcy, Thomas Ishoey, Roger S Lasken, Timothy B Stockwell, Brian P Walenz, Aaron L Halpern, Karen Y Beeson, Susanne M D Goldberg, and Stephen R Quake. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS genetics*, 3(9):1702–1708, September 2007.

[89] Yusi Fu, Chunmei Li, Sijia Lu, Wenxiong Zhou, Fuchou Tang, X Sunney Xie, and Yanyi Huang. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proceedings of the National Academy of Sciences*, 112(38):11923–11928, September 2015.

[90] Todd Thorsen, Richard W Roberts, Frances H Arnold, and Stephen R Quake. Dynamic pattern formation in a vesicle-generating microfluidic device. *Physical review letters*, 86(18):4163–4166, April 2001.

[91] Benjamin J Hindson, Kevin D Ness, Donald A Masquelier, Phillip Belgrader, Nicholas J Heredia, Anthony J Makarewicz, Isaac J Bright, Michael Y Lucero, Amy L Hiddessen, Tina C Legler, Tyler K Kitano, Michael R Hodel, Jonathan F Petersen, Paul W Wyatt, Erin R Steenblock, Pallavi H Shah, Luc J Bousse, Camille B Troup, Jeffrey C Mellen, Dean K Wittmann, Nicholas G Erndt, Thomas H Cauley, Ryan T Koehler, Austin P So, Simant Dube, Klint A Rose, Luz Montesclaros, Shenglong Wang, David P Stumbo, Shawn P Hodges, Steven Romine, Fred P Milanovich, Helen E White, John F Regan, George A Karlin-Neumann, Christopher M Hindson, Serge Saxonov, and Bill W Colston. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, 83(22):8604–8610, November 2011.

[92] Leanna S Morinishi and Paul Blainey. Simple Bulk Readout of Digital Nucleic Acid Quantification Assays. *Journal of Visualized Experiments*, (103):e52925–e52925, 2015.

[93] Martin Bengtsson, Anders Ståhlberg, Patrik Rorsman, and Mikael Kubista. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome research*, 15(10):1388–1392, October 2005.

[94] Quin F Wills, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature Biotechnology*, 31(8):748–752, July 2013.

[95] Kun Zhang, Adam C Martiny, Nikos B Reppas, Kerrie W Barry, Joel Malek, Sallie W Chisholm, and George M Church. Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology*, 24(6):680–686, May 2006.

[96] Linas Mazutis, John Gilbert, W Lloyd Ung, David A Weitz, Andrew D Griffiths, and John A Heyman. Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, 2013.

[97] Christopher M Hindson, John R Chevillet, Hilary A Briggs, Emily N Gallichotte, Ingrid K Ruf, Benjamin J Hindson, Robert L Vessella, and Muneesh Tewari. Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nature Methods*, 10(10):1003–1005, September 2013.

[98] Barbara J Trask. Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Reviews Genetics*, 3(10):769–778, October 2002.

[99] Tanja Woyke, Alexander Sczyrba, Janey Lee, Christian Rinke, Damon Tighe, Scott Clingenpeel, Rex Malmstrom, Ramunas Stepanauskas, and Jan-Fang Cheng. Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. *PLoS ONE*, 6(10):e26161, October 2011.

[100] Paul C Blainey and Stephen R Quake. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Research*, 39(4):e19, March 2011.

[101] Soohong Kim, Joachim De Jonghe, Anthony B Kulesa, David Feldman, Tommi Vatanen, Roby P Bhattacharyya, Brittany Berdy, James Gomez, Jill Nolan, Slava Epstein, and Paul C Blainey. High-throughput automated microfluidic sample preparation for accurate microbial genomics. *Nature Communications*, 8:13919, January 2017.

[102] Arumugham Raghunathan, Harley R Ferguson, Carole J Bornarth, Wanmin Song, Mark Driscoll, and Roger S Lasken. Genomic DNA amplification from a single bacterium. *Applied and environmental microbiology*, 71(6):3342–3347, June 2005.

[103] Brandon K Swan, Manuel Martinez-Garcia, Christina M Preston, Alexander Sczyrba, Tanja Woyke, Dominique Lamy, Thomas Reinthaler, Nicole J Poulton, E Dashiell P Masland, Monica Lluesma Gomez, Michael E Sieracki, Edward F DeLong, Gerhard J Herndl, and Ramunas Stepanauskas. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science*, 333(6047):1296–1300, September 2011.

[104] Emily J Fleming, Amy E Langdon, Manuel Martinez-Garcia, Ramunas Stepanauskas, Nicole J Poulton, E Dashiell P Masland, and David Emerson. What's New Is Old: Resolving the Identity of Leptothrix ochracea Using Single Cell Genomics, Pyrosequencing and FISH. *PLoS ONE*, 6(3):e17769, March 2011.

[105] Ramunas Stepanauskas. Single cell genomics: an individual look at microbes. *Current Opinion in Microbiology*, 15(5):613–620, October 2012.

[106] M Fernandia Nobre, Hans G Truper, and M S DA Costa. Transfer of Thermus ruber (Loginova et al. 1984), Thermus silvanus (Tenreiro et al. 1995), and Thermus chliarophilus (Tenreiro et al. 1995) to Meiothermus gen. nov. as Meiothermus ruber comb. nov., Meiothermus silvanus comb. nov., and Meiothermus chliarophilus comb. nov., Respectively, and Emendation of the Genus Thermus. *International Journal of Systematic Bacteriology*, 46(2):604–606, April 1996.

[107] Kaston Leung, Anders Klaus, Bill K Lin, Emma Laks, Justina Biele, Daniel Lai, Ali Bashashati, Yi-Fei Huang, Radhouane Aniba, Michelle Moksa, Adi Steif, Anne-Marie Mes-Masson, Martin Hirst, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Robust high-performance nanoliter-volume single-cell multiple displacement amplification on planar substrates. *Proceedings of the National Academy of Sciences*, 113(30):8484–8489, July 2016.

[108] Chongyi Chen, Dong Xing, Longzhi Tan, Heng Li, Guangyu Zhou, Lei Huang, and X Sunney Xie. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science*, 356(6334):189–194, April 2017.

[109] Frank B Dean, Seiyu Hosono, Linhua Fang, Xiaohong Wu, A Fawad Faruqi, Patricia Bray-Ward, Zhenyu Sun, Qiuling Zong, Yuefen Du, Jing Du, Mark Driscoll, Wanmin Song, Stephen F Kingsmore, Michael Egholm, and Roger S Lasken. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5261–5266, April 2002.

[110] Lixin Chen, Pingfang Liu, Thomas C Evans, Jr., and Laurence M Ettwiller. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, 355(6326):752–756, February 2017.

[111] Nadin Rohland, Eadaoin Harney, Swapan Mallick, Susanne Nordenfelt, and David Reich. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1660):20130624, January 2015.

[112] Sünje J Pamp, Eoghan D Harrington, Stephen R Quake, David A Relman, and Paul C Blainey. Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome research*, 22(6):1107–1119, June 2012.

[113] Jeremy A Dodsworth, Paul C Blainey, Senthil K Murugapiran, Wesley D Swingley, Christian A Ross, Susannah G Tringe, Patrick S G Chain, Matthew B Scholz, Chien-Chi Lo, Jason Raymond, Stephen R Quake, and Brian P Hedlund. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nature Communications*, 4:1854, 2013.

[114] Matthias Hess, Alexander Sczyrba, Rob Egan, Tae-Wan Kim, Harshal Chokhawala, Gary Schroth, Shujun Luo, Douglas S Clark, Feng Chen, Tao

Zhang, Roderick I Mackie, Len A Pennacchio, Susannah G Tringe, Axel Visel, Tanja Woyke, Zhong Wang, and Edward M Rubin. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–467, January 2011.

[115] Jane Kuypers and Keith R Jerome. Applications of Digital PCR for Clinical Microbiology. *Journal of clinical microbiology*, pages JCM.00211–17, March 2017.

[116] Philip J Johnson and Y M Dennis Lo. Plasma nucleic acids in the diagnosis and management of malignant disease. *Clinical chemistry*, 48(8):1186–1193, August 2002.

[117] R.D. Mitra and G.M. Church. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research*, 27(24):e34–e39, 1999.

[118] Lisa Zeigler Allen, Thomas Ishoey, Mark A Novotny, Jeffrey S McLean, Roger S Lasken, and Shannon J Williamson. Single virus genomics: a new tool for virus discovery. *PLoS ONE*, 6(3):e17722, 2011.

[119] G P Raeber, M P Lutolf, and J A Hubbell. Molecularly Engineered PEG Hydrogels: A Novel Model System for Proteolytically Mediated Cell Migration. *Biophysical journal*, 89(2):1374–1388, August 2005.

[120] Yanbin Wu, Sony Joseph, and N R Aluru. Effect of cross-linking on the diffusion of water, ions, and small molecules in hydrogels. *The Journal of Physical Chemistry B*, 113(11):3512–3520, March 2009.

[121] Edward A Phelps, Nduka O Enemchukwu, Vincent F Fiore, Jay C Sy, Niren Murthy, Todd A Sulchek, Thomas H Barker, and Andres J Garcia. Maleimide Cross-Linked Bioactive PEG Hydrogel Exhibits Improved Reaction Kinetics and Cross-Linking for Cell Encapsulation and In Situ Delivery. *Advanced Materials*, 24(1):64–70, December 2011.

[122] Jennifer Elisseeff. Hydrogels: Structure starts to gel. *Nature Materials*, 7(4):271–273, April 2008.

[123] Alexey Atrazhev, Dammika P Manage, Alexander J Stickel, H John Crabtree, Linda M Pilarski, and Jason P Acker. In-gel technology for PCR genotyping and pathogen detection. *Analytical Chemistry*, 82(19):8079–8087, October 2010.

[124] Huaping Tan, Alicia J DeFail, J Peter Rubin, Constance R Chu, and Kacey G Marra. Novel multiarm PEG-based hydrogels for tissue engineering. *Journal of biomedical materials research. Part A*, 92(3):979–987, March 2010.

[125] Yan Li, Chuan Yang, Majad Khan, Shaoqiong Liu, James L Hedrick, Yi-Yan Yang, and Pui-Lai R Ee. Nanostructured PEG-based hydrogels with tunable

physical properties for gene delivery to human mesenchymal stem cells. *Biomaterials*, 33(27):6533–6541, September 2012.

[126] Thomas P Kraehenbuehl, Prisca Zammaretti, André J Van der Vlies, Ronald G Schoenmakers, Matthias P Lutolf, Marisa E Jaconi, and Jeffrey A Hubbell. Three-dimensional extracellular matrix-directed cardioprogenitor differentiation: systematic modulation of a synthetic cell-responsive PEG-hydrogel. *Biomaterials*, 29(18):2757–2766, June 2008.

[127] Glenn K Fu, Jing Hu, Pei-Hua Wang, and Stephen P A Fodor. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences*, 108(22):9026–9031, May 2011.

[128] Rae M Robertson, Stephan Laib, and Douglas E Smith. Diffusion of isolated DNA molecules: Dependence on length and topology. *Proceedings of the National Academy of Sciences*, 103(19):7310–7314, April 2006.

[129] Laney M Weber, Christina G Lopez, and Kristi S Anseth. Effects of PEG hydrogel crosslinking density on protein diffusion and encapsulated islet survival and function. *Journal of biomedical materials research. Part A*, 90(3):720–729, September 2009.

[130] Roger S Lasken and Timothy B Stockwell. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC biotechnology*, 7:19, 2007.

[131] N Segata, L Waldron, A Ballarini, and V Narasimhan. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature*, 2012.

[132] Martin Wu and Alexandra J Scott. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics (Oxford, England)*, 28(7):1033–1034, April 2012.

[133] J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, May 2010.

[134] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120, August 2014.

[135] Kirill Rotmistrovsky. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets, 2011.

[136] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5):455–477, May 2012.

[137] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, July 2015.

[138] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W Ussery. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9):3100–3108, 2007.

[139] Paola N Perrat, Shamik DasGupta, Jie Wang, William Theurkauf, Zhiping Weng, Michael Rosbash, and Scott Waddell. Transposition-driven genomic heterogeneity in the Drosophila brain. *Science*, 340(6128):91–95, April 2013.

[140] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, October 2009.

[141] Mark N Lee, Chun Ye, Alexandra-Chloé Villani, Towfique Raj, Weibo Li, Thomas M Eisenhaure, Selina H Imboywa, Portia I Chipendo, F Ann Ran, Kamil Slowikowski, Lucas D Ward, Khadir Raddassi, Cristin McCabe, Michelle H Lee, Irene Y Frohlich, David A Hafler, Manolis Kellis, Soumya Raychaudhuri, Feng Zhang, Barbara E Stranger, Christophe O Benoist, Philip L De Jager, Aviv Regev, and Nir Hacohen. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343(6175):1246980, March 2014.

[142] Annapurna Poduri, Gilad D Evrony, Xuyu Cai, and Christopher A Walsh. Somatic Mutation, Genomic Variation, and Neurological Disease. *Science*, 341(6141):1237758–1237758, July 2013.

[143] Jennifer A Erwin, Maria C Marchetto, and Fred H Gage. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nature reviews. Neuroscience*, 15(8):497–506, August 2014.

[144] E Vanneste, C Melotte, T Voet, C Robberecht, S Debrock, A Pexsters, C Staessen, C Tomassetti, E Legius, T D'Hooghe, and J R Vermeesch. PGD for a complex chromosomal rearrangement by array comparative genomic hybridization. *Human Reproduction*, 26(4):941–949, March 2011.

[145] Yu Hou, Wei Fan, Liying Yan, Rong Li, Ying Lian, Jin Huang, Jinsen Li, Liya Xu, Fuchou Tang, X Sunney Xie, and Jie Qiao. Genome analyses of single human oocytes. *Cell*, 2013.

[146] Anver Kuliev and Svetlana Rechitsky. Polar body-based preimplantation genetic diagnosis for Mendelian disorders. *Molecular human reproduction*, 17(5):275–285, May 2011.

[147] Thierry Voet, Parveen Kumar, Peter Van Loo, Susanna L Cooke, John Marshall, Meng-Lay Lin, Masoud Zamani Esteki, Niels Van der Aa, Ligia Mateiu, David J McBride, Graham R Bignell, Stuart McLaren, Jon Teague, Adam Butler, Keiran Raine, Lucy A Stebbings, Michael A Quail, Thomas D'Hooghe, Yves Moreau, P Andrew Futreal, Michael R Stratton, Joris R Vermeesch, and Peter J Campbell. Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Research*, 41(12):6119–6138, July 2013.

[148] Gilad D Evrony, Eunjung Lee, Bhaven K Mehta, Yuval Benjamini, Robert M Johnson, Xuyu Cai, Lixing Yang, Psalm Haseley, Hillel S Lehmann, Peter J Park, and Christopher A Walsh. Cell Lineage Analysis in Human Brain Using Endogenous Retroelements. *Neuron*, 85(1):49–59, January 2015.

[149] Ángel J Picher, Bettina Budeus, Oliver Wafzig, Carola Krüger, Sara García-Gómez, María I Martínez-Jiménez, Alberto Díaz-Talavera, Daniela Weber, Luis Blanco, and Armin Schneider. TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. *Nature Communications*, 7:13296, November 2016.

[150] Robert C Edgar, Brian J Haas,. Jose C Clemente, Christopher Quince, and Rob Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)*, 27(16):2194–2200, August 2011.

[151] Brian J Haas, Dirk Gevers, Ashlee M Earl, Mike Feldgarden, Doyle V Ward, Georgia Giannoukos, Dawn Ciulla, Diana Tabbaa, Sarah K Highlander, Erica Sodergren, Barbara Methé, Todd Z DeSantis, Human Microbiome Consortium, Joseph F Petrosino, Rob Knight, and Bruce W Birren. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, 21(3):494–504, March 2011.

[152] Atray Dixit. Correcting Chimeric Crosstalk in Single Cell RNA-seq Experiments. *bioRxiv*, page 093237, December 2016.

[153] Jing Tu, Jing Guo, Junji Li, Shen Gao, Bei Yao, and Zuhong Lu. Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis. *PLoS ONE*, 10(10):e0139857, 2015.

[154] Jing Tu, Na Lu, Mengqin Duan, Mengting Huang, Liang Chen, Junji Li, Jing Guo, and Zuhong Lu. Hotspot Selective Preference of the Chimeric Sequences

Formed in Multiple Displacement Amplification. *International journal of molecular sciences*, 18(3), February 2017.

[155] Martin Krzywinski, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, September 2009.

[156] Nicola Crosetto, Magda Bienko, and Alexander Van Oudenaarden. Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics*, 2015.

[157] Hironori Niki and Sota Hiraga. Subcellular distribution of actively partitioning F plasmid during the cell division cycle in E. coli. *Cell*, 90(5):951–957, September 1997.

[158] N A Yamada, L S Rector, P Tsang, E Carr, A Scheffer, M C Sederberg, M E Aston, R A Ach, A Tsalenko, N Sampas, B Peter, L Bruhn, and A R Brothman. Visualization of Fine-Scale Genomic Structure by Oligonucleotide-Based High-Resolution FISH. *Cytogenetic and Genome Research*, 132(4):248–254, 2011.

[159] Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4):360–361, April 2014.

[160] Richard A White, Paul C Blainey, H CHRISTINA Fan, and Stephen R Quake. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics*, 10(1):116, 2009.

[161] L Blanco and M Salas. Replication of phage phi 29 DNA with purified terminal protein and DNA polymerase: synthesis of full-length phi 29 DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 82(19):6404–6408, October 1985.

[162] Claudia Spits, Cédric Le Caignec, Martine De Rycke, Lindsey Van Haute, André Van Steirteghem, Inge Liebaers, and Karen Sermon. Whole-genome multiple displacement amplification from single cells. *Nature Protocols*, 1(4):1965–1970, November 2006.

[163] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods*, January 2017.