UNIVERSITY OF CALIFORNIA, SAN DIEGO

Massively Parallel Polymerase Cloning and Genome Sequencing of Single Cells Using the Microwell Displacement Amplification System (MIDAS)

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Jeffrey Gole

Committee in charge:

Professor Kun Zhang, Chair
Professor Vineet Bafna
Professor Michael Heller
Professor Xiaohua Huang
Professor Yu-Hwa Lo

2013

The Dissertation of Jeffrey Gole is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2013

DEDICATION

For my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank everyone who helped and supported me throughout my graduate school experience. Without the support of the following people, this dissertation would not be possible.

First, my advisor, Kun Zhang, decided to take a chance on me five years ago. Though I still believe he just picked the first naïve graduate student who asked him about rotating in his lab, his trust and faith in me over the past five years has allowed me to succeed. He taught me how to be a successful researcher and scientist. More importantly, he taught me to keep a level mind as research does not always go as planned. Although I know I was most likely very frustrating to talk to at times, he never exhibited a negative attitude towards my research or myself. Furthermore, he helped me secure a job post graduation, as well as the prestigious Siebel Scholarship.

I would also like to thank the rest of my committee members, Dr. Xiaohua Huang, Dr. Michael Heller, Dr. Vineet Bafna, and Dr. Yu-Hwa Lo. They were always available for advice when asked. Specifically, Dr. Lo originally helped with some very important design aspects, and gave full access to all of his lab equipment.

Furthermore, I would like to thank several member of my lab. Athurva Gore was a great help with regards to data analysis and algorithm development. His honesty and wise insights helped me to progress my research to new levels. I would also like to thank Andrew Richards, who

helped tremendously by sharing in some of the experimental load, ultimately giving me time to think critically about my results. I would also like to thank Alan Fung for his consistent sequencer maintenance, Alice Li for teaching me some experimental techniques, Chris Wei for help with data analysis, Sam Chiang for help with bacterial culture, and Noi Plongthongkum for help with the robotic system.

I would also like to thank our major collaborators. Specifically, Dr. Jerold Chun and his student, Diane Bushman, for always having freshly sorted neurons available for us. They also gave great advice regarding data analysis. Also, from the Lo lab, Randy Chen and Roger Chiu were an enormous help with microfabrication and clean room technology.

Finally, I would like to thank my family and fiancée. They have been there for me throughout graduate school, continually supporting me through good times and bad, including three surgeries. Without their continual support, I would have greatly struggled through graduate school.

Chapter 2, in part, is a reprint of the material as it appears in: Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.* Used with permission. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, is a reprint of the material as it appears in: Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman

Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.* Used with permission. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in: Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.* Used with permission. The dissertation author was the primary investigator and author of this paper.

VITA

2008 Bachelor of Science, Biomedical Engineering, Washington University in St. Louis

2013 Doctor of Philosophy, Bioengineering, University of California, San Diego

PUBLICATIONS

Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.*

FIELDS OF STUDY

Major Field: Bioengineering

Studies in Single Cell Genomics
Professor Kun Zhang

ABSTRACT OF THE DISSERTATION

Massively Parallel Polymerase Cloning and Genome Sequencing of Single Cells Using the Microwell Displacement Amplification System (MIDAS)

by

Jeffrey Gole

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2013

Professor Kun Zhang, Chair

Genome sequencing of single cells has a variety of applications, including characterizing difficult-to-culture microorganisms and identifying somatic mutations in single cells from mammalian tissues. A major hurdle in

this process is the bias in amplifying the genetic material from a single cell, a procedure known as polymerase cloning. Here we describe the microwell displacement amplification system (MIDAS), a massively parallel polymerase cloning method in which single cells are randomly distributed into hundreds to thousands of nanoliter wells and simultaneously amplified via multiple displacement amplification for shotgun sequencing. MIDAS reduces amplification bias because polymerase cloning occurs in physically separated nanoliter-scale reactors, thus increasing the template concentration by reducing the reaction volume.

MIDAS is first applied to single *E. coli* cells, facilitating the *de novo* assembly of near-complete microbial genomes to unprecedented levels. In addition, MIDAS allowed us to detect single-copy number changes in primary human adult neurons at 1–2 Mb resolution, resolutions not possible with standard whole genome amplification. MIDAS will further the characterization of genomic diversity in many heterogeneous cell populations.

Chapter 1: Introduction

1.1: Next Generation Sequencing

Over the last six years, the cost of sequencing has dropped precipitously, at a rate orders of magnitude greater than Moore's law[1]. In addition to a cost reduction, next generation sequencing allows for hundreds of millions of molecules to be accurately sequenced in parallel. Thus, an entire human genome can be sequenced at high depth in less than 2 weeks, a vast improvement from the original, Sanger sequenced human genome[2]. Commercialization of high throughput sequencers by companies such as Illumina and 454 have given scientists the opportunity to analyze enormous amounts of sequencing data. Specifically, the Illumina HiSeq can generate over 600 billion bases of sequencing data in a single run[3].

Due to the ease of which sequencing data can be obtained and the subsequent assembly of numerous genomes from various organisms, researchers have begun to study genomic complexities, such as point mutations[4], epigenetics[5], and copy number variations on bulk populations of cells[6]. More recently, scientists have noticed that cells exist in heterogeneous populations. An obvious example is cancer[7]. Cancer cells have distinct genomes from healthy cells of the same organism, and scientists continue to decipher these ever changing genomes. However, non-cancerous single cells

from the same tissue of an individual mammal could also possess varying genomes, and these individual cells are just now being studied[8]. Additionally, many bacteria live in heterogeneous populations, making it impossible to sequence a single organism[9]. Important questions remain regarding the genomes of individual cells. As will be described later in this chapter, technology is now becoming available to help decipher single cell genomes.

1.2: Single Cell Isolation

Isolation of single cells cleanly has paramount importance for single cell sequencing. Single cells must be physically separated and contained without cell free contamination. The technology for isolating single cells has greatly improved the throughput in recent years.

Originally, cells were separated through simple dilutions[9]. One can assume that cells seed according to the Poisson Distribution:

$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ is the expected value, *k* is the desired number, and *e* is the base of the natural logarithm. Thus, if cells are diluted in PCR tubes such that there is one cell for every ten tubes, about 99.5% of the tubes will contain either zero or one cell. Although a large percentage of tubes will have zero cells (90.5%), this dilution essentially ensures that there will be no more that one cell in every

tube. This method results in very low throughput isolation. To improve upon isolation efficiency, scientists began to use micromanipulation to physically pipette single cells into individual PCR tubes[10]. Thus, upon proper micromanipulation, every tube is guaranteed to have a single cell. This process, however, is completely manual and very tedious. Another manual, laborious process includes laser microdissection of single cells from tissues[11].

Fluorescent activated cell sorting (FACS) proved to be a major improvement for cell isolation[12]. Cells can be loaded into a FACS machine, and quickly sorted into a 96 or 384 well plate such that each tube has only a single cell. This technique continues to be commonly used as thousands of cells can be sorted in less than one hour. Although FACS sorting greatly increases the throughput of single cell isolation, technical challenges still exist. FACS machines are expensive and bulky, making it difficult for smaller labs to maintain their own machines. Furthermore, contamination of cell free DNA abounds in FACS machines. Thus, they must be prepped and cleaned for several hours prior to usage. This fact also makes the use of a communal FACS machine very difficult, if not impossible.

Perhaps one of the best methods for single cell isolation is the use of microfluidic devices[13-16]. The advent of releasable microvalves allows for the physical isolation of cells. These devices can be made in a microfabrication facility at low cost, and are about the size of a standard microscope slide. Furthermore, they are disposable so cell free contamination is less of an issue. These devices can use simple diffusion, pressure activated flows, or optical

tweezers to isolate cells into small chambers for further processing. However, these devices require special equipment to run. Additionally, not everyone has access to microfabrication facilities to design these devices.

1.3: Single Cell Amplification

For next generation sequencing, generally micrograms of DNA are required as input[17]. Although not an issue for bulk sequencing as often the DNA from millions of cells is used as input, obtaining this amount of DNA poses a large problem for single cell sequencing. The amount of DNA from a single cell ranges from 5 femtograms for a single bacterium, to 5 picograms for a single mammalian cell. These cells often cannot be cultured to obtain millions of clonal duplicates for sequencing. Thus, they must be efficiently and uniformly amplified in a process call whole genome amplification, or WGA[9, 12, 13, 15, 18, 19]. Although PCR based WGA methods exist, the most common WGA is multiple displacement amplification, or MDA[20]. In this method, cells are first lysed and the DNA is denatured. Then, random hexamer primers anneal completely stochastically to different regions of the single stranded DNA. Phi29 polymerase, which has a very long processivity and a very high strand displacement activity, then polymerizes the second strand from each random hexamer. When the polymerase reaches a downstream primer, it displaces the strand. This process continues on the displaced strands resulting in

exponential amplification. The final product is a condensed, hyberbranching structure[12].

The major advantages to MDA are the long processivity and ability to displace strands with high efficiency. This results in long, continuous strands up to 12 kilobases in length. Additionally, the process occurs isothermally at 30 C so thermocycling is not necessary. These properties make it a far better choice to PCR for single cell WGA[9].

Two major disadvantages of MDA have plagued researchers for over ten years, and have prevented significant advancements in single cell genomics. The first major disadvantage is cell free and environmental contamination. Since the mechanism relies on random priming, the hexamers can anneal just as easily to the target template as it can to contaminating DNA. Often from the dust or the environment, cell free DNA can effortlessly float into any given reaction[9]. To complicate matters further, the template DNA mass from single cells is often several orders of magnitude less than the mass of contaminating DNA. After exponential amplification, the contaminating DNA will dominate the amplicon product. Thus, great care must be taken when preparing any MDA reaction. Using ultraviolet radiation on reagents and cleaning equipment with ethanol and other chemicals is necessary prior to any reaction[9, 12, 13, 15, 18, 19].

The second major disadvantage of MDA is the biased amplification of the template genome[12]. Again, the reaction relies on random hexamers annealing completely stochastically to the template. The annealing of primers,

however, does not occur in a uniform and timely manner. For example, at time t=0, primers can anneal to a certain region of the template which begins to exponentially amplify. This region will exist in a much higher concentration than another region where primers begin to anneal at time t=1 hour, and even more so than a region where primers anneal at time t=5 hours. Thus, certain regions of the template might have >1000x coverage, where other regions will have virtually no coverage. This makes downstream analyses, such as genome assembly and CNV calling, extremely difficult[7, 9, 12]. Therefore, many single cell sequencing applications have been limited up until very recently.

The technology of WGA has improved significantly over the past 10 years to account for cell free contamination and amplification bias. Originally, through the use of dilutions in a clean and controlled environment, the amount of contamination was limited to sub femtogram levels. This allowed, for the first time, the sequencing and assembly of a single bacterial cell. The bias was still significant, allowing for only 60% of the genome to be accurately assembled[9]. With the implementation of clean FACS sorting, cell free DNA contamination could be removed from the samples, again greatly improving the contamination rates. Since this was still just an improvement on isolation, the bias during amplification remained relatively high[12]. To improve upon the bias, small volume reactions were implemented. This increased the concentration of the template, allowing for more uniform annealing of the primers in the initial stages of MDA[21]. Microfluidics resulted in easy implementation of small volumes. The amplicon mass was still too small for

sequencing due to the minimal reagents, so the amplicon was reamplified in a PCR tube. For the first time, bias was significantly reduced, though still too high to produce draft quality assemblies[13-16].

Most recently, a non-MDA based WGA technique, known as MALBAC, was implemented, reducing amplification bias to unprecedented levels[22, 23]. This reaction uses several rounds of linear amplification in the initial stages to increase the genomic copy number. After the initial amplification, PCR is used to amplify the DNA to microgram levels. The linear amplification still relies on random priming, so small biases still exist. Furthermore, the enzymes used are error prone, which creates false positive mutations. Despite these small drawbacks, MALBAC proved a strong alternative to MDA for whole genome amplification.

1.4: Library Construction

In order to sequence any given DNA template, a sequencing library must be constructed. Illumina sequencers can only sequence up to 150 base pairs per molecule[3]. Thus, the template DNA must be fragmented to 200-800 base pair fragments prior to sequencing. Furthermore, template DNA For Illumina sequencers, the template must anneal to oligonucleotides attached to a flow cell. Consequently, Illumina adapters must be present on the ends of each template molecule. The process of fragmenting template DNA and adding adapters is referred to as Illumina library construction.

Classically, template DNA is sheared using a Covaris machine to produce high-energy waves to fragment the DNA to a very tight size range. After end repair, A-tailing, and adapter ligation, PCR is used to amplify each individual molecule prior to sequencing. Prior to each step, purification is necessary, resulting in the compounding loss in template DNA. Therefore, micrograms of DNA prior to shearing are necessary to properly construct a library.

More recently, a library construction technique based on random transposition was developed. Called Nextera, complex transposase molecules with Illumina adapters attached randomly insert into and fragment template DNA[24, 25]. The results of this reaction are template molecules, ranging in size from 200-1000 bp, each with attached Illumina adapters. After the transposases are inactivated, PCR can be directly performed on the molecules to create a sequencing library without any purification steps. Although originally created to create sequencing libraries from 50-200 nanograms of genomic DNA, researchers have use Nextera to generate libraries with as little as 10 picograms of DNA. Thus, Nextera proves a great alternative for classical library construction when only small template masses are available, such as those derived from small volume whole genome amplification[24].

1.5: *De Novo* Assembly of Microbial Genomes from Single Cells

As previously mentioned, many bacteria coexist in heterogeneous populations. Prominent examples include sea-water[26], soil[27], and the human microbiome[28]. Less than one percent of bacteria have been sequenced and analyzed, namely because they are unculturable[29-32]. Scientists do not know exactly the types of environments that these bacteria can grow in or the nutrients that they feed on. Moreover, many of these bacteria require symbiotic relationships with differing bacteria to grow. Thus, the only clear way to genomically analyze these bacteria is through the use of single cell amplification and sequencing.

Previously, researchers have studied environmentally heterogeneous bacteria through metagenomics and 16S rRNA sequencing[33, 34]. This involves taking an environmental sample, lysing, creating a sequencing library, and sequencing. Accordingly, the DNA from every organism will be analyzed, and generally resolved down to a phylum taxonomic level. Specific species of bacteria cannot be determined, though recently updated algorithms can separate species in low complexity environments[35]. The16S rRNA gene is universal in bacteria. By PCR amplifying and sequencing this gene, phyla can also be determined.

Many of the unculturable bacteria could prove important in various applications. Those from seawater or soil could possible be used in chemical processing or alternative forms of energy[33]. However, those most interesting lie in the gut of humans[36-39].These can have a direct impact on metabolism.

Recently, scientists have shown through metagenomics that overall populations of bacteria in lean and obese identical twins are completely different. By feeding the bacteria from the lean twin to mice, the mice remained lean. On the other hand, mice fed bacteria from the obese twin became obese[40]. These important studies demonstrate the importance of the human microbiome with regards to metabolism. However, scientists still do not know exactly which genes or bacteria cause the varying metabolic rates.

With the implementation of single cell amplification and sequencing on bacteria, new assembly algorithms were created to account for the large amounts of biases in the sequencing data. Specifically, Velvet-SC[10] and SPAdes[41] were designed specifically for *de novo* assembly of single cell genomes. Both algorithms correct for chimeric reads created during MDA. Velvet-SC relies on De Brujin graphs for assembly. These are optimal for the short reads length generated from Illumina sequencers. It also uses a sliding minimum coverage cutoff, such that regions of low coverage can still be incorporated into the assembly. Thus, up to 70% of the genomes can be assembled. SPAdes relies on paired De Brujin graphs, and stitches together many assemblies into scaffolds. It also uses an initial error correction step to remove low quality reads prior to assemblies, and so misassemblies are minimized. Therefore, SPAdes shows improved results compared to Velvet-SC.

1.6: Copy Number Variation

Copy number variation (CNV) refers to the change in copy number of specific genes or regions. These can be as large as a chromosome and as small as part of a single gene. One of the most well known CNVs is trisomy 21, where chromosome 21 has 3 copies instead of 2, in Down's Syndrome patients. Many smaller germline CNVs, often found in bulk sequencing samples, have been characterized using both microarrays and sequencing[7, 42].

Recently, researchers have begun to analyze CNVs from single cancer cells[7, 42]. Cancer cells often exhibit very large CNVs in terms of size and are often much more than a single copy number gain or loss[43]. Additionally, each cancer cell from a given tumor generally has a unique CNV profile, displaying the need for single cell CNV analysis. With amplification using both MDA and MALBAC, scientists have discovered distinctive CNVs[22].

As with *de novo* assembly, complex algorithms were developed to accurately call CNVs in single cells. One such algorithm was developed by the Cold Spring Harbor Labarotory[42]. After sequencing, unique reads are binned into previously well-defined bins approximately 60 kilobases in size. The bins were chosen to remove mapping biases. After Lowess smoothing based on GC content, copy numbers are called using circular binary segmentation. This algorithm is the most widely used CNV calling algorithm. Another algorithm has been developed which is based on a Hidden Markov Model[22]. By comparing a cancerous cell to a healthy cell, the algorithm

removes background noise and calls unique CNVs. Sequencing data from a healthy cell are necessary for this algorithm to work properly.

In addition to cancer cells, CNVs have recently been called on single neurons from the same brain. Higher cognitive functions of the complex require a complex network of neurons. New evidence has shown that neurons contain non-identical genomes, and the DNA content can vary up to 10% in between neurons from the same brain based on FACS fluorescent quantification. Diseased neurons, such as those from Alzheimer's patients, display even great variations[44-50]. Unlike CNVs in tumor cells, CNVs in neurons are small in size and amplitude, often a single copy number gain or loss. Accordingly, these CNVs are more difficult to accurately characterize than larger CNVs. Due to biased whole genome amplification techniques, convincingly calling small CNVs has become a major hurdle. Nonetheless, this has become a hot topic in genomics, and could possibly elucidate several neurological disorders.

1.7: Scope of the Dissertation

The purpose of this dissertation was to develop a technique to unbiasedly amplify whole genomes of single cells. Additionally, we wanted to amplify many cells in parallel to decrease costs and increase throughput.

In Chapter 2, we describe the technology development portion of the dissertation, which is named the Microwell Displacement Amplification

System, or MIDAS. We describe the design and implementation of a microwell array consisting of thousands of nanoliter microwells. We also demonstrate extraction using micromanipulators followed by an adapted low input library construction technique. Finally, we show that the bias level of amplicons using MIDAS is far less than any previously published technique.

In chapter 3, we apply MIDAS to single *E. coli* cells for *de novo* assembly. We show that due to the low amount of bias, over 90% of the genome can be accurately assembled, close to 50% more assembly than previously published data. We then demonstrate that sequencing to a high depth can possibly result in even more assembled genome.

Lastly, in chapter 4, we apply MIDAS to single neuronal nuclei. First, we demonstrate that trisomy 21 can accurately be called using MIDAS, where it cannot using standard MDA. We then show that CNVs can be accurately called at the one megabase level using a computational spike in method, an unprecedented level from single cell amplicons. Furthermore, we present data that many CNVs overlap between single neuronal nuclei.

Chapter 2: A Technique for Generating Unbiased Whole Genome Amplification Sequencing Libraries from Single Cells

2.1: Abstract

We describe a method to uniformly amplify whole genomes of single cells and generate low input sequencing libraries. Single cell amplification occurs in nanoliter volume microwells, which increase the concentration of the template such that primer annealing can occur evenly. After extracting with a micromanipulator, low input sequencing libraries are generated with sub nanogram inputs. Finally, we show that this amplification method is superior to previously published methods in terms of reducing amplification bias.

2.2: Introduction

Amplification bias has proved to be the most challenging obstacle to overcome in single cell whole genome amplification. During multiple displacement amplification (MDA), random hexamer primers anneal to the template genome. This annealing is not completely stochastic, and due to exponential amplification, certain regions of the genome become

overrepresented in the respective sequencing library[20]. The bias causes great difficulty in many downstream analyses, such as *de novo* assembly and copy number variation calling.

Over the past several years, researchers have attempted to reduce amplification biases. One such technique includes double stranded nuclease normalization[12]. After amplification, amplicons are denatured and slowly reannealed. After a double stranded nuclease treatment, the overrepresented regions are degraded, resulting in a more uniform library. Furthermore, supplementing the reactions with single stranded binding proteins has been used to diminish biases[19, 51]. Another technique to minimize bias is reducing the amplification volume[14, 21]. The template concentration is subsequently increased, improving the primer annealing efficiency. Although bias is reduced, this technique typically does not produce enough amplicon mass for standard Illumina library construction. A third technique employed to curtail bias is using pseudo-linear amplification with the MALBAC method[22, 23]. The initial linear amplification increases the genome copy number, such that the subsequent exponential amplification occurs more uniformly. MALBAC has resulted vast improvements in single cell genomics.

As described previously, most Illumina sequencing libraries require micrograms of input DNA for proper construction. Because low volume amplifications often result in nanogram level amplicons, classical library construction remains unfeasible. With the advent of Nextera transposase based library construction, input DNA mass has decreased significantly.

Libraries can be prepared with as little as 10 picograms of DNA, making possible the use of small volume amplifications[24, 25].

In this chapter, we describe a microwell array device that can amplify hundreds of cells in parallel. The device, the size of a microscope slide, has 16 arrays of 255 nanoliter scale microwells. After random cell seeding, amplification can be seen in real time using an incubated microscope. Positive wells can then be extracted using a micromanipulator with fine glass pipettes, followed by low input library construction using Nextera tagmentation. The resulting libraries are then compared to other single cell amplification methods in addition to bulk libraries, and the bias proves minimal.

2.3 Methods

2.3.1 Microwell Array Fabrication

Microwell arrays were fabricated from polydimethylsiloxane (PDMS). Each array was 7 mm x 7 mm, with 2 rows of 8 arrays per slide and 255 microwells per array. The individual microwells were 400 μm in diameter and 100 μm deep (~12 nL volume), and were arranged in honeycomb patterns in order to minimize space in between the wells. To fabricate the arrays, first, an SU-8 mold was created using soft lithography at the Nano3 facility at UC San Diego. Next, a 10:1 ratio of polymer to curing agent mixture of PDMS was poured over the mold. Finally, the PDMS was degassed and cured for 3 hours at 65 ^{o}C.

2.3.2 Cell Seeding, Lysis, and Multiple Displacement Amplification

All reagents not containing DNA or enzymes were first exposed to ultraviolet light for 10 minutes prior to use. The PDMS slides were treated with oxygen plasma to make them hydrophilic and ensure random cell seeding. The slides were then treated with 1% bovine serum albumin (BSA) (EMD Chemicals, Billerica, MA) in phosphate buffered saline (PBS) (Gibco, Grand Island, NY) for 30 minutes and washed 3x with PBS to prevent DNA from sticking to the PDMS. The slides were completely dried in a vacuum prior to cell seeding. Cells were diluted in 1x PBS to a concentration of 0.1 cells per well per array, and 3 μL of cell dilution was added to each array. This dilution ensures that approximately 99.5% of the wells have no more than one cell.

Initially, to verify that cell seeding adhered to the Poisson distribution, cells were stained with 1x SYBR green and viewed under a fluorescent microscope. Proper cell distribution was further confirmed with SEM imaging. For SEM imaging, chromium was sputtered onto the seeded cells for 6 seconds to increase conductivity. Note that the imaging of cell seeding was only used to confirm the theoretical Poisson distribution and not performed during actual amplification and sequencing experiments due to the potential introduction of contamination.

After seeding, cells were left to settle into the wells for 10 minutes. The seeded cells were then lysed either with 300 U ReadyLyse lysozyme at 100

U/μL (Epicentre, Madison, WI) and incubation at room temperature for 10 minutes, or with five 1 minute freeze/thaw cycles using a dry ice brick and room temperature in a laminar flow hood. After lysis, 4.5 μL of alkaline lysis (ALS) buffer (400 mM KOH, 100 mM DTT, 10 mM EDTA) was added to each array and incubated on ice for 10 minutes. Then, 4.5 μL of neutralizing (NS) buffer (666 mM Tris-HCl, 250 mM HCL) was added to each array. 11.2 μL of MDA master mix (1x buffer, 0.2x SYBR green I, 1 mM dNTP's, 50 μM thiolated random hexamer primer, 8U phi29 polymerase, Epicentre, Madison, WI) was added and the arrays were then covered with mineral oil. The slides were then transferred to the microscope stage enclosed in a custom temperature controlled incubator set to 30 ^{o}C. Images were taken at 30-minute intervals for 10 hours using a 488 nm filter.

2.3.4 Image Analysis

Images were analyzed with a custom Matlab script to subtract background fluorescence. Because SYBR Green I was added to the MDA master mix, fluorescence under a 488 nm filter was expected to increase over time for positive amplifications. If a digital profile of fluorescent wells with increasing fluorescence over time was observed (approximately 10-20 wells per array), the array was kept. If no wells fluoresced, amplification failed and further experiments were stopped. Alternatively, if a majority of the wells fluoresced, the array was considered to have exogenous contamination from

environmental DNA and subsequent analysis was similarly stopped. If 2 abutting wells fluoresced, neither was extracted due to the higher likelihood of more than one cell in each well existing (as in this case, seeding was potentially non-uniform). Finally, only wells with amplicons originating from a single point were extracted, ensuring that only single-cell derived amplicons were processed; thus, any potential cross-well contamination was prevented.

2.3.5 Amplicon Extraction

1 mm outer diameter glass pipettes (Sutter, Novato, CA) were pulled to ~30 μm diameters, bent to a 45 degree angle under heat, coated with SigmaCote (Sigma, St. Louis, MO), and washed 3 times with dH_20. Wells with positive amplification were identified using the custom Matlab script described above. A digital micromanipulation system (Sutter, Novato, CA) was used for amplicon extraction. The glass pipette was loaded into the micromanipulator and moved over the well of interest. The microscope filter was switched to bright field and the pipette was lowered into the well. Negative pressure was slowly applied, and the well contents were visualized proceeding into the pipette. The filter was then switched back to 488 nm to ensure the well no longer contained any fluorescent material. Amplicons were deposited in 1 μL dH_20.

2.3.5 Amplicon Quantification

For quantification of microwell amplification, 0.5 µL of amplicon was amplified a second time using MDA in a 20 µL PCR tube reaction (1x buffer, 0.2x SYBR green I, 1 mM dNTP's, 50 mM thiolated random hexamer primer, 8U phi29 polymerase). After purification using Ampure XP beads (Beckman Coulter, Brea, CA), the 2nd round amplicon was quantified using a Nanodrop spectrophotometer. The 2nd round amplicon was then diluted to 1 ng, 100 pg, 10 pg, 1 pg, and 100 fg to create an amplicon ladder. Subsequently, the remaining 0.5 µL of the 1st round amplicon was amplified using MDA along with the amplicon ladder in a quantitative PCR machine. The samples were allowed to amplify to completion, and the time required for each to reach 0.5x of the maximum fluorescence was extracted. The original amplicon concentration could then be interpolated. This 2nd round of MDA was only performed during amplicon quantification in order to determine approximately how much DNA was produced in each microwell. Amplicons that were sequenced were only subjected to the initial round of MDA, and thus did not have any secondary MDA or quantification performed.

2.3.6 Low Input Library Construction

1.5 µL of ALS buffer was added to the extracted amplicons to denature the DNA followed by a 3-minute incubation at room temperature. 1.5 µL of NS buffer was added on ice to neutralize the solution. 10 U of DNA Polymerase I

(Invitrogen, Carlsbad, CA) was added to the denatured amplicons along with 250 nanograms of unmodified random hexamer primer, 1 mM dNTPs, 1x Ampligase buffer (Epicentre, Madison, Wi), and 1x NEB buffer 2 (NEB, Cambridge, MA). The solution was incubated at 37 ^{o}C for 1 hour, allowing second strand synthesis. 1 U of Ampligase was added to seal nicks and the reaction was incubated first at 37 ^{o}C for 10 minutes and then at 65 ^{o}C for 10 minutes. The reaction was cleaned using standard ethanol precipitation and eluted in 4 µL water.

Nextera transposase enzymes (Epicentre, Madison, WI) were diluted 100 fold in 1x TE buffer and glycerol. 10 µL transposase reactions were then conducted on the eluted amplicons after addition of 1 µL of the diluted enzymes and 1x tagment DNA buffer. The reactions were incubated for 5 minutes at 55 ^{o}C for mammalian cells and 1 minute at 55 ^{o}C for bacterial cells. 0.05 U of protease (Qiagen, Hilden, Germany) was added to each sample to inactivate the transposase enzymes; the protease reactions were incubated at 50 ^{o}C for 10 minutes followed by 65 ^{o}C for 20 minutes. 5 U Exo minus Klenow (Epicentre, Madison, WI) and 1 mM dNTP's were added and incubated at 37 ^{o}C for 15 minutes followed by 65 ^{o}C for 20 minutes. Two stage quantitative PCR using 1x KAPA Robust 2G master mix (Kapa Biosystems, Woburn, MA), 10 µM Adapter 1, 10 µM barcoded Adapter 2 in the first stage, and 1x KAPA Robust 2G master mix, 10 µM Illumina primer 1, 10 µM Illumina primer 2, and 0.4x SYBR Green I in the second stage was performed and the reaction was stopped before amplification curves reached their plateaus. The reactions

were then cleaned up using Ampure XP beads in a 1:1 ratio. A 6% PAGE gel verified successful tagmentation reactions.

2.3.7 Bulk Library Construction

Genomic DNA was extracted from approximately 4,000 neuronal nuclei using the DNeasy blood and tissue kit (Qiagen, Hilden, Germany). The genomic DNA was incubated with 1 µL undiluted Nextera transposase enzymes and 1x tagment DNA buffer for 5 minutes at 55 ^{o}C. The reactions were cleaned with MinElute columns (Qiagen, Hilden, Germany) and eluted in 20 µL water. 5 U Exo minus Klenow (Epicentre, Madison, WI) and 1 mM dNTP's were added and incubated at 37 ^{o}C for 15 minutes followed by 65 ^{o}C for 20 minutes. Two stage quantitative PCR using 1x KAPA Robust 2G master mix (Kapa Biosystems, Woburn, MA), 10 µM Adapter 1, 10 µM barcoded Adapter 2 in the first stage, and 1x KAPA Robust 2G master mix, 10 µM Illumina primer 1, 10 µM Illumina primer 2, and 0.4x SYBR Green I in the second stage was performed and the reaction was stopped before amplification curves reached their plateaus. The reactions were then cleaned up using Ampure XP beads in a 1:1 ratio. A 6% PAGE gel verified successful tagmentation reactions.

2.3.8 Data Analysis

Bacterial libraries were size selected into the 300-600 bp range and sequenced in an Illumina MiSeq using 100 bp paired end reads. *E. coli* data was both mapped to the reference genome. For the mapping analysis, libraries were mapped as single end reads to the reference *E. coli* K12 MG1655 genome using default Bowtie[52] parameters with removal of any reads with multiple matches. Contamination was analyzed, and clonal reads were removed using SAMtools'[53] rmdup function. The reads were then binned into 4,600 equally sized bins for bias analysis.

Mammalian single-cell libraries from neuronal nuclei were sequenced in an Illumina Genome Analyzer IIx or Illumina HiSeq using 36 bp single end reads. For each sample, reads were mapped to the genome using Bowtie. Clonal reads resulting from Polymerase Chain Reaction artifacts were removed using samtools, and the remaining unique reads were then assigned into 49,891 genomic bins of approximately 60 kb in size that were previously determined such that each would contain a similar number of reads after mapping[42]. The binning was used for bias analysis.

In-tube MDA sperm data[54], microfluidic based amplification sperm data[6], MALBAC sperm data[23], and MALBAC cancer data[22] were downloaded from the SRA database corresponding to previously published data. For the sperm samples, 2 random sperm libraries were combined to create a diploid

sample, and the X and Y chromosomes were excluded from further analysis. For the cancer sample, only the diploid chromosomes were included in analysis. The data was analyzed in a similar fashion as the mammalian cells.

2.4: Results

We designed and fabricated microwell arrays of a size comparable to standard microscope slides. The format of the arrays, including well size, pattern and spacing, was optimized to achieve efficient cell loading, optimal amplification yield and convenient DNA extraction. Each slide consisted of 16 arrays, each containing 255 microwells 400 μm in diameter, allowing for parallel amplification of 16 separate heterogeneous cell populations (**Figure 2.1a**). After testing several materials such as SU-8 on glass, which had a very weak bonding, we found that PDMS was the optimal material choice, as array fabrication was very reproducible and cost was minimal. Furthermore, PDMS could obtain hydrophilic properties with the use of oxygen plasma for several hours. All liquid handling procedures (cell seeding, lysis, DNA denaturation, neutralization and addition of amplification master mix) required one pump of a pipette per step per array, minimizing the labor required for hundreds of amplification reactions. This system requires less of each amplification and library construction reagent than conventional methods, as each microwell spatially confines the reaction to 12 nL in volume. Thus, cost was minimized when performing MIDAS (**Table 2.1**).

We tested multiple cell-loading densities to ensure that each well would contain only one single cell, and initially loaded the microwells at densities of roughly 1 cell per well and 1 cell per 10 wells. By the Poisson distribution, described by the equation:

$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ is the expected value, *k* is the desired number, and *e* is the base of the natural logarithm, in the 1 cell per well case, 63% should have at least one cell, but 26% could have more than one cell. Wells containing greater than 1 cell cannot be used for single cell amplification. In the 1 cell per 10 well case, no more than 0.5% of the wells should contain more than 1 cell. We confirmed that the cells were indeed being seeded at the theoretical distribution using fluorescent microscopy after staining cells with SYBR Green I (**Figure 2.2**). We thus decided to load cells at a density of 1 cell per 10 wells, ensuring that 99.5% of generated amplicons would arise from a single cell. The remaining empty wells served as internal negative controls, allowing easy detection and elimination of contaminated samples. We further confirmed proper microbial and mammalian cell seeding in microwells at the 1 cell per 10 well level by scanning electron microscopy (**Figure 2.1b, Figure 2.3**). Scanning electron microscopy proved that each “fluorescent dot” arose from a single cell, and not multiple cells sticking to each other.

After seeding of cell populations into each microwell array, we performed limited multiple displacement amplification on the seeded single cells in the partitioned microwells, each with a physically separated (save for a thin aqueous layer atop the arrays) volume of ~12 nL, in a temperature and humidity controlled chamber (**Figure 2.1c, Figure 2.2**). We used SYBR Green I to visualize the amplicons growing in real-time using an epifluorescent microscope (**Figure 2.4**). A random distribution of amplicons across the arrays was observed with ~10% of the wells containing amplicons, further confirming the parallel and localized amplification within individual microwells as well as the stochastic seeding of single cells. After amplification in the microwells, we used a micromanipulation system to extract amplicons from individual wells for sequencing (**Figure 2.1c**). We estimated that the masses of the extracted amplicons ranged from 500 picograms to 3 nanograms. These masses correspond to expected values of amplicons generated based on the minimal reaction volume, as there is a linear correlation between reaction volume and final amplicon mass.

When performing a single-cell amplification experiment, there are two potential sources of contamination that could result in an inaccurate characterization of the genome of the sample of interest. These are exogenous contamination, in which samples are exposed to cell-free DNA from environmental sources or reagents, and cross-well contamination, in which DNA from one microwell diffuses into other microwells. We ensured that neither form of contamination was occurring. To detect arrays that

contained exogenous contamination, we checked for a uniform increase of fluorescent signal across all microwells. Any samples that showed this high fluorescence across all wells were removed; thus, any samples exposed to cell-free DNA were simply not analyzed. To ensure that cross-well contamination was not occurring, we performed real-time fluorescent monitoring during the amplification procedure. Only single wells with single amplicons originating from a single point were extracted for analysis, preventing any cross-well contamination or selection of any wells containing more than one cell (**Figure 2.5**). If even a miniscule amount of DNA was diffusing out of a microwell, an increased fluorescence would be observed in adjacent wells owing to amplification occurring in every well[16]; this diffusion was not observed in any cases. To further confirm that cross-well contamination was not occurring, we loaded a mixture of human neuronal nuclei with two separate genomic backgrounds, one healthy line and one trisomy 21 line. After extraction, we confirmed that all extracted cells corresponded only to one background by looking for distinct copy numbers of chromosome 21 (**Table 2.2**). Any variation in copy numbers was primarily due batch effects.

Following amplification in the microwells, the amplicons could be easily extracted using a micromanipulator. As mentioned previously, we could monitor the amplification in real time, and correlate the imaging results to the initial cell seedings. The micromanipulator was moved over wells that both initially contained cells and showed an increase in fluorescence over time.

After the removal of mineral oil, the pipette was lowered into the well, and negative pressure was applied to remove the ~12 nL of amplicon. To test efficient extraction, we overloaded the wells with genomic DNA such that every well would amplify. We then extracted from a single well. While the fluorescence from this specific well was removed, the fluorescence of the surrounding wells remained constant (**Figure 2.5**). This proved that extraction was localized and cross contamination during extraction was minimal.

To construct Illumina sequencing libraries from the extracted nanogram-scale DNA amplicons, we used a modified in-tube method based on the Nextera Tn5 transposase. Previous studies have shown that Nextera transposase-based libraries can be prepared using as little as 10 picograms of genomic DNA[25]. However, the standard Nextera protocol was unable to generate high-complexity libraries from MDA amplicons, resulting in poor genomic coverage. Clonal PCR duplicates remained in very high proportion in the sequencing data. We reasoned that the transposases were having difficulty accessing the complex, three-dimensional structure of the MDA amplicons. We tried several strategies to reduce the amplicons to linear DNA. These included physical shearing, debranching with S1 nuclease followed by nick translation, denaturing followed by second strand synthesis with phi29 polymerase, and denaturing followed by second strand synthesis with DNA polymerase I. Ultimately, to address this issue, we used random hexamers and DNA Polymerase I to first convert the hyperbranched amplicons into unbranched double-stranded DNA molecules, which allowed effective library

construction using *in vitro* transposition (**Figure 2.1d**). Enzymatic inactivation of the transposases removed a purification step, further increasing the efficiency of library construction. In addition, we used a small reaction volume to further reduce the biases of library preparation.

Once the libraries were sequenced, we analyzed the biases of the libraries using the pipeline described in the methods section of this chapter. Both bacterial and mammalian single cell libraries amplified with MIDAS were compared to single cell libraries amplified using standard, in-tube MDA. Furthermore, an additional limited in-tube MDA library, in which MDA was limited to 3 hours, was analyzed. For mammalian cells, a bulk library derived from approximately 4,000 cells was used as a gold standard. One can easily determine the extreme reduction in biases from the MIDAS derived libraries when compared to all in-tube MDA libraries, including the limited MDA library (**Figure 2.6**). Where the in tube MDA libraries displayed many spikes throughout the genome, indicating regions of extreme amplification, in addition to regions with very little coverage, the coverage in the MIDAS derived libraries remained relatively constant throughout the entire genome. In fact, the uniformity rivaled that seen in the bulk library. The biases can be more easily visualized in a histogram. For the in-tube MDA libraries, one can see a spike around 0, indicating that most bins have almost no coverage. Furthermore, a very long right tail exists indicating that that a few bins have very high coverages. In contrast, the MIDAS derived libraries display a very tight, normal distribution of coverages, implicating that most regions of the

genome contain the same coverage. Additionally, as the bin sizes increase, the MIDAS derived libraries' distribution of coverages looks almost identical to the distribution in the bulk library (**Figure 2.7**)

We also desired to compare our data to previously published single cell whole genome amplification data as other single cell sequencing methods that reduce amplification bias and increase genomic coverage have been reported. One such method utilizes a microfluidic device to isolate single cells and perform whole genome amplification in a 60nL volume[6]. Another method, MALBAC, incorporates a novel enzymatic strategy to amplify single DNA molecules initially through quasi-linear amplification to a limited magnitude prior to exponential amplification and library construction[22]. MALBAC has been performed in microliter reactions in conventional reaction tubes. MIDAS represents an orthogonal strategy that adapts MDA to a microwell array. We compared data generated from single neurons amplified with MIDAS to previously published data from combined (and therefore diploid) pools of two single sperm cells amplified using standard in-tube MDA[54], the microfluidic device[6] and MALBAC[22, 23]. To ensure a fair comparison, we normalized sequencing depth to an equal amount for each method and processed the raw sequencing data for each sample using an identical computational pipeline. We also compared MIDAS to a single SW480 cancer cell amplified by MALBAC. In this case, to ensure a fair comparison to the primarily diploid cell analyzed using MIDAS, we limited our analysis to regions consistently identified as diploid in the cancer cell (parts of chromosomes 1, 4, 6, 8, 10 and

15)[22]. MIDAS compares favorably to each amplification method (**Figure 2.8, Figure 2.9)**, generating the lowest levels of bias across the genome. In the above-mentioned figures, the in-tube MDA libraries show the most bias, followed by the low volume microfluidic MDA derived libraries. The microfluidic MDA based libraries underwent a second round of amplification to obtain enough template for standard Illumina library construction, which explains the high bias levels. The MALBAC derived libraries display comparatively little bias, however, the MIDAS libraries are superior in terms of uniformity. Thus, the MIDAS libraries prove to be the least biased single cell libraries to date.

2.5: Conclusions

We have shown that we can reproducibly fabricate PDMS microwell arrays for use in MIDAS amplification. Furthermore, we can randomly seed cells such that there are consistently less than two cells per well. By adding in SYBR Green I to our amplification master mix, we can fluorescently monitor the amplicons growing in real time using an incubated epifluorescent microscope. Additionally, this allows us to discern any exogenous or cross well contamination, as a large number of well would show fluorescence. We can then reliably extract positive amplicons from the microwells using a micromanipulation system, and generate low input libraries using a transposase-based method.

The resulting data proves superior to traditional, in-tube MDA. While in-tube MDA shows many regions of the genome with little to no coverage, and other regions with very high coverage, the MIDAS derived libraries display a uniform coverage throughout the genome, comparable to a gold standard bulk library. Additionally, we compared MIDAS to previously published data derived from in-tube MDA, microfluidic MDA, and MALBAC based libraries. MIDAS compares favorably to each method, exhibiting very high coverage uniformity across the genome.

2.6: Acknowlegements

We would like to thank Dr. Yuhwa Lo, Randy Chen, and Roger Chiu for helping with the design of the microwell arrays, and for helping with scanning electron microscopy and oxygen plasma treatments of the arrays. We would like to thank Sam Chiang for providing the limited in-tube MDA data. We would like to thank Andrew Richards for assisting with micromanipulation and pipette fabrication. This research was supported by NIH grants R01HG004876, R01GM097253, U01MH098977 and P50HG005550.

Chapter 2, in part, is a reprint of the material as it appears in: Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.* Used with permission.

The dissertation author was the primary investigator and author of this paper.

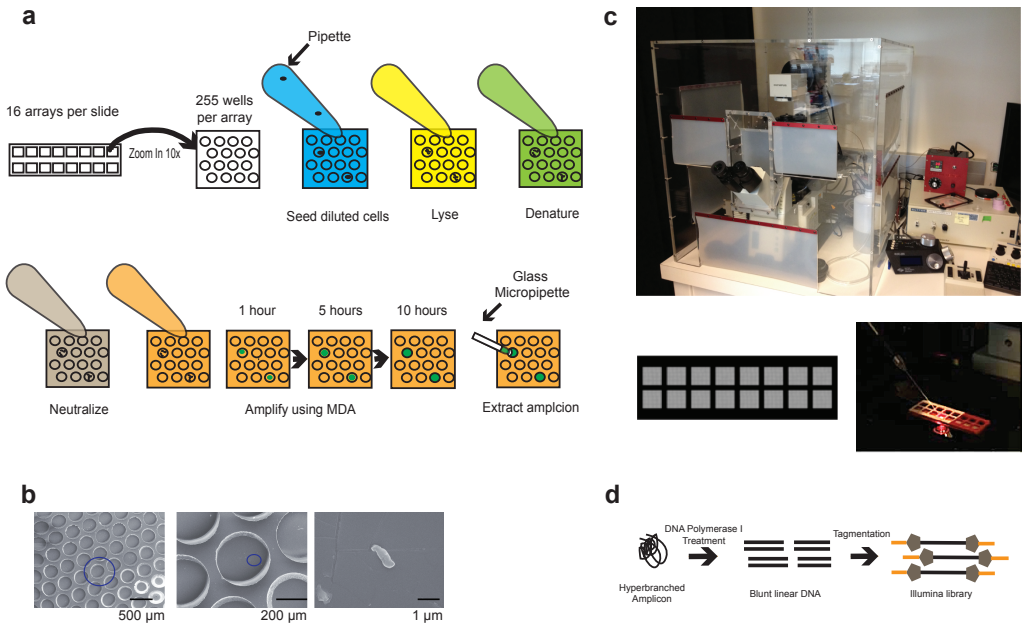

Figure 2.1: Overall Process of MIDAS
Microwell displacement amplification system. **(a)** Each slide contains 16 arrays of 255 microwells each. Cells, lysis solution, denaturing buffer, neutralization buffer and MDA master mix were each added to the microwells with a single pipette pump. Amplicon growth was then visualized with a fluorescent microscope using a real-time MDA system. Microwells showing increasing fluorescence over time were positive amplicons. The amplicons were extracted with fine glass pipettes attached to a micromanipulation system. **(b)** Scanning electron microscopy of a single *E. coli* cell displayed at different magnifications. This particular well contains only one cell, and most wells observed also contained no more than one cell. **(c)** A custom microscope incubation chamber was used for real time MDA. The chamber was temperature and humidity controlled to mitigate evaporation of reagents. Additionally, it prevented contamination during amplicon extraction by self-containing the micromanipulation system. An image of the entire microwell array is also shown, as well as a micropipette probing a well. **(d**) Complex three-dimensional MDA amplicons were reduced to linear DNA using DNA polymerase I and Ampligase. This process substantially improved the complexity of the library during sequencing.

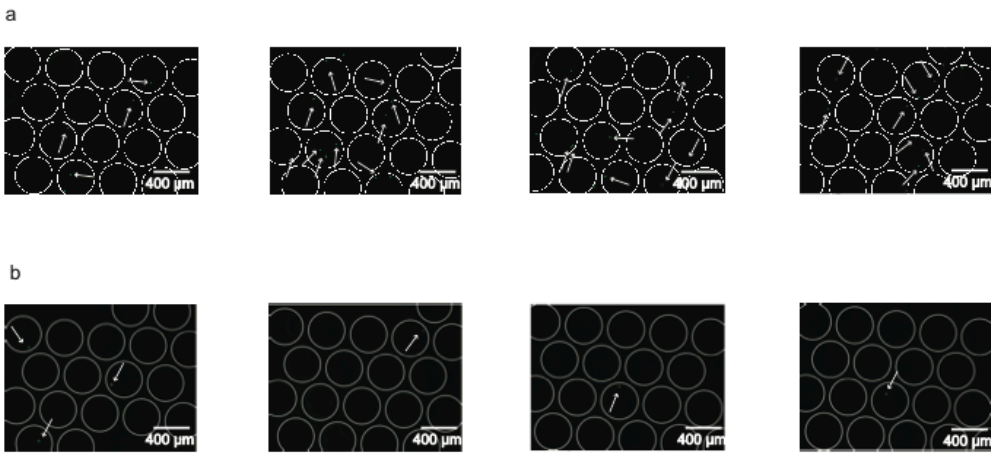

Figure 2.2: Cell seeding in microwells
Cells were stained with SYBR green and visualized under a microscope. Arrows point to single cells (green). Each image is a different position within an array. **(a)** Cells were seeded at 1 cell per well. Most wells contained only 1 cell, with some containing more than 1 cell. **(b)** Cells were seeded at 0.1 cell per well. Most wells contained 0 cells, while a few contained 1 cell. No wells contained more than one cell.

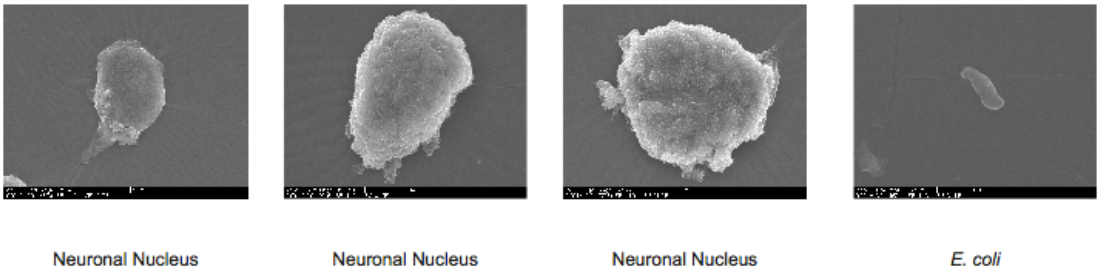

Figure 2.3: SEM images of single cells
SEM images were taken to confirm proper cell seeding. Cells clearly did not stick to each other, further confirming that the Poisson predictions for cell seeding density were accurate.

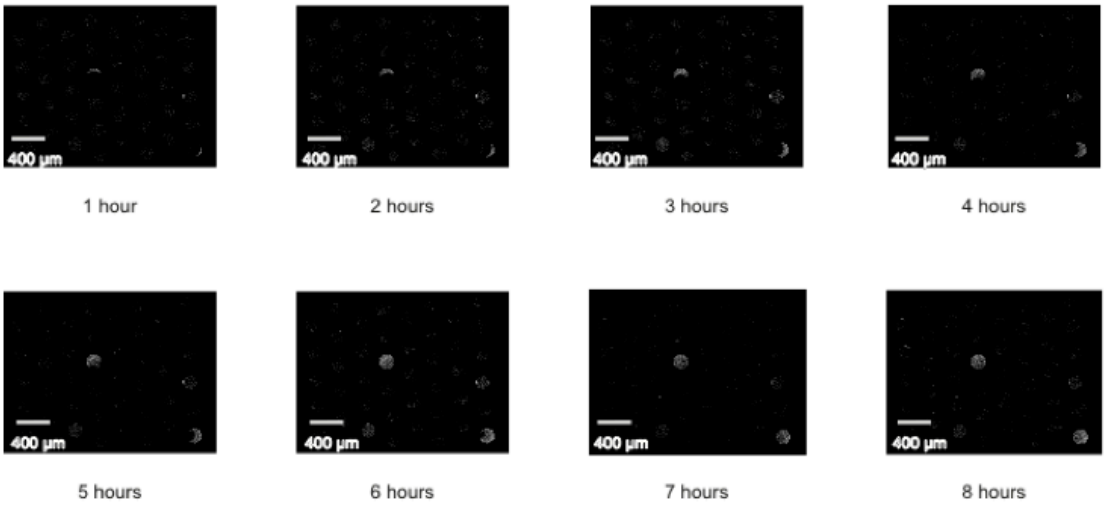

Figure 2.4: Real time MDA
Images were taken every hour using a 488 nm filter. Amplicons were visualized beginning to grow at 1 hour and continuing to grow until they could not amplify due to limited space in the microwell. This saturation point usually occured within 5 to 6 hours. The amplicons appeared to be randomly distributed, further demonstrating random cell seeding, and no amplicons were in abutting wells.

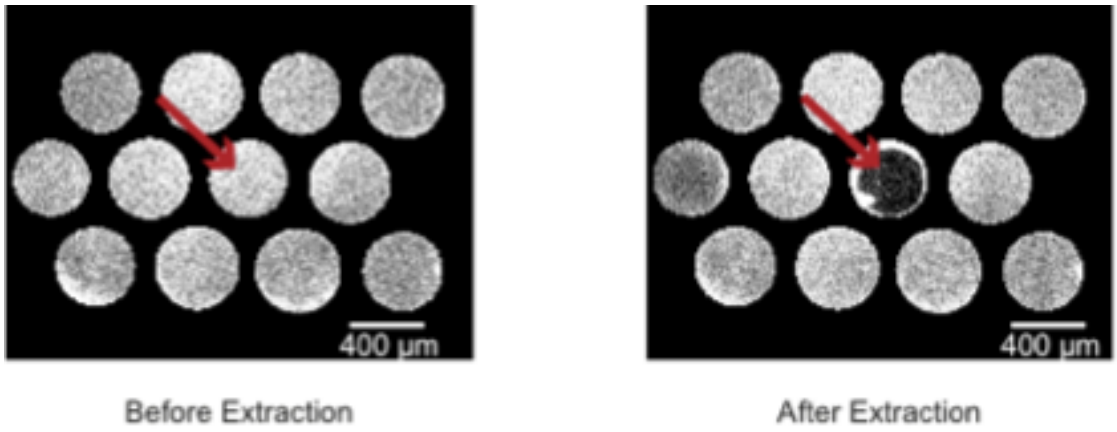

Figure 2.5: Amplicon extraction
Microwells were saturated with genomic DNA and MDA was performed such that every well contained an MDA amplicon. The fluorescence in the left image displays successful amplification. After amplification, a micropipette was lowered into a single well, designated by the arrow, and the amplicon was extracted. The right image shows a successful removal of the amplicon due to loss of fluorescence, without any disturbances in the contents of the nearby microwells.

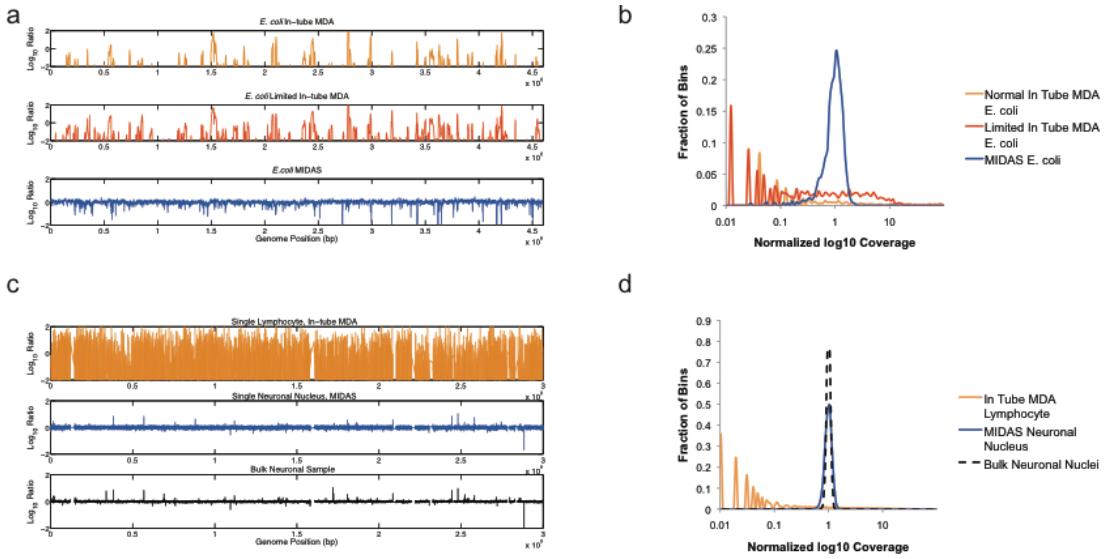

Figure 2.6: Genomic coverage of single bacterial and mammalian cells post MDA and MIDAS

Genomic coverage of single bacterial (a,b) and mammalian (c,d) cells amplified by MDA in a tube and by MIDAS. The observed multi-peak profile for the MDA reactions implies that certain regions may have been amplified with exponentially greater bias compared to the majority of the genome. **(a)** Comparison of single *E. coli* cells amplified in a PCR tube for 10 hours (top), 2 hours (middle) and in a microwell (MIDAS) for 10 hours (bottom). Genomic positions were consolidated into 1 kb bins (x-axis), and were plotted against the $\log_{10}$ ratio (y-axis) of genomic coverage (normalized to the mean). **(b)** Distribution of coverage of amplified single bacterial cells. The x-axis shows the $\log_{10}$ ratio of genomic coverage normalized to the mean. **(c)** Comparison of single human cells amplified using traditional MDA in a PCR tube for 10 hours (top) or in a microwell (MIDAS) for 10 hours (middle) to a pool of unamplified human cells (bottom). Genomic positions were consolidated into variable bins of approximately 60 kb in size previously determined to contain a similar read count[42], and were plotted against the $\log_{10}$ ratio (y-axis) of genomic coverage (normalized to the mean). **(d)** Distribution of coverage of amplified single mammalian cells. The x-axis shows the $\log_{10}$ ratio of genomic coverage normalized to the mean.

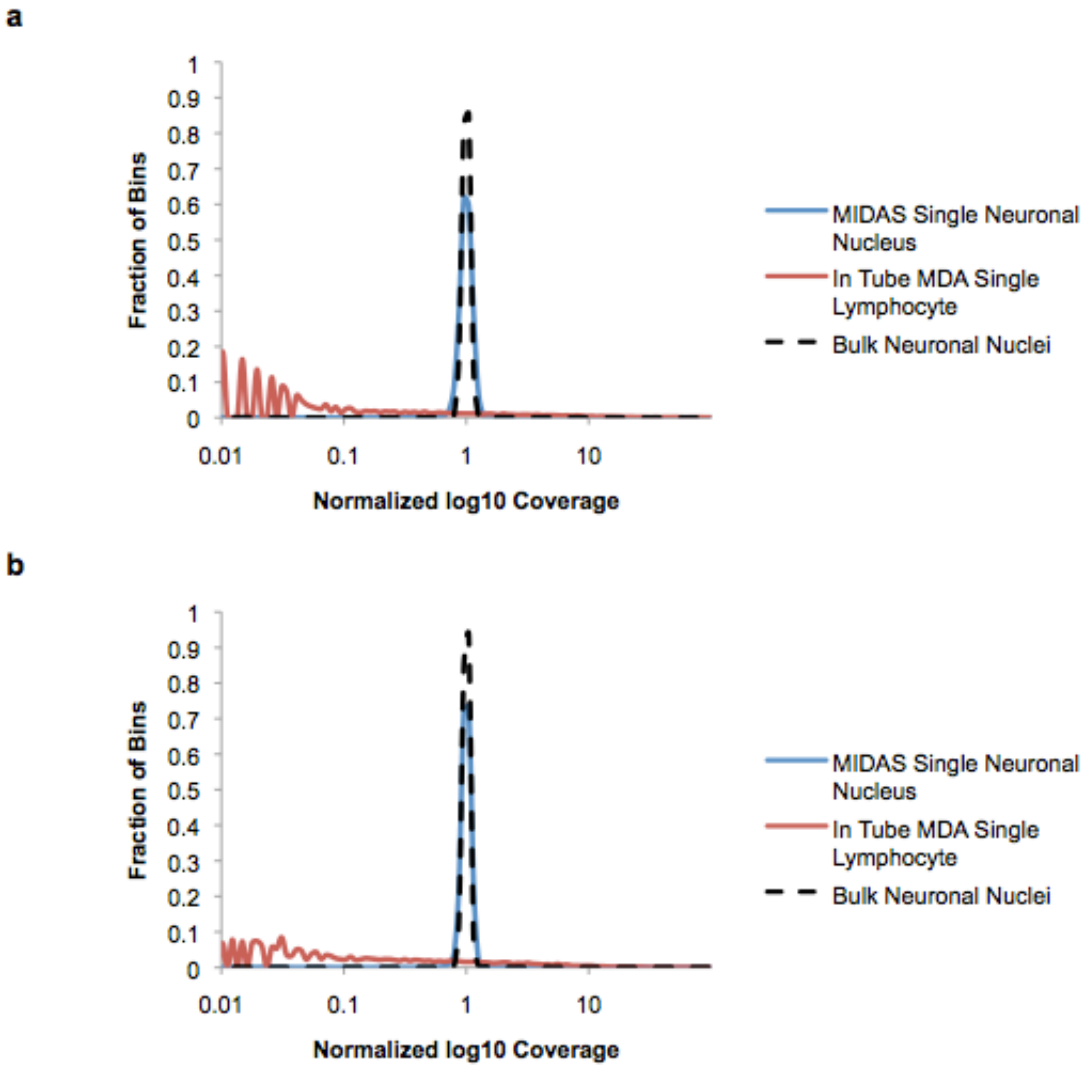

Figure 2.7: Distribution of coverage of amplified single mammalian cells with larger bin size.
The x-axis shows the log10 ratio of genomic coverage normalized to the mean. MIDAS (blue) showed a much tighter coverage distribution than an in-tube MDA library (orange), regardless of bin size. MIDAS even approached the bias level of unamplified genomic DNA from multiple cells when using larger bin sizes of **(a)** ~120 kb, generated by merging pairs of adjacent bins together and **(b)** ~240 kb, generated by merging four adjacent bins together.

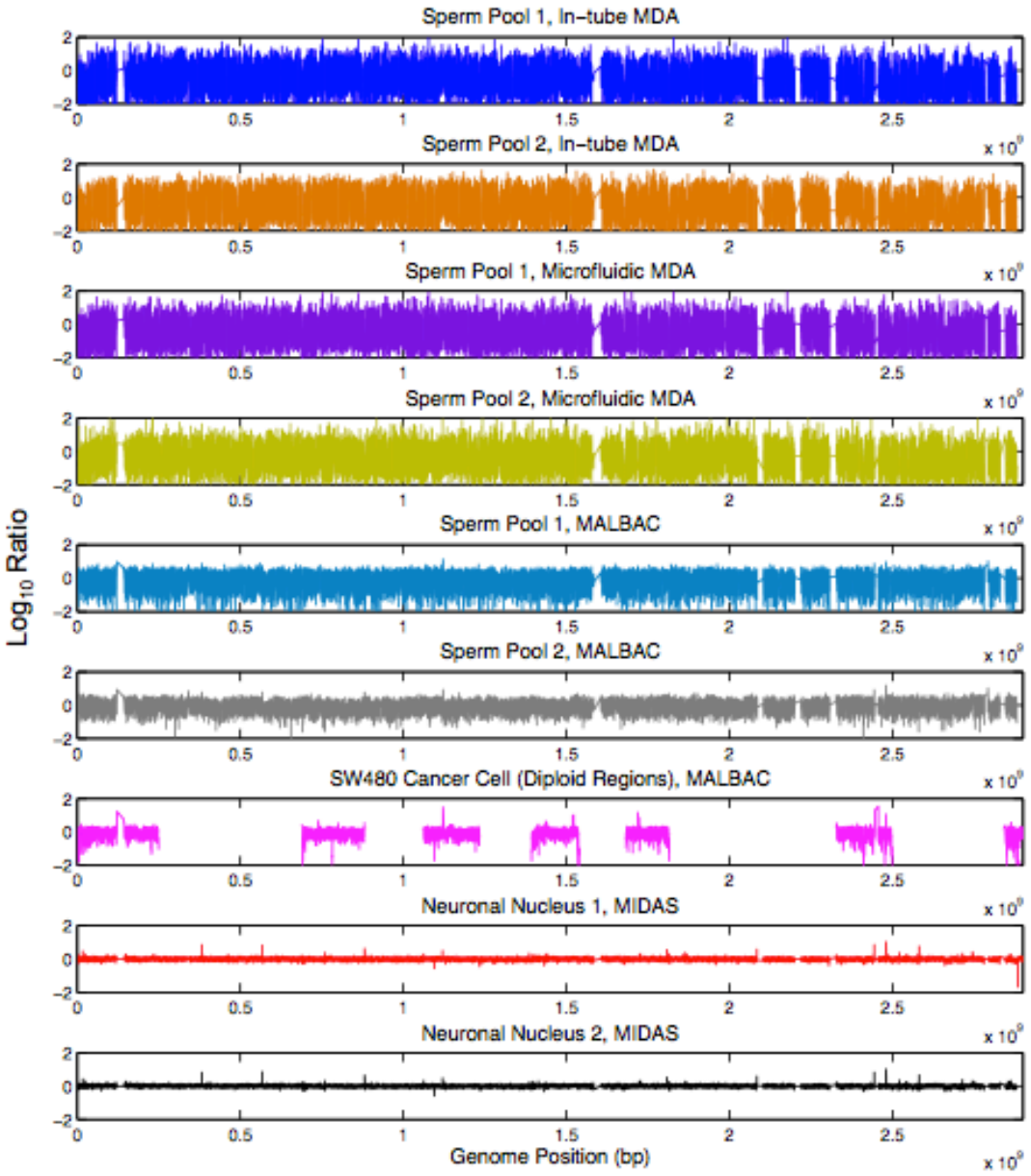

Figure 2.8: Comparison of MIDAS to in-tube MDA, microfluidic MDA, and MALBAC

Comparison of MIDAS to previously published data for in-tube MDA[35], microfluidic MDA[10] and MALBAC[36].for diploid regions of pools of two sperm cells and diploid regions of a single SW480 cancer cell processed using MALBAC[34]. Genomic positions were consolidated into variable bins of approximately 60 kb in size previously determined to contain a similar read count[42], and were plotted against the $\log_{10}$ ratio (y-axis) of genomic coverage (normalized to the mean). For the cancer cell data, non-diploid regions have been masked out (white gaps between pink) to remove the bias generated by comparing a highly aneuploid cell to a primarily diploid cell.

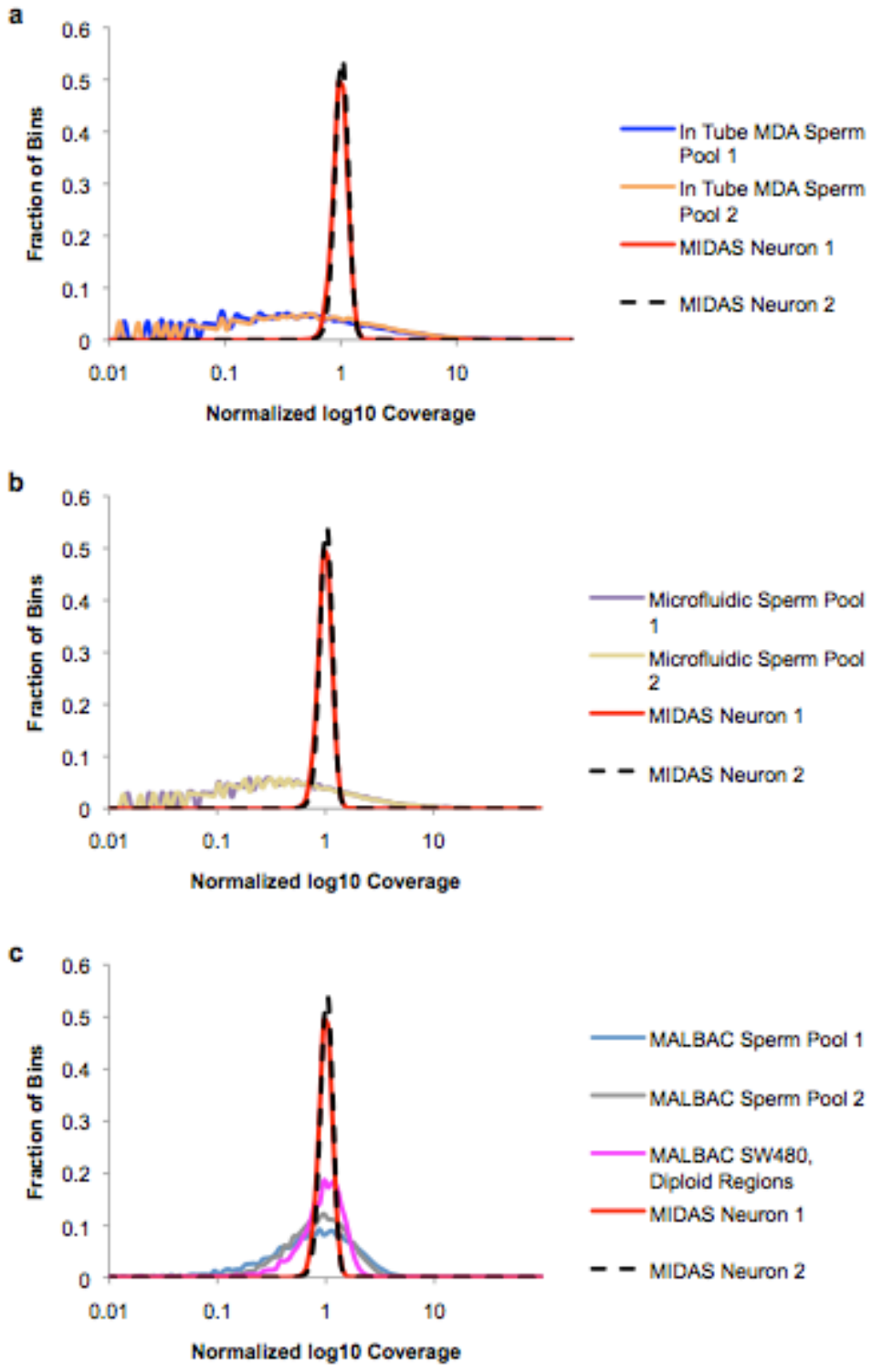

Figure 2.9: Comparison of distributions of coverage
MIDAS neuron libraries compared to **(a)** in-tube MDA, **(b)** microfluidic MDA, and **(c)** MALBAC.

Table 2.1: Cost of library preparation per sample
Costs were based on publicly available reagent prices found online.

Process	MIDAS cost per array (15 libraries)	MDA cost per 15 in-tube reactions
Microwell Fabrication	0.65	N/A
MDA	15.80	237
Amplicon Extraction	2.10	N/A
Library Construction	148	302.40
Bead Purifications	27.60	110.40
Total	$194.15	$649.80

Table 2.2: Cross-well contamination is not present in mixed-sample MIDAS Microwells were loaded with a 1:1 mixture of down syndrome single neurons and healthy single neurons. After whole genome amplification, sequencing libraries were generated from four amplicons. The fraction of reads originating from chromosome 21 in each cell was measured and compared to that obtained during single sample MIDAS; the ploidy of chromosome 21 in each mixed-sample cell was estimated by interpolation. Out of the four tested cells, two were identified as clearly containing trisomy 21 while the other two were clearly diploid; thus, no cross-well contamination occurred.

Sample	% of Reads from Chromosome 21	Known Ploidy	Estimated Ploidy
Healthy Neuron, Cell 1	1.19%	2	
Healthy Neuron, Cell 2	1.20%	2	
Down Syndrome Neuron, Cell 1	1.74%	3	
Down Syndrome Neuron, Cell 2	1.70%	3	
Down Syndrome Neuron, Cell 3	1.63%	3	
Down Syndrome Neuron, Cell 4	1.74%	3	
Mixed-Sample Neuron, Cell 1	1.67%		2.94 → 3
Mixed-Sample Neuron, Cell 2	1.09%		1.79 → 2
Mixed-Sample Neuron, Cell 3	1.23%		2.07 → 2
Mixed-Sample Neuron, Cell 4	1.59%		2.78 → 3

Chapter 3: *De Novo* Assembly of Single *E. Coli* Cell Genomes

3.1: Abstract

We apply the MIDAS method for single cell whole genome amplification to single *E. coli* cells. Following low input library construction and sequencing, we perform single cell *de novo* assembly using the updated SPAdes algorithm, which relies on paired De Brujin graphs. We present data on 3 individual single cell assemblies, in which around 98% of the genome is covered by mapping to the reference *E. coli* genome, and over 90% is correctly assembled. The assembled genomes are as much as 50% greater than any other previously assembled single cells genome. Additionally, we show that increasing the sequencing efforts will most likely results in greater amounts of assembly.

3.2: Introduction

Bacteria remain prevalent in various environments, including sea-water[33], soil[27], and the human body[28]. Unknown genes from rare bacteria continue to be explored as ways to increase productivity in chemical processes and improve human metabolism. Many of these species exist in

heterogeneous environments, symbiotically relying on other microbes to prosper. The codependent relationships make it extremely difficult, if not impossible, to resolve the bacterial origins of these important and interesting genes from standard sequencing.

16S rRNA sequencing has allowed for a quick, inexpensive manner to determine the phyla of various environments[33, 34]. The 16S rRNA gene exists in most bacterial species, and PCR amplification of this region followed by Sanger sequencing can easily determine the phyla residing in different environmental samples. In addition to 16S rRNA sequencing, metagenomic analysis has allowed researchers to delve into vast amounts of sequencing data from environmental microbial samples[33, 34]. By shotgun sequencing an environmental sample, consisting of thousands or more of different bacterial species, without any preprocessing, scientists can explore various genes residing in these specific locations. Genomic differences amongst different phyla allow for grouping sequences to specific phyla, though species are often too similar to be distinguished. Recent algorithms, based on tetranucleotide frequency, allow for the resolution of metagenomic data to a species level for non-complex environments[35]. These algorithms should continue to improve, though it may never be possible to determine individual genomes from very complex metagenomic data.

Single cell sequencing and *de novo* genome assembly allow for environmental microbial samples to be perused at the single species level. Single cells from an environmental sample can be isolated into individual

reactors to remove any exogenous contamination[9, 12]. Whole genome amplification allows for the 5 femtograms of DNA to be amplified such that enough DNA exists for library construction and sequencing. Although this method appears to be a viable alternative to metagenomic analysis, the complexities of single cell whole genome amplification create hurdles for *de novo* assembly. First, the low amount of template allows for a great amount of contamination[9]. Second, the extreme amplification biases generate obstacles in *de novo* assembly, as often certain regions of the genome have little to no coverage following amplification[9, 12]. Third, the hyperbranching mechanism can create chimeric reads[12], resulting in misassembiles. Thus, despite the best previous efforts, only as much as 60% of a single cell genome has been properly assembled[10].

Assembly software has recently been adapted for single cell genomes, and thus greatly improved assembly. The algorithms now account for biases, reducing the minimum base coverage for assembly and allowing for smaller contigs to be scaffolded together. Additionally, the algorithms can properly detect and remove chimeric junctions. One such algorithm, known as SPAdes[41], has greatly improved assembly statistics for single cell genomes. Namely, these are N50, the minimum contig size such that 50% of the total assembled bases are assembled after sorting the contigs based on size, maximum contig length, and total assembled bases. Better assemblies involve greater number in all three of these statistics. Additionally, QUAST[55],

a program developed to help analyze assemblies, allows for easy quality control of individual assemblies by flagging any misassemblies.

In this chapter, we apply MIDAS to single *E. Coli* cells for use in *de novo* assembly. Using the same techniques described in chapter 2, we amplify, create libraries, and sequence 3 single cells. Following data processing, we use SPAdes to assemble the genomes of these cells, and QUAST to analyze the assemblies. The assemblies prove a significant improvement on previously published data, with over 90% of the genome assembled. These same techniques can be applied to environmental samples such that new genomes can be properly assembled with draft quality.

3.3: Methods

3.3.1 Bacterial Preparation

E. coli K12 MG1655 was cultured overnight in LB Broth, collected in log-phase, and washed 3x in PBS. After quantification using a Nanodrop spectrophotometer, the solution was diluted to 10 cells/μL in PBS.

3.3.2 Amplification and Library preparation

The cells were loaded into the microwells and amplified using the MIDAS method described in chapter 2. Extraction and library construction was performed similaray as in chapter 2 as well.

3.3.3 Mapping and *De Novo* Assembly of Bacterial Genomes

Bacterial libraries were size selected into the 300-600 bp range and sequenced in an Illumina MiSeq using 100 bp paired end reads. Approximately 2-8 million reads were sequenced for each cell. *E. coli* data was both mapped to the reference genome and *de novo* assembled. For the mapping analysis, libraries were mapped as single end reads to the reference *E. coli* K12 MG1655 genome using default Bowtie[52] parameters with removal of any reads with multiple matches. Contamination was analyzed by determining the mapping rate to the reference genome. Contamination was further analyzed by selecting approximately 1000 reads, mapping the nt database using BLAST[56], and visualizing the clusters of organisms using MEGAN[57]. For non-contaminated libraries, clonal reads were removed using SAMtools[53,] rmdup function. Chimeras were analyzed by flagging paired reads on the same strand or paired reads with a mismatched orientation. Chimeric junctions were defined as the number of chimeric reads divided by the total number of mapped bases.

For the *de novo* assembly, paired end reads with a combined length less than 200 bp were first joined and treated as single end reads. All remaining paired end reads and newly generated single end reads were then quality trimmed, and any remaining Illumina adapter sequences were removed. *De novo* assembly was performed using SPAdes[41] v. 2.4.0. Corrected reads were assembled with kmer values of 21, 33, and 55. The assembled scaffolds were mapped to the NCBI nt database with BLAST[56], and the organism distribution was visualized using MEGAN[57]. Obvious contaminants (e.g., human) were removed from the assembly and the assembly was analyzed using QUAST[55]. The remaining contigs were annotated using RAST[58] and KAAS[59].

3.3.4 Data Analysis

Mapped and assembled data was further analyzed and visualized in CIRCOS. For the mapped data, mapped reads were imported into CIRCOS for visualization. The assembled scaffolds were also imported into CIRCOS to picture assembled regions of the genome. Last, histograms were made after binning the reads as described in chapter 2 for bacterial libraries, using the average coverage across a given bin, and this was correlated to the assembled and unassembled regions of the genome. Histograms were also made using the average mapped coverage of each scaffold.

3.4: Results

We used MIDAS to sequence three single *E. coli* cells separated by diluting to one cell for every ten wells and seeding into the microwells. Single cells were confirmed with fluorescent microscopy, and a contamination was discounted by observing amplification in distinct wells, ultimately seeing a digital amplification profile. We then generated 2-8 million paired-end reads of 100 base pairs in length for each library using an Illumina MiSeq. The MiSeq was chosen due to its fast turnaround time. As the *E. coli* genome is only approximately 4.6 million bases, many reads were not needed to obtain significant coverage. Thus, the final genomic coverage ranged from 87-364x. Comparatively, previous single cell genome assemblies were performed with coverages much greater than 1,000x[10].

We first mapped the reads to the reference *E. coli* genome to determine the amount of the genome covered. From this data, we found that 98-99% of the genome was recovered at >1x coverage. We then sought to test the limits of minimum sequencing coverage needed to still obtain a significant percentage of the genome. We therefore down-sampled the coverage to much lower depths, and found that even with as little as 10x coverage, we could recover around 90% of the genome for each library (**Figure 3.1)**. Consequently, the amplification proved extremely uniform such that much of the genome was represented in the sequencing data.

Next, we sought to determine the chimeric rate of the sequenced reads. Chimeras occur often during multiple displacement amplification. The single

stranded displaced DNA can anneal to other single strands from completely different regions of the genome. Hence, a false bridge between the regions was created, and this can lead to incorrect assemblies. We also found, based on personal experience, that increasing the ligation time following Polymerase I treatment greatly increased the number of chimeric junctions. Although Ampligase should be limited to nick sealing, disparate fragments were also ligating to each other. Thus, ligation time was limited to a minimum. Previously published data found approximately 1 chimeric junction per every 4-10 kilobases[10]. Since we were performing reactions in small volumes, the rate could be higher since differing fragments would be closer in special location. We found, however, the chimeric rate to be approximately 1 junction per every 5 kilobases, a range consistent with previously published data (**Table 3.1)**.

After we determined the samples were viable for assembly, we desired to assemble the libraries *de novo.* First, many preprocessing steps were necessary to avoid misassemblies and limit computational power necessary. The Nextera method is prone to adding adapter sequences to different areas of the reads, and these are not always removed by the sequencing analysis scripts. Thus, they first were computationally removed. Furthermore, low quality sequences can easily result in false scaffolds, and hence low quality bases were computationally removed as well. Additionally, Nextera library construction creates a very broad range of fragments, which cannot be controlled. Thus, even after size selection, many paired end reads are less

than a combined 200 bases. The resulting paired reads were joined if they were less than a total of 200 bases into a single long read, as these often cause the assembler to crash due to lack of available memory. If any additional single read less than 100 bases remained, it was also removed from further analysis.

After attempting *de novo* assembly with several algorithms, SPAdes was determined to be the most efficient. In brief, both general assemblers, including SOAPdenovo[60], ABySS[61], and Velvet[62], and single cell assemblers, including Velvet-SC and SPAdes, were used to assemble the single cells libraries. The single cell assemblers were found to perform much better than the general assemblers. The single cell assemblers allowed for a sliding, and not fixed, minimum coverage, allowing for more reads to be used in assembly. This implement is extremely important for any biased library. Therefore, Velvet-SC and SPAdes were used for a majority of the assemblies. Both assemblers directly or indirectly require error correction, further removing and correcting low quality bases. Again, this step was necessary for efficient assembly. SPAdes was found to produce superior assemblies than Velvet-SC, as it implements paired De Brujin graphs compared to singular De Brujin graphs. These graphs combined mate pair reads into the actual De Brujin graphs, instead of into post processing steps. Furthermore, SPAdes efficiently combined assemblies from several kmers into unique scaffolds, ultimately increasing the important assembly statistics such as N50 and max contig length.

SPAdes was thus used for assembly for the single cell libraries following preprocessing and error correction. With the same sequencing coverage as mentioned previously (2-8 million reads), we assembled 88-94% of the *E. coli* genome (**Figure 3.2**). We determined an N50 of 2,654-27,882 base pairs, and a max contig length of 18,645-132,037 base pairs (**Table 3.2**). The disparate values directly correlated to the sequencing coverages, as the higher coverage libraries resulted in larger N50s and max contig lengths. These relationships were further investigated. By comparing the assembled contigs to the average read depth throughout the genome, we found that areas with less overall coverage did not assemble as well as areas with greater coverage (**Figure 3.3**). This resulted in very small gaps in the assembly, often less than one kilobase. When the sequencing depth increased, these smaller gaps tended to be assembled, resulting in an order of magnitude greater N50 and max contig length (**Figure 3.2, Table 3.2**). Therefore, we believe that sequencing these genomes to an even higher coverage will help assemble an even larger percentage of the genome.

In comparison to previously published data sets[41], we were able to assemble much larger percentages of the genome. Namely, our assemblies constructed greater than ninety percent of the genome, compared to just over sixty percent. The previous data, however, reported much greater N50's and max contig lengths. We thus decided to investigate this further. We reasoned that the bias during standard MDA resulted in regions with very high coverage, which could be easily assembled, and these regions could be hundreds of

kilobases in length. Furthermore, if a smaller amount of total bases assembled existed, the N50's could be skewed towards larger values. Therefore, an assembly with large contigs in addition to large gaps was observed. In comparison, MIDAS derived assemblies had very small gaps with moderately sized contigs (**Figure 3.4**).

Following assembly, the contigs were analyzed for contamination using BLAST. The resulting mapped organisms were visualized in MEGAN. Even though many different organisms are displayed, an overwhelming majority of the assembled bases, over 80%, map to *E. coli.* Much of the remaining bases map to small regions in known MDA contaminants, such as *Delftia* and *Acidovorax*[63]. Additionally, many very small contigs were unmappable, and most likely were the result of junk DNA. Even though some human contigs were assembled, these could be easily removed from future analysis. Thus, the assemblies proved relatively clean.

Finally, we annotated the genome using the RAST and KAAS annotation servers. Scaffolds mapping strictly to *E. coli* were extracted and uploaded to the individual servers. Following annotation, over 96% of *E. coli* genes were either partially or fully covered in the assembly. Major biosynthetic pathways, including glycolysis and the citric acid cycle, were also present. Furthermore, pathways for amino acid synthesis and tRNA development were covered. MIDAS was thus able to assemble an extremely large portion of the genome from a single cell with comparatively minimal sequencing.

3.5: Conclusions

We have shown that we can assemble genome from single microbial cells at unprecedented levels. Over 90% of the *E. coli* genome was assembled, a 50% improvement over previously published data. Since around 98% of the genome was covered in mapping, we believe that larger sequencing efforts will fill in some of the small gaps in assembly in order to increase the total bases assembled, N50, and max contig sizes. The resulting assemblies were relatively contamination free, and contained chimeric junction rates similar to previously published data. The assemblies contained all major biosynthetic pathways, and well as pathways for amino acid development.

In the scope of this thesis, we only applied MIDAS to single *E. coli* cells. However, MIDAS can be easily implemented with environmental samples. Clean harvesting of cells remains crucial, as cell free contaminant DNA can easily permeate individual reactors. FACS sorting of the cells into clean PBS should help with this issue. Furthermore, cell integrity is a looming issue, as many bacterial cells, especially those from sea-water, lyse easily in solution. Care must be taken to avoid cell lysis during storage. Several strategies, including storing in glycerol, performing MIDAS immediately after cell sorting, and visually checking cellular integrities, must be implemented. Last, lysis varies from cell to cell, but a combination of freeze/thaw and lysozyme treatment should effectively lyse all cell types. As long as a methodical approach is applied to environmental samples, MIDAS should help to

assemble many previously unknown, and potentially important, single cell genomes of rare organisms.

3.6: Acknowledgements

We would like to thank Dr. Yuhwa Lo, Randy Chen, and Roger Chiu for helping with oxygen plasma treatments of the arrays. We would like to thank Sam Chiang for advice on cell culture, maintenance and harvest of environment samples, and assistance with lysis. We would like to thank Andrew Richards for assisting with micromanipulation and pipette fabrication. We would like to thank Chris Wei for help with comparative assembly. We would like to thank Dr. Pavel Pevzner and the SPAdes team for help with assembly and SPAdes troubleshooting. This research was supported by NIH grants R01HG004876, R01GM097253, U01MH098977 and P50HG005550.

Chapter 3, in part, is a reprint of the material as it appears in: Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.* Used with permission. The dissertation author was the primary investigator and author of this paper.

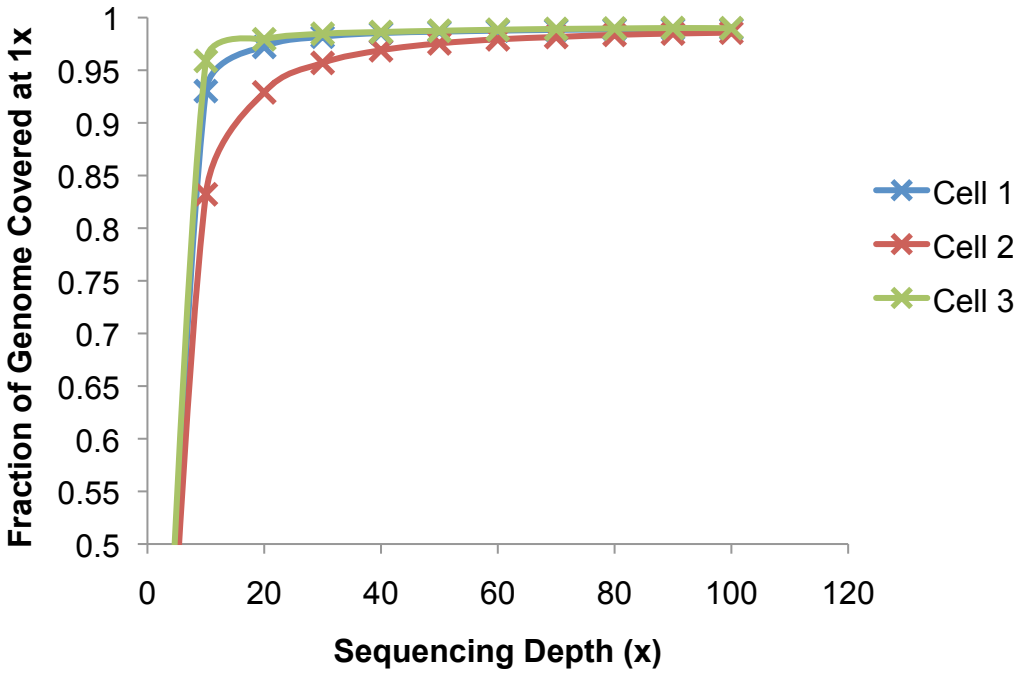

Figure 3.1: Coverage vs depth plots.
Sequencing data was downsampled to various depths for each of the *E. coli* libraries. A significant portion of the genome was covered even at low sequencing depth in each sample.

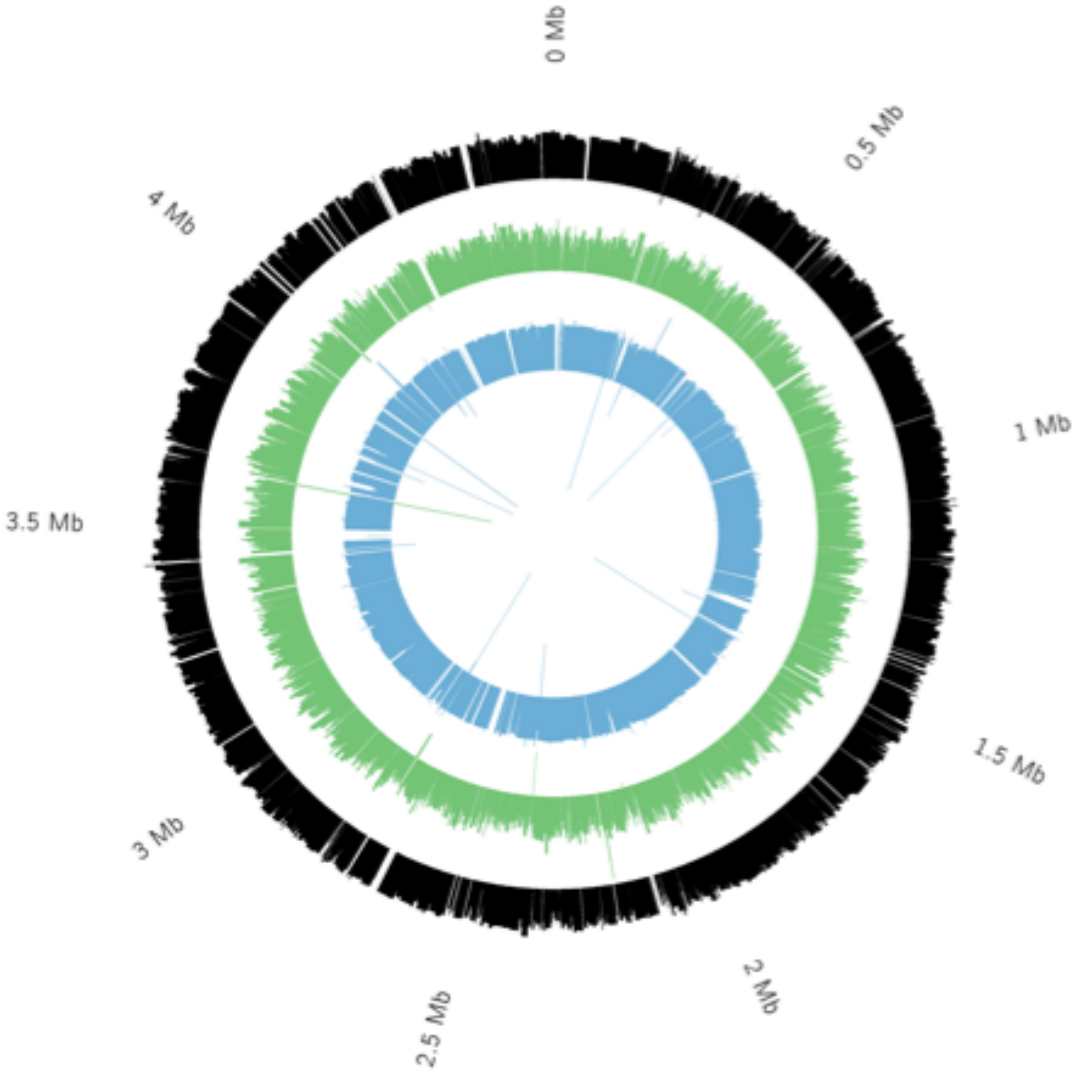

Figure 2: Depth of coverage of assembled contigs aligned to the reference *E. coli* genome

Three single *E. coli* cells were analyzed using MIDAS. Between 88% and 94% of the genome was assembled from 2–8M paired-end 100bp reads. Each colored circle is a histogram of the $\log_2$ of average depth of coverage across each assembled contig for one cell. Gaps are represented by blank whitespace in between colored contigs.

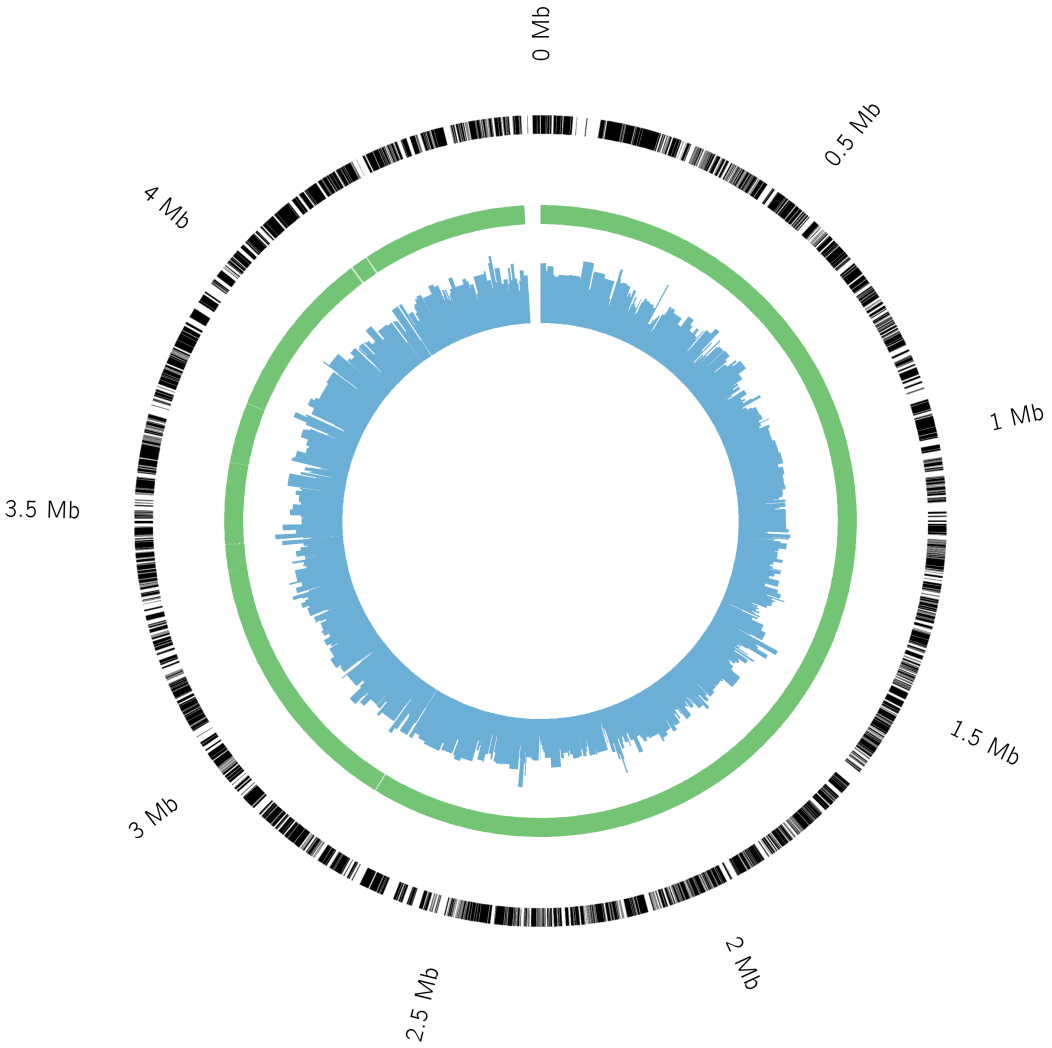

Figure 3.3: Comparison of assembly to mapped reads across genome
The outer track displays the assembled contigs mapping to *E. coli.* The middle track shows the raw reads mapping to *E. coli.* The inner track presents the coverage of the reads. Lower coverage is present in mapped regions where contigs were not assembled, indicating that additional sequencing depth could fill in gaps between contigs

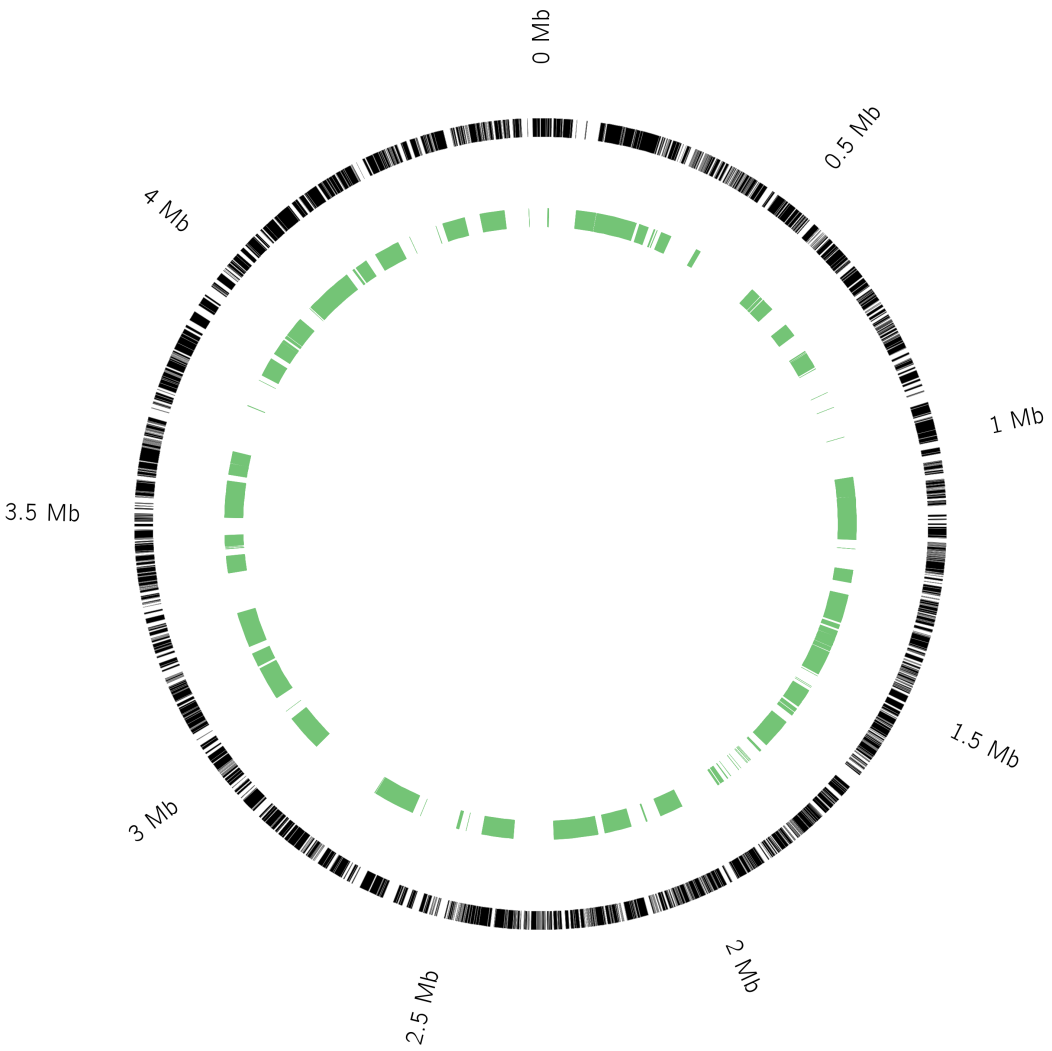

Figure 3.4: Assembly Comparison to an In-tube MDA Derived Library
The outer black track represents the assembled contigs of a MIDAS derived library. Most of the gaps are very small in the assembly. The inner green track represents an in-tube MDA derived library[10]. Although the contigs are large, so are the gaps, indicating a highly biased amplification.

Table 3.1: **Chimera statistics**

Chimeric reads and non-paired reads are reported for each library. Some non-paired reads may have been chimeric, while others may have been contamination. The *prochlorococcus* sample is a previously published in-tube MDA data set, and was used as a control[2]. The chimeric junction rate is defined as number of chimeric reads divided by total number of mapped bases. MIDAS had a similar chimeric junction rate as traditional in-tube MDA.

Sample	**Cell 1**	**Cell 2**	**Cell 3**	***Prochlorococcus***
Correct	1,826,328	1,970,804	9,636,520	12,211,036
Chimeric	71,840	86,508	476,250	665,952
Single	100,307	118,898	773055	484,641
Total Mapped	1,998,475	2,176,210	10,885,825	13,361,629
% Correct	91.39	90.56	88.52	91.39
% Chimeric	3.59	3.98	4.37	4.98
% Single	5.02	5.46	7.1	3.63
Junctions/bp	1 junction/5,238 bp	1 junction/5,383 bp	1 junction/4,477 bp	1 junction/4,301 bp

Table 3.2: Single *E. coli* assembly statistics

Total number of reads, number of contigs mapping to *E. coli*, N_{50}, maximum contig length, total base pairs assembled to *E. coli* K12 MG1655 genome, percent of *E. coli* K12 MG1655 covered in assembly, complete and partial genes covered, and percent of genome covered by mapped reads are reported for each library. Total number of reads refers to all sequencing reads, including non-mapping and clonal reads.

Cell #	Total # of reads	# contigs greater than 500 bp	N_{50} (bp)	Max contig (bp)	Total (bp)	% Genome Covered In Assembly	Complete/ Partial Genes Covered	% Genome Covered by Mapping
1	2,019,892	1,172	6,416	32,552	4,283,777	92.33%	3,308/775	98.91%
2	3,884,950	2,102	2,654	18,465	4,065,096	87.62%	2,313/1,683	98.57%
3	8,482,573	765	27,882	132,037	4,368,254	94.15%	3871/185	98.71%

Chapter 4: Identification of Copy Number Variants in Single Neurons

4.1: Abstract

We applied MIDAS to single neuronal nuclei to investigate copy number variations (CNVs). We began by looking at single neuronal nuclei from patients with Down's Syndrome, which thus contain three copies of chromosome 21. After positively identifying the trisomy, computational spike-ins of small regions from chromosome 21 into diploid chromosomes were performed to test the limits of the CNV calling algorithms. We found an identification rate of 99% with 2 megabase spike-ins, and 70% with 1 megabase spike-ins. No spike-ins at these levels were correctly identified using traditional, in-tube MDA. We next compared the CNVs called on the single neuronal nuclei to those called in a bulk sample, and found that 75% of the CNVs called in the bulk were also called in the single neurons. Additionally, many "somatic" CNVs were called in the individual single cell libraries that were not called in the bulk library.

4.2: Introduction

Human genomes contain two copies of each autosomal chromosomes, one copy from each parent, in addition to two X chromosomes for females or

one X and one Y chromosome for males. Any aberration from the diploid copy number is referred to as a copy number variation, or CNV. CNVs can be as large as an entire chromosome, or as small as a single gene. One of the most well characterized CNVs is three copies of chromosome 21 in people with Down's Syndrome[64]. This CNV causes devastating effects such as severe mental retardation, round faces, and heart issues. Other such large chromosomal CNVs include trisomy 13, trisomy 18, and XXY syndrome[65]. With the exception of CNVs of the sex chromosomes, most other trisomies result in very short life spans. Smaller CNVs can also create severe consequences in humans. A copy number increase of the APP gene, an approximately 300 kb segment in chromosome 21, occurs with high frequency in Alzheimer's disease patients[66].

Until recently, scientists have only studied CNVs on a bulk cell population level. This had led to many technical variations and inconsistencies when calling CNVs using both microarray and sequencing techniques. Many of the CNVs, perhaps existing on smaller populations of cells, were averaged out, and thus many false negative occurred. If scientists could probe into and sequence single cells, CNVs could be called more accurately.

Therefore, researchers began to devise methods for calling CNVs from single cells[7, 22, 42]. Cancer cells were first chosen, as they were known to have different populations of cells originating from a single tumor. Each population of cells would have a distinct genome with distinct CNVs, which could possibly

be distinguished at the single cell level. Furthermore, due to a cancer cell's inability to correct errors, these CNVs would be large in terms of both number of bases and copy number, thus making them easier to accurately call. Scientists at the Cold Spring Harbor National Lab thus dissected tumors into small cell populations[7]. Following FACS sorting into individual tubes, whole genome amplification, and library construction, the genomes were sequenced to determine specific copy number variations. An ingenious algorithm to bin reads, removing all mapping biases, and combine bins with similar numbers of reads was designed to accurately call CNVs of the single cancer cells[42].

More recently, scientists have been interested in accurately calling CNVs in non-cancerous single cells. Most interestingly, FACS sorting has shown DNA content variation in single neuronal nuclei of up to 10% from neurons residing in the same regions of human brains[44-50]. These variations were found to be more prominent in neurons derived from brains of neurologically diseased patients, such as Alzheimer's disease and schizophrenia. These CNVs, however, have yet to be accurately called, mainly arising from the amplification biases in whole genome amplification techniques. Unlike CNVs in cancer cells, these CNVs are thought to be small in terms of number of base pairs and copy. The biases in MDA lead to many false positive and false negative CNV calls, extremely hampering the ability to accurately call CNVs in non-cancerous cells.

In this chapter, we apply MIDAS to single neuronal nuclei to generate unbiased sequencing libraries. We present data on both neurons from healthy

patients and neurons from Down's syndrome patients. Following modifications to the CNV calling script, we show accurately called CNVs at the whole chromosome level. Using a computational spike-in method, we test the limits of the CNV calling algorithm to determine the minimum CNV size that can be accurately called with a single copy number increase. Finally, we compare CNVs called in single cell libraries to CNVs called in bulk libraries, and establish a concordance rate on very small CNVs.

4.3: Methods

4.3.1 Neuron Preparation

Human neuronal nuclei were isolated as previously described[8, 50] and fixed in ice-cold 70% ethanol. Nuclei were labeled with a monoclonal mouse antibody against NeuN (1:100 dilution) (Chemicon, Temecula, CA) and an AlexaFluor 488 goat anti-mouse IgG secondary antibody (1:500 dilution) (Life Technologies, San Diego, CA). Nuclei were counterstained with propidium iodide (50ug/ml) (Sigma, St. Louis, MO) in PBS solution containing 50 µg/ml RNase A (Sigma) and chick erythrocyte nuclei (Biosure, Grass Valley, CA). Nuclei in the G1/G0 cell cycle peak, determined by propidium iodide fluorescence, were electronically gated on a Becton Dickinson FACS-Aria II (BD Biosciences, San Jose, CA) and selectively collected based on NeuN+ immunoreactivity.

4.3.2 Amplification and Library preparation

The cells were loaded into the microwells and amplified using the MIDAS method described in chapter 2. Extraction and library construction was performed similaray as in chapter 2 as well.

4.3.3 Identification of CNVs in MIDAS and MDA data

Mammalian single-cell libraries were sequenced in an Illumina Genome Analyzer IIx or Illumina HiSeq using 36 bp single-end reads. The CNV algorithm previously published by Cold Spring Harbor Laboratories[7] was used to call copy number variation on each single neuron, with modifications to successfully analyze non-cancer cells. Briefly, for each sample, reads were mapped to the genome using default parameters in Bowtie, while removing reads mapping to multiple places in the genome. Clonal reads resulting from Polymerase Chain Reaction artifacts were removed using samtools, and the remaining unique reads were then assigned into 49,891 genomic bins of approximately 60 kb in size that were previously determined such that each would contain a similar number of reads after mapping[42]. Each bin's read count was then expressed as a value relative to the average number of reads per bin in the sample, and then normalized by GC content of each bin using a weighted sum of least squares algorithm (LOWESS). Circular binary segmentation was then used to divide each chromosome's bins into adjacent segments with similar means. Unlike the previously published algorithm, in

which a histogram of bin counts was then plotted and the second peak chosen as representing a copy number of two, it was assumed, due to samples not being cancerous and thus being unlikely to contain significant amounts of aneuploidy, that the mean bin count in each sample would correspond to a copy number of two. Each segment's normalized bin count was thus multiplied by two and rounded to the nearest integer to call copy number. MIDAS data clearly showed a CNV call designating Trisomy 21 in all Down Syndrome single cells, while the traditional MDA-based method was not able to call Trisomy 21.

4.3.4 Identification of Artificial CNVs in MDA and MIDAS data

In order to test the ability of the CNV algorithm described above to call small CNVs, artificial CNVs were computationally constructed. Prior to circular binary segmentation, in each Down Syndrome sample, one hundred random genomic regions across chromosomes 1-22 were chosen, each consisting of either 17 or 34 bins of approximately 60 kb in size, thus corresponding to either 1 megabase or 2 megabases. Each region was replaced with an equivalently sized region from chromosome 21, to represent copy number 3, or chromosome 4, to be used as a control (**Supplementary Table 5**). The above algorithm was then run on each "spiked-in" sample, and the number of new CNV calls in each sample that matched each spike-in was tallied. For the chromosome 21 spike-ins, MIDAS was able to accurately call up to 99% of

spiked-in CNVs at the 2 Mb level and 80% of spiked-in CNVs at the 1 Mb level, while the traditional MDA-based method was not able to call any spiked-in CNVs. As expected, spike-ins of chromosome 4 did not result in any additional CNV calls.

4.3.5 Identification of true CNVs in MIDAS data

The above algorithm was run on the original MIDAS derived single neuronal nucleus libraries in addition to a bulk library created with approximately 4,000 neuronal nuclei. CNVs were directly compared between the bulk and single cell libraries. CNVs less than 1 megabase could not be called with confidence, however, if they overlapped between the bulk and single cell libraries, they were called as CNVs. Any further discrepancies were not clarified, as it is unclear whether these are true somatic CNVs, false positive, or false negative CNV calls. Following CNV calling, the CNVs were analyzed and sorted based on function using the UCSC genome browser.

4.4: Results

We applied MIDAS to the characterization of copy number variation in single mammalian cells. The higher cognitive function of the human brain is supported by a complex network of neurons and glia. It has long been thought that all cells in a human brain share the same genome. Recent evidence suggests that individual neurons could have non-identical genomes owing to

aneuploidy[44-47], active retrotransposons[48, 49] and other DNA content variation[50]. However, the presence of somatic genetic variation in individual neurons has not been conclusively demonstrated at the single-genome scale.

To demonstrate the viability of MIDAS as a tool for investigating copy number variation in single primary human neurons, we prepared nuclei from one post-mortem brain sample from a healthy female donor and a second post-mortem brain sample from a female individual with Down Syndrome. We purified cortical neuronal nuclei by flow sorting based on neuron-specific NeuN antibody staining. We generated six sequencing libraries (two disease-free and four Down Syndrome) from individual nuclei using MIDAS, and analyzed the data using a method based on circular binary segmentation to call copy number variation (CNV)[42] (**Table 4.1**). Raw sequencing reads were divided into 49,891 genomic bins ~60 kb in size, each of which had been previously determined to contain a similar number of sequencing reads in a fully diploid cell[42]. Although clonal read counts arising from PCR duplication appeared relatively high, this is a consequence of the low-input Nextera library construction protocol; because the amplification is limited, the amount of initial molecules is smaller, leading to more duplicates. However, the reduction in bias compensated for the apparent decrease in usable read count. We similarly observed a marked reduction of amplification bias in the MIDAS libraries when compared to the conventional in-tube MDA-based method (**Fig. 2.6c,d**). However, both MIDAS and in-tube MDA had higher levels of sequencing bias and variability than data generated from unamplified genomic

DNA from 4,000 mammalian cells, though the bias in MIDAS was only slightly higher. We desired to determine the minimum bin size such that MIDAS derived single neuronal could approximated bulk libaries. Using a larger bin size of ~240kb (which results in a lower-resolution analysis) allowed MIDAS to match the level of bias from unamplified genomic DNA.

We next sought to characterize the sensitivity of detecting single copy-number changes. It was not possible to distinguish true copy number differences from random amplification bias for the conventional single-cell MDA data, even with aggressive binning into large genomic regions. However, the uniform genome coverage in the MIDAS libraries allowed clear detection of Trisomy 21 in each of the Down Syndrome nuclei, where trisomy could not be detected in any in-tube MDA based library (**Figure 4.1a, b**). The extreme biases, indicated by large spikes followed by regions of little coverage, additionally lead to many small false positive CNV calls.

Rigorous validation of single-cell sequencing methods has been extremely challenging, primarily because any single cell might have genomic differences that are not detectable in the bulk cell population. Hence, there is no reference genome that single-cell data can be compared to. To determine the CNV detection limit of MIDAS, we computationally simulated sequencing data sets containing reference CNV events 1 or 2 Mb in size. We randomly selected 1 or 2 Mbps regions of either chromosome 21 (to simulate the gain of a single copy, the smallest possible copy number change) or chromosome 4 (as a negative control), and computationally transplanted these regions into

100 other random genomic locations (**Table 4.2)**. This computational approach, similar to a strategy previously used for assessing sequencing errors[67], yielded data sets containing reference CNVs at known positions without affecting the inherent technical noise in the data. We identified 99/100 of 2 Mb T21 insertions and 80/100 of 1 Mb T21 insertions in the simulated data set from Down Syndrome Cell 1, indicating that MIDAS is able to call copy number events at the megabase-scale with high sensitivity (**Figure 4.1c, Table 4.2**). As expected, detection levels in the other data sets were similar for libraries with sufficient sequencing depth (80/100 for Down Syndrome Cell 2, 99/100 for Down Syndrome Cell 4), while libraries with insufficient sequencing depth could not be used for accurate small CNV calling (32/100 for Down Syndrome Cell 3). However, this issue can be easily solved by increasing the sequencing depth. As expected, the insertion of diploid chromosome 4 regions did not generate any copy number calls. High-fidelity CNV calling (96%) at the 2 Mb level was retained even when 20% additional random technical noise was applied to the read count results (**Figure 4.2**). Therefore, batch effects in library preparation should cause little changes in accurately calling CNVs. When the same simulation was performed with data from traditional in-tube MDA libraries, no T21 insertions were detected due to the large amounts of biases and variations in the libraries, indicating that at this level of sequencing depth, traditional MDA-based methods are unable to call small CNVs (**Figure 4d**), and thus prove insufficient in for non-cancerous single cell CNV calling.

We next performed CNV calling on each individual neuron using the parameters calibrated by the T21 transplantation simulation. MIDAS called 9–18 copy number events in each neuron (**Table 4.3**). Only 8/60 called CNV events were larger than 2 Mb, and only 13/60 were larger than 1 Mb. It remained unclear whether the remaining events represented true copy number changes or whether they were false positives owing to the small size of most of the calls. It was also unclear which CNV calls represented somatic copy number variation and which represented germline CNV calls that might have been missed in one sample.

To address these issues and further probe the ability of MIDAS to identify germline and *de novo* CNV events, we performed library construction and sequencing on unamplified genomic DNA from two pools of ~4,000 neuronal nuclei from the healthy donor, and compared the results to those obtained from the same donor's single neuronal nuclei (**Table 4.3**). We identified 22 CNV events in the unamplified libraries, of which only two were not shared between the two pools, further confirming the accuracy of the CNV calling algorithm. The discordant CNVs are likely false positive or false negative CNV calls in one sample. However, no CNV events identified in the pools were larger than 1 Mb. This finding is not surprising, as germline CNV events with size greater than 1 Mb do not commonly occur[43]. Although MIDAS does not have sufficient specificity when calling CNVs smaller than 1 Mb, we investigated how many small germline CNVs could be identified in the single cell libraries, and found that 75% were detected. Overall, based on the

T21 computational transplantation results, it appears that the six human neurons contain an average of 2.2 regions each with a somatic gain of one copy at the megabase scale, and that several smaller CNV events might also be present. Following analysis using the UCSC genome browser[68], it was that many of the genes found to have CNVs in the single neuronal nuclei are involved in protease inhibition, vesicle formation, and coagulation **(Table 4.4)**. At this point, it is unclear whether the smaller, sub-megabase, CNVs are true somatic CNVs, or false positive/false negative CNVs. Higher depth sequencing might further improve the resolution of the CNV calling algorithm such that CNVs smaller than 1 megabase can be accurately called.

It should also be noted that a Hidden Markov Model (HMM) was also employed to call CNVs, which found little success. Although previously used to call CNVs on single cancer cell data[22], we found several issues with the HMM method. First, the model required a control cell for comparison. As cancer samples can easily use non-cancerous cells from the sample subject as a control, we did not have that luxury, as only brain cells were available. Secondly, the numbers reported in the transmissions and emission matrices seemed somewhat arbitrary, and could vary greatly amongst samples. We did not have any method to accurately determine these numbers, and by varying the numbers slightly, the results drastically changed. Last, due to the nature of HMMs, a slight increase in read count for a given bin could result in a copy number change. As a 33% change is needed for a copy number increase from 2 to 3, and a 50% change is needed for a copy number decrease from 2

to 1, we found that as little as a 10% change in read count for a given bin resulted in a false copy number increase or decrease.

4.5: Conclusions

We have shown that the minimal biases created in MIDAS derived libraries result in the ability to undoubtedly call CNVs at the whole chromosome level. Comparisons to a standard in-tube MDA derived libraries clearly display vast improvements in CNV calling, since these in-tube MDA libraries cannot even be used to call whole trisomy in chromosome 21. Furthermore, the amplification biases resulted in several small false positive CNV calls, which were minimized in the MIDAS derived libraries.

The computational spike-ins allowed us to determine the specificity of the CNV calling algorithm for both the MIDAS and in-tube derived libraries. Positive CNVs were simulated by swapping bins with copy number 3 into those with copy number 2 in random chromosomal positions. Due to minimal biases and noise throughout the whole genome, the CBS based algorithm was able to accurately call up to 99% of 2 megabase spike-ins, and up to 80% of 1 megabase spike-ins in the MIDAS derived libraries. The diploid spike-in (from chromosome 4) caused no false positive CNV calls. Furthermore, technical noise was computationally added, and with up to 20% technical variability, CNVs could accurately be called. Thus, amplification and library construction could be conceivably performed multiple times with little change in the results.

Although the modified CBS algorithm could accurately call CNVs greater than 1 megabase, most of the CNVs called in the single neuronal libraries were smaller than 1 megabase, and thus could not be called with any certainty. To address this issue, we compared CNVs across the single neuronal libraries to those found in bulk libraries. As much as 75% of the CNVs found in the bulk libraries were also found in the single cell libraries. As previously mentioned, it is unclear whether the remaining 25% are true somatic CNVs or false negatives. Additionally, many novel CNVs were observed in each single cell library, and it is also unclear whether these are somatic CNVs or false positives. Further analysis and sequencing is necessary to accurately and efficiently determine true CNVs that are less than 1 megabase.

4.6: Acknowledgements

We would like to thank Dr. Yuhwa Lo, Randy Chen, and Roger Chiu for helping with oxygen plasma treatments of the arrays. We would like to thank Andrew Richards for assisting with micromanipulation and pipette fabrication. We would like to thank Dr. Jerold Chun, Diane Bushman, and Gwen Kaeser for harvesting neuronal nuclei, FACS sorting the nuclei, providing advice for cell storage, and continually providing us with fresh and diverse neuronal samples. This research was supported by NIH grants R01HG004876, R01GM097253, U01MH098977 and P50HG005550.

Chapter 4, in part, is a reprint of the material as it appears in: Jeff Gole, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, Kun Zhang. "Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells." Accepted for publication in *Nature Biotech.* Used with permission. The dissertation author was the primary investigator and author of this paper.

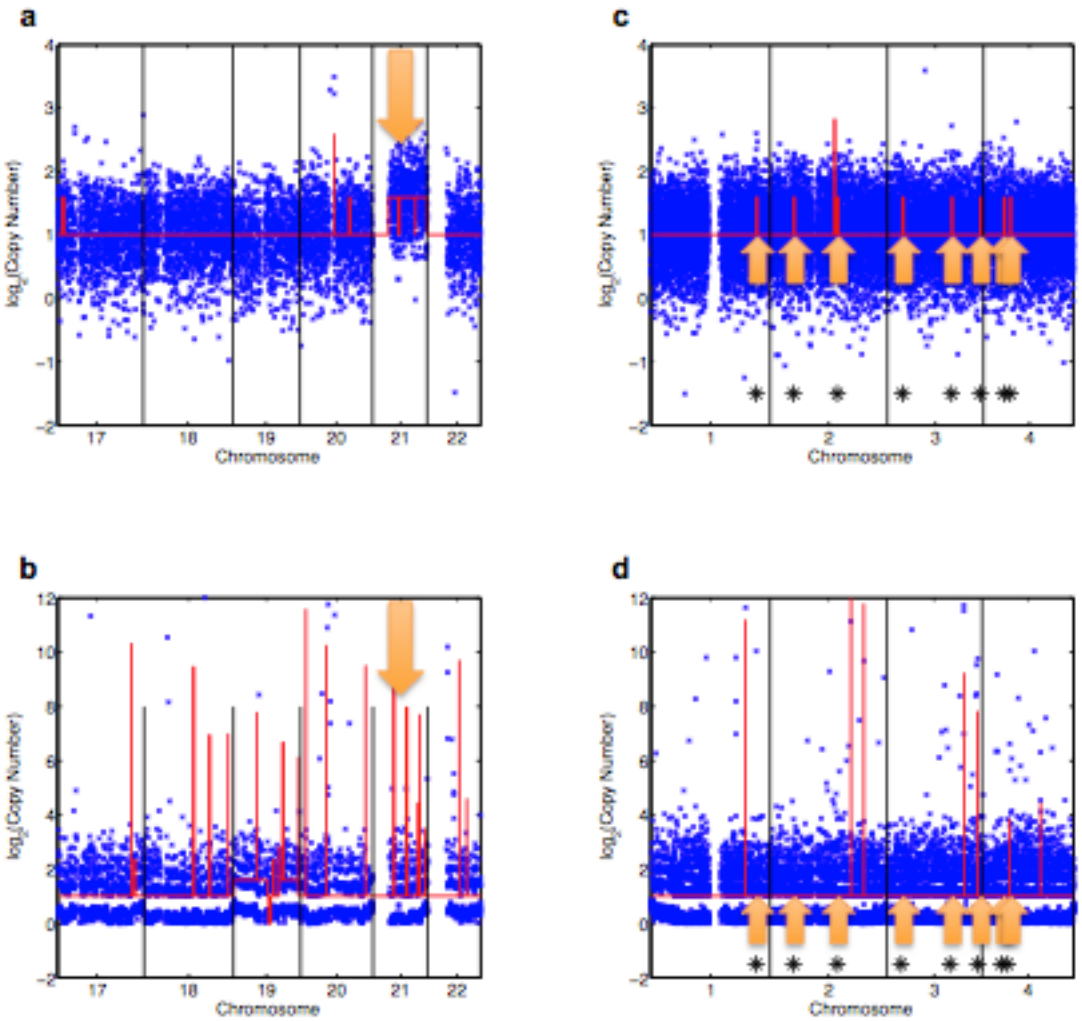

Figure 4.1: Detection of copy number variants using MIDAS and in-tube MDA.
(a) Copy number variation in a Down Syndrome single cell analyzed with MIDAS. The x-axis shows genomic position, while the y-axis shows (on a $\log_2$ scale) the estimated copy number as a red line. **(b)** Copy number variation in a Down Syndrome single cell analyzed with traditional in-tube MDA. The x-axis shows genomic position, while the y-axis shows (in a $\log_2$ scale) the estimated copy number as a red line. **(c)** Copy number variation in a Down Syndrome single cell with Trisomy 21 "spike-ins." The x-axis shows genomic position, while the y-axis shows (in a $\log_2$ scale) the estimated copy number as a red line. At each arrow, prior to CNV calling, data from a randomly determined 2 Mb section of Trisomy chromosome 21 was computationally inserted into the genome, simulating a small gain of single copy event. At each location, a copy number variant was called, showing that MIDAS can detect 2 Mb copy number variation accurately. **(d)** Copy number variation in a Down Syndrome single cell with Trisomy 21 "spike-ins." The x-axis shows genomic position, while the y-axis shows (on a $\log_2$ scale) the estimated copy number as a red line.

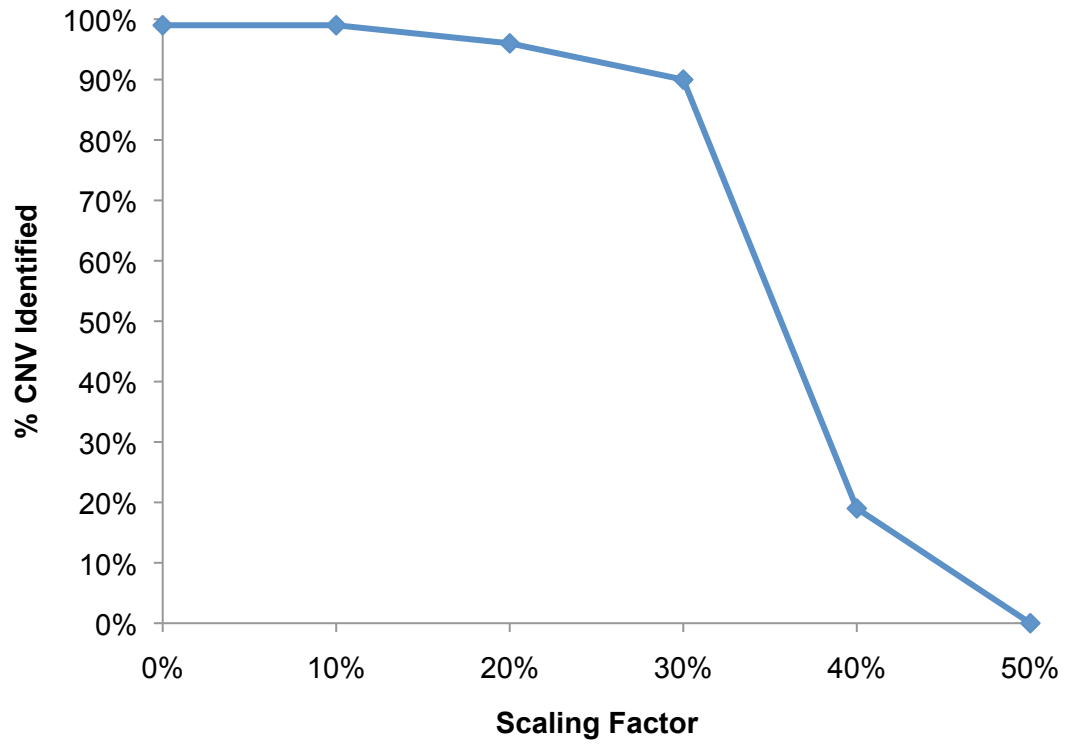

Figure 4.2: MIDAS identifies 2 Mb spike-in CNVs even with 20% additional technical noise

Random noise was generated using a uniform distribution ranging between ±(scaling factor × mean), where the scaling factor was varied. MIDAS was able to tolerate approximately 20% additional random technical noise in terms of read counts and still accurately call 2 Mb spike-in CNVs.

Table 4.1: **Single neuron and unamplified neuron cellular pool sequencing statistics**

Total number of reads, uniquely mapped reads, clonal reads, and usable reads are reported for each neuronal library. Reads mapping to more than one position in the human genome were excluded from analysis.

Cell	Total Number of Reads	# of Uniquely Mapped Reads (%)	# of Clonal Reads (%)	# of Usable Reads
Normal 1	9,624,192	7,013,494 (72.87)	3,122,574 (44.52)	3,890,920
Normal 2	11,975,698	9,121,607 (76.17)	2,770,985 (30.38)	6,350,622
Down Syndrome 1	33,467,888	21,917,127 (65.49)	16,517,878 (75.37)	5,399,249
Down Syndrome 2	31,036,314	19,834,388 (63.91)	17,559,742 (88.53)	2,274,646
Down Syndrome 3	17,251,349	10,740,824 (62.26)	9,406,037 (87.57)	1,334,787
Down Syndrome 4	25,273,979	15,349,190 (60.73)	11,803,624 (76.9)	3,545,566
Normal Pool 1	24,008,671	18,192,924 (75.78)	348,668 (1.92)	17,844,256
Normal Pool 2	18,924,248	14,391,747 (76.05)	240,103 (1.67)	14,151,644

Table 4.2: Artificial CNV transplantation statistics
Each genomic location used for calling of artificial CNVs is shown, along with whether or not MIDAS was able to call the artificial CNV. Only spike-ins of Trisomy Chromosome 21 from MIDAS samples generated CNV calls; spiking in either MIDAS Chromosome 4 or Trisomy Chromosome 21 from the traditional MDA-based method did not result in any artificial CNV calls.

2 Mb Spike-in Region	1 Mb Spike-in Region	chr21 2 Mb Spike-in	chr21 1 Mb Spike-in	2 Mb Spike-in Detected?	1 Mb Spike-in Detected?
chr1:35,953,938-37,889,989	chr1:35,953,938-36,992,975	chr21:15,869,057-17,759,721	chr21:15,869,057-16,841,316	Yes	Yes
chr1:91,042,930-93,048,451	chr1:91,042,930-92,070,940	chr21:35,733,857-37,620,466	chr21:35,733,857-36,687,022	Yes	Yes
chr1:98,284,802-100,167,143	chr1:98,284,802-99,236,989	chr21:31,329,048-33,234,529	chr21:31,329,048-32,284,116	Yes	Yes
chr1:101,720,184-103,622,384	chr1:101,720,184-102,680,542	chr21:15,549,571-17,439,036	chr21:15,549,571-16,523,267	Yes	Yes
chr1:158,948,121-160,904,574	chr1:158,948,121-159,956,586	chr21:43,947,454-45,973,419	chr21:43,947,454-45,032,873	Yes	Yes
chr1:180,612,063-182,538,641	chr1:180,612,063-181,604,263	chr21:18,144,565-20,109,045	chr21:18,144,565-19,193,040	Yes	No
chr1:219,167,316-221,099,504	chr1:219,167,316-220,161,764	chr21:37,382,817-39,338,993	chr21:37,382,817-38,415,585	Yes	No
chr1:241,304,468-243,539,334	chr1:241,304,468-242,305,206	chr21:45,256,736-47,160,835	chr21:45,256,736-46,250,492	Yes	Yes
chr2:47,279,743-49,257,602	chr2:47,279,743-48,315,884	chr21:43,895,354-45,919,088	chr21:43,895,354-44,976,485	Yes	Yes
chr2:51,016,978-52,900,279	chr2:51,016,978-51,977,475	chr21:28,485,490-30,475,416	chr21:28,485,490-29,548,591	Yes	Yes
chr2:120,917,453-122,818,393	chr2:120,917,453-121,860,157	chr21:20,701,039-22,557,692	chr21:20,701,039-21,663,245	Yes	Yes
chr2:139,284,812-141,151,575	chr2:139,284,812-140,248,942	chr21:21,715,029-23,559,945	chr21:21,715,029-22,661,749	Yes	No
chr2:151,537,791-153,484,681	chr2:151,537,791-152,544,463	chr21:19,016,794-20,913,230	chr21:19,016,794-20,003,839	Yes	No
chr2:175,346,199-177,271,180	chr2:175,346,199-176,315,579	chr21:17,384,295-19,361,384	chr21:17,384,295-18,362,470	Yes	No
chr2:204,550,336-206,420,645	chr2:204,550,336-205,511,965	chr21:46,087,520-47,989,191	chr21:46,087,520-47,051,166	Yes	No
chr2:240,935,763-242,873,950	chr2:240,935,763-241,911,231	chr21:45,534,869-47,430,139	chr21:45,534,869-46,517,024	Yes	Yes
chr3:21,457,475-23,388,030	chr3:21,457,475-22,429,886	chr21:17,384,295-19,361,384	chr21:17,384,295-18,362,470	Yes	No
chr3:29,794,211-31,649,290	chr3:29,794,211-30,766,280	chr21:32,609,078-34,563,159	chr21:32,609,078-33,620,107	Yes	Yes
chr3:64,759,471-66,748,898	chr3:64,759,471-65,729,410	chr21:25,797,240-27,686,265	chr21:25,797,240-26,755,429	Yes	Yes
chr3:94,728,396-96,639,499	chr3:94,728,396-95,718,372	chr21:29,548,591-31,432,266	chr21:29,548,591-30,529,736	Yes	Yes
chr3:131,350,124-133,321,647	chr3:131,350,124-132,355,013	chr21:20,380,269-22,240,402	chr21:20,380,269-21,341,117	Yes	Yes
chr3:169,532,039-171,494,322	chr3:169,532,039-170,559,996	chr21:43,312,554-45,314,840	chr21:43,312,554-44,279,997	Yes	Yes
chr3:190,659,728-192,553,633	chr3:190,659,728-191,648,743	chr21:45,699,913-47,592,766	chr21:45,699,913-46,673,526	Yes	Yes
chr4:26,717,043-28,591,712	chr4:26,717,043-27,687,706	chr21:34,563,159-36,473,184	chr21:34,563,159-35,575,904	Yes	Yes

Table 4.2: Artificial CNV transplantation statistics (continued)

2 Mb Spike-in Region	1 Mb Spike-in Region	chr21 2 Mb Spike-in	chr21 1 Mb Spike-in	2 Mb Spike-in Detected?	1 Mb Spike-in Detected?
chr4:41,807,132-43,732,453	chr4:41,807,132-42,806,118	chr21:46,195,641-48,129,895	chr21:46,195,641-47,160,835	Yes	No
chr4:47,152,041-52,814,683	chr4:47,152,041-48,136,122	chr21:41,811,953-43,737,990	chr21:41,811,953-42,776,109	Yes	Yes
chr4:55,036,501-56,991,161	chr4:55,036,501-56,032,590	chr21:38,137,201-40,021,631	chr21:38,137,201-39,127,553	Yes	Yes
chr4:59,922,675-61,851,268	chr4:59,922,675-60,922,058	chr21:41,100,954-43,045,476	chr21:41,100,954-42,076,268	Yes	Yes
chr4:62,174,303-64,098,264	chr4:62,174,303-63,153,820	chr21:39,759,243-41,648,348	chr21:39,759,243-40,718,810	Yes	Yes
chr4:68,752,406-70,866,031	chr4:68,752,406-69,818,131	chr21:20,965,576-22,821,144	chr21:20,965,576-21,926,708	Yes	Yes
chr4:120,492,349-122,402,672	chr4:120,492,349-121,477,609	chr21:26,012,981-27,901,092	chr21:26,012,981-26,970,320	Yes	Yes
chr4:122,895,270-124,846,586	chr4:122,895,270-123,899,166	chr21:31,013,874-32,887,206	chr21:31,013,874-31,960,666	Yes	Yes
chr4:147,655,266-149,580,558	chr4:147,655,266-148,648,700	chr21:33,344,236-35,307,614	chr21:33,344,236-34,350,216	Yes	Yes
chr5:42,837,955-44,872,058	chr5:42,837,955-43,916,598	chr21:25,960,034-27,848,628	chr21:25,960,034-26,918,132	Yes	Yes
chr5:75,108,161-77,082,347	chr5:75,108,161-76,129,009	chr21:23,559,945-25,429,916	chr21:23,559,945-24,529,659	Yes	Yes
chr5:87,692,054-89,588,135	chr5:87,692,054-88,655,990	chr21:16,576,891-18,476,000	chr21:16,576,891-17,549,158	Yes	Yes
chr5:103,446,408-105,320,978	chr5:103,446,408-104,415,213	chr21:37,328,729-39,286,544	chr21:37,328,729-38,356,392	Yes	No
chr5:112,313,258-114,258,965	chr5:112,313,258-113,297,312	chr21:36,161,668-38,084,299	chr21:36,161,668-37,109,297	Yes	Yes
chr5:138,855,793-140,840,449	chr5:138,855,793-139,867,090	chr21:17,264,941-19,246,363	chr21:17,264,941-18,248,286	Yes	No
chr5:143,451,255-145,392,608	chr5:143,451,255-144,448,400	chr21:37,972,811-39,864,943	chr21:37,972,811-38,971,124	Yes	Yes
chr5:151,514,438-153,460,134	chr5:151,514,438-152,525,509	chr21:40,127,624-42,024,470	chr21:40,127,624-41,100,954	Yes	Yes
chr6:42,942,145-44,895,345	chr6:42,942,145-43,948,491	chr21:27,952,868-29,915,446	chr21:27,952,868-28,908,233	Yes	Yes
chr6:46,028,017-47,924,616	chr6:46,028,017-46,996,571	chr21:24,257,335-26,118,999	chr21:24,257,335-25,216,223	Yes	Yes
chr6:55,721,707-58,501,072	chr6:55,721,707-56,691,195	chr21:43,947,454-45,973,419	chr21:43,947,454-45,032,873	Yes	Yes
chr6:79,811,072-81,774,818	chr6:79,811,072-80,802,629	chr21:34,402,142-36,316,410	chr21:34,402,142-35,416,887	Yes	No
chr6:137,141,139-139,051,493	chr6:137,141,139-138,109,997	chr21:25,271,009-27,132,790	chr21:25,271,009-26,223,918	Yes	Yes
chr7:27,905,786-29,810,788	chr7:27,905,786-28,850,665	chr21:45,145,316-47,051,166	chr21:45,145,316-46,142,620	Yes	Yes
chr7:37,524,768-39,414,023	chr7:37,524,768-38,513,924	chr21:41,212,176-43,151,199	chr21:41,212,176-42,183,940	Yes	Yes
chr7:56,531,510-63,813,298	chr7:56,531,510-62,248,786	chr21:29,759,604-31,643,582	chr21:29,759,604-30,743,560	Yes	Yes

Table 4.2: Artificial CNV transplantation statistics (continued)

2 Mb Spike-in Region	1 Mb Spike-in Region	chr21 2 Mb Spike-in	chr21 1 Mb Spike-in	2 Mb Spike-in Detected?	1 Mb Spike-in Detected?
chr8:78,929,030-80,820,600	chr8:78,929,030-79,895,407	chr21:18,248,286-20,218,830	chr21:18,248,286-19,304,425	Yes	No
chr9:1,055,886-2,968,704	chr9:1,055,886-2,053,043	chr21:45,145,316-47,051,166	chr21:45,145,316-46,142,620	Yes	Yes
chr9:26,653,725-28,577,848	chr9:26,653,725-27,652,274	chr21:45,200,702-47,107,423	chr21:45,200,702-46,195,641	Yes	Yes
chr9:78,438,145-80,383,381	chr9:78,438,145-79,427,903	chr21:29,602,048-31,485,060	chr21:29,602,048-30,583,583	Yes	Yes
chr9:85,352,360-87,347,753	chr9:85,352,360-86,341,914	chr21:36,473,184-38,415,585	chr21:36,473,184-37,439,834	Yes	Yes
chr9:108,943,484-110,868,831	chr9:108,943,484-109,924,645	chr21:30,027,091-31,908,111	chr21:30,027,091-31,013,874	Yes	Yes
chr9:136,054,731-138,033,870	chr9:136,054,731-137,134,437	chr21:27,357,146-29,249,735	chr21:27,357,146-28,326,314	Yes	No
chr9:138,851,087-140,884,555	chr9:138,851,087-139,905,741	chr21:41,432,175-43,366,620	chr21:41,432,175-42,401,776	Yes	Yes
chr10:37,235,643-43,386,249	chr10:37,235,643-38,367,612	chr21:29,438,499-31,329,048	chr21:29,438,499-30,420,106	Yes	Yes
chr10:50,517,442-53,244,584	chr10:50,517,442-52,259,071	chr21:41,265,140-43,202,177	chr21:41,265,140-42,236,377	Yes	Yes
chr10:84,733,418-86,669,139	chr10:84,733,418-85,709,427	chr21:30,027,091-31,908,111	chr21:30,027,091-31,013,874	Yes	Yes
chr11:1,155,181-3,103,945	chr11:1,155,181-2,197,444	chr21:32,662,136-34,618,144	chr21:32,662,136-33,675,918	Yes	Yes
chr11:60,102,184-62,149,657	chr11:60,102,184-61,169,073	chr21:33,454,741-35,416,887	chr21:33,454,741-34,454,555	Yes	Yes
chr11:87,270,689-89,838,290	chr11:87,270,689-88,321,483	chr21:27,572,323-29,548,591	chr21:27,572,323-28,537,710	Yes	Yes
chr11:90,997,546-92,889,388	chr11:90,997,546-91,975,942	chr21:38,031,056-39,916,751	chr21:38,031,056-39,023,029	Yes	Yes
chr11:109,670,892-111,573,136	chr11:109,670,892-110,675,346	chr21:32,174,455-34,132,773	chr21:32,174,455-33,174,429	Yes	Yes
chr11:116,094,960-118,005,159	chr11:116,094,960-117,087,728	chr21:21,017,554-22,872,356	chr21:21,017,554-21,979,621	Yes	Yes
chr11:125,026,409-126,921,222	chr11:125,026,409-126,022,156	chr21:28,695,036-30,689,040	chr21:28,695,036-29,759,604	Yes	Yes
chr12:49,227,100-51,283,876	chr12:49,227,100-50,281,689	chr21:45,145,316-47,051,166	chr21:45,145,316-46,142,620	Yes	Yes
chr12:52,063,249-54,036,936	chr12:52,063,249-53,042,314	chr21:44,467,407-46,465,237	chr21:44,467,407-45,534,869	Yes	Yes
chr12:93,762,836-95,714,509	chr12:93,762,836-94,755,100	chr21:22,083,898-23,935,965	chr21:22,083,898-23,028,045	Yes	No
chr12:105,555,128-107,478,702	chr12:105,555,128-106,536,849	chr21:34,509,021-36,421,675	chr21:34,509,021-35,521,899	Yes	Yes
chr12:125,199,005-127,114,720	chr12:125,199,005-126,193,374	chr21:41,265,140-43,202,177	chr21:41,265,140-42,236,377	Yes	Yes
chr13:20,265,706-22,292,811	chr13:20,265,706-21,295,812	chr21:21,286,351-23,135,668	chr21:21,286,351-22,240,402	Yes	Yes
chr13:72,352,465-74,251,208	chr13:72,352,465-73,311,492	chr21:32,067,733-34,023,166	chr21:32,067,733-33,061,883	Yes	No
chr13:111,996,474-114,071,865	chr13:111,996,474-113,130,054	chr21:30,475,416-32,337,385	chr21:30,475,416-31,432,266	Yes	Yes

Table 4.2: Artificial CNV transplantation statistics (continued)

2 Mb Spike-in Region	1 Mb Spike-in Region	chr21 2 Mb Spike-in	chr21 1 Mb Spike-in	2 Mb Spike-in Detected?	1 Mb Spike-in Detected?
chr14:40,383,400-42,297,026	chr14:40,383,400-41,368,453	chr21:16,421,666-18,309,003	chr21:16,421,666-17,384,295	Yes	No
chr14:47,205,765-49,129,974	chr14:47,205,765-48,164,135	chr21:36,842,693-38,809,659	chr21:36,842,693-37,864,928	Yes	No
chr14:49,785,308-51,764,266	chr14:49,785,308-50,827,716	chr21:34,618,144-36,526,570	chr21:34,618,144-35,629,603	Yes	Yes
chr14:54,010,611-55,940,946	chr14:54,010,611-54,992,372	chr21:16,317,320-18,197,321	chr21:16,317,320-17,264,941	Yes	Yes
chr14:58,280,794-60,215,080	chr14:58,280,794-59,282,961	chr21:20,808,528-22,661,749	chr21:20,808,528-21,768,367	Yes	Yes
chr14:65,798,341-67,757,790	chr14:65,798,341-66,798,253	chr21:45,087,470-46,993,898	chr21:45,087,470-46,087,520	Yes	Yes
chr14:69,993,123-71,950,871	chr14:69,993,123-70,999,014	chr21:40,718,810-42,616,819	chr21:40,718,810-41,704,149	Yes	Yes
chr14:79,855,300-81,757,060	chr14:79,855,300-80,828,708	chr21:26,970,320-28,855,428	chr21:26,970,320-27,952,868	Yes	Yes
chr15:34,986,155-36,878,959	chr15:34,986,155-35,976,285	chr21:31,643,582-33,561,518	chr21:31,643,582-32,609,078	Yes	Yes
chr15:79,683,541-81,583,565	chr15:79,683,541-80,679,599	chr21:40,551,918-42,454,621	chr21:40,551,918-41,542,378	Yes	Yes
chr16:47,776,123-49,696,183	chr16:47,776,123-48,790,979	chr21:40,127,624-42,024,470	chr21:40,127,624-41,100,954	Yes	No
chr16:73,420,082-75,485,559	chr16:73,420,082-74,472,527	chr21:23,935,965-25,797,240	chr21:23,935,965-24,896,759	Yes	Yes
chr17:4,089,647-6,142,294	chr17:4,089,647-5,192,301	chr21:36,421,675-38,356,392	chr21:36,421,675-37,382,817	Yes	Yes
chr17:13,450,820-15,379,070	chr17:13,450,820-14,460,546	chr21:46,195,641-48,129,895	chr21:46,195,641-47,160,835	Yes	Yes
chr17:37,373,102-39,387,990	chr17:37,373,102-38,433,119	chr21:35,307,614-37,163,621	chr21:35,307,614-36,264,921	Yes	Yes
chr17:42,946,627-45,535,886	chr17:42,946,627-44,092,717	chr21:39,023,029-40,887,012	chr21:39,023,029-39,970,006	Yes	Yes
chr17:62,034,684-64,166,903	chr17:62,034,684-63,235,801	chr21:23,770,888-25,639,505	chr21:23,770,888-24,739,139	Yes	Yes
chr18:42,248,309-44,139,666	chr18:42,248,309-43,192,657	chr21:36,900,623-38,861,882	chr21:36,900,623-37,919,721	Yes	No
chr19:24,366,238-29,680,778	chr19:24,366,238-28,755,127	chr21:20,965,576-22,821,144	chr21:20,965,576-21,926,708	Yes	Yes
chr20:47,231,329-49,203,437	chr20:47,231,329-48,237,944	chr21:16,630,136-18,541,644	chr21:16,630,136-17,604,225	Yes	Yes
chr22:49,539,479-51,909,988	chr22:49,539,479-50,562,673	chr21:30,361,934-32,231,305	chr21:30,361,934-31,329,048	Yes	Yes

Table 4.3: Copy number events called in each single neuron or pooled sample
All identified copy number events in each sample are listed, along with the size of the CNV in actual base pairs and number of base pairs in the CNV that were non-repetitive according to a previously published algorithm[8]. Unique CNVs are presented in black text, while CNVs shared between one or more samples are presented in red (if a CNV call was partially identified in another sample) or **bold** (if a CNV call was fully identified in another sample). Aside from Trisomy 21 (identified in all three Down Syndrome cells), most CNV calls were fairly small in both size and non-repetitive size. No germline CNVs above 1 Mb in size were detected in the pools of unamplified cells from the healthy donor. Despite this, MIDAS was able to call 75% of these smaller germline CNVs in the single cells. DS refers to Down Syndrome.

Sample	CNV #	Chr.	Start	End	Copy #	Size	Size (Valid Genomic Regions)	CNV Type
Normal 1	**1**	**1**	**16,949,551**	**17,257,431**	**5**	**307,881**	**120,000**	**Germline**
Normal 1	**3**	**1**	**147,802,093**	**149,049,044**	**3**	**1,246,952**	**120,000**	**Germline**
Normal 1	**4**	**2**	**133,000,723**	**133,135,043**	**4**	**134,321**	**120,000**	**Germline**
Normal 1	7	3	75,275,861	76,035,772	3	759,912	420,000	Germline
Normal 1	**11**	**4**	**190,664,845**	**191,154,276**	**4**	**489,432**	**240,000**	**Germline**
Normal 1	**16**	**6**	**32,526,395**	**32,645,736**	**1**	**119,342**	**120,000**	**Germline**
Normal 1	**18**	**8**	**39,308,029**	**39,363,306**	**1**	**55,278**	**60,000**	**Germline**
Normal 1	**24**	**10**	**47,008,316**	**47,538,599**	**4**	**530,284**	**180,000**	**Germline**
Normal 1	**30**	**11**	**48,858,583**	**48,959,202**	**4**	**100,620**	**60,000**	**Germline**
Normal 1	32	11	122,887,817	123,010,937	1	123,121	120,000	Somatic
Normal 1	**37**	**15**	**34,761,777**	**34,873,738**	**1**	**111,962**	**60,000**	**Germline**
Normal 1	38	16	3,762,009	3,818,563	1	56,555	60,000	Somatic
Normal 1	39	16	32,340,630	34,746,226	3	2,405,597	1,140,000	Germline
Normal 1	**40**	**16**	**71,141,287**	**71,246,392**	**7**	**105,106**	**60,000**	**Germline**
Normal 1	**44**	**17**	**21,257,685**	**21,374,155**	**3**	**116,471**	**120,000**	**Germline**
Normal 1	46	17	77,452,319	77,652,085	4	199,767	60,000	Somatic
Normal 1	54	20	29,449,066	29,811,435	4	362,370	120,000	Germline
Normal 2	**1**	**1**	**16,949,551**	**17,257,431**	**4**	**307,881**	**120,000**	**Germline**
Normal 2	2	1	34,347,191	34,666,699	3	319,509	360,000	Somatic
Normal 2	**3**	**1**	**147,802,093**	**149,049,044**	**4**	**1,246,952**	**120,000**	**Germline**
Normal 2	4	2	132,846,449	133,135,043	3	288,595	180,000	Germline
Normal 2	7	3	75,803,231	75,901,346	4	98,116	60,000	Germline
Normal 2	8	3	195,457,070	195,525,025	3	67,956	60,000	Somatic
Normal 2	14	6	0	358,119	3	358,120	180,000	Germline

Table 4.3: Copy number events called in each single neuron or pooled sample (continued)

Sample	CNV #	Chr.	Start	End	Copy #	Size	Size (Valid Genomic Regions)	CNV Type
Normal 2	39	16	34,410,499	34,746,226	3	335,728	360,000	Somatic
Normal 2	**40**	**16**	**71,141,287**	**71,246,392**	**9**	**105,106**	**60,000**	**Germline**
Normal 2	**44**	**17**	**21,257,685**	**21,374,155**	**3**	**116,471**	**120,000**	**Germline**
Normal 2	47	18	59,103,041	59,431,597	3	328,557	360,000	Somatic
Normal 2	53	20	25,753,877	29,868,184	3	4,114,308	420,000	Somatic
Normal 2	55	20	35,971,800	36,129,265	3	157,466	180,000	Somatic
DS 1	**11**	**4**	**190,664,845**	**191,154,276**	**5**	**489,432**	**240,000**	**Germline**
DS 1	**18**	**8**	**39,308,029**	**39,363,306**	**0**	**55,278**	**60,000**	**Germline**
DS 1	19	8	133,002,438	133,213,163	3	210,726	240,000	Somatic
DS 1	20	8	133,849,747	134,259,686	4	409,940	480,000	Somatic
DS 1	**23**	**10**	**38,869,769**	**42,858,972**	**7**	**3,989,204**	**240,000**	**Germline**
DS 1	24	10	47,008,316	50,466,755	3	3,458,440	2,040,000	Germline
DS 1	26	10	69,854,431	70,514,102	1	659,672	660,000	Somatic
DS 1	28	11	42,988,580	43,047,328	1	58,749	60,000	Somatic
DS 1	29	11	47,457,209	48,298,148	1	840,940	840,000	Somatic
DS 1	**40**	**16**	**71,141,287**	**71,246,392**	**12**	**105,106**	**60,000**	**Germline**
DS 1	41	16	75,196,555	75,609,280	1	412,726	420,000	Somatic
DS 1	42	16	82,709,545	82,811,813	1	102,269	120,000	Somatic
DS 1	43	17	10,708,923	12,583,225	3	1,874,303	2,040,000	Somatic
DS 1	**54**	**20**	**29,449,066**	**29,868,184**	**6**	**419,119**	**180,000**	**Germline**
DS 1	56	20	42,392,899	43,933,855	3	1,540,957	1,680,000	Somatic
DS 1	57	21	14,432,540	22,240,402	3	7,807,863	7,800,000	Germline
DS 1	57	21	23,455,005	37,620,466	3	14,165,462	15,600,000	Germline
DS 1	57	21	37,684,932	48,129,895	3	10,444,964	11,340,000	Germline
DS 2	5	2	229,504,954	230,098,069	4	593,116	660,000	Somatic
DS 2	**11**	**4**	**190,664,845**	**191,154,276**	**6**	**489,432**	**240,000**	**Germline**
DS 2	**23**	**10**	**38,869,769**	**42,858,972**	**11**	**3,989,204**	**240,000**	**Germline**
DS 2	25	10	65,820,124	67,011,981	3	1,191,858	1,320,000	Somatic
DS 2	**39**	**16**	**34,002,234**	**34,220,262**	**11**	**218,029**	**60,000**	**Germline**
DS 2	49	19	29,082,056	29,963,452	3	881,397	960,000	Somatic
DS 2	51	19	53,713,097	54,039,720	1	326,624	300,000	Germline
DS 2	54	20	29,449,066	29,927,709	7	478,644	240,000	Germline
DS 2	57	21	14,432,540	17,707,137	3	3,274,598	2,820,000	Germline
DS 2	57	21	17,817,443	17,434,016	3	-383,426	33,600,000	Germline
DS 3	**4**	**2**	**133,000,723**	**133,135,043**	**12**	**134,321**	**120,000**	**Germline**
DS 3	15	6	30,059,760	31,464,153	1	1,404,394	1,440,000	Somatic
DS 3	21	9	26,272,740	26,876,123	1	603,384	660,000	Somatic
DS 3	22	9	108,142,068	108,196,729	0	54,662	60,000	Somatic
DS 3	**23**	**10**	**38,869,769**	**42,858,972**	**11**	**3,989,204**	**240,000**	**Germline**
DS 3	27	10	79,878,358	87,104,049	3	7,225,692	7,680,000	Somatic
DS 3	39	16	34,002,234	34,344,606	7	342,373	180,000	Germline

Table 4.3: Copy number events called in each single neuron or pooled sample (continued)

Sample	CNV #	Chr.	Start	End	Copy #	Size	Size (Valid Genomic Regions)	CNV Type
DS 3	40	16	71,141,287	71,518,003	5	376,717	360,000	Germline
DS 3	52	20	21,609,652	23,522,517	3	1,912,866	2,100,000	Somatic
DS 3	**54**	**20**	**29,449,066**	**29,868,184**	**10**	**419,119**	**180,000**	**Germline**
DS 3	**57**	**21**	**14,432,540**	**48,129,895**	**3**	**33,697,356**	**36,180,000**	**Germline**
DS 3	59	22	27,293,111	30,701,252	3	3,408,142	3,600,000	Somatic
DS 4	6	3	38,811,680	38,866,011	21	54,332	60,000	Somatic
DS 4	9	4	0	145,219	4	145,220	120,000	Somatic
DS 4	**11**	**4**	**190,664,845**	**191,154,276**	**5**	**489,432**	**240,000**	**Germline**
DS 4	12	5	1,048,280	1,262,100	1	213,821	240,000	Somatic
DS 4	13	5	55,307,969	55,487,243	1	179,275	180,000	Somatic
DS 4	18	8	39,251,762	39,363,306	0	111,545	120,000	Germline
DS 4	**23**	**10**	**38,869,769**	**42,858,972**	**8**	**3,989,204**	**240,000**	**Germline**
DS 4	**24**	**10**	**47,008,316**	**47,538,599**	**5**	**530,284**	**180,000**	**Germline**
DS 4	33	12	840,332	1,126,445	1	286,114	300,000	Somatic
DS 4	34	12	2,120,721	4,014,749	3	1,894,029	2,100,000	Somatic
DS 4	36	14	23,787,120	25,915,695	3	2,128,576	2,220,000	Somatic
DS 4	**39**	**16**	**34,002,234**	**34,220,262**	**7**	**218,029**	**60,000**	**Germline**
DS 4	**40**	**16**	**71,141,287**	**71,246,392**	**5**	**105,106**	**60,000**	**Germline**
DS 4	48	19	21,989,226	22,253,600	1	264,375	240,000	Somatic
DS 4	50	19	42,581,327	42,914,555	4	333,229	360,000	Somatic
DS 4	51	19	53,115,972	54,939,873	1	1,823,902	1,680,000	Germline
DS 4	**54**	**20**	**29,449,066**	**29,868,184**	**7**	**419,119**	**180,000**	**Germline**
DS 4	**57**	**21**	**14,432,540**	**48,129,895**	**3**	**33,697,356**	**36,180,000**	**Germline**
Bulk 1	**1**	**1**	**16,949,551**	**17,257,431**	**4**	**307,881**	**120,000**	**Germline**
Bulk 1	**3**	**1**	**147,802,093**	**149,049,044**	**3**	**1,246,952**	**120,000**	**Germline**
Bulk 1	**4**	**2**	**133,000,723**	**133,135,043**	**3**	**134,321**	**120,000**	**Germline**
Bulk 1	7	3	75,803,231	76,035,772	3	232,542	180,000	Germline
Bulk 1	**10**	**4**	**69,452,451**	**69,520,814**	**1**	**68,364**	**60,000**	**Germline**
Bulk 1	**11**	**4**	**190,720,786**	**191,154,276**	**7**	**433,491**	**180,000**	**Germline**
Bulk 1	**14**	**6**	**0**	**410,962**	**3**	**410,963**	**240,000**	**Germline**
Bulk 1	**16**	**6**	**32,526,395**	**32,645,736**	**1**	**119,342**	**120,000**	**Germline**
Bulk 1	17	7	152,002,940	152,125,223	3	122,284	120,000	Somatic
Bulk 1	**18**	**8**	**39,308,029**	**39,363,306**	**1**	**55,278**	**60,000**	**Germline**
Bulk 1	**23**	**10**	**38,869,769**	**42,858,972**	**4**	**3,989,204**	**240,000**	**Germline**
Bulk 1	**24**	**10**	**47,008,316**	**47,538,599**	**4**	**530,284**	**180,000**	**Germline**
Bulk 1	**31**	**11**	**50,720,119**	**51,180,004**	**4**	**459,886**	**60,000**	**Germline**
Bulk 1	**35**	**12**	**38,103,488**	**38,173,312**	**3**	**69,825**	**60,000**	**Germline**
Bulk 1	**37**	**15**	**34,761,777**	**34,873,738**	**1**	**111,962**	**60,000**	**Germline**
Bulk 1	**39**	**16**	**32,499,141**	**34,220,262**	**3**	**1,721,122**	**540,000**	**Germline**
Bulk 1	**40**	**16**	**71,141,287**	**71,246,392**	**7**	**105,106**	**60,000**	**Germline**
Bulk 1	**44**	**17**	**21,257,685**	**21,374,155**	**3**	**116,471**	**120,000**	**Germline**
Bulk 1	**54**	**20**	**29,449,066**	**29,868,184**	**4**	**419,119**	**180,000**	**Germline**

Table 4.3: Copy number events called in each single neuron or pooled sample (continued)

Sample	CNV #	Chr.	Start	End	Copy #	Size	Size (Valid Genomic Regions)	CNV Type
Bulk 1	58	22	0	17,024,456	3	17,024,457	120,000	Germline
Bulk 2	**1**	**1**	**16,949,551**	**17,257,431**	**4**	**307,881**	**120,000**	**Germline**
Bulk 2	**3**	**1**	**147,802,093**	**149,049,044**	**3**	**1,246,952**	**120,000**	**Germline**
Bulk 2	**4**	**2**	**133,000,723**	**133,135,043**	**3**	**134,321**	**120,000**	**Germline**
Bulk 2	**10**	**4**	**69,452,451**	**69,520,814**	**1**	**68,364**	**60,000**	**Germline**
Bulk 2	**11**	**4**	**190,664,845**	**190,720,786**	**6**	**55,942**	**60,000**	**Germline**
Bulk 2	**11**	**4**	**190,720,786**	**191,154,276**	**7**	**433,491**	**180,000**	**Germline**
Bulk 2	**14**	**6**	**0**	**410,962**	**3**	**410,963**	**240,000**	**Germline**
Bulk 2	**16**	**6**	**32,526,395**	**32,645,736**	**1**	**119,342**	**120,000**	**Germline**
Bulk 2	18	8	39,251,762	39,415,337	1	163,576	180,000	Germline
Bulk 2	**23**	**10**	**38,869,769**	**42,858,972**	**4**	**3,989,204**	**240,000**	**Germline**
Bulk 2	**24**	**10**	**47,008,316**	**47,538,599**	**4**	**530,284**	**180,000**	**Germline**
Bulk 2	**30**	**11**	**48,858,583**	**48,959,202**	**4**	**100,620**	**60,000**	**Germline**
Bulk 2	**31**	**11**	**50,720,119**	**51,180,004**	**4**	**459,886**	**60,000**	**Germline**
Bulk 2	**35**	**12**	**38,103,488**	**38,173,312**	**3**	**69,825**	**60,000**	**Germline**
Bulk 2	**37**	**15**	**34,761,777**	**34,873,738**	**1**	**111,962**	**60,000**	**Germline**
Bulk 2	**39**	**16**	**32,499,141**	**34,220,262**	**3**	**1,721,122**	**540,000**	**Germline**
Bulk 2	**40**	**16**	**71,141,287**	**71,246,392**	**7**	**105,106**	**60,000**	**Germline**
Bulk 2	**44**	**17**	**21,257,685**	**21,374,155**	**3**	**116,471**	**120,000**	**Germline**
Bulk 2	45	17	44,204,426	44,350,738	3	146,313	240,000	Somatic
Bulk 2	**54**	**20**	**29,449,066**	**29,868,184**	**4**	**419,119**	**180,000**	**Germline**
Bulk 2	58	22	0	17,289,008	3	17,289,009	240,000	Germline

Table 4.4: List of genes identified to possess somatic copy number changes in single neurons

Each gene was identified as part of a somatic CNV detected by MIDAS in a single neuron; gene expression of each could be affected by the copy number gain. A majority of gains were identified in Down Syndrome Cell 3. Manual annotation revealed several genes involved in protease inhibition, vesicle formation, and coagulation.

Cell	CNV #	Gene	Copy Number
Down Syndrome 1	43	TMEM220-AS1	3
Down Syndrome 1	43	PIRT	3
Down Syndrome 1	43	SHISA6	3
Down Syndrome 1	43	DNAH9	3
Down Syndrome 1	43	ZNF18	3
Down Syndrome 1	43	MAP2K4	3
Down Syndrome 1	43	MIR744	3
Down Syndrome 1	43	LINC00670	3
Down Syndrome 1	43	MYOCD	3
Down Syndrome 1	56	TOX2	3
Down Syndrome 1	56	JPH2	3
Down Syndrome 1	56	OSER1	3
Down Syndrome 1	56	OSER1-AS1	3
Down Syndrome 1	56	GDAP1L1	3
Down Syndrome 1	56	FITM2	3
Down Syndrome 1	56	R3HDML	3
Down Syndrome 1	56	HNF4A	3
Down Syndrome 1	56	MIR3646	3
Down Syndrome 1	56	TTPAL	3
Down Syndrome 1	56	SERINC3	3
Down Syndrome 1	56	PKIG	3
Down Syndrome 1	56	ADA	3
Down Syndrome 1	56	LOC79015	3
Down Syndrome 1	56	WISP2	3
Down Syndrome 1	56	KCNK15	3
Down Syndrome 1	56	RIMS4	3
Down Syndrome 1	56	YWHAB	3
Down Syndrome 1	56	PABPC1L	3
Down Syndrome 1	56	TOMM34	3
Down Syndrome 1	56	STK4-AS1	3
Down Syndrome 1	56	STK4	3
Down Syndrome 1	56	KCNS1	3
Down Syndrome 1	56	WFDC5	3
Down Syndrome 1	56	WFDC12	3
Down Syndrome 1	56	PI3	3
Down Syndrome 1	56	SEMG1	3
Down Syndrome 1	56	SEMG2	3
Down Syndrome 1	56	SLPI	3

Table 4.4: List of genes identified to possess somatic copy number changes in single neurons (continued)

Cell	CNV #	Gene	Copy Number
Down Syndrome 1	56	MATN4	3
Down Syndrome 2	25	ANXA2P3	3
Down Syndrome 3	15	TRIM31	1
Down Syndrome 3	15	TRIM40	1
Down Syndrome 3	15	TRIM10	1
Down Syndrome 3	15	TRIM15	1
Down Syndrome 3	15	TRIM26	1
Down Syndrome 3	15	HCG17	1
Down Syndrome 3	15	HLA-L	1
Down Syndrome 3	15	HCG18	1
Down Syndrome 3	15	TRIM39	1
Down Syndrome 3	15	TRIM39-RPP21	1
Down Syndrome 3	15	RPP21	1
Down Syndrome 3	15	HLA-E	1
Down Syndrome 3	15	GNL1	1
Down Syndrome 3	15	PRR3	1
Down Syndrome 3	15	ABCF1	1
Down Syndrome 3	15	MIR877	1
Down Syndrome 3	15	PPP1R10	1
Down Syndrome 3	15	MRPS18B	1
Down Syndrome 3	15	ATAT1	1
Down Syndrome 3	15	C6orf136	1
Down Syndrome 3	15	DHX16	1
Down Syndrome 3	15	PPP1R18	1
Down Syndrome 3	15	NRM	1
Down Syndrome 3	15	MDC1	1
Down Syndrome 3	15	TUBB	1
Down Syndrome 3	15	FLOT1	1
Down Syndrome 3	15	IER3	1
Down Syndrome 3	15	DDR1	1
Down Syndrome 3	15	MIR4640	1
Down Syndrome 3	15	GTF2H4	1
Down Syndrome 3	15	VARS2	1
Down Syndrome 3	15	SFTA2	1
Down Syndrome 3	15	DPCR1	1
Down Syndrome 3	15	MUC21	1
Down Syndrome 3	15	MUC22	1
Down Syndrome 3	15	HCG22	1
Down Syndrome 3	15	C6orf15	1
Down Syndrome 3	15	PSORS1C1	1
Down Syndrome 3	15	CDSN	1
Down Syndrome 3	15	PSORS1C2	1
Down Syndrome 3	15	CCHCR1	1
Down Syndrome 3	15	TCF19	1

Table 4.4: List of genes identified to possess somatic copy number changes in single neurons (continued)

Cell	CNV #	Gene	Copy Number
Down Syndrome 3	15	POU5F1	1
Down Syndrome 3	15	PSORS1C3	1
Down Syndrome 3	15	HCG27	1
Down Syndrome 3	15	HLA-C	1
Down Syndrome 3	15	HLA-B	1
Down Syndrome 3	15	MICA	1
Down Syndrome 3	15	HCP5	1
Down Syndrome 3	15	HCG26	1
Down Syndrome 3	27	LINC00856	3
Down Syndrome 3	27	LINC00595	3
Down Syndrome 3	27	ZMIZ1-AS1	3
Down Syndrome 3	27	ZMIZ1	3
Down Syndrome 3	27	PPIF	3
Down Syndrome 3	27	ZCCHC24	3
Down Syndrome 3	27	EIF5AL1	3
Down Syndrome 3	27	SFTPA2	3
Down Syndrome 3	27	SFTPA1	3
Down Syndrome 3	27	BEND3P3	3
Down Syndrome 3	27	NUTM2B	3
Down Syndrome 3	27	LOC642361	3
Down Syndrome 3	27	LOC100288974	3
Down Syndrome 3	27	MBL1P	3
Down Syndrome 3	27	SFTPD	3
Down Syndrome 3	27	TMEM254-AS1	3
Down Syndrome 3	27	TMEM254	3
Down Syndrome 3	27	PLAC9	3
Down Syndrome 3	27	ANXA11	3
Down Syndrome 3	27	LINC00857	3
Down Syndrome 3	27	MAT1A	3
Down Syndrome 3	27	DYDC1	3
Down Syndrome 3	27	DYDC2	3
Down Syndrome 3	27	FAM213A	3
Down Syndrome 3	27	TSPAN14	3
Down Syndrome 3	27	SH2D4B	3
Down Syndrome 3	27	NRG3	3
Down Syndrome 3	27	GHITM	3
Down Syndrome 3	27	C10orf99	3
Down Syndrome 3	27	CDHR1	3
Down Syndrome 3	27	LRIT2	3
Down Syndrome 3	27	LRIT1	3
Down Syndrome 3	27	RGR	3
Down Syndrome 3	27	LINC00858	3
Down Syndrome 3	27	CCSER2	3
Down Syndrome 3	52	PAX1	3

Table 4.4: List of genes identified to possess somatic copy number changes in single neurons (continued)

Cell	CNV #	Gene	Copy Number
Down Syndrome 3	52	LOC100270679	3
Down Syndrome 3	52	LOC284788	3
Down Syndrome 3	52	LINC00261	3
Down Syndrome 3	52	FOXA2	3
Down Syndrome 3	52	SSTR4	3
Down Syndrome 3	52	THBD	3
Down Syndrome 3	52	CD93	3
Down Syndrome 3	52	LINC00656	3
Down Syndrome 3	52	NXT1	3
Down Syndrome 3	52	GZF1	3
Down Syndrome 3	52	NAPB	3
Down Syndrome 3	52	CSTL1	3
Down Syndrome 3	52	CST11	3
Down Syndrome 3	52	CST8	3
Down Syndrome 3	52	CST13P	3
Down Syndrome 3	59	MN1	3
Down Syndrome 3	59	PITPNB	3
Down Syndrome 3	59	TTC28-AS1	3
Down Syndrome 3	59	MIR3199-1	3
Down Syndrome 3	59	MIR3199-2	3
Down Syndrome 3	59	TTC28	3
Down Syndrome 3	59	CHEK2	3
Down Syndrome 3	59	HSCB	3
Down Syndrome 3	59	CCDC117	3
Down Syndrome 3	59	XBP1	3
Down Syndrome 3	59	ZNRF3	3
Down Syndrome 3	59	ZNRF3-AS1	3
Down Syndrome 3	59	C22orf31	3
Down Syndrome 3	59	KREMEN1	3
Down Syndrome 3	59	EMID1	3
Down Syndrome 3	59	RHBDD3	3
Down Syndrome 3	59	EWSR1	3
Down Syndrome 3	59	GAS2L1	3
Down Syndrome 3	59	RASL10A	3
Down Syndrome 3	59	AP1B1	3
Down Syndrome 3	59	MIR3653	3
Down Syndrome 3	59	SNORD125	3
Down Syndrome 3	59	RFPL1S	3
Down Syndrome 3	59	RFPL1	3
Down Syndrome 3	59	NEFH	3
Down Syndrome 3	59	THOC5	3
Down Syndrome 3	59	NIPSNAP1	3
Down Syndrome 3	59	NF2	3
Down Syndrome 3	59	CABP7	3

Table 4.4: List of genes identified to possess somatic copy number changes in single neurons (continued)

Cell	CNV #	Gene	Copy Number
Down Syndrome 3	59	ZMAT5	3
Down Syndrome 3	59	UQCR10	3
Down Syndrome 3	59	ASCC2	3
Down Syndrome 3	59	MTMR3	3
Down Syndrome 3	59	HORMAD2	3
Down Syndrome 3	59	LIF	3
Down Syndrome 3	59	OSM	3
Down Syndrome 3	59	GATSL3	3
Down Syndrome 3	59	TBC1D10A	3
Down Syndrome 4	34	CACNA1C	3
Down Syndrome 4	34	CACNA1C-AS4	3
Down Syndrome 4	34	CACNA1C-IT3	3
Down Syndrome 4	34	CACNA1C-AS1	3
Down Syndrome 4	34	LOC283440	3
Down Syndrome 4	34	FKBP4	3
Down Syndrome 4	34	ITFG2	3
Down Syndrome 4	34	NRIP2	3
Down Syndrome 4	34	LOC100507424	3
Down Syndrome 4	34	FOXM1	3
Down Syndrome 4	34	RHNO1	3
Down Syndrome 4	34	TULP3	3
Down Syndrome 4	34	TEAD4	3
Down Syndrome 4	34	TSPAN9	3
Down Syndrome 4	34	PRMT8	3
Down Syndrome 4	34	EFCAB4B	3
Down Syndrome 4	34	PARP11	3
Down Syndrome 4	36	BCL2L2-PABPN1	3
Down Syndrome 4	36	PABPN1	3
Down Syndrome 4	36	SLC22A17	3
Down Syndrome 4	36	EFS	3
Down Syndrome 4	36	IL25	3
Down Syndrome 4	36	CMTM5	3
Down Syndrome 4	36	MYH6	3
Down Syndrome 4	36	MIR208A	3
Down Syndrome 4	36	MYH7	3
Down Syndrome 4	36	MIR208B	3
Down Syndrome 4	36	NGDN	3
Down Syndrome 4	36	THTPA	3
Down Syndrome 4	36	ZFHX2	3
Down Syndrome 4	36	AP1G2	3
Down Syndrome 4	36	JPH4	3
Down Syndrome 4	36	DHRS2	3
Down Syndrome 4	36	DHRS4-AS1	3
Down Syndrome 4	36	DHRS4	3

Table 4.4: List of genes identified to possess somatic copy number changes in single neurons (continued)

Cell	CNV #	Gene	Copy Number
Down Syndrome 4	36	DHRS4L2	3
Down Syndrome 4	36	DHRS4L1	3
Down Syndrome 4	36	LRRC16B	3
Down Syndrome 4	36	CPNE6	3
Down Syndrome 4	36	NRL	3
Down Syndrome 4	36	PCK2	3
Down Syndrome 4	36	DCAF11	3
Down Syndrome 4	36	FITM1	3
Down Syndrome 4	36	PSME1	3
Down Syndrome 4	36	EMC9	3
Down Syndrome 4	36	PSME2	3
Down Syndrome 4	36	RNF31	3
Down Syndrome 4	36	IRF9	3
Down Syndrome 4	36	REC8	3
Down Syndrome 4	36	IPO4	3
Down Syndrome 4	36	TM9SF1	3
Down Syndrome 4	36	TSSK4	3
Down Syndrome 4	36	CHMP4A	3
Down Syndrome 4	36	MDP1	3
Down Syndrome 4	36	NEDD8-MDP1	3
Down Syndrome 4	36	NEDD8	3
Down Syndrome 4	36	GMPR2	3
Down Syndrome 4	36	TINF2	3
Down Syndrome 4	36	TGM1	3
Down Syndrome 4	36	RABGGTA	3
Down Syndrome 4	36	DHRS1	3
Down Syndrome 4	36	NOP9	3
Down Syndrome 4	36	CIDEB	3
Down Syndrome 4	36	LTB4R2	3
Down Syndrome 4	36	LTB4R	3
Down Syndrome 4	36	ADCY4	3
Down Syndrome 4	36	RIPK3	3
Down Syndrome 4	36	NFATC4	3
Down Syndrome 4	36	NYNRIN	3
Down Syndrome 4	36	CBLN3	3
Down Syndrome 4	36	KHNYN	3
Down Syndrome 4	36	SDR39U1	3
Down Syndrome 4	36	CMA1	3
Down Syndrome 4	36	CTSG	3
Down Syndrome 4	36	GZMH	3
Down Syndrome 4	36	GZMB	3
Down Syndrome 4	36	STXBP6	3

Chapter 5: Discussion and Future Directions

5.1: Discussion

Owing to the extreme bias caused by whole-genome amplification from a single DNA molecule, genomic analysis of single cells has remained a challenging task. A large amount of sequencing resources is required to produce a draft-quality genome assembly or determine a low-resolution copy number variation profile owing to amplification bias and coverage dropout. MIDAS addresses this issue through the use of nanoliter-scale spatially confined volumes to generate nanogram-scale amplicons and the use of a low-input transposon-based library construction method. Compared to the conventional single-cell library construction and sequencing protocol, MIDAS provides a more-uniform, higher-coverage approach to analyze single cells from a heterogeneous population.

We applied MIDAS to single *E. coli* cells and resolved nearly the entire genome with relatively low sequencing depth. Additionally, using *de novo* assembly, greater than 90 percent of the genome was assembled with far less sequencing effort than traditional MDA-based methods. These results suggest that applying MIDAS to an uncultivated organism would provide a draft quality assembly. Currently, a majority of unculturable bacteria are analyzed using metagenomics, as part of a mixed population rather than individually.

Metagenomics has only recently allowed for the assembly of genomes frosingle cells, and doing so requires a sample with limited strain heterogeneity[35]. Through the use of MIDAS on heterogeneous environmental samples, novel single-cell organisms and genes can be easily discovered and characterized in a high-throughput manner, allowing a much higher-resolution and more complete analysis of single microbial cells than is possible through previous methods.

We also applied MIDAS to the analysis of copy number variation in single human neuronal nuclei. With less than 0.4x sequencing coverage, we used MIDAS to call single copy number changes of 1–2 million base pairs or larger in size. It has been shown recently that, in human adult brains, post-mitotic neurons in different brain regions exhibit various levels of DNA content variation (DCV)[50]. The exact genomic regions that associate with DNA content variation have been difficult to map to single neurons because of the amplification bias with existing MDA-based methods. CNVs in single tumor cells have been successfully characterized with a PCR-based whole-genome amplification method[7]. However, tumor cells tend to be highly aneuploid and exhibit copy number changes of larger magnitude, which are more easily detected. The applicability of a PCR-based strategy to other primary cell types with more subtle CNV events remains unclear. We have demonstrated that MIDAS greatly reduces the variability of single-cell analysis to a level such that a 1–2 Mb single-copy change is detectable, allowing characterization of much more subtle copy number variation. With additional improvements in

sequencing methods, the use of MIDAS might enable the identification of even smaller CNVs, as currently 75% of smaller germline CNVs below the detection limit of MIDAS are still identifiable. Thirteen somatic gain of single copy events at the megabase level were identified in single neurons, and it appeared that several protease inhibitors, genes involved in vesicle formation, and genes involved in coagulation could be affected. A majority of gene copy changes occurred in one single cell, indicating that gene copy number might greatly vary across individual neurons. MIDAS can be used to simultaneously probe the individual genomes of many cells from patients with neurological diseases, and thus will allow identification of a range of structural genomic variants and eventually allow accurate determination of the influence of somatic CNVs on brain disorders in a high-throughput manner.

Last, we compared MIDAS to other single cell sequencing methods that reduce amplification bias and increase genomic coverage. As described previously, one such method utilizes a microfluidic device to isolate single cells and perform whole genome amplification in a 60nL volume[6]. Another method, MALBAC, incorporates a novel enzymatic strategy to amplify single DNA molecules initially through quasi-linear amplification and reports unprecedented levels of uniformity. MIDAS represents an orthogonal strategy that adapts MDA to a microwell array. We demonstrated that data generated from single neurons amplified with MIDAS compares very favorably to previously published data from combined (and therefore diploid) pools of two single sperm cells amplified using standard in-tube MDA[54], the microfluidic

device[6] and MALBAC[22, 23], as well as a single cancer cell amplified with MALBAC. To ensure a fair comparison, we normalized sequencing depth to an equal amount for each method and processed the raw sequencing data for each sample using an identical computational pipeline. MIDAS generates the lowest levels of bias across the genome.

5.2: Future Directions

Several aspects of MIDAS could be technologically improved to increase success rate and efficiency. First, the current efficiency of amplification is limited to 10%, owing to the use of a low cell loading density to avoid having more than one cell per microwell. This efficiency could be improved 3 to 5 fold by increasing the cell loading density, imaging the microwell arrays containing fluorescently stained cells prior to amplification, and excluding the wells with more than one cell from further analyses. Thus, a much greater percentage of the microwells would result in positive single sell amplifications. Second, amplicon extraction by micromanipulation is currently performed manually at a speed of ~10 amplicons per hour. This number could be improved by at least one order of magnitude by implementing robotic automation. Several automated micromanipulation systems are currently employed by several labs throughout the country[69, 70], and many companies have made these commercially available. Directly relating to the automation is the improvement of visually calling positively amplified wells. The calling

scripts can be further optimized, and then implanted into the automated micromanipulator such that the system can automatically call and extract positive amplicons once the fluorescence reaches a certain threshold. Third, the PDMS microwell arrays used for cell loading are highly customizable but require access to a microfabrication facility for standard oxygen plasma treatments. Routine practice of MIDAS will depend on the commercial availability of hydrophilic microwell arrays. Finally, although each single cell is physically segregated into one microwell, the cells are not in total fluidic isolation. Thus, there may be the potential for cross-contamination between wells, and fluorescent imaging is required throughout amplification to ensure only single cell amplicons are used. Ideally, an adaptation of MIDAS could use only a single initial image for background subtraction, and a final image. This implementation would limit the need for a custom microscope incubation chamber and an automated fluorescent microscope.

In addition to technological improvements, the future of MIDAS relies on its implementation using several sample types. First, as previously described, MIDAS can be executed on environmental microbial samples. The same library preparations and data analysis methods used on single *E. coli* cells can be implemented. We are currently collaborating with a lab to provide us with clean, robust bacterial samples. Again, these samples must be sorted to remove any contaminating DNA, and must be stored carefully to prevent any degradation. With great care, researchers can analyze MIDAS derived

libraries to potentially determine rare bacteria residing in the ocean or human gut, opening up vast opportunities for studying energy and metabolism.

As alluded to earlier, researchers can also use MIDAS to study neurons from complex neurological disorders, from Alzheimer's disease to depression. The unbiased nature of amplification leads to great prospects in determining the subtle copy number variations resulting in the 250 megabase DNA content variations observed amongst single neurons in FACS. A great first step is to study the APP gene, which is found to show a copy number increase in Alzheimer's disease. The large gene of about 300 kilobases has potential to be called in MIDAS derived libraries. Thus, many neurological diseases can be studied in new and exciting manners.

Obviously, scientists can study tumor cells using MIDAS. Although much work has already been done in this field, and the large size and copy number changes do not necessitate MIDAS for accurate CNV calling, the reduced cost of MIDAS can help researchers to study many more cancerous cells. Thus, new relationships between tumor cells can be established.

Finally, MIDAS can be employed on single chromosomes for use in human haplotyping. Many complex diseases involve a multiple of genes interacting with each other[71]. Since every human contains 2 copies of each autosomal chromosome, specific chromosomal locations (i.e., on the maternal or paternal chromosome) of mutations prove important, since both copies of a gene could potentially be inactivated. Thus, scientists can use MIDAS to seed condensed, metaphase chromosomes into the microwells. More than

one chromosome can be in each well as long as it is not the same chromosome. The amplification and library construction procedures would be exactly the same, with the exception of a protease step for histone removal. Since MIDAS produces the most uniform libraries, researchers could potentially construct end-to-end haplotypes, thus assisting in complex disease study.

References

1. Wetterstrand, K., Vol. 2013 (2013).

2. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198-203 (2011).

3. Boland, J.F., Chung, C.C., Roberson, D., Mitchell, J., Zhang, X., Im, K.M., He, J., Chanock, S.J., Yeager, M. & Dean, M. The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Human genetics* **132**, 1153-1163 (2013).

4. Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., Lee, J.H., Loh, Y.H., Manos, P.D., Montserrat, N., Panopoulos, A.D., Ruiz, S., Wilbert, M.L., Yu, J., Kirkness, E.F., Izpisua Belmonte, J.C., Rossi, D.J., Thomson, J.A., Eggan, K., Daley, G.Q., Goldstein, L.S. & Zhang, K. Somatic coding mutations in human induced pluripotent stem cells. *Nature* **471**, 63-67 (2011).

5. Diep, D., Plongthongkum, N., Gore, A., Fung, H.L., Shoemaker, R. & Zhang, K. Library-free methylation sequencing with bisulfite padlock probes. *Nature methods* **9**, 270-272 (2012).

6. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* **150**, 402-412 (2012).

7. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J. & Wigler, M. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).

8. Westra, J.W., Peterson, S.E., Yung, Y.C., Mutoh, T., Barral, S. & Chun, J. Aneuploid mosaicism in the developing and adult cerebellar cortex. *The Journal of comparative neurology* **507**, 1944-1951 (2008).

9. Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W. & Church, G.M. Sequencing genomes from single cells by polymerase cloning. *Nature biotechnology* **24**, 680-686 (2006).

10. Chitsaz, H., Yee-Greenbaum, J.L., Tesler, G., Lombardo, M.J., Dupont, C.L., Badger, J.H., Novotny, M., Rusch, D.B., Fraser, L.J., Gormley, N.A., Schulz-Trieglaff, O., Smith, G.P., Evers, D.J., Pevzner, P.A. & Lasken, R.S. Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature biotechnology* **29**, 915-921 (2011).

11. Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K. & Song, Q. Direct determination of molecular haplotypes by chromosome microdissection. *Nature methods* **7**, 299-301 (2010).

12. Rodrigue, S., Malmstrom, R.R., Berlin, A.M., Birren, B.W., Henn, M.R. & Chisholm, S.W. Whole genome amplification and de novo assembly of single bacterial cells. *PloS one* **4**, e6864 (2009).

13. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. *Nature biotechnology* **29**, 51-57 (2011).

14. Marcy, Y., Ishoey, T., Lasken, R.S., Stockwell, T.B., Walenz, B.P., Halpern, A.L., Beeson, K.Y., Goldberg, S.M. & Quake, S.R. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS genetics* **3**, 1702-1708 (2007).

15. Marcy, Y., Ouverney, C., Bik, E.M., Losekann, T., Ivanova, N., Martin, H.G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D.A. & Quake, S.R. 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**, 11889-11894 (2007).

16. Blainey, P.C. & Quake, S.R. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic acids research* **39**, e19 (2011).

17. Diguistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S.K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., Mardis, E., Marra, M.A., Hamelin, R.C., Bohlmann, J., Breuil, C. & Jones, S.J. De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome biology* **10**, R94 (2009).

18. Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N.,

Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X. & Wang, J. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873-885 (2012).

19. Pan, X., Urban, A.E., Palejev, D., Schulz, V., Grubert, F., Hu, Y., Snyder, M. & Weissman, S.M. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 15499-15504 (2008).

20. Lasken, R.S. Single-cell genomic sequencing using Multiple Displacement Amplification. *Current opinion in microbiology* **10**, 510-516 (2007).

21. Hutchison, C.A., 3rd, Smith, H.O., Pfannkoch, C. & Venter, J.C. Cell-free cloning using phi29 DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 17332-17336 (2005).

22. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622-1626 (2012).

23. Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A.R., Zhu, P., Hu, X., Xu, L., Yan, L., Bai, F., Qiao, J., Tang, F., Li, R. & Xie, X.S. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627-1630 (2012).

24. Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X. & Shendure, J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119 (2010).

25. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome research* **22**, 1139-1143 (2012).

26. Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S. & Bhattacharya, D. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714-717 (2011).

27. Kvist, T., Ahring, B.K., Lasken, R.S. & Westermann, P. Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Applied microbiology and biotechnology* **74**, 926-935 (2007).

28. Fitzsimons, M.S., Novotny, M., Lo, C.C., Dichosa, A.E., Yee-Greenbaum, J.L., Snook, J.P., Gu, W., Chertkov, O., Davenport, K.W., McMurry, K., Reitenga, K.G., Daughton, A.R., He, J., Johnson, S.L., Gleasner, C.D., Wills, P.L., Parson-Quintana, B., Chain, P.S., Detter, J.C., Lasken, R.S. & Han, C.S. Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome research* **23**, 878-888 (2013).

29. DeLong, E.F. & Pace, N.R. Environmental diversity of bacteria and archaea. *Systematic biology* **50**, 470-478 (2001).

30. Riesenfeld, C.S., Schloss, P.D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annual review of genetics* **38**, 525-552 (2004).

31. Torsvik, V., Ovreas, L. & Thingstad, T.F. Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science* **296**, 1064-1066 (2002).

32. Whitman, W.B., Coleman, D.C. & Wiebe, W.J. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6578-6583 (1998).

33. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. & Smith, H.O. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74 (2004).

34. Bik, E.M., Eckburg, P.B., Gill, S.R., Nelson, K.E., Purdom, E.A., Francois, F., Perez-Perez, G., Blaser, M.J. & Relman, D.A. Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 732-737 (2006).

35. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W. & Nielsen, P.H. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* **31**, 533-538 (2013).

36. Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D. & Gordon, J.I. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11070-11075 (2005).

37. Ley, R.E., Turnbaugh, P.J., Klein, S. & Gordon, J.I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022-1023 (2006).

38. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R. & Gordon, J.I. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027-1031 (2006).

39. Rawls, J.F., Mahowald, M.A., Ley, R.E. & Gordon, J.I. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**, 423-433 (2006).

40. Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R., Muehlbauer, M.J., Ilkayeva, O., Semenkovich, C.F., Funai, K., Hayashi, D.K., Lyle, B.J., Martini, M.C., Ursell, L.K., Clemente, J.C., Van Treuren, W., Walters, W.A., Knight, R., Newgard, C.B., Heath, A.C. & Gordon, J.I. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).

41. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. 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**, 455-477 (2012).

42. Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B., Wigler, M., Navin, N. & Hicks, J. Genome-wide copy number analysis of single cells. *Nature protocols* **7**, 1024-1041 (2012).

43. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. & McVean, G.A. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

44. Rehen, S.K., McConnell, M.J., Kaushal, D., Kingsbury, M.A., Yang, A.H. & Chun, J. Chromosomal variation in neurons of the developing

and adult mammalian nervous system. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 13361-13366 (2001).

45. Rehen, S.K., Yung, Y.C., McCreight, M.P., Kaushal, D., Yang, A.H., Almeida, B.S., Kingsbury, M.A., Cabral, K.M., McConnell, M.J., Anliker, B., Fontanoz, M. & Chun, J. Constitutional aneuploidy in the normal human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **25**, 2176-2180 (2005).

46. Yang, A.H., Kaushal, D., Rehen, S.K., Kriedt, K., Kingsbury, M.A., McConnell, M.J. & Chun, J. Chromosome segregation defects contribute to aneuploidy in normal neural progenitor cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **23**, 10454-10462 (2003).

47. Yurov, Y.B., Iourov, I.Y., Vorsanova, S.G., Liehr, T., Kolotii, A.D., Kutsev, S.I., Pellestor, F., Beresheva, A.K., Demidova, I.A., Kravets, V.S., Monakhov, V.V. & Soloviev, I.V. Aneuploidy and confined chromosomal mosaicism in the developing human brain. *PloS one* **2**, e558 (2007).

48. Muotri, A.R. & Gage, F.H. Generation of neuronal variability and complexity. *Nature* **441**, 1087-1093 (2006).

49. Singer, T., McConnell, M.J., Marchetto, M.C., Coufal, N.G. & Gage, F.H. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in neurosciences* **33**, 345-354 (2010).

50. Westra, J.W., Rivera, R.R., Bushman, D.M., Yung, Y.C., Peterson, S.E., Barral, S. & Chun, J. Neuronal DNA content variation (DCV) with regional and individual differences in the human brain. *The Journal of comparative neurology* **518**, 3981-4000 (2010).

51. Inoue, J., Shigemori, Y. & Mikawa, T. Improvements of rolling circle amplification (RCA) efficiency and accuracy using Thermus thermophilus SSB mutant protein. *Nucleic acids research* **34**, e69 (2006).

52. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).

53. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing,

S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

54. Kirkness, E.F., Grindberg, R.V., Yee-Greenbaum, J., Marshall, C.R., Scherer, S.W., Lasken, R.S. & Venter, J.C. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome research* **23**, 826-832 (2013).

55. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).

56. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).

57. Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. MEGAN analysis of metagenomic data. *Genome research* **17**, 377-386 (2007).

58. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. The RAST Server: rapid annotations using subsystems technology. *BMC genomics* **9**, 75 (2008).

59. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**, W182-185 (2007).

60. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.W. & Wang, J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).

61. Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J. & Birol, I. ABySS: a parallel assembler for short read sequence data. *Genome research* **19**, 1117-1123 (2009).

62. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829 (2008).

63. Woyke, T., Tighe, D., Mavromatis, K., Clum, A., Copeland, A., Schackwitz, W., Lapidus, A., Wu, D., McCutcheon, J.P., McDonald, B.R., Moran, N.A., Bristow, J. & Cheng, J.F. One bacterial cell, one complete genome. *PloS one* **5**, e10314 (2010).

64. Hulten, M.A., Jonasson, J., Iwarsson, E., Uppal, P., Vorsanova, S.G., Yurov, Y.B. & Iourov, I.Y. Trisomy 21 mosaicism: we may all have a touch of Down syndrome. *Cytogenetic and genome research* **139**, 189-192 (2013).

65. Wu, J., Springett, A. & Morris, J.K. Survival of trisomy 18 (Edwards syndrome) and trisomy 13 (Patau Syndrome) in England and Wales: 2004-2011. *American journal of medical genetics. Part A* (2013).

66. Galimberti, D., Ghezzi, L. & Scarpini, E. Immunotherapy against amyloid pathology in Alzheimer's disease. *Journal of the neurological sciences* (2013).

67. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. & Church, G.M. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732 (2005).

68. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, D. The human genome browser at UCSC. *Genome research* **12**, 996-1006 (2002).

69. Choi, J.H., Ogunniyi, A.O., Du, M., Du, M., Kretschmann, M., Eberhardt, J. & Love, J.C. Development and optimization of a process for automated recovery of single cells identified by microengraving. *Biotechnology progress* **26**, 888-895 (2010).

70. Wang, Y., Shah, P., Phillips, C., Sims, C.E. & Allbritton, N.L. Trapping cells on a stretchable microwell array for single-cell analysis. *Analytical and bioanalytical chemistry* **402**, 1065-1072 (2012).

71. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., Robasky, K., Zaranek, A.W., Lee, J.H., Ball, M.P., Peterson, J.E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M.I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K.P., Ebert, J.C., Nilsen, G.B., Baccash,

J., Halpern, A.L., Church, G.M. & Drmanac, R. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190-195 (2012).