UNIVERSITY OF CALIFORNIA SAN DIEGO

Characterizing Genomic Mosaicism in Single Neurons from Adult Human Brains

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

in

Bioengineering

by

Andrew Gordon Richards

Committee in charge:

Professor Kun Zhang, Chair
Professor Sheng Zhong, Co-chair
Professor Jerold Chun
Professor Xiaohua Huang
Professor Jin Zhang

2018

The Dissertation of Andrew Gordon Richards is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-chair

_____
Chair

University of California San Diego

2018

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AD: Alzheimer's Disease

APS: Ammonium Persulfate

bp: Base Pair

BSA: Bovine Serum Albumin

CNV: Copy Number Variation

cDNA: Complementary DNA

DMEM: Dulbecco's Modified Eagle's Medium

DMSO: Dimethyl Sulfoxide

DNA: Deoxyribonucleic Acid

DOP-PCR: Degenerate Oligonucleotide Primed Polymerase Chain Reaction

EDTA: Ethylenediaminetetraacetic Acid

FACS: Fluorescence Activated Cell Sorting

FBS: Fetal Bovine Serum

FOTS: Perfluorooctyltrichlorosilane

gDNA: Genomic DNA

HFE: Hydrofluoroether

kb: Kilobase

MAPD: Median Absolute Pairwise Distance

Mb: Megabase

MDA: Multiple Displacement Amplification

mRNA: Messenger RNA

NGS: Next Generation Sequencing

PBS: Phosphate Buffered Saline

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PDMS: Polydimethylsiloxane

PFO: Perfluorooctanol

QC: Quality Control

qPCR: Quantitative PCR

RNA: Ribonucleic Acid

rRNA: Ribosomal RNA

TBE: Tris Borate EDTA

TEMED: Tetramethylethylenediamine

TPM: Transcripts Per Kilobase Per Million

tRNA: Transfer RNA

UV: Ultraviolet

WGA: Whole-Genome Amplification

WGS: Whole-Genome Sequencing

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Kun Zhang for his support as the chair of my committee. His mentorship and encouragement has been essential during my Ph.D. work. I would also like to acknowledge the entire Zhang lab for their invaluable help and camaraderie, as well as the lab of Jerold Chun for their assistance in providing samples and useful discussions. I would especially like to thank Suzanne Rohrback for lending her expertise in CNV calling during the drafting of Chapter 2.

Chapter 1, in full, is a reprint of the material as it appears in Lab on a Chip (Royal Society of Chemistry) (Hoople, Gordon D.*, Andrew Richards*, Yan Wu, Kota Kaneko, Xiaolin Luo, Gen-Sheng Feng, Kun Zhang, and Albert P. Pisano. 2017.). The dissertation author was a primary author of this paper.

VITA

2011 Bachelor of Science, University of California San Diego

2018 Doctor of Philosophy, University of California San Diego

PUBLICATIONS

"IL-17 regulates adipogenesis, glucose homeostasis, and obesity" Journal of Immunology 185(11), 2010.

"Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells" Nature Biotechnology 31, 1126-1132, November 2013.

"Gel-seq: whole-genome and transcriptome sequencing by simultaneous low-input DNA and RNA library preparation using semi-permeable hydrogel barriers" Lab on a Chip, Issue 15, June 2017.

"Gel-Seq: A Method for Simultaneous Sequencing Library Preparation of DNA and RNA Using Hydrogel Matrices" Journal of Visualized Experiments, 2018.

FIELDS OF STUDY

Major Field: Bioengineering

Professors Kun Zhang and Sheng Zhong

ABSTRACT OF THE DISSERTATION

Characterizing Genomic Mosaicism in
Single Neurons from Adult Human Brains

by

Andrew Gordon Richards

Doctor of Philosophy in Bioengineering

University of California San Diego, 2018

Professor Kun Zhang, Chair
Professor Sheng Zhong, Co-Chair

DNA copy number variations (CNVs) have previously been reported in human cortical neurons from non-diseased patients, but these alterations do not appear to be consistent from cell to cell and appear to be rare among neurons overall. Interestingly, Alzheimer's disease patients appear to have a higher prevalence of CNVs than non-diseased, although the biological significance of this observation is still largely unknown. Single-cell whole-genome next-generation sequencing holds promise to investigate these variations and the regions in which

they occur in an unbiased manner. Unlike recent advances in single-cell RNA-seq, however, library preparation for single-cell DNA-seq suffers from extremely limited throughput. Furthermore, it is difficult to assess the significance of individual variations from whole-genome sequencing alone, particularly when control samples from non-diseased patients also show some variation at lower frequency. A potential solution is a multi-omics approach, in which information is collected about multiple species of biomolecules simultaneously from each sample, which taken together aid the interpretation of individual observations with respect to biological significance.

This dissertation describes the design and development of a technology to physically separate DNA and RNA and to prepare sequencing libraries from each in parallel from limited starting samples without splitting, which we called Gel-seq. Thirty-two paired DNA and RNA sequencing libraries were successfully prepared from a variety of human and mouse cells lines and from mouse liver tissue using Gel-seq. Sample types could be clearly distinguished from each other based on either genomic copy number or transcriptomic profiles. This dissertation also describes the design and development of a technology to prepare a thousand single-cell whole-genome sequencing libraries in a single run. A proof-of-concept was performed with 87 cells from human and mouse lines. Copy number profiles agreed with bulk, and 96% and 92% of human and mouse cells, respectively, clustered correctly within their cell line based on copy number profile alone. These technologies will help to enable the unbiased characterization of genomic alterations not only in neurodegenerative disorders, but potentially also in other conditions associated with mosaic genomic backgrounds, such as cancer, microbiome disorders, or infectious diseases.

INTRODUCTION

Genomic mosaicism describes a situation in which different cells of the same tissue or organism contain different versions of the genetic code, i.e., DNA, and can encompass a broad range of DNA variations. Although variations in DNA sequence or content between cells in a tissue or organism are often thought of in the context of disease-causing mutations, there are several examples of genomic mosaicism that can be found in developmentally typical and non-diseased examples. These variations can be patch-work mosaic, such as the patch-work skin coloration of a calico cat. Examples of normally occurring genomic mosaicism can also be found at the single-cell level, such as V(D)J recombination in T- and B-lymphocytes to expand the repertoire of antigen recognition in the immune system ("The Nobel Prize in Physiology or Medicine 1987" n.d.).

Genomic mosaicism is also associated with a variety of disease states, however, and in some cases, can drive the disease itself. Early mutations in fetal development, and the stage at which they occur, can underlie both the presentation and the severity of a variety of neurological disorders, such as megacephaly (McConnell et al. 2017; Cai et al. 2014). Tumor development is also characterized by the emergence of clonally mosaic genomic mutations (Gao et al. 2016). Observations such as these have motivated a great deal of interest in studying genomic mosaicism down to the single-cell level, much like the study of transcriptional variation within has spurred technology development towards extremely high-throughput single-cell RNA-seq methods that can both elucidate relationships between known cell populations while simultaneously identifying previously unknown populations (Klein et al. 2015; Macosko et al. 2015).

Patients with certain neurodegenerative disorders, such as Alzheimer's disease, have been shown to have increased prevalence of DNA alterations in the pre-frontal cortex (Westra et al. 2010), but the shortcomings of existing technologies to characterize the type, scale, locations, and subsequent impact of these alterations across sufficient numbers of single neurons has hampered investigation into the potential biological significance of these genomic events. A promising method recently available is single-cell next generation sequencing (NGS), which has generated new possibilities for profiling a wide range of mosaic genomic alterations, but existing methods for single-cell DNA library prep are still too slow, expensive, and labor intensive to meet the required throughput. A second critical limitation of current NGS approaches is that multiple data types cannot typically be investigated in the same single cells, which places a severe limitation on the ability of investigators to link DNA alterations to any functional outcomes, such as RNA expression. There exists an unmet engineering need, therefore, to scale-up the throughput at which individual cells process for NGS library preparation, reduce the cost and effort required to prepare those libraries, and to expand the repertoire of technical approaches that can be applied to the same single-cell.

This work seeks to address the aforementioned needs by showing proof-of-principle for two concepts: First, by demonstrating successful DNA and RNA sequencing from the same low-input (100 to 1000 cell) starting samples at negligible added cost by using a novel semi-permeable hydrogel design; and Second, the development and implementation of a novel microfluidic single-cell hydrogel encapsulation device to allow a single investigator to generate thousands of single-cell whole-genome sequencing libraries in a single experiment.

CHAPTER 1. DEVELOPMENT AND PROOF-OF-CONCEPT OF A LOW-INPUT MULTIOMIC HYDROGEL SEPARATION TECHNOLOGY FOR SIMULTANEOUS PREPARATION OF BOTH DNA AND RNA SEQUENCING LIBRARIES FROM THE SAME STARTING SAMPLE

1.1. Abstract of Chapter 1

The advent of next generation sequencing has fundamentally changed genomics research. Unfortunately, standard protocols for sequencing the genome and the transcriptome are incompatible. This forces researchers to choose between examining either the DNA or the RNA for a particular sample. Here we describe a new device and method, collectively dubbed Gel-seq, that enables researchers to simultaneously sequence both DNA and RNA from the same sample. This technology makes it possible to directly examine the ways that changes in the genome impact the transcriptome in as few as 100 cells. The heart of the Gel-seq protocol is the physical separation of DNA from RNA. This separation is achieved electrophoretically using a newly designed device that contains several different polyacrylamide membranes. Here we report on the development and validation of this device. We present both the manufacturing protocol for the device and the biological protocol for preparing genetic libraries. Using cell lines with uniform expression (PC3 and Hela), we show that the libraries generated with Gel-seq are similar to those developed using standard methods for either RNA or DNA. Furthermore, we demonstrate the power of Gel-seq by generating a matched genome and transcriptome library from a sample of 100 cells collected from a mouse liver tumor.

1.2. Introduction

Genomicists strive to understand how the information encoded by our DNA is turned into life. Understanding the way variations in DNA impact RNA expression is critical to decoding cell behavior. Recent advances in sequencing technology have made it possible to examine either the genome or the transcriptome of increasingly small samples (Gole et al. 2013; Sasagawa et al. 2013; Ramskold et al. 2012). Both approaches are extremely powerful, however the protocols are generally incompatible. This presents a challenge for simultaneously investigating both DNA and RNA.

When samples are sufficiently large, they can be split in half and processed for either for DNA or RNA sequencing. Unfortunately, large samples tend to average out interesting variations between cells (Shapiro, Biezuner, and Linnarsson 2013). Researchers are increasingly interested in investigating the variations present in small populations of cells (Shapiro, Biezuner, and Linnarsson 2013). To illustrate the importance of studying small cell populations, consider that tumors are often composed of heterogeneous cell populations (Spratt et al. 2016). Evidence suggests this heterogeneity may be responsible for treatment failure (Sottoriva et al. 2013). In order to understand tumor genomics, it would be useful to profile small groups of cells from different locations. When collecting just a few hundred cells from such a tumor, splitting a sample in half could result in two distinctly different cell populations, making it difficult to establish a causal link between genomic and transcriptomic variations. Gel-seq is our solution to this problem. Rather than splitting the sample, researchers can instead use Gel-seq to generate DNA and RNA libraries from the same starting cells. This method allows for the direct comparison of DNA and RNA data from low input samples.

The ability to sequence either DNA or RNA from low input samples has only been achieved in the last five years (Gole et al. 2013; Sasagawa et al. 2013; Ramskold et al. 2012). Consequently there has been very little work regarding how to sequence both DNA and RNA from the same sample. To date we are only aware of two other publications on this topic, both from 2015, and both having taken very different approaches from our method. Dey et al. have developed a protocol, DR-Seq, for simultaneously amplifying and sequencing DNA and RNA from the same single cell (Dey et al. 2015). DR-Seq takes a computational approach to distinguish between genomic DNA and the cDNA derived from RNA. To calculate DNA coverage in DR-Seq, reads where only exons are present are computationally suppressed, as those could have originated from either DNA or RNA. The genomic profile is instead determined using data based only on sequences containing introns. A drawback of this approach is that it requires a priori knowledge (exons vs. introns) of a reference genome assembly. Furthermore, intron splicing is not always conserved in disease states such as cancer. Macaulay et al. have developed G&T-seq, a method for separating, amplifying, and sequencing DNA and RNA from the same single cell (Macaulay et al. 2015). This approach relies on a physical separation of RNA from genomic DNA by using the 3′ polyadenylated tail as a pull-down target. Messenger RNA is captured on a magnetic bead using a biotinylated oligo-dT primer, allowing it to be separated from genomic DNA.

The novel aspect of Gel-seq is the ability to separate DNA and RNA in hundreds of cells based exclusively on size. Our method requires no a priori knowledge of the genome and is not limited to polyadenylated transcripts. For applications where a researcher can start with a few hundred cells, or where the transcripts of interest are not polyadenylated, Gel-seq provides an alternative approach to existing methods using cheap and widely-available materials.

Our method takes advantage of the vast size differences between DNA and RNA. At the heart of the Gel-seq protocol is the electrophoretic separation of DNA and RNA/cDNA hybrids based on this size difference. Genomic DNA from humans, for example, is tens of millions of base-pairs (bp) long for the shortest chromosomes and will remain megabase-scale if shearing is minimized. Most messenger RNA, on the other hand, are only a few hundred to a few thousand nucleotides. Understanding this size difference, we developed two membranes that could be used to separate DNA from RNA. The first membrane, a low-density polyacrylamide gel, allows RNA molecules to pass through but stops larger genomic DNA. The second membrane, a high-density polyacrylamide gel, traps the RNA molecules. Both membranes allow small fragments (<100 bases) of unwanted artifacts, such as primers, to pass through. The membranes also allow small buffer ions to pass through unimpeded, a necessary condition for electrophoresis. While it is well documented in the literature that ion gradients can form in microfluidic systems in response to applied electric fields (Zangle, Mani, and Santiago 2010), we see no evidence that such gradients are negatively impacting our separation. We theorize that the large size of our buffer reservoir, the high potential difference across the membrane, and the short timespan over which we run the device the mitigates the effects of any ion buildup.

Our basic approach to separating DNA and RNA is shown in Fig. 1. Fig. 1A shows DNA and RNA free floating in solution near a synthetic membrane. When an electric field is applied, as shown Fig. 1B, DNA and RNA experience an electrophoretic force that induces migration through the membrane. By tuning the membrane properties, we created a semi-

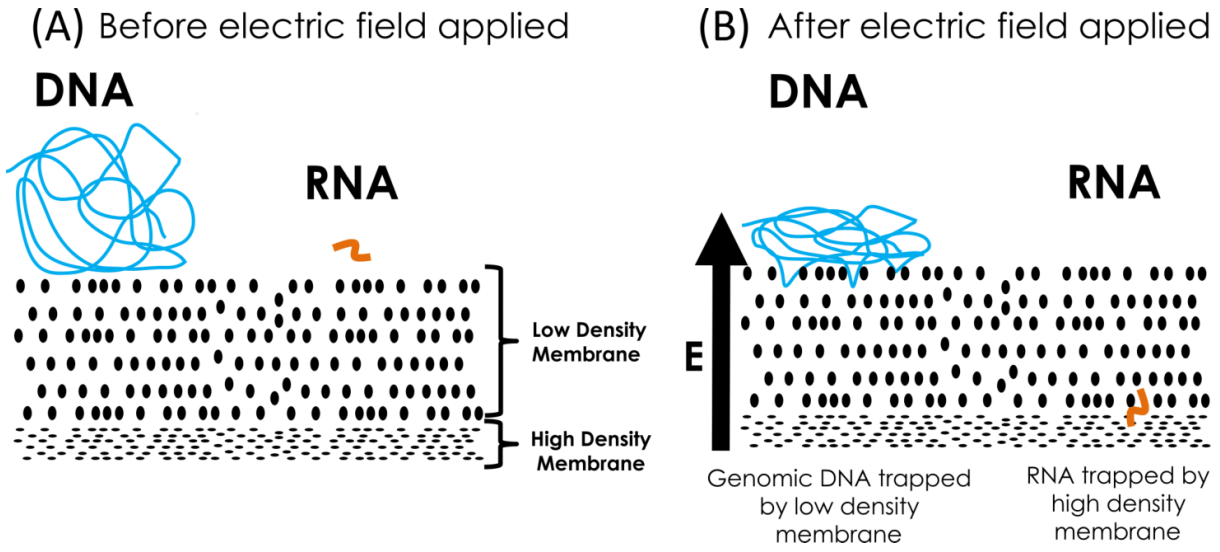

Figure 1: The underlying principle used to physically separate DNA and RNA. In an applied electric field, small RNA molecules migrate through the low-density membrane but large DNA molecules are trapped at the surface.

permeable membrane that separates DNA from RNA. The genomic DNA molecules are pushed against the membrane, but become trapped at the edge due to their large size. Smaller RNA molecules, on the other hand, are able to weave their way through the low-density membrane much like a snake through grass, a process known as reputation (Viovy 2000). These RNA molecules are then stopped by a second, high density membrane. Once they have been physically separated, the DNA and RNA can be recovered and processed into genomic and transcriptomic sequencing libraries.

Though we conceived of the method independently, our approach harkens back to the disc gel electrophoresis invented by Orstein and Davis in the 1960s (Ornstein 1964; Davis 1964). In disc gel electrophoresis, hydrogels with discontinuous pore sizes are used to increase the separation resolution for proteins. Our method differs from traditional disc gel

electrophoresis in that our high-density membrane is designed to stop a species of interest rather than improve the resolution between bands.

1.3. Experimental

1.3.1. Gel-seq Overview

An overview of the Gel-seq protocol is shown in Fig. 2A. We used a protocol adapted with minor modifications from Nextera XT to prepare DNA libraries after separation. To prepare RNA libraries, we first converted RNA to cDNA using a modification of the Smart-Seq protocol developed by Ramskold followed by a modified version of Nextera XT (Ramskold et al. 2012; Illumina 2015). While we can separate DNA and RNA, we have found that converting the RNA to cDNA before separation helps mitigate problems associated with RNAse contamination. We begin the protocol with between 100 and 1000 intact cells, apply a lysis buffer, and perform reverse transcription with template switching. This generates cDNA/RNA hybrids that are more stable than RNA alone. This protocol does not have a measurable impact on the quality of the genomic DNA (gDNA). The resulting cDNA/RNA hybrids are orders of magnitude smaller than the genomic DNA, enabling size-based separation as shown in Fig. 1 using a custom fabricated gel system.

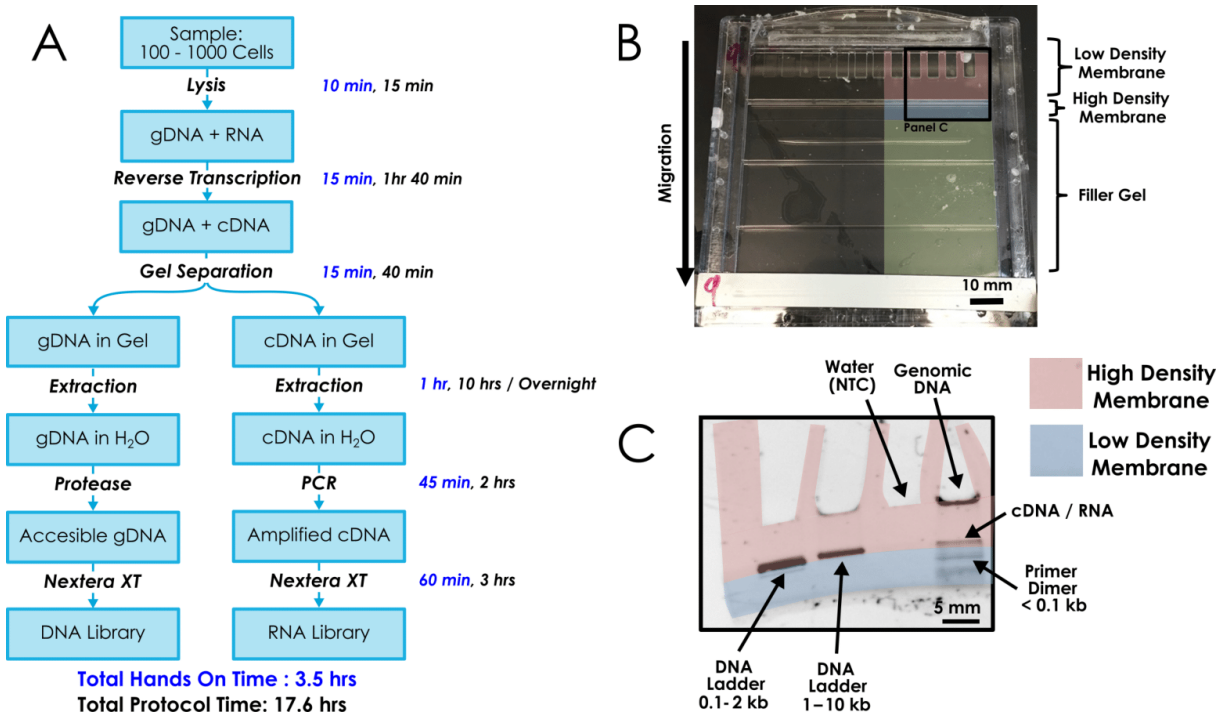

Figure 2: An overview of the Gel-seq protocol (A) and device (B). False color has been added to half of the device to clearly demarcate the different regions of polyacrylamide gel. The third panel (C) is a fluorescent image showing the separation of genomic DNA and cDNA/RNA hybrids. Black bands indicate the presence of nucleic acids. Lanes loaded with only DNA ladder show a single band that has been trapped by the high density membrane. Lanes loaded with genomic DNA and RNA/cDNA show two bands, suggesting that genomic DNA has been separated from RNA/cDNA.

The Gel-seq device shown in Fig. 2B consists of three regions of polyacrylamide gel. The top layer, highlighted with false color in pink, consists of a low-density membrane of 4% total (T) acrylamide and 3% cross-linker (C) bis-acrylamide. A standard gel electrophoresis comb is used to define loading wells. This layer stops genomic DNA but allows transcripts less than 10,000 nucleotides to pass through. The second layer, highlighted in purple, is a high-density membrane of 30% T acrylamide cross linked with 5% C bis-acrylamide. This layer stops RNA/cDNA but permits the passage of ions necessary for electrophoresis. The bottom layer, shown in green, fills the remainder of the gel cassette but is not used in the separation. The filler gel is also a 4% T acrylamide cross-linked with 3% C bis-acrylamide. Using a low-

density filler gel, rather than filling the rest of the cassette with high density gel, ensures that there is a sufficiently large potential drop across the separation region to induce RNA/cDNA migration. The resulting gel cassette is compatible with standard buffer chambers and power supplies commonly found in life science laboratories. The fabrication protocol, described in detail in the next section, is straightforward and utilizes commonly available equipment and materials.

After placing the device into a buffer chamber, we then pipette the DNA and reverse transcription products into the wells. We induce electrophoresis by applying 210 V across the cassette for 30 minutes. Once the genomic DNA and RNA/ cDNA have been separated, we cut out the gel sections to recover the nucleic acids using a modified crush and soak procedure. We prepare a DNA sequencing library directly from the genomic DNA using the Nextera XT protocol. For RNA, we first PCR amplify the cDNA fraction and then prepare a sequencing library by Nextera XT.

1.3.2. Device Fabrication

Many companies sell standard gel electrophoresis systems that come with a power supply, electrophoresis chamber, and empty cassettes. These systems dramatically simplify the process of conducting experiments with gel electrophoresis. End users simply fill the cassette with the desired density polyacrylamide based on their needs. Once the gel has polymerized, the cassette is placed in the electrophoresis buffer chamber, sample is added, and the chamber is connected to a power supply to apply an electric field. In this paper we based our fabrication protocol around the XCell SureLock® Mini-Cell system (Lonza); however, any similar system could be used.

Device fabrication builds on skills that will be familiar to researchers who use standard polyacrylamide gel cassettes. Before fabrication, monomer solutions are made for each layer by combining acrylamide/bis-acrylamide solution, 10X Tris-borate-EDTA (TBE), water, and sucrose solution (50% w/ v) as shown in Table 1. The addition of sucrose to the polyacrylamide precursor solution is key to the formation of a smooth interface layers between the different densities, but has minimal impact on electrophoresis. Stock acrylamide/bis-acrylamide solutions used in these recipes can be made by combining acrylamide (monomer) and bis-acrylamide (crosslinker) powders using the following formulas:

$$\%\,\mathrm{T} = \frac{\text{monomer mass (g)} + \text{crosslinker mass (g)}}{\text{solvent volume (mL)}}$$

$$\%\,\mathrm{C} = \frac{\text{crosslinker mass (g)}}{\text{monomer mass (g)} + \text{crosslinker mass (g)}}$$

The gel precursor solutions are mixed in a tube and vortexed to ensure thorough mixing, and then immersed in a sonicator under house vacuum. This helps to remove dissolved gases that could inhibit the polymerization process. Immediately before transferring the precursor solution to the cassette, a polymerization initiator containing ammonium persulfate (APS) and catalyst (TEMED) are added and the mix is briefly vortex again. Note that the high-density gel does not contain any TBE. While it could be included, we find it easier to mix the precursor solution when it is not included as we are approaching the solubility limit of acrylamide and bis-acrylamide. We have noticed no negative impact on device performance from the omission of TBE in this region.

Table 1: Recipes for mixing polyacrylamide gel precursors

Filler gel precursor (4% T, 3% C)		High-density gel precursor (30% T, 5% C)		Low-density gel precursor (4% T, 3% C)	
40% T, 3.3% C acrylamide/bis-acrylamide solution	0.8 mL	50% T, 5% C acrylamide/bis-acrylamide solution	1.2 mL	40% T, 3.3% C acrylamide/bis-acrylamide solution	0.4 mL
Deionized water	5.12 mL	Deionized water	0.48 mL	Deionized water	3.2 mL
Sucrose (50% w/v)	1.28 mL	Sucrose (50% w/v)	0.32 mL	10X TBE	0.4 mL
10X TBE	0.8 mL	APS (10%)	25 uL	APS (10%)	26 uL
APS (10%)	52 uL	TEMED	0.5 uL	TEMED	1.5 uL
TEMED	3 uL				
Total volume	8.1 mL	Total volume	2 mL	Total volume	4.0 mL

An overview of the protocol is shown in Fig. 3. Layers are fabricated from bottom to the top. We first add 6 mL of filler gel precursor to the cassette. The remainder of the cassette is filled with de-ionized, degassed water. The filler gel is allowed to polymerize for at least one hour or up to overnight. The water overlay ensures the formation of a smooth interface. After polymerization, we remove the water overlay by simply inverting the cassette and shaking. Compressed air can be used to assist in the removal of any trapped water droplets. We then add 350 μL of the high-density precursor to the cassette. Due to the small volume of high density gel, it is important to ensure the precursor is evenly distributed by tilting the cassette back and forth to allow the liquid to uniformly spread out over the filler gel. Once the high-density precursor has been uniformly distributed, we again add a water overlay. In order to obtain the best interface, it is important to add the water slowly to the center of the cassette in order to minimize mixing with the high-density precursor. We allow the high-density gel to polymerize for at least 10 minutes before the water overlay is removed. Finally, we add the low-density precursor to fill the remainder of the cassette, approximately 1.65 mL. In order to define the loading wells, we insert a standard gel comb into the cassette. Cassettes can be fabricated with different numbers and sizes of wells by using different combs. In this work, we fabricated gels

with either 10 or 12 well combs. We allow the low-density gel to polymerize overnight before using the cassettes. Cassettes can be stored immersed in TBE buffer for several weeks.

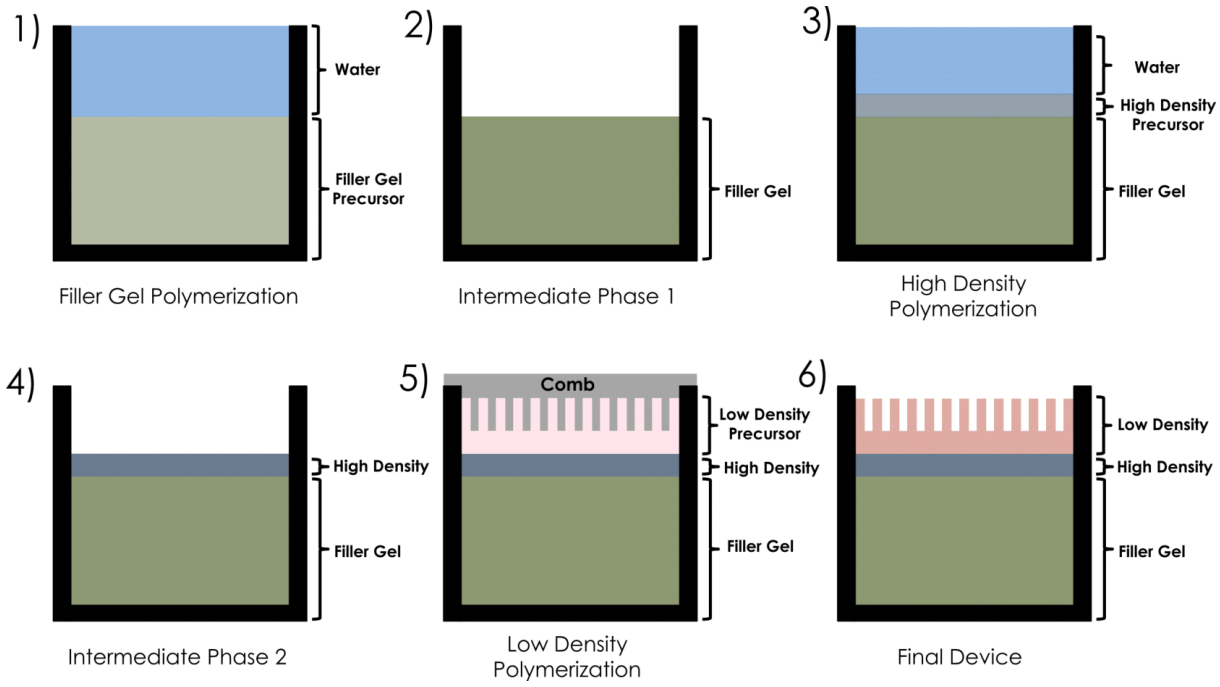

Figure 3: The fabrication protocol for the cassette based devices. Each layer of gel is allowed to polymerize before the next layer of gel is poured on top of it. A water overlay helps to create a smooth interface between layers.

1.3.3. Gel-seq protocol

In addition to device development, there was a need to adapt existing biochemical protocols to be compatible with physical separation of gDNA and RNA and to prepare libraries from both. Recognizing the susceptibility of RNA to degradation, we reverse transcribe RNA to cDNA before separating it from gDNA. Once we separate gDNA and RNA/cDNA, we then prepare a sequencing library from the gDNA using Nextera XT. In parallel, we amplify the cDNA sample by PCR and prepare a sequencing library, also using Nextera XT. In order to minimize the shearing of genomic DNA, which could cause it to enter the separation gel, we avoided vortexing samples. Instead all samples were mixed by gently pipetting up and down

approximately 10 times. While this will shear the chromosomes somewhat, the fragments are still orders of magnitude larger than the RNA/cDNA hybrids.

We begin the protocol by preparing cells in PBS at a concentration of 100 to 1000 cells per μL. Using the reagents provided in the Smart-Seq v4 kit (Clontech Laboratories), we mix 19 μL of lysis buffer and 1 μL of RNase inhibitor to prepare a 10× stock solution of reaction buffer. We then combine 1 μL of the cell suspension, 0.5 μL of 10× reaction buffer and 2.75 μL of nuclease-free water and mix by pipetting up and down 5 times. We then add 1 μL of 3′ SMART-Seq CDS Primer II and 1 μL of 20 μM random hexamer with SMART-Seq adapter (Integrated DNA Technologies (IDT):

5′ AAGCAGTGGTATCAACGCAGAGTACNNNNNN 3′

Each sample is incubated at 72 °C in a preheated thermal cycler with heated lid for 3 minutes to lyse the cells. Note that the addition of random hexamer seemed to have minimal impact and the mapping rates to rRNA remain below 1% (see Fig. S9). After lysis, we add a master mix containing 2 μL of 5× Ultra Low First-Strand Buffer, 0.5 μL of SMART-Seq v4 Oligos, 0.25 μL RNase Inhibitor, and 1 μL SMARTScribe Reverse Transcriptase. We mix the sample by pipetting up and down 5 times and then immediately place it in a preheated thermal cycler at 42 °C with a heated lid for 90 min, followed by a heat inactivation step at 70 °C for 10 min.

Following the completion of reverse transcription, we mix the samples with 2 μL of 6× DNA Gel Loading Dye (ThermoFisher). We load the entire reaction volume into the Gel-seq device (one sample per well) and apply an electric field of 210 V across the device for 30 minutes to separate RNA from DNA. After separation, we stain the gel in 30 mL of 0.5× TBE

with 3 µl SYBR Gold (ThermoFisher) for 5 minutes. We image the gel using a 30 second exposure on a Bio-Rad Gel Doc. We then cut out the regions containing gDNA and cDNA/RNA using a scalpel. Visualizing the cDNA from the 100 cell input samples sometimes presented a challenge due to the small amount of nucleic acids present. Fortunately, the ability to visualize the location of the gDNA or cDNA is not a requirement for recovering it from the gel. We designed the device so the gDNA stops at the start of the well and the cDNA stops at the start of the high-density gel. As these locations are both visible to the naked eye, the gel can be cut without the use of a UV backlight. In practice we found using the UV backlight convenient as most samples could be visualized, but this is not a strict requirement.

Once cut from the gel, each gel section is placed into a separate tube and ground up using the end of a pipette. We add 40 µL of nuclease-free water to the gel containing gDNA and 80 µL of nuclease-free water to the gel containing cDNA/ RNA. We then tape the tubes containing the gel and water to a vortex mixer inside 37 °C incubator and shake them for 8 to 12 hours. This allows the nucleic acids to diffuse out from the gel into the water.

After incubating the samples, we pipette the samples into an 8 µm mesh filter plate (Corning HTS Transwell 96-well permeable support) and spin the plate at 2600 RCF for 5 minutes to strain out the gel fragments. We then pipette the gel-free water into a new 200 µL tube.

For the gDNA sample, we add 1 µL of protease (Qiagen, diluted to 0.9 AU/mL) and incubate at 50 °C for 15 min followed by 70 °C for 15 min. This step is critical for depleting nucleosomes, making the DNA accessible for Nextera XT library preparation. Next, we use an 18-gauge needle to create holes in the caps of all samples tube before spinning them in a vacufuge to reduce sample volume. The cDNA/RNA samples are reduced to 10 µL and the

gDNA samples reduced to 5 μL. This step takes 30–60 minutes, depending on the number of samples in the vacufuge. If samples were found to be below the target volume, 1–2 μL of clean nuclease free water was added to bring them to the correct target volume.

We generate libraries from the gDNA samples by following the standard Nextera XT protocol (Illumina 2015). To conserve reagents, we have found that using half volume reactions does not significantly impact our library quality. The protocol is otherwise identical from this point on.

To generate libraries from the cDNA/RNA samples, we first amplify the sample using PCR. We combine a 10 μL sample with 12.5 μL 2× KAPA SYBR Fast qPCR MasterMix (KAPA Biosystems), 0.5 μL PCR Primer II A (12 μM, from the Smart-Seq kit), and 2 μL nuclease-free water. We perform qPCR in a BioRad thermocycler using the following protocol: hot-start at 95 °C for 3 min, followed by 20–30 cycles of 98 °C for 10 seconds, 65 °C for 30 s, and 72 °C for 3 min. We adjust the number of cycles depending on the amount of starting sample and the shape of the qPCR curves to avoid over-amplification. After amplification, we clean the product using AMPure XP beads following the protocol described in the Smart-Seq Manual (Clontech 2016). Finally, once the amplified cDNA has been purified, we prepare libraries using the Nextera XT protocol with half volume reactions.

The entire protocol requires 3.5 hours of hands on time and can be completed in 17.6 hours. We recommend starting the protocol in the afternoon so that the crush and soak step can take place overnight.

1.4. Results and Discussion

1.4.1. Validation of DNA and RNA/cDNA Separation

To validate our separation approach, we tested the device using four samples: a low mass DNA ladder (0.1–2 kilobases (kb)), a high mass DNA ladder (1–10 kb), water as negative control, and genomic DNA and RNA/cDNA hybrids. Commercially purchased DNA is not generally appropriate as a control for genomic DNA in this case, as it tends to be sheared somewhat during production. The best solution is to use freshly lysed cultured cells in each experiment. After electrophoresis, the device was stained with SYBR Gold and imaged. The resulting fluorescent image is shown in Fig. 2C; false color has been added to distinguish between the different regions of the gel.

The negative control (lane 3) showed no signal, demonstrating that the device is not auto-fluorescent. The first two lanes, loaded with DNA ladder, show the presence of black bands indicating that nucleic acid has been trapped in a specific location. The first lane, which was loaded with the low mass DNA ladder, contains only one band at the interface between the low and high-density gels. This band contains fragments ranging from 100–2000 basepairs. Rather than spreading throughout the gel, as is typical in standard gel electrophoresis, the bands stack on top of each other at the interface. This is exactly the desired behavior; small fragments of cDNA and RNA should move through the low-density gel and collect at the interface of the high-density region. Importantly, this ladder also demonstrates that fragments as small as 100 bp are stopped by the high-density membrane.

The second lane, loaded with the high mass DNA ladder, shows similar behavior. The major difference here is that the ladder fragments range in size up to 10 kb. Again, the ladder has stacked at the interface with the high-density gel, except for a small fraction at the top of

the low-density gel. This suggests a size cut-off somewhere between 2 and 10 kb, and perhaps a range of partial migration efficiency above 2 kb, however the great majority of cDNA/RNA species of interest are below this size (Suzuki et al. 2000).

Finally, the fourth lane demonstrates the separation of genomic DNA and cDNA/RNA hybrids. A clear dark band present at the top of the start of the low-density membrane represents megabase scale genomic DNA, which is unable to enter the gel, while cDNA/RNA hybrids are stacked at the interface of the low and high-density regions. Unlike the lanes loaded with ladder only, however, there are several bands present within the high-density region of the gel. These fragments, smaller than 100 bp, are off-target products generated from primer oligonucleotides during reverse transcription. By allowing these bands to pass through the high-density membrane, we can easily remove them from the experiment by only cutting out the cDNA/RNA hybrids stacked at the membrane interface.

As mentioned previously, there is no commercially purchased genomic DNA control shown in this example, as purified DNA tends to be sheared somewhat during production, and does not accurately represent the full native size of mammalian chromosomes. Furthermore, DNA library preparations in early iterations of Gel-seq failed until the addition of a protease digestion step to the protocol after gel separation, indicating that genomic DNA as loaded into our device is still complexed with nucleosomes. We hypothesize that these protein components of DNA in fact assist in trapping virtually all genomic DNA at the gel surface, aiding recovery by preventing nucleic acids from embedding in the gel during electrophoresis.

In order to validate the conclusions inferred from this image, as well as assess the data quality of sequencing libraries, we cut out sections of the gel with the genomic DNA and cDNA/RNA hybrids and generated sequencing libraries.

1.4.2. Validation of DNA and RNA Libraries

We compared Gel-seq against standard methods common in the genomics field using commercially available kits, which we refer to as "tube controls", to prepare a total of 32 sequencing libraries (see Table 2) from two human cell lines (PC3 prostate cancer and HeLa cervical cancer), a mouse cell line (3T3 fibroblasts), and primary derived hepatocytes from mouse liver. PC3 and HeLa were chosen because they are representative of cancers with extensive copy number variations (CNVs). CNVs are either duplications or deletions of large regions of the genome, and can be detected by coverage density with shallow sequencing. CNVs are known to play a role in many cancers and are a widely studied area in cancer genomics (Lucito et al. 2003; Sebat et al. 2004; Guffanti et al. 2013; Glessner et al. 2009). In addition, CNVs provide a useful signal for genomic data that lends itself to easy comparison between different approaches for whole genome sequencing library preparation. Primary derived hepatocytes from mouse were chosen in order to validate Gel-seq using cells from a complete organ, which presents additional challenges in terms of sample prep and reaction efficiency due to the presence of extracellular matrix and other inhibitory factors. 3T3 fibroblasts were included as a positive control against liver tissue samples.

Table 2: All 16 samples for both human and mouse. For each sample, both DNA and RNA libraries were generated (32 in total). Tube samples (standard method performed in tube as control) were split before lysis for subsequent DNA and RNA library prep protocols in parallel. Gel-seq samples were lysed first, and DNA and RNA were separated in device before library prep

Human			Mouse		
Cell type	Method	Sample name	Cell type	Method	Sample name
HeLa	Gel-seq	HeLa-G1	3T3	Gel-seq	3T3-G1
		HeLa-G2			3T3-G2
	Tube	HeLa-T1		Tube	3T3-T1
		HeLa-T1			3T3-T2
PC3	Gel-seq	PC3-G1	Hepatocytes	Gel-seq	Liver-G1
		PC3-G2			Liver-G2
	Tube	PC3-T1		Tube	Liver-T1
		PC3-T2			Liver-T2

Gel-seq and tube control experiments were performed in parallel for all samples to assess the level of agreement between methods. DNA and RNA libraries were prepared for both human and mouse samples. For Gel-seq samples, RNA data was generated from the exact same cells as the DNA data, because DNA and RNA are separated after lysis, while cells used in the tube controls were split 50/50 before lysis. Technical replicates were generated for all samples in order to assess reproducibility of both genomic and transcriptomic profiles from Gel-seq data. Finally, we compared transcriptomic profiles between the different samples types within each species to assess whether Gel-seq can distinguish cell type on the basis of RNA expression.

Fig. 4A shows a comparison of genome-wide CNV profiles generated from PC3 using either Gel-seq or a standard tube reaction. Each point is a mean normalized bin count; bins are defined from reference genome data such that each bin has equal expected count in a healthy diploid cell, i.e., a flat line, representing equal copies for each region of all autosomal (excluding

X and Y) chromosomes. In PC3, many CNVs can be seen as spikes above a background copy number of two, and Gel-seq yields a qualitatively similar CNV profile as standard tube reaction. Agreement between the two plots can be assessed quantitatively by linear regression in Fig. 4B. A Pearson correlation of R = 0.90 indicates that genomic data gathered from either method is functionally equivalent. Fig. 4C shows maximum predicted library coverage at saturation sequencing depth, indicating that Gel-seq yields high coverage libraries similar to standard methods. Full coverage extrapolations as a function of depth are shown in Fig. S2.

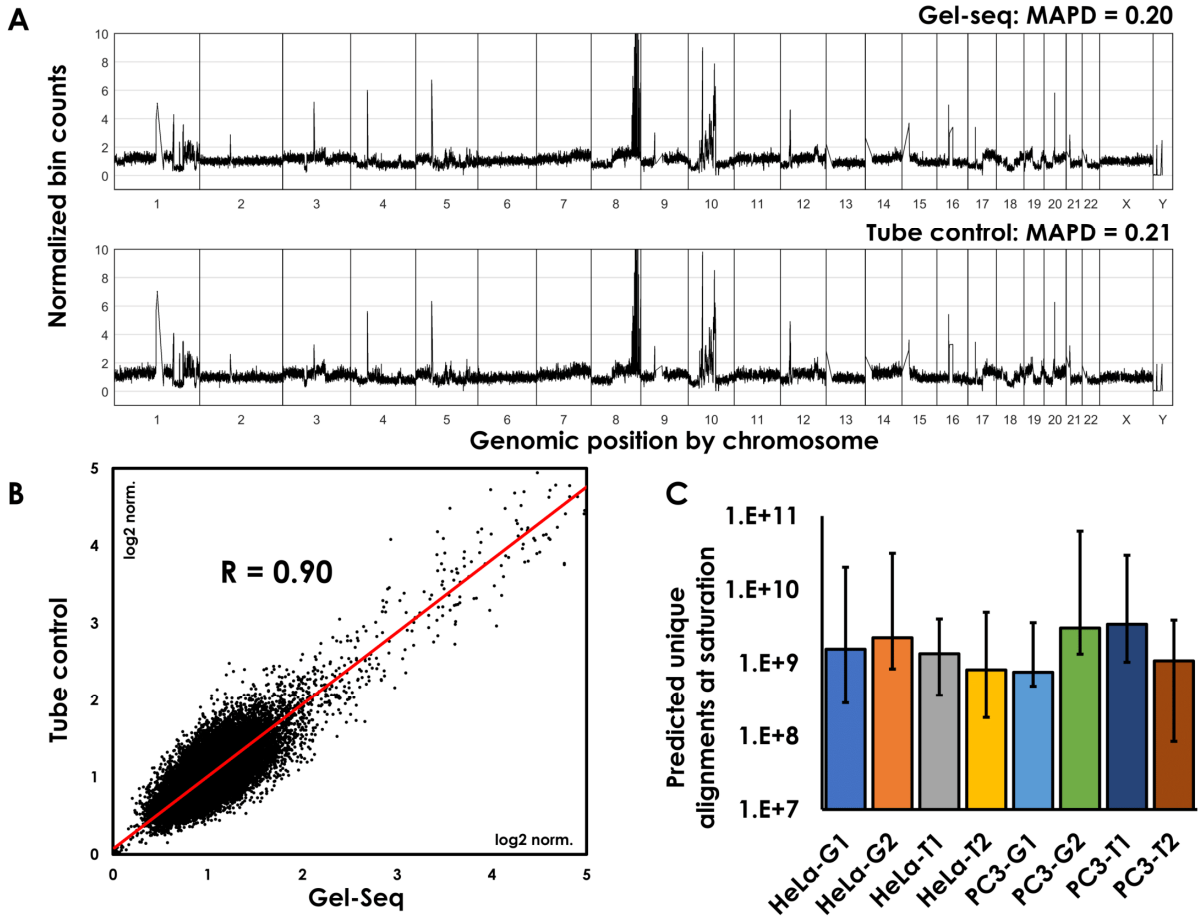

Figure 4: Comparing genomic data generated using the Gel-seq protocol to tube control. (A) Mean normalized bin counts for Gel-seq (top) and a tube control (bottom). Random noise is quantified by median absolute pairwise difference (MAPD, upper right). A MAPD of ~0.2 indicates very low noise. (B) Pearson correlation between two representative libraries. Full pairwise correlations are shown in Fig. S4.† (C) Maximum predicted genomic coverage for all human DNA libraries extrapolated to saturation sequencing depth. Error bars are 95% confidence intervals. Suffixes indicate Gel-seq data (−G) or tube controls (−T), numbers indicate technical replicates (1 or 2).

Similarly, we compared the transcriptome data from our Gel-seq protocol to the standard in-tube Smart-Seq protocol. Fig. 5A shows the correlation between both Gel-seq technical replicates and between Gel-seq and the standard method. Each point is a count in transcripts per kilobase per million (TPM) for each gene detected at TPM > 5 in both dataset. The linear regressions are shown as red lines, and the Pearson correlation coefficient is shown in the upper left corner. Technical replicates from Gel-seq agree with each other ($R \sim 0.8$), but correlate less well with the standard method ($R < 0.7$). This suggests that Gel-seq introduces a bias in gene counts, but that the bias is systematic and meaningful conclusions are still possible between different biological samples. We performed linear regression for all pair-wise combinations of the 8 human RNA datasets: PC3 and HeLa, Gel-seq and tube, two technical replicates each (Fig. S5). Pearson correlation coefficients for all 28 pairs are condensed in Fig. 5B by comparison type. The green bars represent correlations between pairs of datasets from the same cell type generated from either standard tube reactions, Gel-seq, or Gel-seq versus standard. The red bars represent correlations between pairs of datasets from different cell types, which are expected to have lower R values due to biologically different transcriptomic profiles. Although Gel-seq does not agree well with the standard method ($R = 0.66$ for matched samples), it shows similar difference in correlation between matched and mis-matched samples ($R = 0.81$ versus $R = 0.70$, respectively) to the standard method ($R = 0.97$ versus $R = 0.86$), suggesting that Gel-seq still provides powerful insight into transcriptional variation between different cell types. Indeed, Fig. 5C demonstrates that RNA-seq data generated from Gel-seq (left plot) discriminates well between HeLa and PC3 cell types based on principal component analysis (PCA), as does the standard in-tube method (right plot). Fig. 5C shows that samples separate by method on the first principal component with 96.3% variance explained, confirming that

Gel-seq introduces a systematic bias, but that different cell types (HeLa and PC3, red and blue clusters) still separate well on principal component 2.

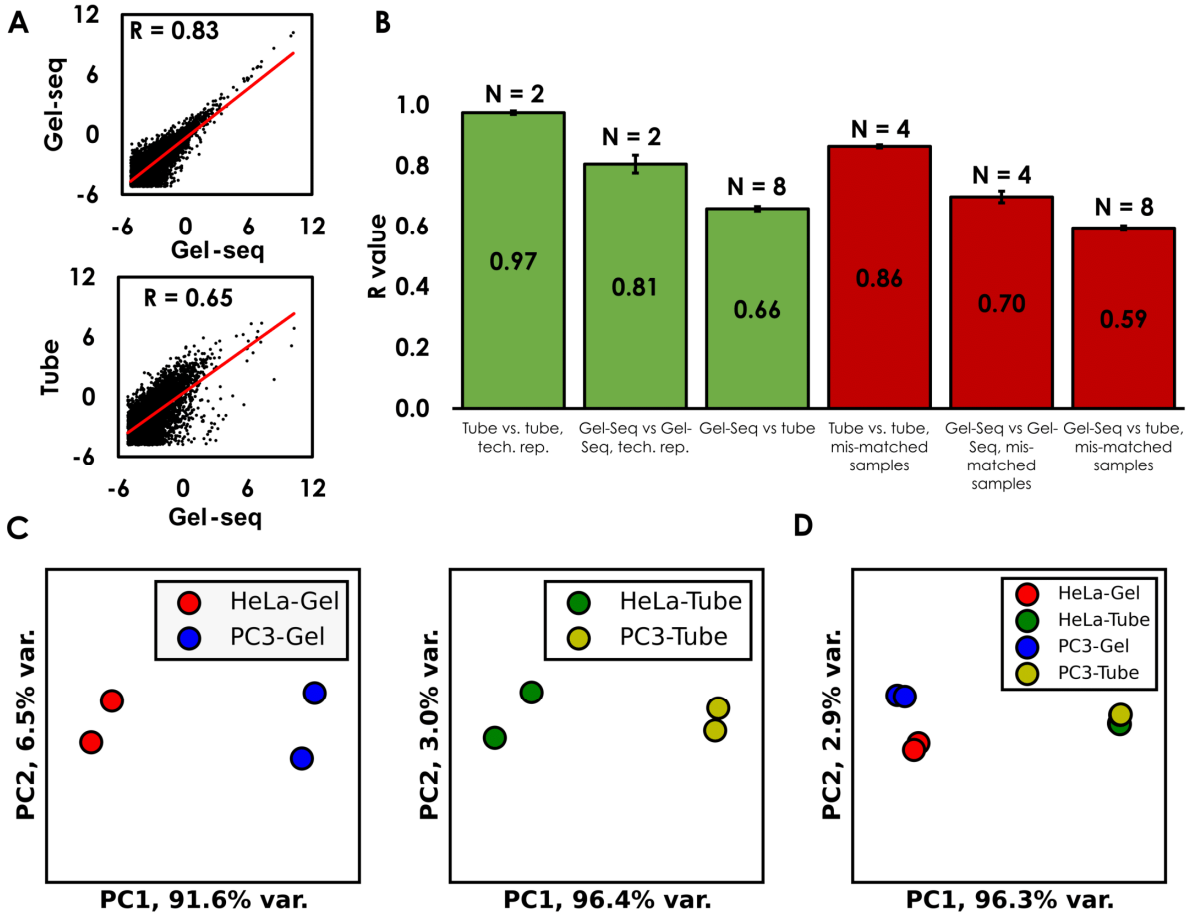

Figure 5: Comparing transcriptomic data generated using the Gel-seq protocol to tube controls. (A) Two representative scatter plots of TPM per gene (above threshold of TPM > 5) with an overlaid linear regression and Pearson correlation coefficient. The plot on the top compares two technical replicates using Gel-seq, while the plot on the bottom shows a comparison between a Gel-seq sample and a standard method performed in a tube as control. (B) Pearson coefficients from all 28 pair-wise linear regressions for all 8 HeLa and PC3 RNA datasets generate from with Gel-seq and tube controls. Full table of scatter-plots and regressions are shown in Fig. S5.† (C) PCA for Gel-seq datasets on the left and tube controls on the right. First two principal components are plotted for each, with a total of 98.1% and 99.4% of variance explained for Gel-seq and tube controls, respectively. (D) PCA for all 8 human samples, with total of 99.2% variance explained by the first two principal components.

As reported by the SEQC/MAQC-III Consortium, all RNA-seq methods show some gene specific bias (Consortium and others 2014). The key for any new approach is to demonstrate reproducibility so that differences observed between samples can be attributed to

a biologically relevant phenomenon. While Gel-seq does not perfectly replicate the results from Smart-Seq, it gives reproducible results and can be used to identify differences between samples.

1.4.3. Generating Paired Libraries From Tissue

Gel-seq allows researchers to generate both genome and transcriptome data from the same limited sample using commonly available materials. This is useful in scarce samples, such as those collected from living tissue in a biopsy. As mentioned previously, preparing next-gen sequencing libraries from tissue rather than cell lines presents substantial additional challenges. Cell lines divide rapidly, typically doubling in number in 24 to 48 hours, and tend to be highly transcriptionally active, expressing a broader set of genes at high levels compared to an adult tissue under homeostasis. Tissue samples are also subject to the presence of additional extracellular matrix, which can severely inhibit enzymatic reactions. Several iterations of both our device and accompanying biochemical methods were tested before establishing Gel-seq as a robust protocol that works in tissue as well as cultured cell lines. We also lowered the input to 100 cells (0.61 ng DNA). Gel-seq libraries from mouse tissue displayed high quality statistics in terms of unique DNA alignments and genes detected by RNA (Table S1). Genomic coverage for DNA data and library complexity for RNA data were extrapolated to high sequencing depth (saturation) in Fig. S2 based on bootstrapping simulations, indicating that Gel-seq yields high-quality libraries with coverage similar to standard methods for cells from both cultured lines and complex tissue.

1.5. Conclusion

One of our goals in developing Gel-seq was to create a protocol that could be easily implemented by other researchers. We therefore decided to fabricate devices within the standard form factor of a polyacrylamide gel cassette. While the technique we used to define our different membranes is novel, most genetics labs already have all of the necessary equipment to fabricate the Gel-seq device. Furthermore, the cost of the device is trivial – just $5.25 for a device that can process 12 samples. We believe researchers will find it straightforward to implement Gel-seq in their own labs and hope this will facilitate the rapid adoption of the technology.

As with any library preparation protocol using commercial reagents, the overall cost for generating libraries with Gel-seq remains high. Our reagent cost per sample was $28 for Nextera XT and $50 for Smart-Seq. As cheaper alternatives for library preparation are developed, however, our protocol can be adapted to work with these new techniques. We focused on creating a device that could be adapted for different applications. While in this paper we demonstrated the Gel-seq protocol using Nextera XT and a modified Smart-Seq, the device itself can be used with a wide range of library preparation approaches. For example, during development we successfully tested the device using an older RNA library amplification protocol CellAmp (Kurimoto et al. 2007). The core innovation in this technology, separating DNA and RNA based on size using polyacrylamide membranes, is agnostic to the library preparation approach. We anticipate that future biological innovations in library preparation could be integrated into our work flow.

We were successful in generating RNA libraries from cell lines regardless of whether we generated the cDNA either before or, as in earlier iterations using Cell Amp, after separation

from the genomic DNA. An unforeseen aspect in the development of the Gel-seq protocol, however, was the challenge of starting from whole tissue. We found that it was important to adhere strictly to the Smart-seq protocol to generate cDNA from tissue samples as soon as possible. We also experimented with freezing tissue or cell suspensions from tissue in liquid nitrogen, but we found that the best results were obtained when processing fresh samples. We suspect that the extracellular matrix in our tissue samples may have contained RNases, proteases, or other inhibitory factors. Fortunately, Gel-seq is a flexible protocol and proved to be adaptable to liver samples. Although Gel-seq showed generally higher random noise in technical replicates compared with our tube controls, the ability to include genomic data from the same cells in the downstream analysis may justify the trade-off in many applications. Newly developed RNA library preparation methods or optimization of separation and recovery may improve the precision of the RNA data in the future.

An interesting phenomenon observed in the RNA data was that in all 4 samples types (HeLa, PC3, 3T3, and primary hepatocytes) Gel-seq technical replicates agreed with each other, but did not have high correlations with the standard in-tube method. This suggests an underlying systematic difference between methods, which some day might be corrected with either additional optimization of separation and recovery, or accounted for computationally based on known parameters. Our first suspicion was exonic transcript length, with the assumption that very long or very short genes could be lost or trapped in the device. While we did observe a weak relationship between RNA gene counts and gene length in Gel-seq data, with medium length genes showing the highest gene counts, we observed an identical effect in tube control data. Attempting to normalize by a lowess fitted correction function did not improve the correlation between Gel-seq and tube (not shown). This could suggest that additional factors

beyond gene length are affecting the data. For many applications the addition of synthetic RNA spike-ins at a range of known concentrations (e.g., ERCC control (Lemire et al. 2011)) could be used to quantify systematic biases in sample data. This is already a common approach in the field for correcting systematic biases introduced by different kits. Future work will focus on addressing these challenges and improving the Gel-seq method. For the time being, however, Gel-seq is already a powerful and sensitive tool for finding differences in expression between samples.

Unfortunately, Gel-seq cannot be used in this embodiment to generate data at the single cell level. The geometry and low throughput of the device presented here makes it infeasible to process meaningful numbers of single cell datasets, although it is possible to fabricate qualitatively similar devices on the micron scale that could achieve this goal (Lee et al. 2013). While the sample loss in Gel-seq is variable and hard to accurately quantify, we have observed that anywhere from 10% to 50% of the nucleic acids cannot be recovered from the gel after separation. This number agrees with the literature for similar crush and soak extraction protocols from polyacrylamide gel (Sambrook and Russell 2006). When working with 100 to 1000 cells, these losses do not appear to substantially change the resulting libraries. To analyze samples below this limit, however, we will need to modify our protocol.

One approach to improve the protocol could be the use of dissolvable gels to increase sample recovery. We made several attempts at using dissolvable gels during development of the device, but none were successful. Agarose is too porous to be used for the high-density gel region and a hybrid device with a separation layer made from agarose and a high-density layer made from polyacrylamide was too fragile to handle. We tried using BAC crosslinked polyacrylamide following protocols developed by Hansen (Hansen 1981), but found low-

density BAC gels for the separation layer were more fragile than their standard BIS counterparts. For the high-density region, we found that the gels could not be dissolved, a result Hansen also reported in his work. That said, there are many other dissolvable polymer chemistries, such as DHEBA, that might improve device performance.

We explored the use of a Phi-29 MDA whole-genome amplification, but found it was not necessary, as we were able to recover sufficient starting material from our target input of approximately 100 cells for the Nextera XT protocol. A preamplification step before library prep could be added either before or after separation. This could potentially reduce the required cell input, but scaling down cell inputs in our experiments introduced substantial inconsistencies in performance, most likely due to a large coefficient of variation in input when attempting to load small numbers of cells. Even with pre-amplification, we suspect that this issue would hamper meaningful comparisons between samples. Alternatively, recent work has shown that with optimization of lysis conditions, high-quality sequencing libraries can be prepared directly from single cells using Tn5 without pre-amplification (Zhan et al. 2017).

Although the protocol we adapted from Smart-Seq relies on a poly-T primer, we also added primers with random binding sequences early in our experiments in an attempt to improve performance based on previous work on RNA sequencing from nuclei. We saw no effect, but kept the protocol unchanged for consistency.

As Gel-seq relies on hydrogel immobilization of sample material, it offers interesting possibilities when applied to new methods, such as the potential to change buffer between incompatible protocols without loss of sample material, or to amplify material inside the gel before attempting to extract. Future work in both device fabrication and protocol development could decrease input into the single cell range. A very recent publication from Adam Abate's

group shows that single bacterial cells can be encapsulated in agarose hydrogels and uniquely barcoded, allowing 50 000 single-cell whole-genome libraries to be generated in a few hours (Lan et al. 2017). The fundamental concepts of separation and library preparation demonstrated in Gel-seq via bulk-scale 100 to 1000 cell experiments are also relevant at the single-cell level, and many of the challenges that we faced in developing Gel-seq likely also apply at smaller scales. We believe that the solutions we present in this manuscript are a valuable resource for future work in single-cell genomics using hydrogels.

Since Gel-seq does not require a poly-A tail to achieve separation, it is also uniquely positioned for microbial studies, as prokaryotes typically do not polyadenylate their coding transcripts. A modification to the library prep would be required, as we relied primarily on a poly-T Smart-Seq primer, but Gel-seq benefits from an inherent flexibility in terms of different biochemical approaches. Gel-immobilized material can be washed or transferred, for example, into buffers suitable for either a poly-A tailing step or some other total RNA prep method, as long as RNAses are inhibited.

As for input, with microbial studies it might not be necessary to start with the same total mass of DNA as with mammalian genomes. While typical bacteria have only about 0.1% the nucleic acid content of mammalian cells, this also means that far less sequencing effort is needed to reconstruct either the genome or transcriptome. Previous work in the Zhang lab has shown 90% complete de novo assembly from a single E. coli bacterium after MDA pre-amplification in 12 picoliter PDMS microwells.1 Even one million paired end 100-base reads yields 200 million bases, which, for a single E. coli with 6 million bases total, gives 33× coverage. Assuming sufficiently uniform coverage, this is enough reads to perform de novo assembly. Even the smallest visible colony of E. coli that a researcher might pick from a plate

using a toothpick may contain more than enough material for Gel-seq. The question that remains to be answered is what amount of material is irrecoverable from the gel barrier. We suspect that the amount of irrecoverable material is likely a function of surface area. Reducing the device geometry to suit a toothpick sized sample might achieve the same goal as preamplification when working with microbes.

We have shown in this paper that Gel-seq can be used to generate high quality libraries from vanishingly small populations of cells. It is a flexible protocol that can be used to quickly process samples with an inexpensive and easy-to-fabricate device. The development of a gel based method for preparing next-generation DNA and RNA sequencing libraries from the same cells opens news doors for genomics, allowing researchers to ask if DNA mutations in small numbers of cells affect RNA expression in those same cells. It is also our hope that the physical principals described here might someday be translated to a single-cell technique to allow simultaneous profiling of tens of thousands of single-cell genomes and transcriptomes. Such a device would provide a more general approach for linking DNA variation to RNA expression in complex samples such as tumors or microbial populations.

1.5 Appendix to Chapter 1

1.5.1 Additional Experimental Techniques

Positive Control Library Preparation: Standard protocols were used to generate reference libraries as a comparison to our Gel-seq protocol. To generate libraries from RNA, we followed the Smart-Seq and Nextera XT manuals (Clontech 2016; Illumina 2015). The only modification we made to these protocols was to use half volume reactions and the addition of random priming to the reverse transcription step of Smart-Seq. To generate libraries from genomic DNA, we lysed cells using a simple lysis buffer developed by Shatzkes (Shatzkes, Teferedegne, and Murata 2014). Once the cells were lysed, we followed the standard Nextera XT manual using half volume reactions (Illumina 2015).

Cell Culture: PC3 was cultured in F-12K media (Gibco) supplemented with 10% heat-inactivated (HI) FBS (Gibco) and 1% penicillin/streptomycin (P/S) (Gibco). HeLa was cultured in Eagle's Minimum Essential Medium (ATCC) supplemented with 10% HI FBS and 1% P/S. 3T3 was cultured in high-glucose Dulbecco's Modified Eagle's Medium (4.5 g/L glucose and L-glutamine) supplemented 10% HI FBS and 1% P/S and 3T3 cell lines were cultured in DMEM with 4.5 g/ml glucose and 1 mM sodium pyruvate (ThermoFisher Scientific) supplemented with 10% heat-inactivated FBS (ThermoFisher Scientific) and 1% penicillin-streptomycin (ThermoFisher Scientific).

Mouse Primary Hepatocyte Collection: Mouse livers were perfused with a classic two-step method. Briefly, livers were perfused via the portal vein with 20 ml of pre-warmed wash buffer followed by 20 ml of digestion buffer containing 5000 U collagenase Type IV (Gibco) and 5000 U collagenase type I (Worthington). After perfusion, tissue was cut as small as possible, passed through 100-µm cell strainer, and centrifuged at 50g for 5 min to pellet

hepatocytes. The animal protocols (s09108) for all procedures were approved by the UCSD Institutional Animal Care and Use Committee (IACUC). All methods were performed in accordance with the relevant guidelines and regulations.

1.5.2 Analysis of Sequencing Data

Sequencing and De-multiplexing: Libraries were sequenced on a MiSeq (Illumina) using v3 kits and standard sequencing primers. Libraries were loaded at 27-30 pM and at least 50 cycles were obtained for read 1 for each experiment, plus 8 cycles for Index 1 and 8 cycles for Index 2. Base calls were de-multiplexed to fastq using bcl2fastq.

Extrapolation Simulations: Library complexity and genomic coverage simulations were performed with preseq (Daley and Smith 2013) using extrac_lc and extrap_gc, respectively, with 100 bootstrapping iterations each.

DNA Mapping and CNV Calling: Copy number profiling on DNA libraries was performed as described in Baslan et al. with minor modifications (Baslan et al. 2012). Briefly, reads were trimmed to 36 bases using fastx (14 bases from the start and all bases after 50) and mapped to GRCh38 or mm10 with bowtie (Langmead et al. 2009). For both mouse and human, alignments were counted across 25,000 bins whose boundaries were calculated such that mapping their respective reference genomes would generate equal counts per bin. Bin counts were then normalized to mean for each sample and GC corrected in matlab by lowess regression based on GC content. No segmentation was performed. Pearson correlations were performed in Python using scipy (Jones et al. 2001) and plotted using matplotlib (Hunter 2007).

RNA Mapping and PCA: RNA fastq was mapped to a pre-annotated index constructed from either GRCh38.87 (human, hg38) or GRCm38.87 (mouse, mm10) using STAR (Dobin et

al. 2013) and read counts for each gene were converted to TPM. A threshold of TPM > 5 was applied. Pearson correlations and PCA were performed in Python using scipy (Jones et al. 2001) and scikit-learn (Pedregosa et al. 2011), respectively, and plotted using matplotlib (Hunter 2007). Base-wise percentage of reads mapping to exons was calculated from GenCode GTF files using bedtools coverage. Base-wise percentage of reads mapping to ribosomal and transfer RNA was calculated from Ensembl BioMart BED files, also using bedtools coverage. Percentage of RNA reads mapping to mitochondrial genes was calculated from alignments using grep to search for the MT in the chromosome field.

Sequencing pool	Type	Sample
A	DNA	3T3-G1
		3T3-G2
		Liver-G1
		Liver-G2
		3T3-T1
		3T3-T2
		Liver-T1
		Liver-T2
B	RNA	3T3-G1
		3T3-G2
		3T3-T1
		3T3-T2
C	RNA	Liver-G1
		Liver-G2
		Liver-T1
		Liver-T2
D	DNA	HeLa-G1
		HeLa-G2
		HeLa-T1
		HeLa-T2
		PC3-G1
		PC3-G2
		PC3-T1
		PC3-T2
E	RNA	HeLa-G1
		HeLa-G2
		HeLa-T1
		HeLa-T2
		PC3-G1
		PC3-G2
		PC3-T1
		PC3-T2

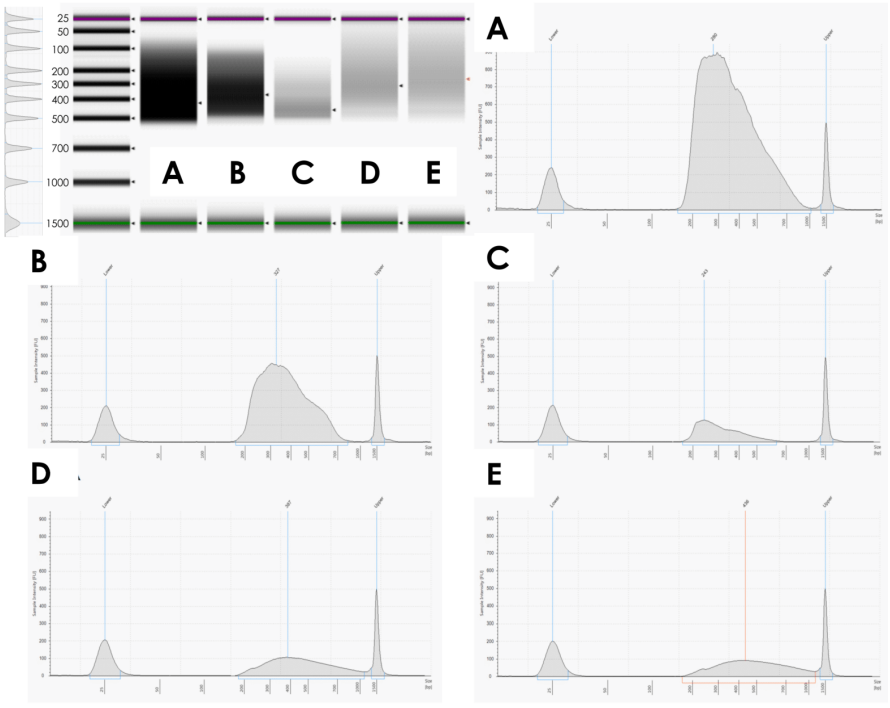

Figure 6: TapeStation traces for all sequencing library pools.

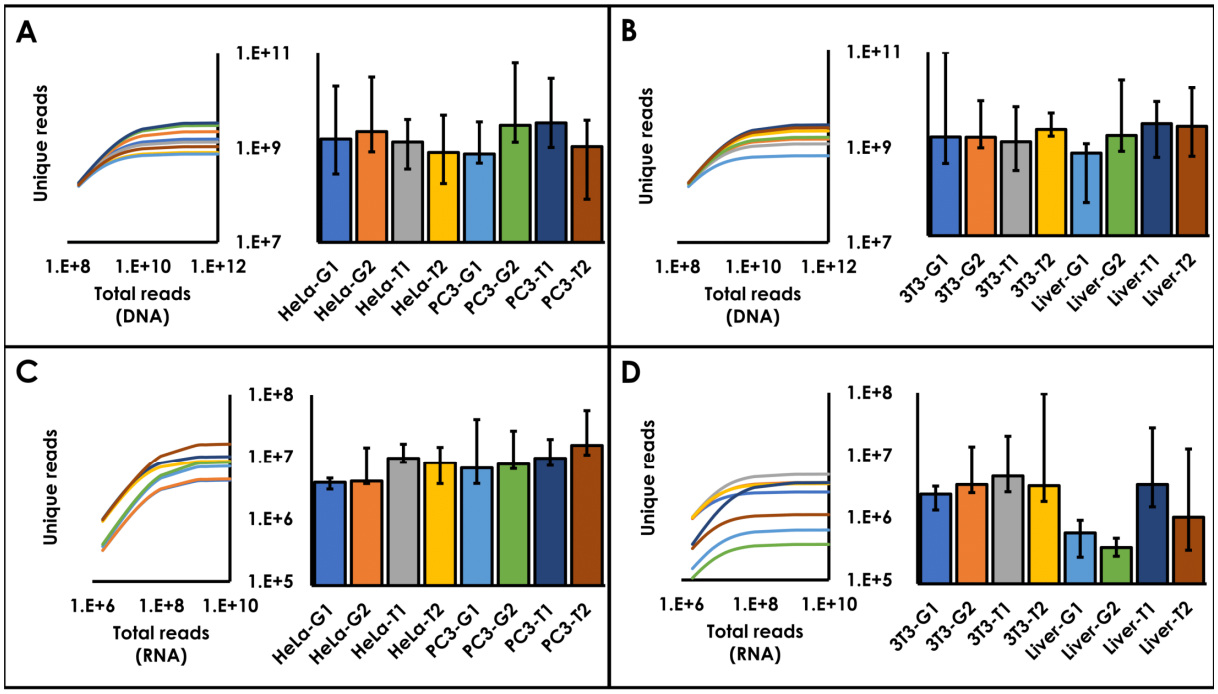

Figure 7: Panels A and B show predicted genomic coverage as a function of depth of coverage in DNA libraries from human and mouse samples, respectively. Bar graphs represent maximum predicted coverage at saturation. Panels C and D show predicted library complexity for RNA libraries from human and mouse samples, respectively. Error bars are 95% confidence intervals created from 100 bootstrapping simulations.

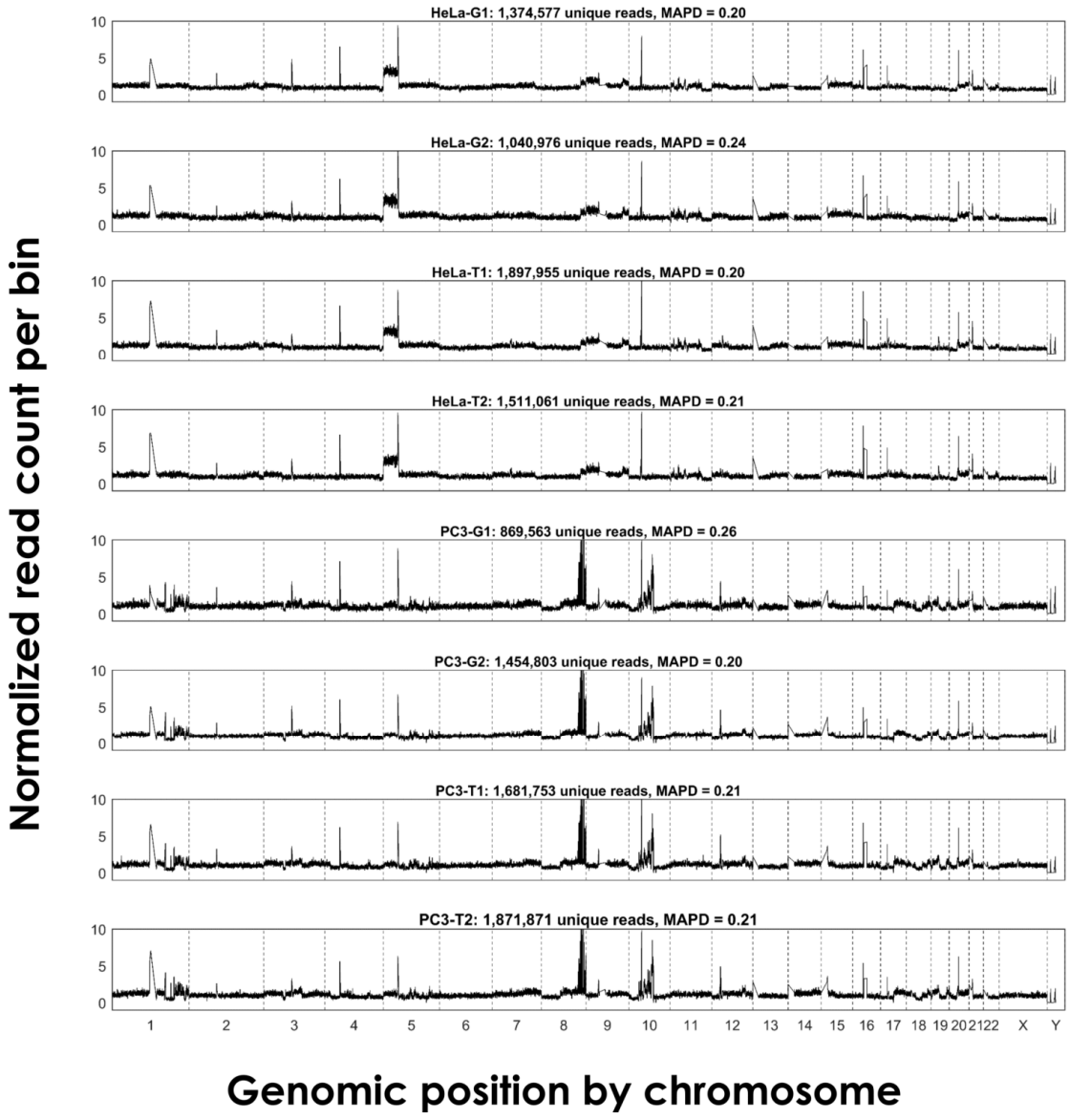

Figure 8: Copy number profiles across 25,000 bins for human DNA libraries from HeLa and PC3 cell lines. G and T in sample names indicate Gel device and Tube controls, respectively, while numbering indicates technical replicates. Horizontal axis denotes chromosome and bin position, vertical axis denotes mean normalized bin count (corrected for GC content).

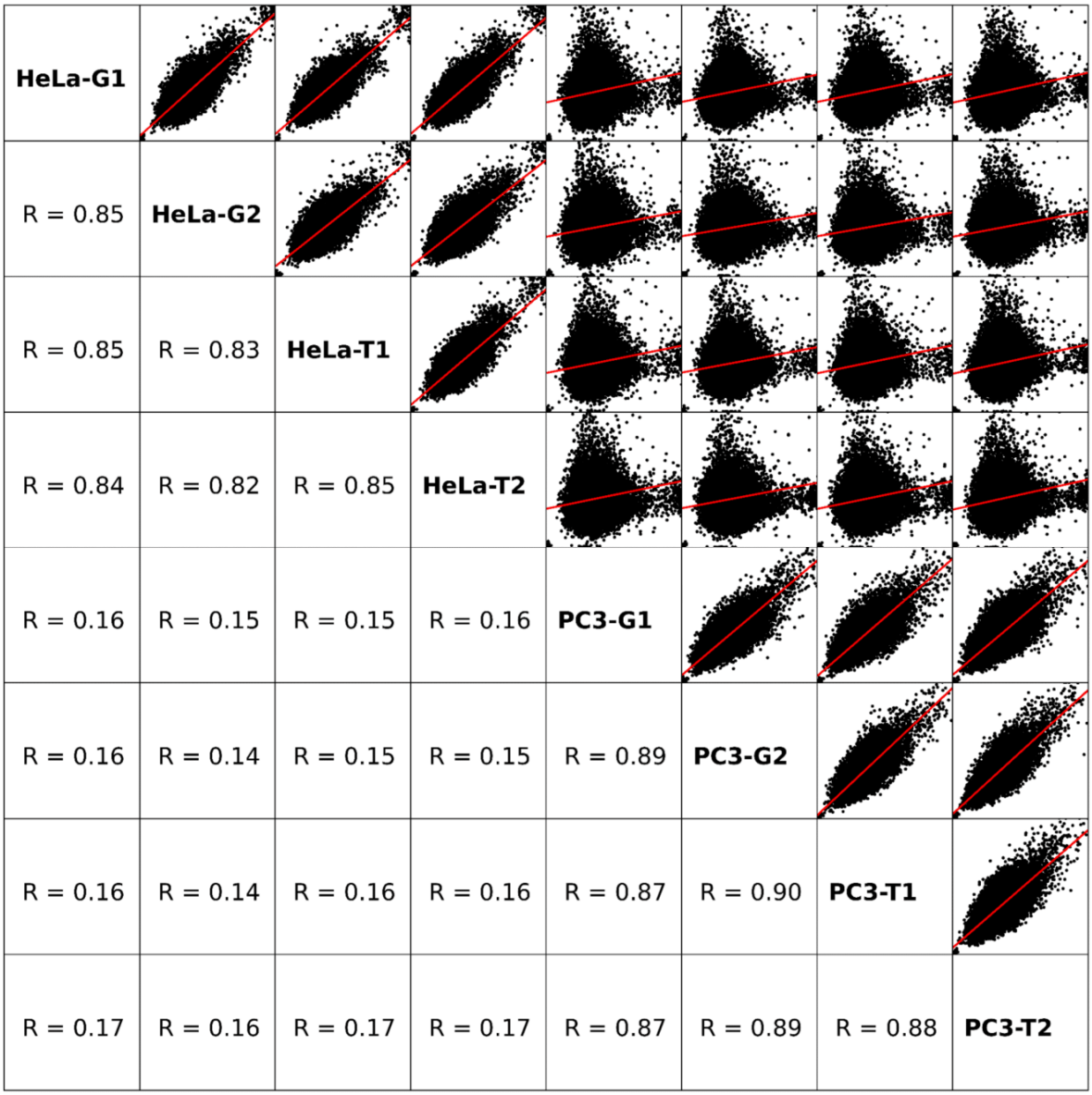

Figure 9: Pairwise correlations between bin counts for human DNA libraries from HeLa and PC3 cell lines. Main diagonal entries are sample names, lower diagonal entries are Pearson correlation coefficients.

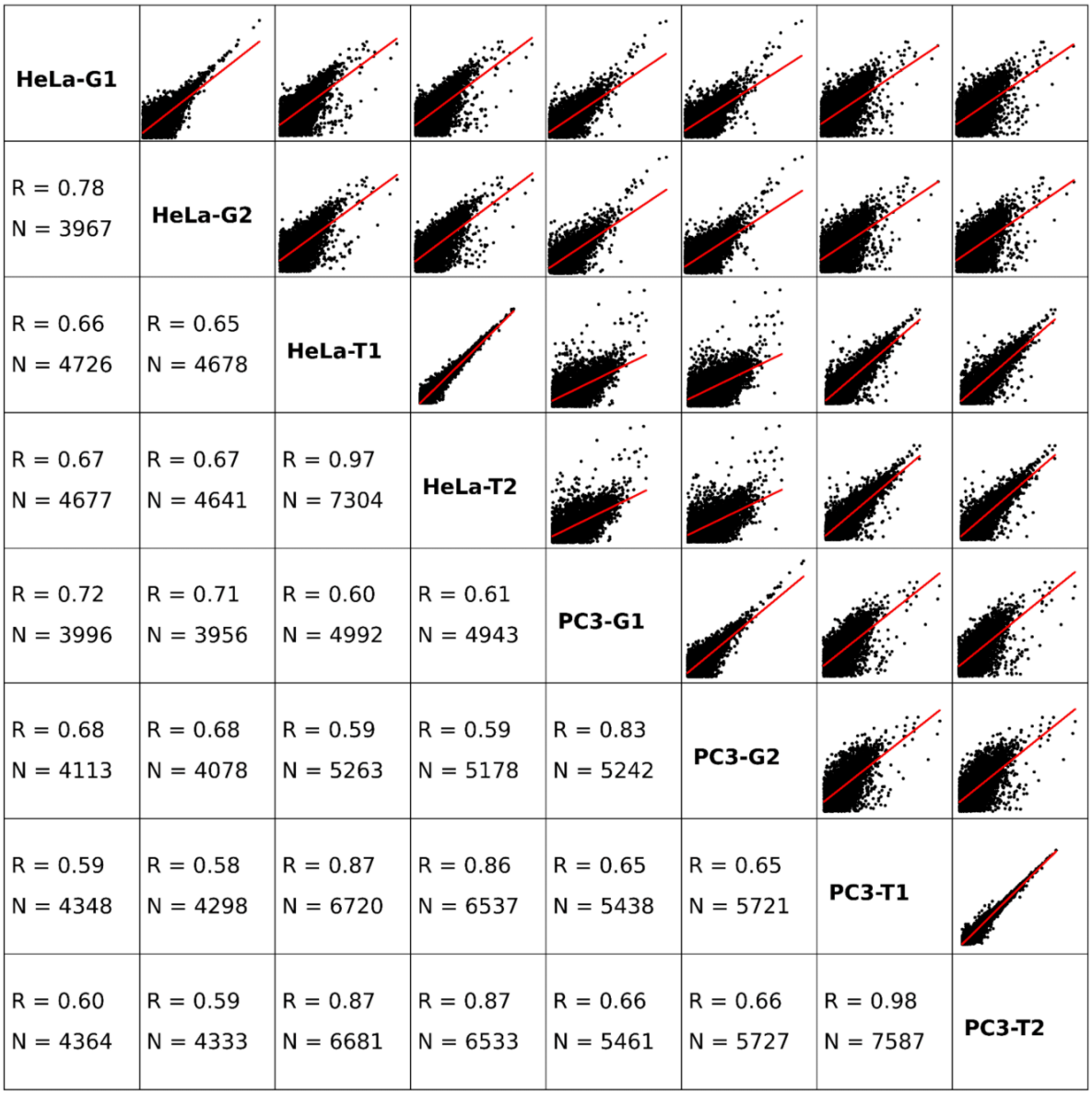

Figure 10: Pairwise correlations between detected gene counts for all human RNA libraries from HeLa and PC3. Lower diagonal entries are Pearson correlation coefficients (R value) and number of detected genes (N values, genes with non-zero counts) in common for each comparison.

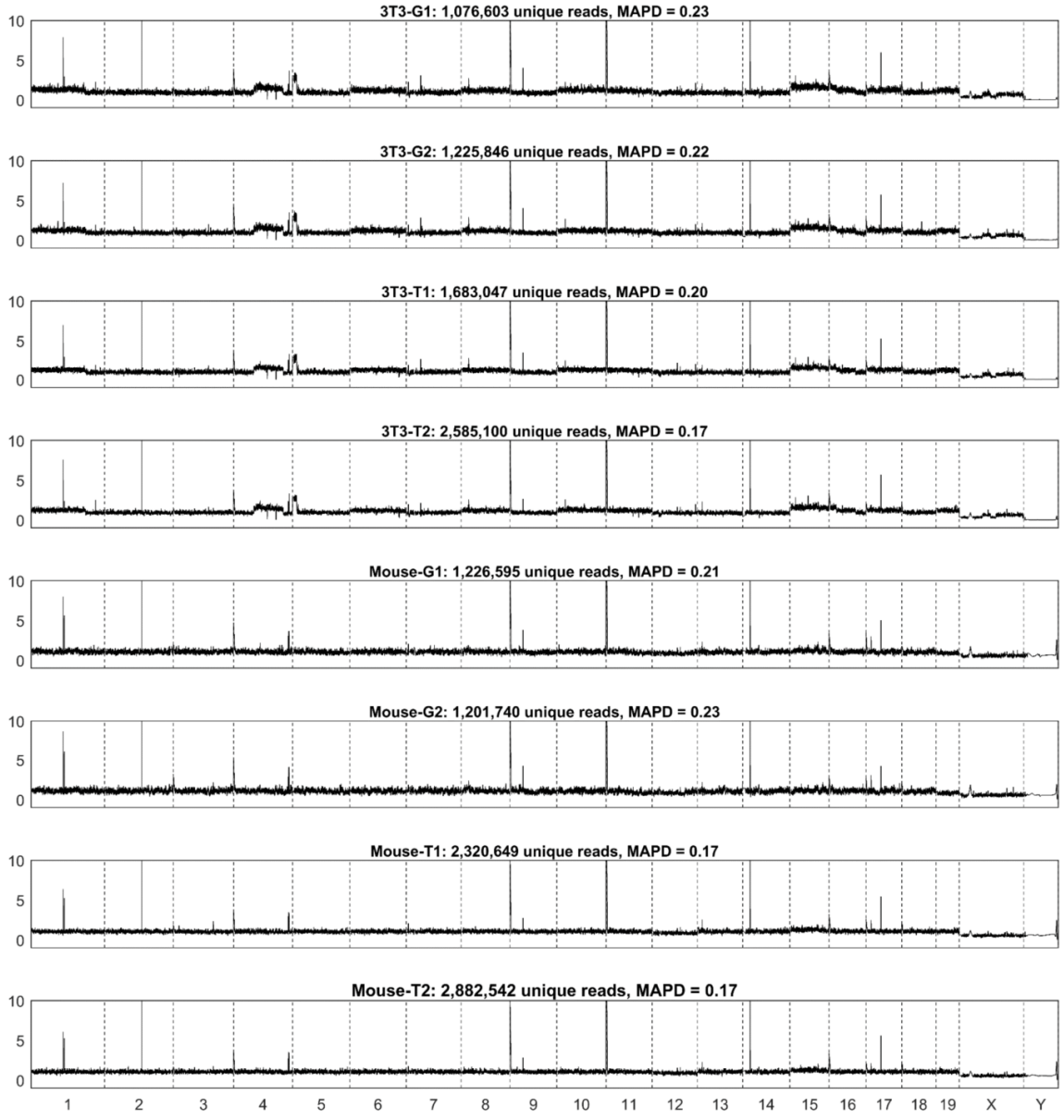

Figure 11: Copy number profiles across 25,000 bins for mouse DNA libraries from a 3T3 cell line and mouse primary tissue. G and T in sample names indicate Gel device and Tube controls, respectively, while numbering indicates technical replicates. Horizontal axis denotes chromosome and bin position, vertical axis denotes mean normalized bin count (corrected for GC content). Extreme peaks in mouse primary samples are due to "bad bins" in the reference genome, in which repetitive sequences present in the true genome are not included in the reference genome, leading to a false pile up of reads from experimental data.

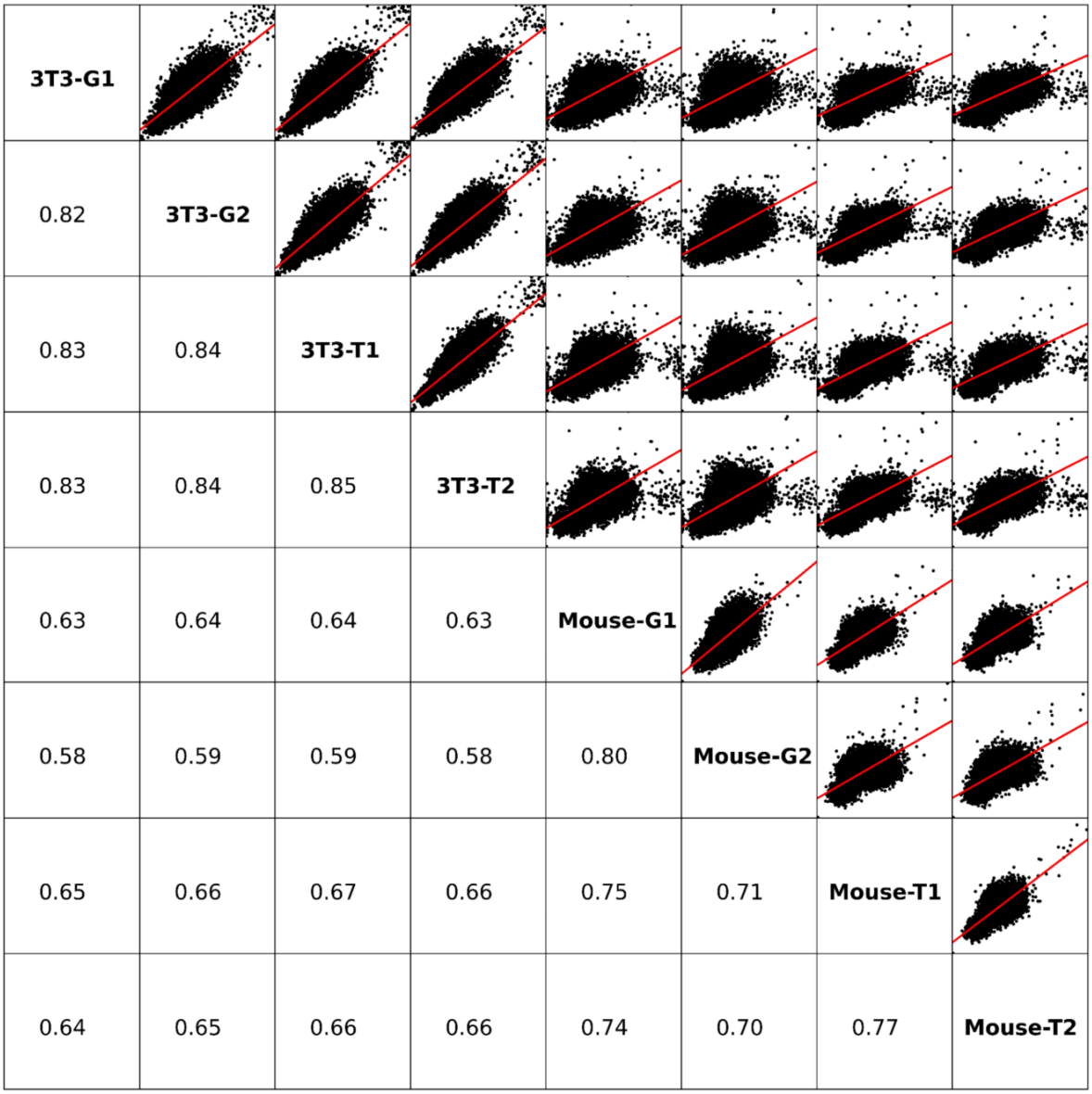

Figure 12: Pairwise correlations between bin counts for mouse DNA libraries from 3T3 and mouse tissue. Main diagonal entries are sample names, lower diagonal entries are Pearson correlation coefficients. Correlations are weaker than for PC3 versus HeLa due to less extreme copy number variation in 3T3 and almost none in the mouse primary sample, leading to little dynamic range in bin counts.

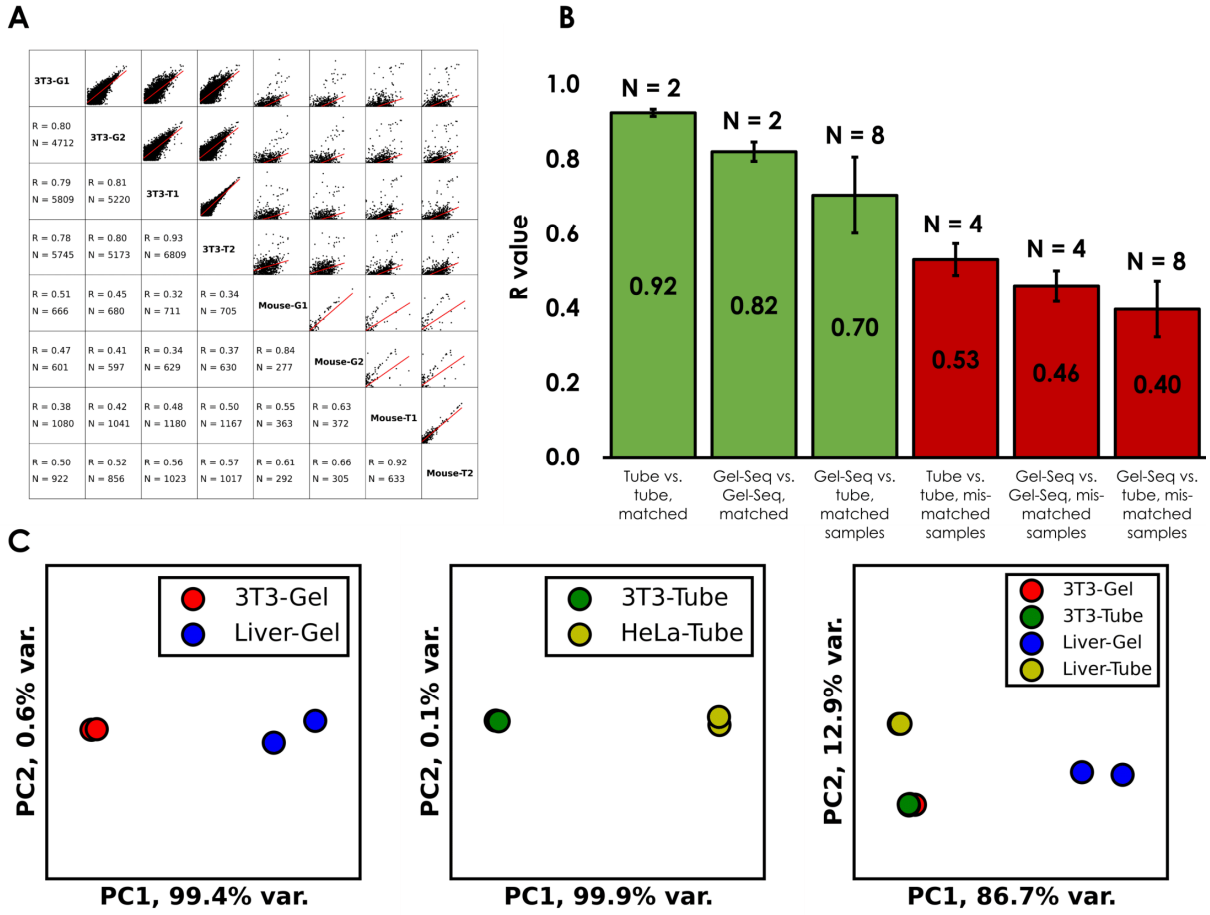

Figure 13: Pairwise correlations and PCA for Gel-seq and tube samples. Panel A shows all 28 pair-wise correlations between gene counts (TPM > 5) for RNA libraries from mouse 3T3 fibroblast cell line and mouse liver. Lower diagonal entries are Pearson correlation coefficients (R value) and number of detected genes (N values, genes with non-zero counts) in common for each comparison. Panel B shows average R values for different comparison types (Error bars are standard deviation, N is number of comparisons in each type.) Panel C shows PCA separation for Gel-seq, standard tube method controls, and all 8 together.

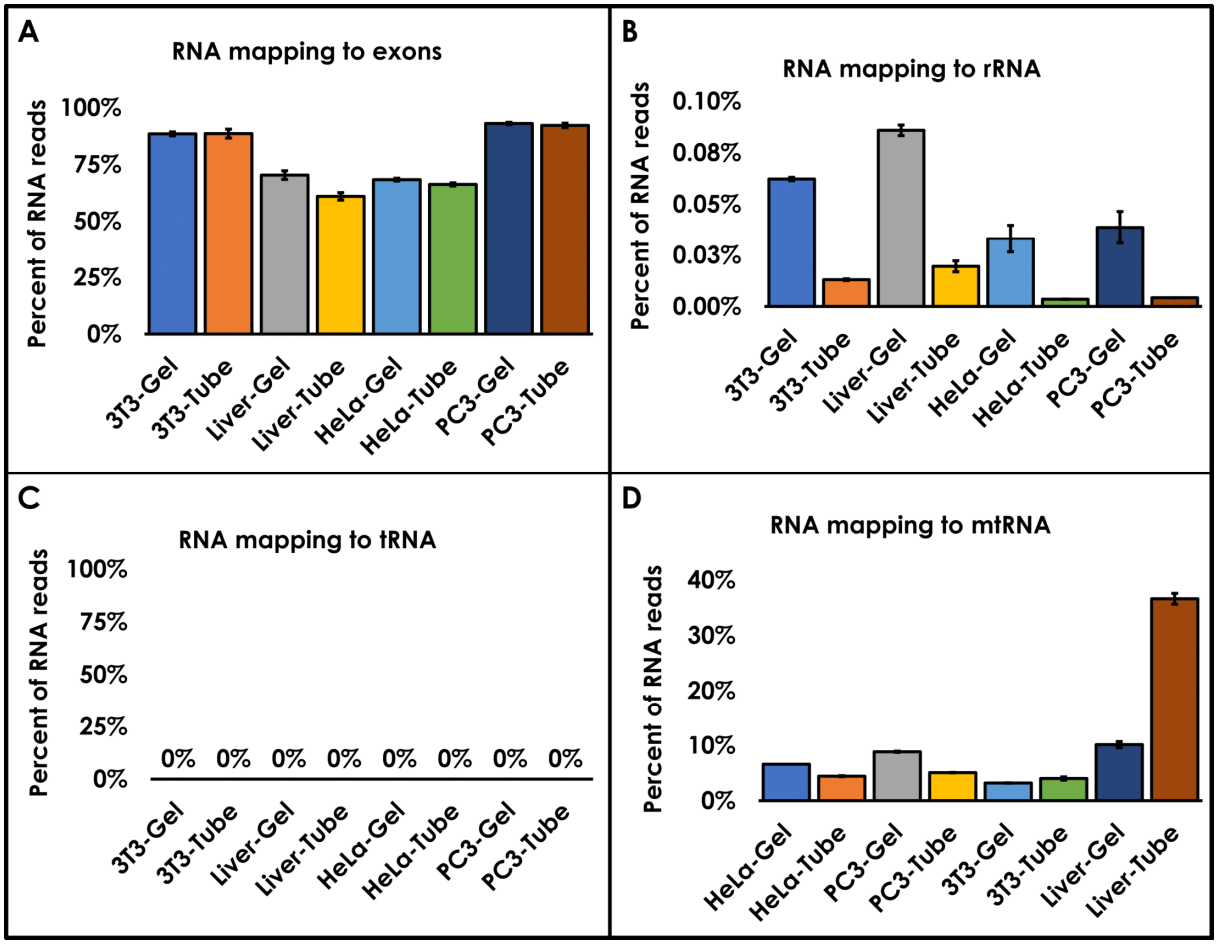

Figure 14: RNA mapping features. Panel A shows base-wise percentage of RNA alignments in annotated exons in either GRCh38 or mm10 reference genomes. Only mouse liver samples show any significant difference, with Gel-seq data mapping to exonic regions at a slightly higher rate ($p = 0.03$, two-tailed t-test, unequal variance). Panel B shows base-wise percentage of RNA alignments in ribosomal RNA genes (rRNA). There is evidence that polyadenylation of rRNA acts as a degradation signal, 11 and short rRNA degradation fragments (less than 1000 bases) would be expected to migrate faster than the majority of mRNA, which might explain why Gel-seq detects a higher proportion of rRNA than tube controls. Panel C shows base-wise percentage of RNA alignments in transfer RNA genes (tRNA). tRNA is short and not polyadenylated, which likely explains why we see no mapping when using the Smart-Seq poly-T primers. It seems that the random priming also failed to detect tRNA in either Gel-seq or tube controls, possibly due to the short length of tRNA. Panel D shows the percentage of RNA reads mapping to the mitochondrial chromosome. Hepatocytes contain very large number of mitochondria, so a high RNA mapping rate to mitochondrial genes is not necessarily surprising in liver, although we cannot fully explain why Gel-seq detected less than the tube controls in this comparison.

1.6 Acknowledgement for Chapter 1

Chapter 1, in full, is a reprint of the material as it appears in Lab on a Chip (Royal Society of Chemistry) (Hoople, Gordon D.*, Andrew Richards*, Yan Wu, Kota Kaneko, Xiaolin Luo, Gen-Sheng Feng, Kun Zhang, and Albert P. Pisano. 2017.). The dissertation author was a primary author of this paper.

CHAPTER 2. DEVELOPMENT AND PROOF-OF-CONCEPT OF A MICROFLUIDIC HYDROGEL ENCAPSULATION TECHNOLOGY FOR SINGLE-CELL WHOLE-GENOME SEQUENCEING LIBRARY PREPARATION

2.1 Abstract of Chapter 2

Although genomic mosaicism has been shown to occur in the human brain in the form of copy number variations (CNVs), changes occur less frequently on a per cell basis than in tumors. Rarer variants require larger number of cells to accurately determine the underlying distribution, and existing whole-genome sequencing library preparation methods are limited in throughput. We designed a microfluidic device to encapsulate single neuronal nuclei in hydrogel droplets to facilitate combinatorial library prep of a thousand neurons in a single two-day experiment with no special equipment necessary, e.g., a flow cytometer or commercial microfluidic system. We showed proof-of-concept by mixing mouse and human cells and demonstrating strict mapping specificity with very shallow depth of 87 cells, with potential for scaling into the many thousands. The copy number profiles generated agreed with ground truth observations from down-sampled bulk sequencing libraries of the same cell lines. This technology will not only enable the unbiased copy number characterization of human neuronal genomes, but can also be applied to tumor and microbiome profiling.

2.2 Introduction

Single-cell RNA-seq especially has experienced a revolution in the last three years. Experimental techniques have increased throughput from the order of tens of cells prior to 2015 (cite) to many thousands (Macosko et al. 2015; Klein et al. 2015) per run, with over a hundred

thousand cells per run in some recent publications (Cao et al. 2017; Rosenberg et al. 2018) with demonstrated potential for scaling well into the hundreds of thousands. These changes came very quickly, and were enabled largely by two key concepts: Large sequencing barcode spaces created by combinatorial indexing and novel methods of physical compartmentalization.

There are obvious parallels between the design goals for developing single-cell genomics, transcriptomics, epigenomics, etc., such as the need to minimize cost and time per cell, achieve sufficient throughput per sample, and avoid the introduction of bias in the data, but the study of genomics entails specific technical challenges that have held the field back compared to epigenomics and transcriptomics. Subsequently, most highly-parallel single-cell methods have focused on targets such as RNA or methylation patterns, while whole-genome sequencing (WGS) of DNA from single cells has seen much less of the spotlight. Much of this discrepancy is simply due to the fact that mammalian genomes are so large and incompletely annotated, which creates two substantial problems: First, it is technically challenging to generate even coverage; and Second, the majority of variants called are likely to have unknown significance. In terms biologically meaningful inferences, therefore, modern day single-cell WGS is relatively information poor compared to RNA-seq, with a correspondingly lower ratio of "signal-to-sequencing-dollar."

A second factor that has held back single-cell WGS is the technical challenge of accessing the genomic DNA (gDNA) of a single cell for library prep. DNA in mammalian systems is heavily protected by both the double-membrane of the nucleus as well the protein components of chromatin, primarily composed of millions of nucleosome complexes, an octameric assembly of 4 different histone proteins connected by linkers, which are spaced about 121 bases apart in closed chromatin.

Our group has previously demonstrated an approach called MIDAS (Gole et al. 2013) to overcome some of these limitations, whereby a very large number of 12 nL PDMS microwells are loaded randomly with a limiting dilution of cells, such that an average of only one in ten wells is loaded with a very low rate of multiplets. Our lysis method used alkaline conditions to lyse membranes and denature proteins, which can then be neutralized by an acidic solution. The major objective at the time was reducing the random bias incurred during multiple displacement amplification (MDA) by limiting the size of the reaction, which we accomplished, although with limited throughput. Figure 20 shows a fluorescent image of the microwells with neuronal nuclei deposited, and illustrates some of the difficulty of this approach in practice. Two years of continuous experimentation yielded nearly 200 single-cell libraries from human cortical neuron, yet only 60 of these passed quality control filters suitable for CNV calling. Of these, 27 cells were from 3 AD patients, a further 27 came from 3 ND patients, and 6 came from a DS patient.

To address this limitation, we explored a commercial microfluidic system using a variety of custom protocol. The automatic loading, known locations of capture site, and automated harvesting promises higher throughputs, but high complexity of valve-based microfluidics places a cap on scalability. Furthermore, while there is some flexibility in the types of reagents used, the geometry is essentially fixed, which limits the extent to which this approach could be adapted for future assays. We ultimately found that the method did not scale adequately, and after a year of protocol optimization we could not match the data quality of MIDAS.

Motivated by the above challenges and limitations, we subsequently developed a hydrogel droplet encapsulation technique using an in-house designed and fabricated

microfluidic device to prepare a thousand single-cell libraries for parallel sequencing in a single run. As shown in Figure 15, cells or nuclei are lysed on-chip to release DNA into the droplet such that the solution can be well-mixed by the serpentine outlet. This ensure a homogeneous distribution of molecules throughout the resulting hydrogels, which we have observed to be critical to the efficient capture of DNA by entangling in hydrogel, thorough removal of proteins and lipids during washing, and uniform accessibility of the entangled DNA. Approximately 1000 droplets can be produced per second at a monodisperse diameter of 60 um (volume = 113 pL). Samples are loaded at limiting dilution, such that nine out of ten droplets are empty, one in ten are loaded, and a low rate of loaded droplets contain more than one cell. The hydrogels are allowed to polymerize overnight, after which the emulsions are broken and hydrogel beads are recovered into an aqueous phase and washed extensively. Remaining protein complexes, such as nucleosomes, are then enzymatically digested and washed out. Library prep is performed on groups of ~2000 PA bead encapsulated nuclei per well in a plate format with combinatorially barcoded transposon oligos enzymatically inserted by Tn5 transposase.

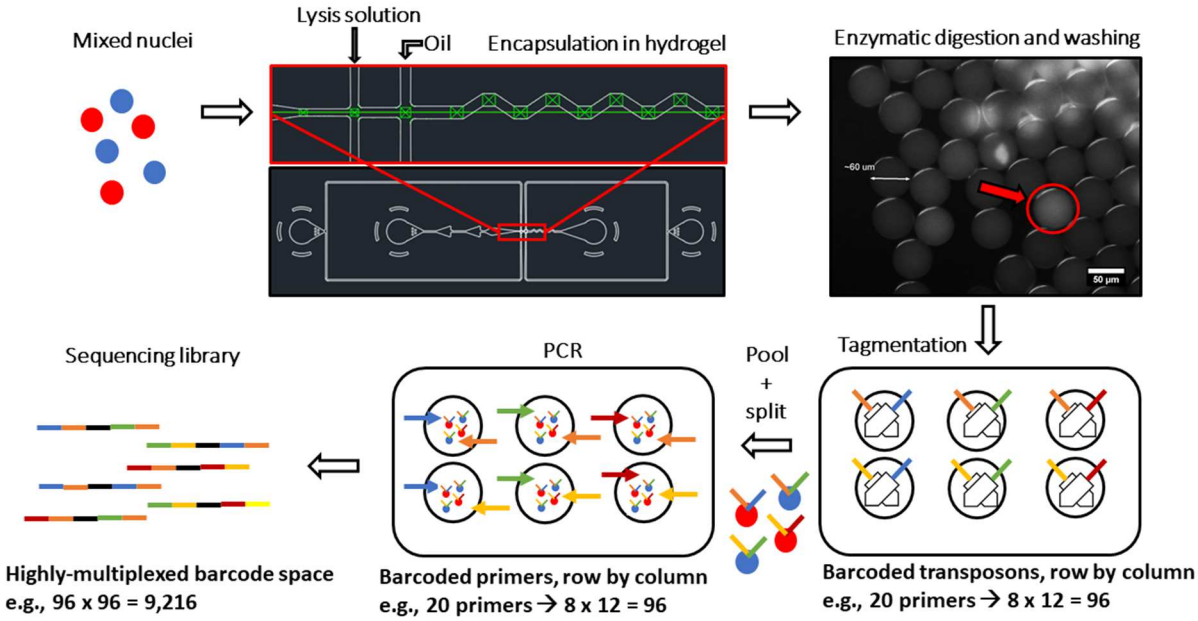

Figure 15: Experimental workflow. A mixed population of cells or nuclei are encapsulated in polyacrylamide (PA) hydrogel droplets using a polydimethylsiloxane (PDMS) an in-house designed microfluidic droplet generator.

2.3 Results

2.3.1 Validation of single cell genomic compartmentalization

Cross-mapping between mouse and human is handled by aligning to a combined genome and setting a mapping quality threshold for each read to ensure a high rate of unique alignment. Sequencing barcodes corresponding to both mouse and human DNA appear clearly in the plot away from the axis, whereas orthogonal mapping against the axis indicates single-cell data. In this case, the collision rate can be calculated by observing the events along the diagonal and considering that the total cross-cell collision rate is equal to twice the cross-species collision rate (possible collisions are human-human, mouse-mouse, human-mouse, and mouse-human), as same-species collisions will still appear along the axis (Macosko et al. 2015). A total of approximately 640 nuclei extracted from a human HeLa cervical cancer line and mouse 3T3 fibroblasts were pre-mixed at a 50/50 ratio before polyacrylamide encapsulation and

subsequently sequenced on a single MiSeq at approximately 9 million reads. A cut-off of 20,000 reads yielded 27 HeLa and 60 mouse 3T3 libraries (Fig. 16, panel C).

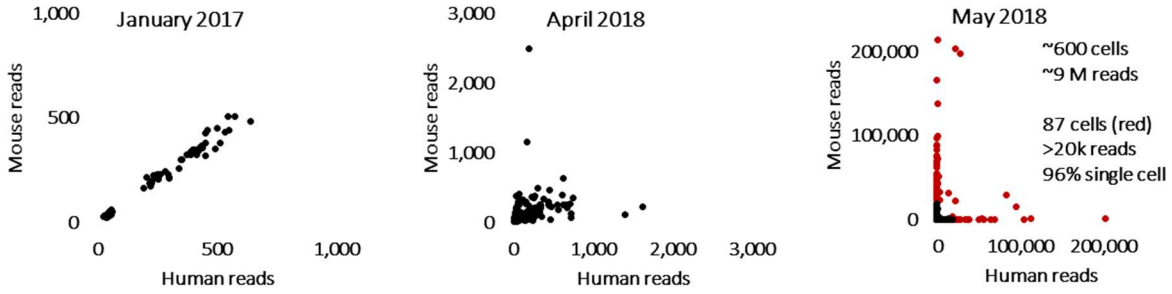

Figure 16: Improvement of mapping orthogonality over time. Mixing mouse and human cell lines prior to the experiment enables a quantitative estimate of the degree cross-contamination between single cells.

2.3.2 Genome-wide copy number uniformity

The next step was to investigate the evenness of coverage to ensure that the hydrogel matrix encapsulating each single genome did not interfere excessively with Tn5 based library preparation. Figure 17 shows genome-wide averages for both bin counts and bin-wise integer copy number estimated using the CSHL varbin circular binary segmentation (CBS) method (Baslan et al. 2012) for 27 HeLa and 60 3T3 cells. Bulk libraries with approximately 2 million reads each for HeLa and 3T3 were downsampled many times to generate distribution of counts from a perfectly uniform genome for each cell line. These downsampled bulk represent a control for each line, showing the "true" copy number profile for comparison to the single cell libraries, assuming a homogenous population in culture.

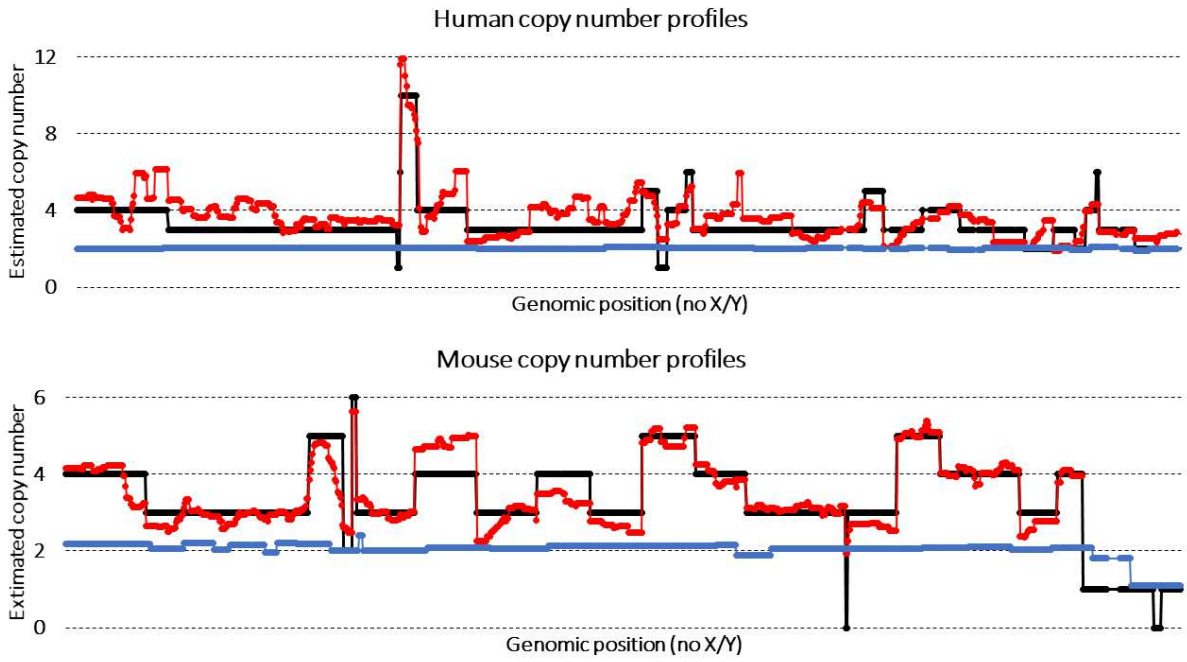

Figure 17: Genome wide coverage uniformity of bin counts for both mouse and human. Reference genomes were binned into 1,000 variable length regions such that each region contained the same number of mappable reads. Bulk libraries shown here are downsampled to 20,000 reads.

While the single cells show more noise on a bin-by-bin basis compared to the bulk, there are visual similarities in called regions which indicates that there could still be signal detectable above noise.

2.3.3 Correlation of cell line copy number profiles

To determine whether the method in its current form can detect signal above noise, we tested whether we could distinguish between cell lines based on copy number alone. Many cell lines, including HeLa and 3T3, are marked by distinct amplifications and deletions are large regions compared to their reference genomes. This dynamic range in copy number on a per-region basis allows for a correlation metric to be calculated for each pair of cells on a bin-wise basis. Pearson correlation coefficients were calculated for all 27 HeLa single-cell copy number

profiles compared to down-sampled bulk libraries for both HeLa and human pre-B lymphoblast line GM12878. Figure 18 shows a scatter plot of correlation coefficients, with 26 out of 27 (96%) of HeLa cells correctly assigned to HeLa reference based on copy number profile, indicated by their position above a diagonal indicating random assignment between cell lines.

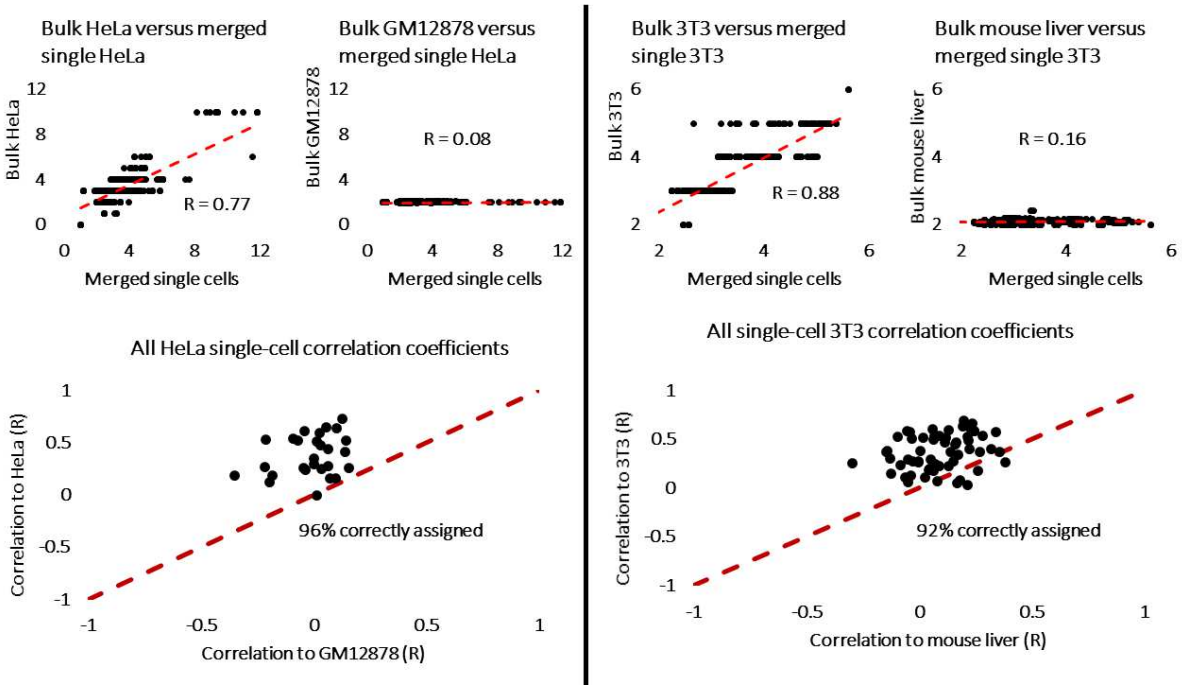

Figure 18: Pearson coefficients for HeLa cells correlated to bulk HeLa versus bulk GM12878.

Figure 30 in the appendix shows similar scatter plots for HeLa cells correlated to HeLa versus 3T3, as well as similar plots for 3T3 cells. Panel A shows HeLa cells correlated to bulk HeLa versus bulk 3T3, with 26 out of 27 (96%) of HeLa cells correctly assigned to HeLa. Panel B shows 3T3 cells correlated to bulk 3T3 versus bulk GM12878, with 55 out of 60 (92%) of 3T3 cells correctly assigned to 3T3. Panel C shows 3T3 cells correlated to bulk 3T3 versus bulk Hela, again with 55 out of 60 (92%) of 3T3 cells correctly assigned to 3T3.

Genome wide copy number profiles were also correlated on a bin-wise basis between all single cells for both HeLa and 3T3, as well as nine bulk sequencing control libraries down-

sampled at varying depths (200,000, 20,000, and 2000 reads) each for both HeLa and 3T3. Figure 31 in the appendix shows that manually sorted single cells by either read depth or noise level (MAPD) did not create any apparent pattern in the heatmap. Figure 32 in the appendix shows a similar map of all pair-wise cell-to-cell correlation coefficients after clustering, indicating that more than 90% of single cells clustered accurately within their respective cell lines along with the bulk controls.

The down-sampled bulk samples in from the two different cell lines have an average correlation of ~0.26, which provides a useful cutoff for correlation as there is no expected relationship between the copy number profiles of HeLa and 3T3. Much of the non-agreement between cells within each line and the down-sampled bulk are likely due to random noise. Only about half of cells in each line have an R of greater than 0.5 when compared to the down-sampled bulk, although the fact that clustering was accurate indicates that there was still signal detected above noise. Further optimization of the protocol will likely noise to levels similar to those seen in another recent hydrogel encapsulation, Tn5-based whole-genome sequencing approach using bacteria (Lan et al. 2017), discussed below. However, some of the disagreement within cell lines could also be explained of sub-clonal populations present within the cell-to-cell correlations, similar to another single-cell library prep method using Mu transposase ((Xi et al. 2017), in which cells in culture developed copy number variations after many passages. If true, this could provide a useful and convenient model for study in the future development of the hydrogel approach described here, as detecting clonal sub-populations of cells is an application of interest for a whole-genome single-cell copy number profiling method.

2.4 Discussion

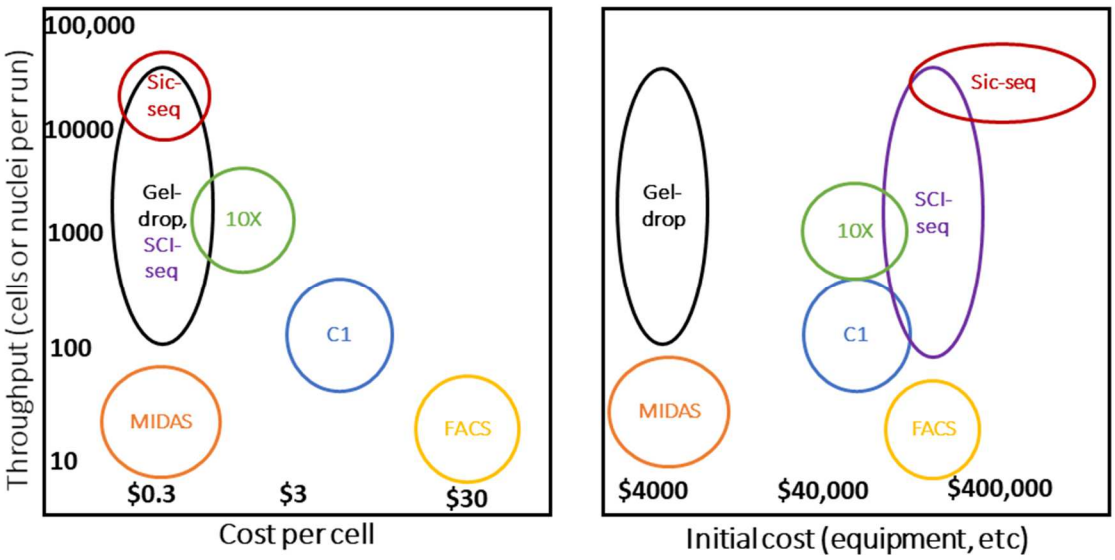

Figure 19: Comparison of scGel-seq to other whole-genome sequencing library preparation methods. FACS (yellow) refers to all methods which rely on single-cell sorting (e.g., sorting into 96- or 384-well plates) for compartmentalization.

Figure 19 shows a qualitative comparison of a variety of methods for single cell whole-genome sequencing library preparation, some of which are not yet commercially available. FACS-based methods have the longest history and highest degree of commercial availability, and can have very good coverage uniformity, but have high cost per cell due to large reaction volumes on a per cell basis. Cell sorters have also historically come with high associated costs, on the order of hundreds of thousands of dollars. While large institutions often have core facilities that can be shared across labs, reliance on FACS could be a barrier for smaller organizations. Finally, single-cell sorting can only be scaled linearly at best by increasing the number of collection plates, and the high cell losses typically incurred during sorting in single-cell purity mode limits the total number of cells able to be sequenced.

An advantage of a microwell array approach such as MIDAS is that sorting is not a technical requirement, meaning that any population of cells, nuclei, or even bacteria can be loaded into each array due to the low chance of a multiplet with such a large number of wells, although many of our mammalian samples had been FACS purified based on cell markers of interest prior to the experiment. While we demonstrated parallel amplification of hundreds of genomes in parallel, and successfully demonstrated a substantially superior coverage uniformity compared to competing methods at the time, the extraction and processing of each amplified genome was labor intensive and required a skilled operator to work through the day to collect 96 samples. After two additional days of processing, only about 25% of libraries would typically pass all QC standards, which included shallow depth sequencing to screen libraries before deeper and more costly sequencing. This led to a long data turnaround time for relatively few samples. We also realized that even substantial automation of the process would likely only increase throughput by less than an order of magnitude.

The Fluidigm C1 DNA-seq chip consumes reagents with substantially higher efficiency due to small reaction volumes, with a correspondingly lower cost per cell, and has nearly an order of magnitude higher throughput per run. The C1 also yields high coverage on a per cell basis, but suffers from high region-to-region coverage noise in our hands for the purpose of CNV calling. Furthermore, the chip requires expensive equipment for loading, running, and harvesting samples, requires the purchase of expensive licensing to run custom protocols to reduce noise, and any valve-based microfluidic compartmentalization approach presents an obstacle to scaling up. Fluidigm has announced an 800 chamber chip, but it is unclear how much farther it can be scaled.

Although some of the costs are still difficult to estimate prior to any commercial implementations of highly-multiplexed single-cell whole-genome methods, it is apparent from the techniques described in the literature and their demonstrated throughputs that the cost per cell prior to sequencing stands to be lowered dramatically, likely below 10 cents per cell. The recent announcement of a CNV calling platform from 10X genomics will be the first commercially available solution for generating approximately one thousand single cell whole-genome sequencing libraries, and the costs of their single cell RNA-seq applications (~$800 - $1000 per run) can be used as an estimate, which puts their approach somewhat in the middle at just under a dollar per cell. The 10X system also requires pricey a microfluidic station to run each chip, although the run time is short and the machine can be shared between many labs more easily than the C1, which has an overnight runtime.

The advantage of combinatorial split-pool methods is the ease of exponential scalability at low additional cost, but there are particular technical challenges in adapting such approaches to mammalian whole-genome assays due to the dense association in chromatin between DNA and proteins such as histones. There exists a need for a method of uniformly accessiblizing the entire genome, thus removing the endogenous structural components of genome organization, while simultaneously maintaining a physical association between all molecules of interest from the same single cell or nucleus. There are a number of ways in which a method can fail to achieve these conflicting objectives. First, wildly uneven coverage can result from incomplete removal of the histone proteins, a process which has been referred to as nucleosome depletion (Vitak et al. 2017). Harsh conditions and reagents are generally required for complete stripping of nucleosomes; lipid structure such as cell and nuclear membranes would not be expected to survive these treatments, are so any method that relies on these structure for

compartmentalization of nuclei acids is unlikely to be adaptable to whole-genome sequencing. Harsh conditions can also inactive enzymes introduced early in the experiment, and carry-over reagents can interfere with downstream reactions. Vitak *et al.* addressed this in SCI-seq by either cross-linking chromatin together prior to nucleosome depletion and combinatorial tagmentation by Tn5, followed by cross-link reversal before PCR, or by a diiodosalycate treatment to strip nucleosomes but preserve nuclear matrix proteins. In our lab, both approaches yield an extremely fragile product, and the combinatorial design requires a high-accuracy flow sorter to deposit precise numbers of cells (~20) per well for PCR barcoding (Vitak et al. 2017). Attempting to distribute such a small number of particles by hand pipette would result in unacceptable losses and uneven deposition due to stickage inside the pipette tip.

Adam Abate's group introduced the use of agarose hydrogels in Sic-seq to obtain high coverage, uniform single-cell libraries from bacteria, showing that at least some hydrogels can be employed for single-cell encapsulation and whole-genome sequencing without impacting the evenness of coverage. SiC-seq requires the use of pre-made barcode beads instead of split-pool barcoding, however, and uses a variety of highly sophisticated microfluidic devices to encapsulate, merge with Tn5 and barcodes, and perform PCR, which would be likely result in a high start-up cost for a potential user.

The approach described in this work demonstrates a rapid polyacrylamide encapsulation technique followed by highly-scalable combinatorial split-pool Tn5 barcoding that does not require any droplet merging or FACS, can be performed by a single operator in 2.5 days, and has a start-up cost of well under $10,000. Because only 1 in 10 hydrogels contain a cell, the other 90% of dropets are empy and act as an inert carrier. This reduces the degree and variability of sample loss due to pipetting, and enables distribution of as little as 20 cells (~200 beads) per

well with sufficient uniformity to control barcode collisions going into PCR without requiring time-consuming sorting with expensive FACS equipment. Hydrogels also perform very well in repeated washing steps due to their high structural integrity compared to cells, nuclei, or cross-linked chromosomes. Finally, hydrogels such as polyacrylamide also have interesting chemical attachment properties that can be exploited in the future, such as the potential for acrydite-modified poly-T primers to capture mRNA in for library prep in parallel with DNA, similar to the concept described in Gel-seq in Chapter 2 of this dissertation. Future work in the Zhang lab will explore the potential of these approaches.

It is our hope that the polyacrylamide encapsulation approach described here will facilitate fast, easy, and highly-scalable whole-genome single-cell library preps on the order of a thousand cells per run, providing an unbiased landscape of the genomic copy number of human neurons in the brain. We also hope that this approach in general will provide a platform to implement multi-omics on the single cell level, using a technique that can implemented in any lab, without large, specialized equipment, at minimal cost per cell prior to sequencing.

2.5 Materials and Methods

2.5.1 Cell Culture

HeLa and 3T3 cell lines were cultured in DMEM, high glucose, pyruvate (ThermoFisher Scientific Cat. 11995-065) supplemented with 10% FBS (ThermoFisher Scientific Cat. 10437010) and 1% Penicillin-Streptomycin (ThermoFisher Scientific Cat. 15070063). Cells were trypsinized with TrypLE Express Enzyme (1X), no phenol red (ThermoFisher Scientific Cat. 12604013) and stored at 1 million cells/mL in DMEM with 50% FBS and 5% DMSO (Sigma Aldrich Cat. D2650) for liquid nitrogen cryopreservation. Cells were barely thawed and washed in cold PBS prior immediately prior to nuclei extraction. Nuclei were isolated using a custom nuclei isolation buffer (NIB) (0.32 M sucrose, 5 mM CaCl, 3 mM Mg-Acetate, 0.1 mM EDTA, 10 mM Tris-HCl, 0.1% Triton X-100, pH 8.0). Cell pellets were resuspended in 1 mL of cold NIB and incubated on ice for 30 minutes, followed by centrifugation at 300 rcf for 5 minutes. Nuclei were washed once in cold PBSEB (PBS plus 1% fatty-acid free BSA and 1 mM EDTA) before use.

2.5.1 Microwell MDA amplification and library prep

All microwell experiments were performed according to the methods described in (Gole et al. 2013), with the exception of the alkaline denaturation of amplicons after MDA followed by second strand synthesis and ethanol purification prior to Tn5 library prep. Amplicons were instead pipetted up and down ~8 times to loosen hyper-branched structure and immediately used as Tn5 template. This simplification reduced per-sample processing time by nearly 6 hours and increased library prep success rate, with no measurable effect on data quality for CNV calling.

2.5.2 Custom library prep using commercial microfluidic chips

Custom protocols for the Fluidigm C1 were generated the Open App Developers Pack (100-8588) with small cell size (100-8134) IFCs. Alkaline lysis was performed in chamber 1 using 400 mM NaOH, while neutralization was performed in chamber 2 with equimolar amount of HCl. Primer-limited MDA pre-amplification was performed on-chip using Phi29 polymerase in 3 steps using random nonamers with double 3' terminal phosphoramidite backbone modifications (IDT; 5' NNNNNNN*N*N 3'). Steps 1 and 2 were carried out at 30 C for 30 minutes with 250 nM primer in microfluidic chambers 3 and 4, respectively, while step 3 was performed at 30 C for 3 hours with 250 uM in chamber 5. Tn5 library prep was performed off-chip using Nextera XT Tn5 according to manufacturer's instructions (Illumina 2015).

2.5.3 Microfluidic Device Fabrication

Microfluidic chips were fabricated by PDMS soft lithography using Sylgard 184 with 10:1 silicone base:curing agent mixed for 3:30 min at 2000 rpm in a Thinky mixer. Bottom layers were cast by pouring 10 g of mixed silicone into a 10 cm diameter petri dish while top layers were cast with 25 g in a 50 x 75 mm area on a custom SU-8 mold fabricated in-house (see below), followed by de-gassing by vacuum for 45 minutes. Samples were then baked at 80 C for 1 hour and devices were cut out using an exacto knife. Input/output ports were punched using a 0.7 mm biopsy punch. All microfluidic channels were then cleaned using 3M Magic tape after casting and cutting. Device top and bottom layers were surface activated by O2 plasma using an Oxford Plasmalab 100 oven at 250W power (0 reflected) with 5 sccm O2 for 15 seconds. Top and bottom layers were immediately bonded after plasma treatment and baked

30 minutes at 80 C. Microfluidic channels were surface treated by Aquapel followed by Fluorinert FC-40 rinse. Aquapel was prepared fresh for each experiment by cutting the glass vial out of a plastic Aquapel applicator, placing it into a plastic bag, shattering it, and straining the solution through a 40 um strainer, and loading into a 3 mL. Microfluidic channels were filled with Aquapel by holding the syringe gently against the inlet of each device and depressing the plunger until the channels were visibly filled. Each device was then flushed out with air using an empty syringe. FC-40 was then flushed through each device using an identical technique. Each device was then thoroughly flushed with air, holding the dish vertically and catching all displaced liquid with a kim-wipe until completely flushed. Devices were then baked at 65 C for 20 minutes and covered with 3M Magic tape, which was then scored by scalpel for convenient individual use. Each casting yielded 2 chips containing 9 devices each, suitable for 18 independent 1000-plex single-cell experiments and could be fabricated within a day for a total cost of less than $100, including time charged for clean room usage.

Microfluidic molds were fabricated in a clean room on 4” test grade silicon wafers. Wafers were solvent cleaned before use in a hot acetone bath at 55 C for 5 min, followed by rinse in methanol at room temperature for 3 min, followed by rinse with DI water and blow-dry with nitrogen. Wafers were then cleaned by oxygen plasma at 5 sccm O2 with 250 W power for 5 min. SU-8 2025 was deposited at 30 um target thickness by spin coating at 500 rpm for 10 seconds at 100 rpm/second acceleration, followed by 3000 rpm for 30 seconds at 300 rpm/second acceleration. Soft bake was performed at 65 C for 2 minutes followed by 95 C for 5 minutes. Exposure was performed on an EVG 620 mask aligner in hard contact mode at 13 mW/cm^2 for 12.3 seconds for a total exposure of 160 mJ/cm^2. Post-exposure bake was performed at 65 C for 1 min followed by 95 C for 5 min. Wafers were developed 5 minutes in

SU-8 developer, followed by 10 second rinse in fresh developer, followed by rinse in isopropanol and blow dry with nitrogen. Wafers were then hard baked at 150 C for 5 minutes and surface treated by FOTS vapor deposition. Wafers were then taped to the bottom of 15 cm petri dishes and covered in a large quantity (~80 g) PDMS for the first casting. Subsequent castings were performed inside the 50 x 75 mm mold space left after cutting out chips.

Custom photomasks for the final device design were chrome on 5" sq. x 0.090" soda lime ordered from a commercial vendor (FrontRange PhotoMask) with 10 micron tolerance, dark field background, and right read (chrome) down.

2.5.4 Pressurization of Microfluidic Devices

A house-air driven constant-pressure microfluidic testing station was constructed using general purpose 1/2" ID 5/8" OD PVC tubing (McMaster-Carr 5233K66), 0.5 to 30 PSI 1/4" NPT regulators (McMaster-Carr 43275K16) and 0 to 30 PSI 1/2" NPT digital pressure gauges (McMaster-Carr 2798K211) for each pressure line. Smaller tubing (McMaster-Carr 55485K72) was used for each line to connect to a needle inserted through a metal washer and a 1/4" rubber stoppers, which was then used to pressurize a disposable 3 mL syringe (Beckton Dickinson 309657) mounted vertically (tip down) on a standard laboratory ring stand with the plunger removed. A No. 18 ball joint clamp were used to secure the rubber stopper by sandwiching it between the metal washer and the flanges of the syringe. Syringes were attached to 22-gauge blunt leur stubs (Instech LS22) inserted into 0.023" ID polyethylene tubing (Instech BTPE-50) with attached 22-gauge right-angle metal adapters (Instech SC22/15RA) to be inserted into microfluidic I/O ports. Connections were made with a variety of barbed, threaded, and leur-lok adapters, depending on the components.

2.5.5 Hydrogel Encapsulation of Nuclei

After washing in cold PBSEB, nuclei were resuspended in hydrogel precursor buffer (PBS plus 20% v/v Optiprep (Sigma Cat. D1556) 18.6% w/v acrylamide, 0.54% w/v N,N'-Methylenebis(acrylamide), 1% w/v fatty-acid free BSA, and 1 mM EDTA) and strained through a 20 um mesh. Suspensions were loaded into 3 mL syringe mounted on a laboratory ring stand and pressurized as described above.

Lysis buffer was composed of buffer G2 with 0.45% w/v APS prepared fresh for each experiment and strained through a 20 um mesh. Novec HFE-7500 was used as carrier oil with 1.5% w/v fluorosurfactant-008 (RAN) and 0.4% v/v TEMED added fresh for each experiment.

All microfluidic experiments were performed on an upright microscope. Visual monitoring of the microfluidic junction between the sample, lysis, and oil lines enabled tuning of on-chip reagent mixing ratios. Because the boundary between lysis reagents, sample, and oil are clearly visible with a light microscope, the mixing ratio can be tuned on-chip. A target mixing ratio of 1:2 sample:lysis was achieved by tuning pressure in each line such that the width of each of the two lysis flows was equal to the width of the sample flow in the middle of the junction. This was confirmed by tracking reagent delivery through microfluidic tubing marked off in 1 cm increments as shown in Figure 26. The ratio of the slopes of the curves are equal to the flow rates and were within 10% of the desired 1:2 mixing ratio.

Samples were collected under 300 ul of mineral oil and allowed to polymerize overnight at room temperature. Emulsions were broken by removing as much HFE and mineral oil from below and above the bead layer as possible and then washing twice with 20% PFO in HFE with 300 rcf centrifugation for 5 minutes at each step, followed by two washes with 1% SPAN-80 in hexane, followed by two washed in TEBST buffer (10 mM Tris-HCl, 10 mM EDTA, 137

mM NaCl, 2.7 mM KCl, 0.1% (v/v) Triton X-100). Samples were then washed 4 times in buffer G2 and digested for 1 hour at 50 C with Qiagen Protease followed by heat inactivation at 70 C for 15 minutes. Samples were then washed 4 times in tagmentation buffer before staining, counting, and aliquoting for Tn5 library prep as described in *Preissl et al.* (Preissl et al. 2018).

2.5.6 Library Preparation and Sequencing

Hyperactive Tn5 was obtained from Berkeley Macrolabs produced using a published protein expression and purification protocol (Picelli et al. 2014). All libraries were sequences on Illumina GAIIx, HiSeq, or Miseq platforms. Combinatorial Tn5 libraries were sequenced with 50 cycles for read 1, 43 cycles for index 1, and 37 cycles for index 2 (SE50+43+37).

2.5.7 Data Processing and Analysis

BCL files were converted to FASTQ for reads and index reads using Illumina's bcl2fastq with a dummy sample sheet and options to generate FASTQ for index reads. Tn5 and PCR barcodes were extracted from the first 8 and last 8 bases of index reads 1 and 2. Demultiplexing was performed with deindexer (https://github.com/ws6/deindexer). Library IDs corresponding to each barcode combination were then assigned to the FASTQ header line of each read using pysam (https://github.com/pysam-developers/pysam) (Li et al. 2009).

For quantification of species mixing, FASTQ was merged and mapped to a bowtie2 (Langmead and Salzberg 2012) index built from a merged reference genome of both hg38 and mm10 and filtered with a MAPQ cutoff of >10 using samtools (Li et al. 2009). Mapping rates to hg38 and mm10 were then plotted with matplotlib (Hunter 2007).

For single cell copy number profiling, a cutoff of 20,000 reads per cell to select 87 single cell libraries. 27 HeLa libraries were identified along with 60 for 3T3, with 2 libraries colliding in barcode space. This corresponds to a collision rate of 2*2/87 = 5%, meaning that 95% of libraries analyzed above cutoff are expected to be from single cells. Libraries were demultiplexed and mapped as described above for species mixing, except that FASTQ was mapped in parallel to either hg38 or mm10 instead of a combined reference.

The varbin CBS approach from CSHL (Baslan et al. 2012) was used to profile genome-wide copy number with 1000 varbins generated from 50 base simulated reads from hg38 and mm10 using bowtie2 with a MAPQ cutoff of 10.

2.6 Appendix to Chapter 2

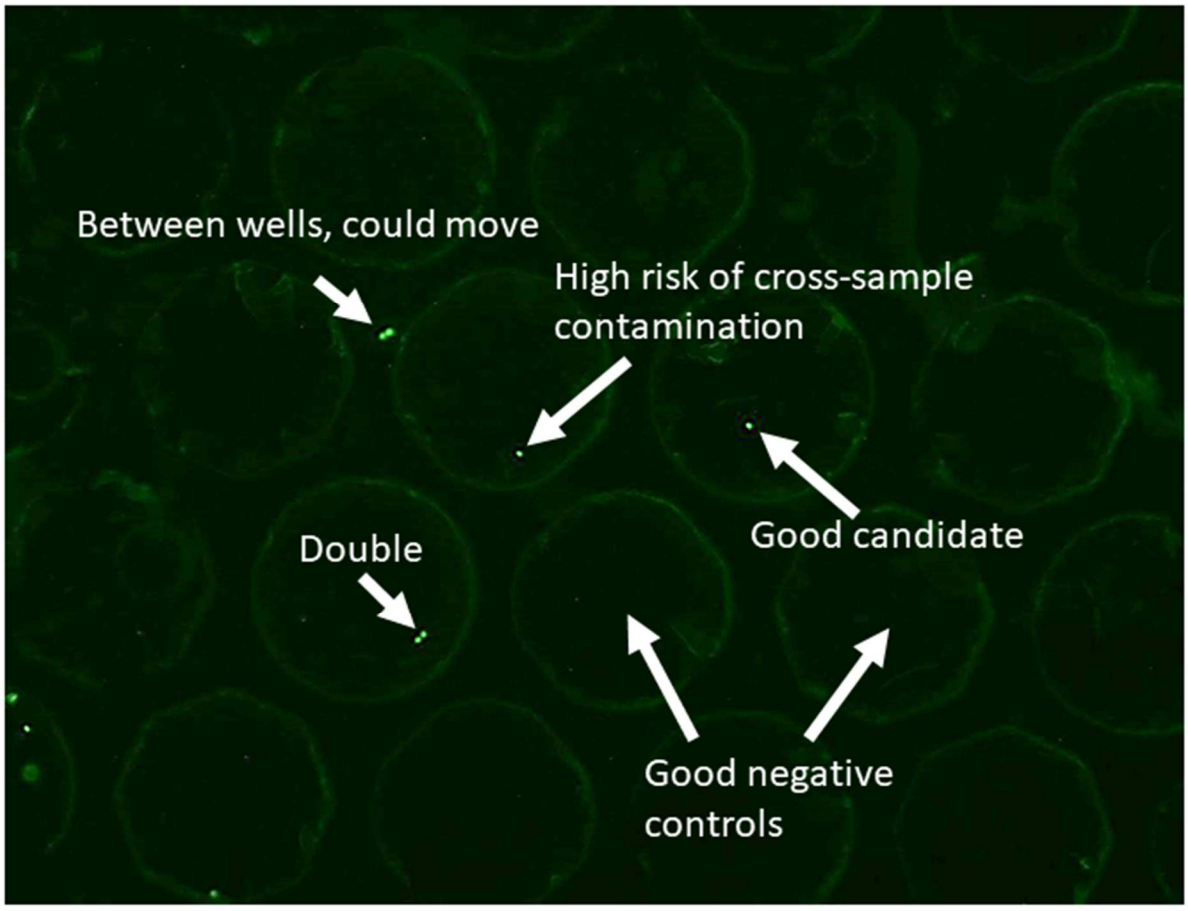

Figure 20: Fluorescent images of microwells containing nuclei stained for DNA prior to MDA amplification (MIDAS).

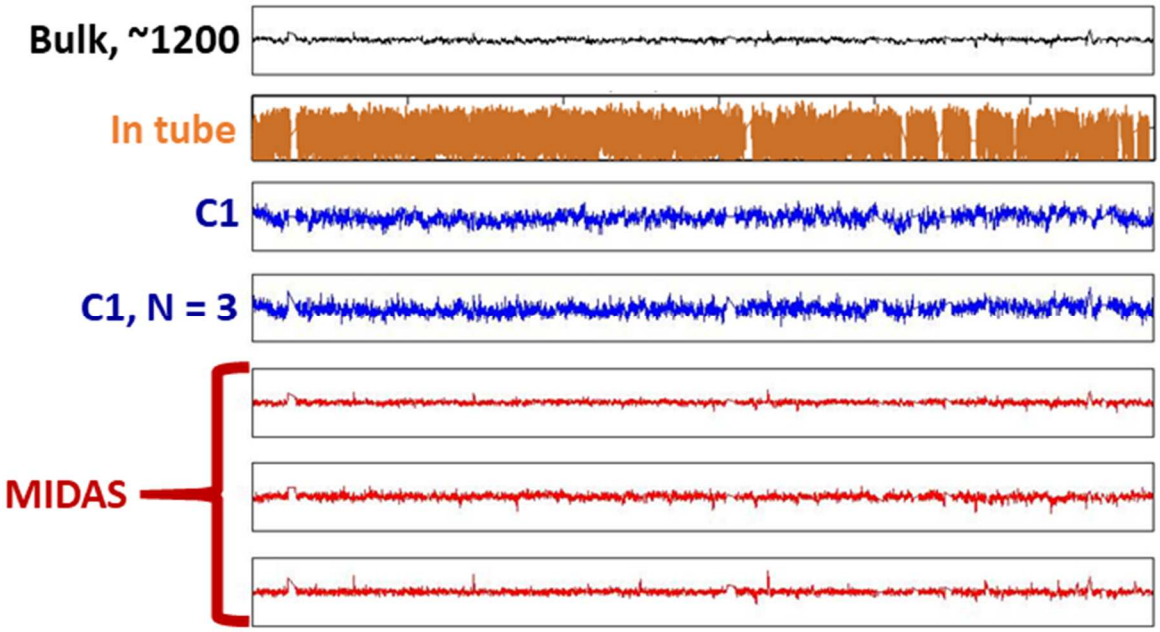

Figure 21: Geneome-wide mapping coverage of assumed euploid samples using different approaches. The flatness of the traces indicates qualitatively the evenness of coverage. MIDAS shows that lowest noise, approaching that of a bulk library.

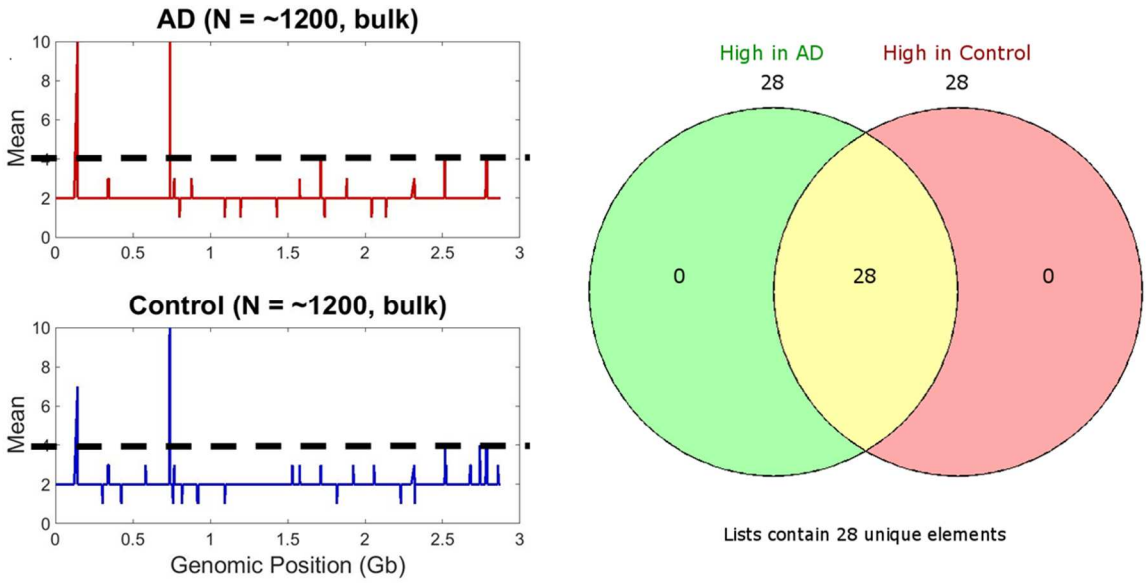

Figure 22: Bulk copy number from Alzheimer's and non-diseased control patients. A threshold of copy number > 4 was used to avoid any calls from chromosome 21 in Down's syndrome in further analysis. Bulk Down's syndrome sample was not available at this time. 28 genes were called as having high copy number, and were assumed to be mapping artifacts due to incomplete reference and excluded from further consideration.

Table 3: List of single-cell samples generated for each patient type using MIDAS

Patient ID	Disease state	#	Sum
1-20	AD	5	
24-01	AD	11	27
25-00	AD	11	
M1864	DS	6	6
60831	ND	22	
1568	ND	2	27
7-03	ND	3	

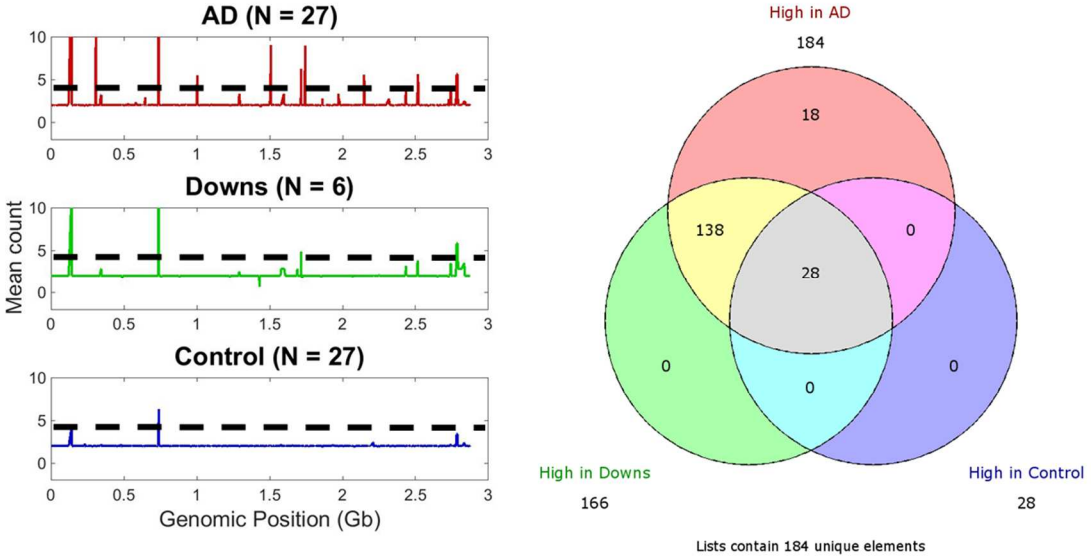

Figure 23: Average copy number calls from MIDAS data across cells of each group. All groups including control shared the same 28 genes called above threshold as discussed in Figure 20. Alzheimer's neurons showed 18 genes with high copy number across cells, as well as a further 138 genes that were also seen at high copy number in Down's syndrome neurons.

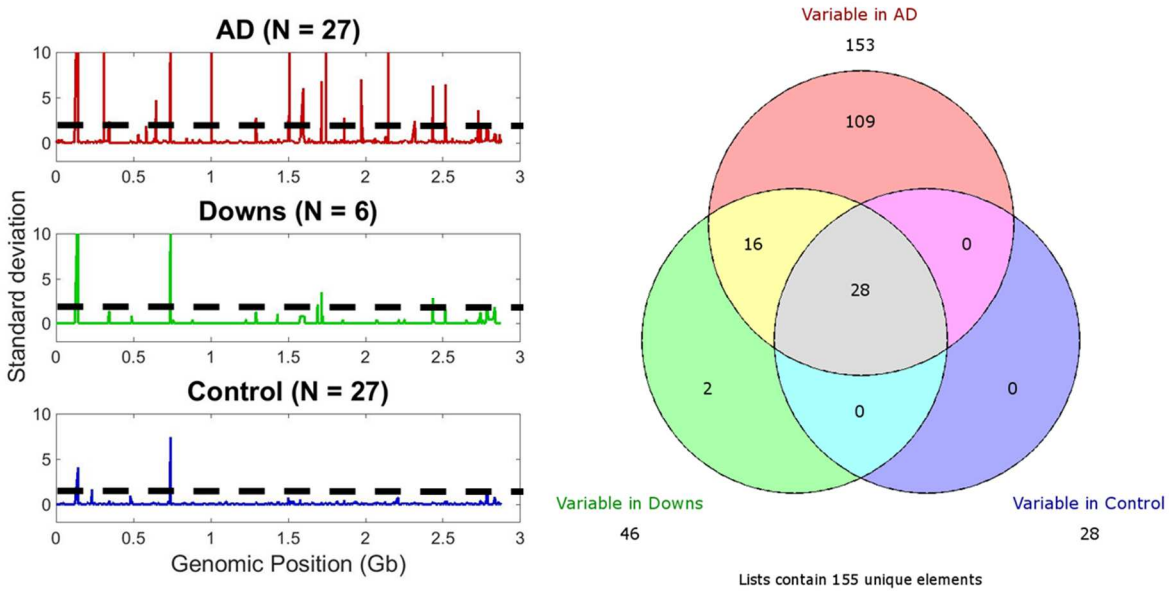

Figure 24: Investigation of variability in copy number states between samples. Setting a threshold of 2 eliminated all but the 28 spurious genes in control ND samples identified previously. AD samples had 109 unique genes variable above this threshold, with another 16 shared with Downs.

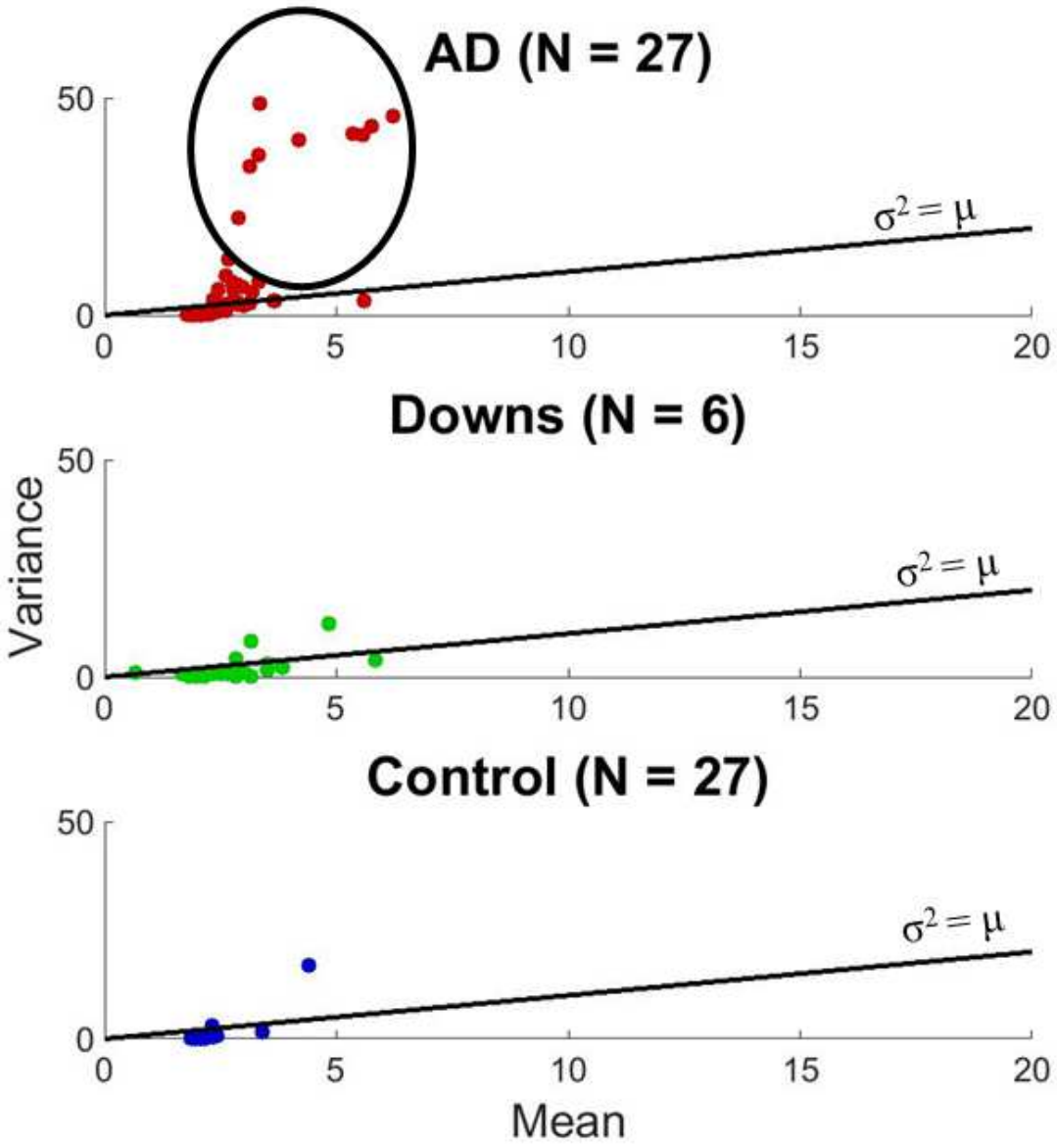

Figure 25: Variance versus mean for MIDAS microwell libraries. When plotting variance versus mean for a variety of calls across cells of each group, random noise is expected to be distributed around a line given by σ^2 = μ, referred to as Fano noise. While this is the case Down's and Control, AD has a subset of calls which are much more variable across cells than expected by chance, indicating a higher degree of non-shared variation.

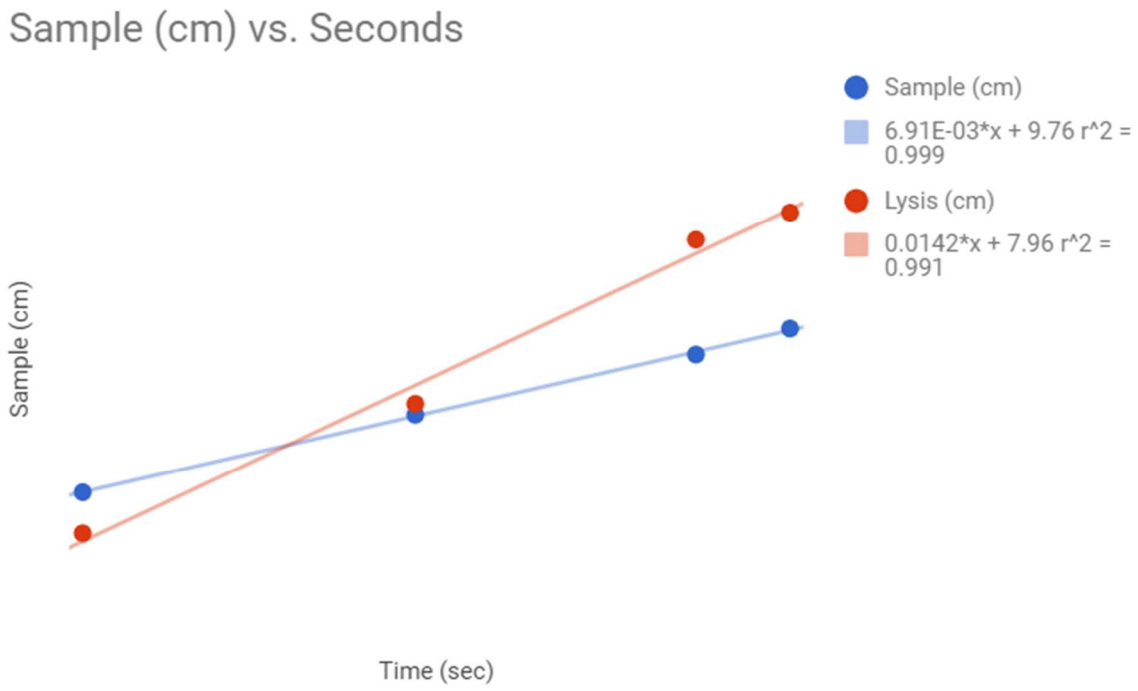

Figure 26: Microfluidic reagent flow versus time.

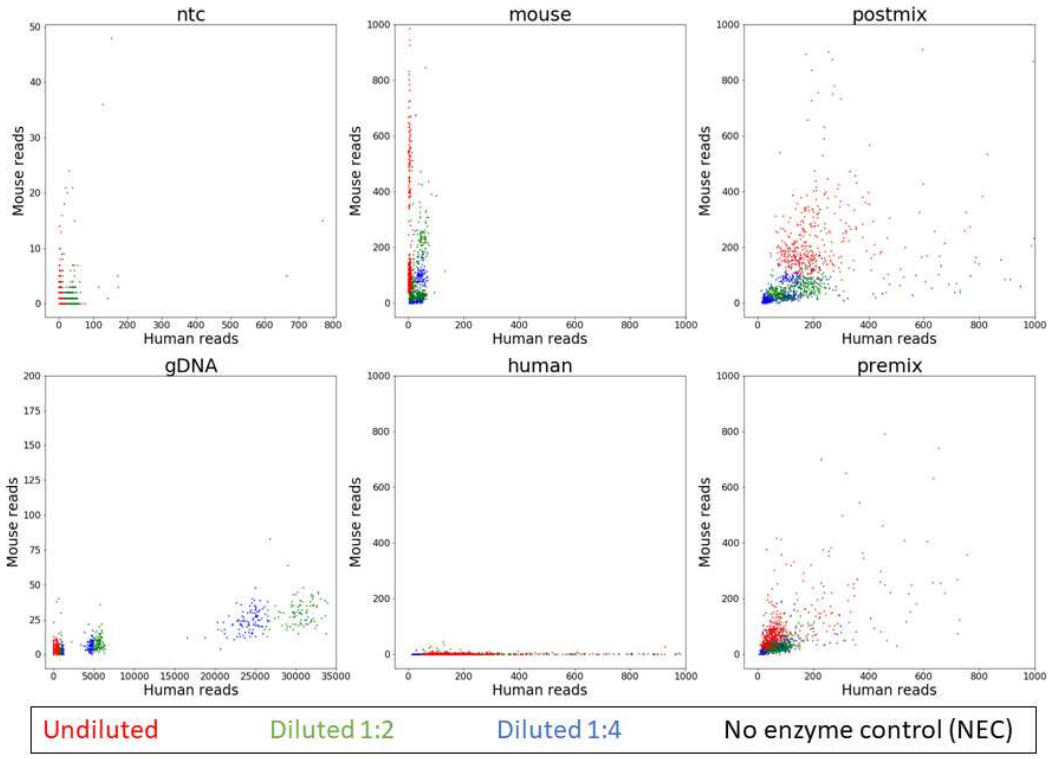

Figure 27: Mouse reads versus human reads for combinatorial gel bead libraries

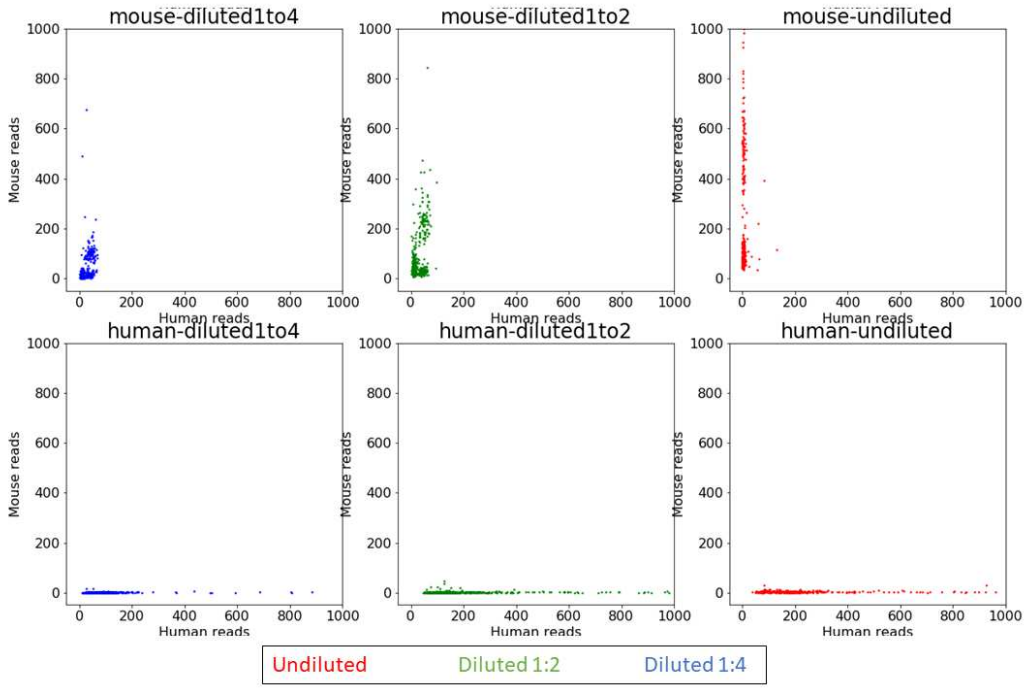

Figure 28: Mouse reads versus human reads for combinatorial gel bead libraries plotted by titration

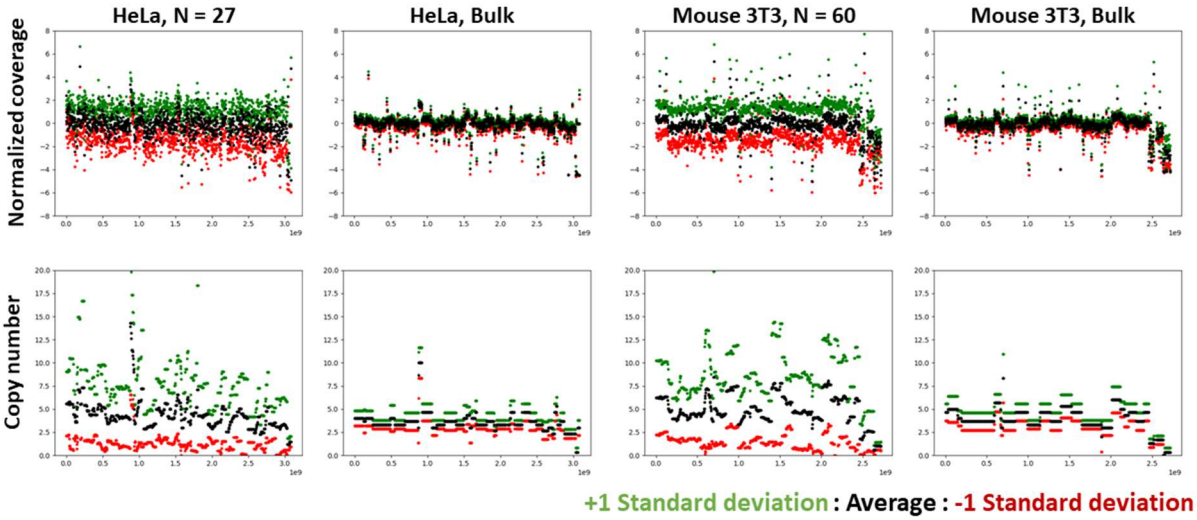

Figure 29: Genome wide bin counts and copy number estimates for HeLa and 3T3 for both single cells and bulk. Green and red points indicate one standard deviation above and below the mean, respectively. Standard deviation for bulk libraries are the result of repeated downsampling. All plots are at 1000 bin resolution.

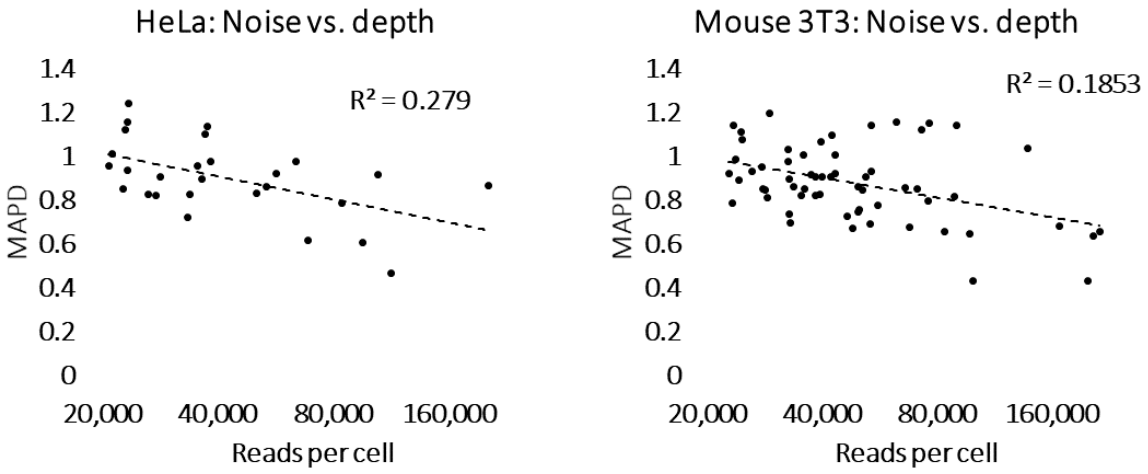

Figure 30: Combinatorial gel bead library noise of coverage versus depth of sequencing.

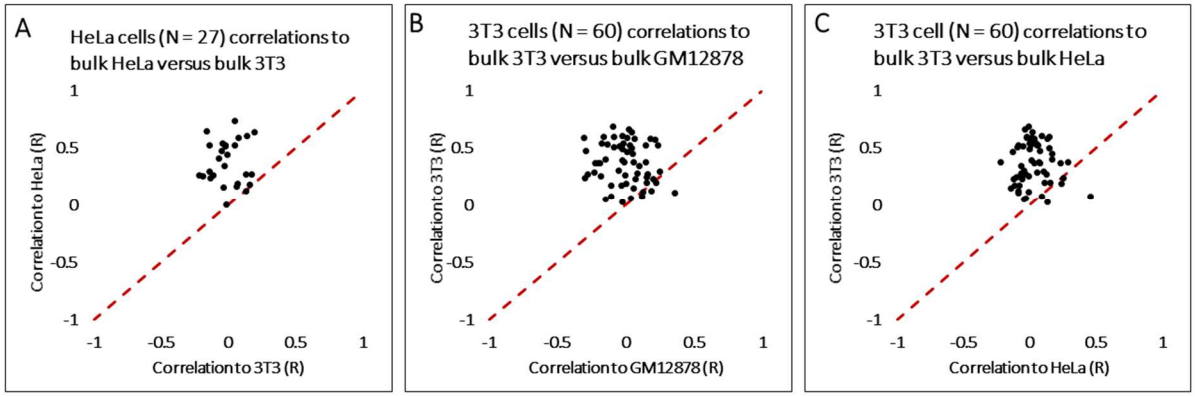

Figure 31: Scatter plots of Pearson correlation coefficients for all single cells compared to bulks.

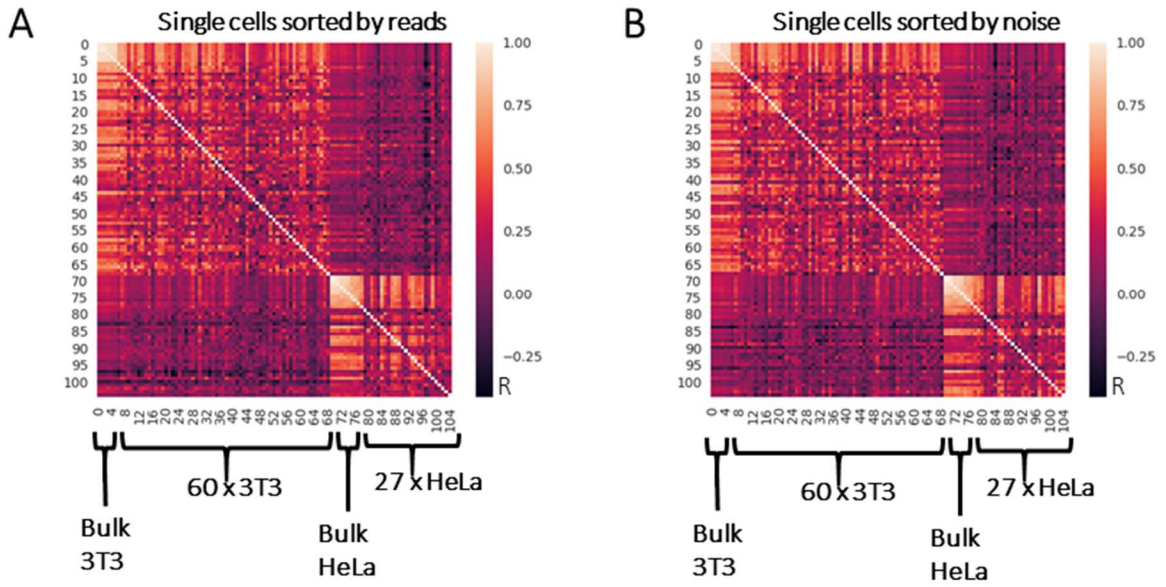

Figure 32: Correlation heatmaps for unclustered samples. Panel A shows single cells sorted by read depth per library. Panel B shows single cells sorted by noise (MAPD).

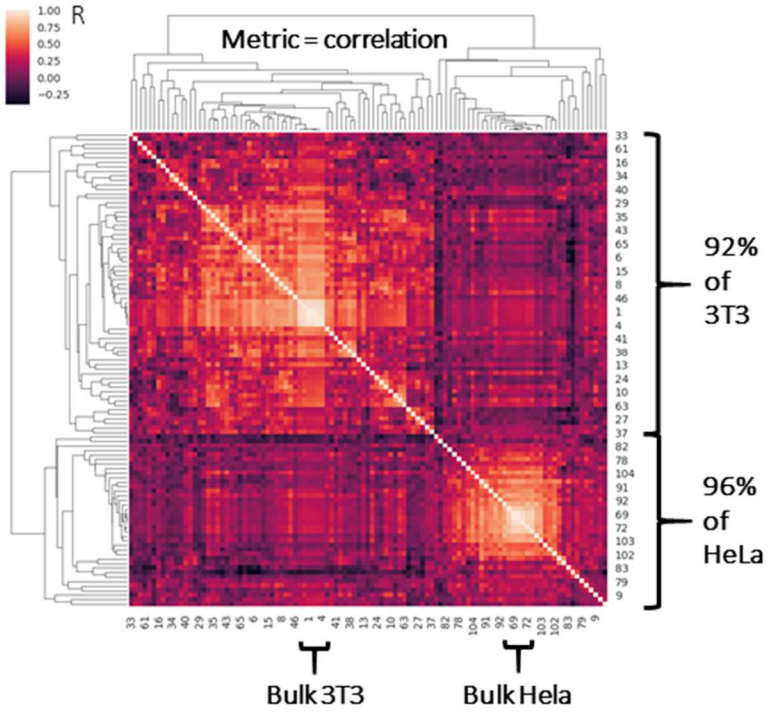

Figure 33: Pairwise correlations for all 87 single cells from both HeLa and 3T3, as well as 9 downsampled bulk for each cell line. Single cell copy number profile correlation clustering. (Bulk 3T3: 1 to 9; Single cell 3T3: 10 to 69; Bulk HeLa: 70 to 79; Single cell HeLa: 80-105).

REFERENCES

Baslan, Timour, Jude Kendall, Linda Rodgers, Hilary Cox, Mike Riggs, Asya Stepansky, Jennifer Troge, Kandasamy Ravi, Diane Esposito, B Lakshmi, Michael Wigler, Nicholas Navin, and James Hicks. 2012. "Genome-Wide Copy Number Analysis of Single Cells." *Nature Protocols* 7 (6): 1024–41. https://doi.org/10.1038/nprot.2012.039.

Cai, Xuyu, Gilad D. Evrony, Hillel S. Lehmann, Princess C. Elhosary, Bhaven K. Mehta, Annapurna Poduri, and Christopher A. Walsh. 2014. "Single-Cell, Genome-Wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain." *Cell Reports* 8 (5): 1280–89. https://doi.org/10.1016/j.celrep.2014.07.043.

Cao, Junyue, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. 2017. "Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism." *Science (New York, N.Y.)* 357 (6352): 661–67. https://doi.org/10.1126/science.aam8940.

Clontech. 2016. "SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing User Manual."

Consortium, Seqc/Maqc-Iii, and others. 2014. "A Comprehensive Assessment of RNA-Seq Accuracy, Reproducibility and Information Content by the Sequencing Quality Control Consortium." *Nature Biotechnology* 32 (9): 903–14.

Daley, Timothy, and Andrew D Smith. 2013. "Predicting the Molecular Complexity of Sequencing Libraries." *Nature Methods* 10: 325–27.

Davis, B J. 1964. "DISC Electrophoresis-II Method and Application to Human Serum Proteins." *Annals of the New York Academy of Sciences* 121: 404–27.

Dey, Siddharth S, Lennart Kester, Bastiaan Spanjaard, Magda Bienko, and Alexander van Oudenaarden. 2015. "Integrated Genome and Transcriptome Sequencing of the Same Cell." *Nature Biotechnology*, January, 1–7.

Dobin, A, C A Davis, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and T R Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29: 15–21.

Gao, Ruli, Alexander Davis, Thomas O McDonald, Emi Sei, Xiuqing Shi, Yong Wang, Pei-Ching Tsai, Anna Casasent, Jill Waters, Hong Zhang, Funda Meric-Bernstam, Franziska Michor, and Nicholas E Navin. 2016. "Punctuated Copy Number Evolution and Clonal Stasis in Triple-Negative Breast Cancer." *Nature Genetics* 48 (10): 1119–30. https://doi.org/10.1038/ng.3641.

Glessner, Joseph T, Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E Kim, Shawn Wood, Haitao Zhang, Annette Estes, Camille W Brune, Jonathan P Bradfield, and others. 2009. "Autism Genome-Wide Copy Number Variation Reveals Ubiquitin and Neuronal Genes." *Nature* 459 (7246): 569–73.

Gole, Jeff, Athurva Gore, Andrew Richards, Yu-Jui Chiu, Ho-Lim Fung, Diane Bushman, Hsin-I Chiang, Jerold Chun, Yu-Hwa Lo, and Kun Zhang. 2013. "Massively Parallel Polymerase Cloning and Genome Sequencing of Single Cells Using Nanoliter Microwells." *Nature Biotechnology* 31 (12): 1126–32.

Guffanti, Guia, Federica Torri, Jerod Rasmussen, Andrew P Clark, Anita Lakatos, Jessica A Turner, James H Fallon, Andrew J Saykin, Michael Weiner, Marquis P Vawter, and others. 2013. "Increased CNV-Region Deletions in Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) Subjects in the ADNI Sample." *Genomics* 102 (2): 112–22.

Hansen, J Norman. 1981. "Use of Solubilizable Acrylamide Disulfide Gels for Isolation of DNA Fragments Suitable for Sequence Analysis." *Analytical Biochemistry* 116 (1): 146–51.

Hoople, Gordon D., Andrew Richards, Yan Wu, Kota Kaneko, Xiaolin Luo, Gen-Sheng Feng, Kun Zhang, and Albert P. Pisano. 2017. "Gel-Seq: Whole-Genome and Transcriptome Sequencing by Simultaneous Low-Input DNA and RNA Library Preparation Using Semi-Permeable Hydrogel Barriers." *Lab on a Chip* 17 (15): 2619–30. https://doi.org/10.1039/C7LC00430C.

Hunter, J D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing In Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Illumina. 2015. "Nextera XT DNA Library Preparation Guide."

Jones, Eric, Travis Oliphant, Pearu Peterson, and others. 2001. "SciPy: Open Source Scientific Tools for Python."

Klein, Allon M., Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. 2015. "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells." *Cell* 161 (5): 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.

Kurimoto, Kazuki, Yukihiro Yabuta, Yasuhide Ohinata, and Mitinori Saitou. 2007. "Global Single-Cell CDNA Amplification to Provide a Template for Representative High-Density Oligonucleotide Microarray Analysis." *Nature Protocols* 2 (3): 739–52.

Lan, Freeman, Benjamin Demaree, Noorsher Ahmed, and Adam R Abate. 2017. "Single-Cell Genome Sequencing at Ultra-High-Throughput with Microfluidic Droplet Barcoding." *Nature Biotechnology*, Advanced online publication.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25.1--R25.10.

Lee, Ho Suk, Wai Keung Chu, Kun Zhang, and Xiaohua Huang. 2013. "Microfluidic Devices

with Permeable Polymer Barriers for Capture and Transport of Biomolecules and Cells." *Lab on a Chip* 13 (17): 3389–97.

Lemire, A, K Lea, D Batten, S Jian Gu, P Whitley, K Bramlett, and L Qu. 2011. "Development of ERCC RNA Spike-in Control Mixes." *Journal of Biomolecular Techniques: JBT* 22 (Suppl): S46.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Lucito, Robert, John Healy, Joan Alexander, Andrew Reiner, Diane Esposito, Maoyen Chi, Linda Rodgers, Amy Brady, Jonathan Sebat, Jennifer Troge, and others. 2003. "Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation." *Genome Research* 13 (10): 2291–2305.

Macaulay, Iain C, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, Miriam Smith, Niels Van der Aa, Ruby Banerjee, Peter D Ellis, Michael A Quail, Harold P Swerdlow, Magdalena Zernicka-Goetz, Frederick J Livesey, Chris P Ponting, and Thierry Voet. 2015. "G&T-Seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes." *Nature Methods* 12 (April): 519.

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

McConnell, Michael J, John V Moran, Alexej Abyzov, Schahram Akbarian, Taejeong Bae, Isidro Cortes-Ciriano, Jennifer A Erwin, Liana Fasching, Diane A Flasch, Donald Freed, Javier Ganz, Andrew E Jaffe, Kenneth Y Kwan, Minseok Kwon, Michael A Lodato, Ryan E Mills, Apua C M Paquola, Rachel E Rodin, Chaggai Rosenbluh, Nenad Sestan, Maxwell A Sherman, Joo Heon Shin, Saera Song, Richard E Straub, Jeremy Thorpe, Daniel R Weinberger, Alexander E Urban, Bo Zhou, Fred H Gage, Thomas Lehner, Geetha Senthil, Christopher A Walsh, Andrew Chess, Eric Courchesne, Joseph G Gleeson, Jeffrey M Kidd, Peter J Park, Jonathan Pevsner, Flora M Vaccarino, and Brain Somatic Mosaicism Brain Somatic Mosaicism Network. 2017. "Intersection of Diverse Neuronal Genomes and Neuropsychiatric Disease: The Brain Somatic Mosaicism Network." *Science (New York, N.Y.)* 356 (6336). https://doi.org/10.1126/science.aal1641.

Ornstein, L. 1964. "DISC Electrophoresis-I Background and Theory." *Annals of the New York Academy of Sciences* 321-- 349.

Pedregosa, F, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher,

M Perrot, and E Duchesnay. 2011. “Scikit-Learn: Machine Learning in {P}ython.” *Journal of Machine Learning Research* 12: 2825–30.

Picelli, Simone, Åsa K. Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. 2014. “Tn5 Transposase and Tagmentation Procedures for Massively Scaled Sequencing Projects.” *Genome Research* 24 (12): 2033–40. https://doi.org/10.1101/gr.177881.114.

Preissl, Sebastian, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U. Gorkin, Yanxiao Zhang, Brandon C. Sos, Veena Afzal, Diane E. Dickel, Samantha Kuan, Axel Visel, Len A. Pennacchio, Kun Zhang, and Bing Ren. 2018. “Single-Nucleus Analysis of Accessible Chromatin in Developing Mouse Forebrain Reveals Cell-Type-Specific Transcriptional Regulation.” *Nature Neuroscience* 21 (3): 432–39. https://doi.org/10.1038/s41593-018-0079-3.

Ramskold, Daniel, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, Gary P Schroth, and Rickard Sandberg. 2012. “Full-Length MRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells.” *Nature Biotechnology* 30 (8): 777–82.

Rosenberg, Alexander B, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic, and Georg Seelig. 2018. “Single-Cell Profiling of the Developing Mouse Brain and Spinal Cord with Split-Pool Barcoding.” *Science (New York, N.Y.)* 360 (6385): 176–82. https://doi.org/10.1126/science.aam8999.

Sambrook, Joseph, and David W Russell. 2006. “Isolation of DNA Fragments from Polyacrylamide Gels by the Crush and Soak Method.” *Cold Spring Harb Protoc*.

Sasagawa, Yohei, Itoshi Nikaido, Tetsutaro Hayashi, Hiroki Danno, Kenichiro D Uno, Takeshi Imai, and Hiroki R Ueda. 2013. “Quartz-Seq: A Highly Reproducible and Sensitive Single-Cell RNA Sequencing Method, Reveals Non-Genetic Gene-Expression Heterogeneity.” *Genome Biology* 14 (4): R31.

Sebat, Jonathan, B Lakshmi, Jennifer Troge, Joan Alexander, Janet Young, Par Lundin, Susanne Månér, Hillary Massa, Megan Walker, Maoyen Chi, and others. 2004. “Large-Scale Copy Number Polymorphism in the Human Genome.” *Science* 305 (5683): 525–28.

Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. 2013. “Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science.” *Nature Reviews Genetics*, July, 1–13.

Shatzkes, Kenneth, Belete Teferedegne, and Haruhiko Murata. 2014. “A Simple, Inexpensive Method for Preparing Cell Lysates Suitable for Downstream Reverse Transcription Quantitative PCR.” *Scientific Reports* 4 (April): 1–7.

Sottoriva, Andrea, Inmaculada Spiteri, Sara G M Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. 2013. “Intratumor

Heterogeneity in Human Glioblastoma Reflects Cancer Evolutionary Dynamics." *Proceedings of the National Academy of Sciences* 110 (10): 4009–14.

Spratt, Daniel E, Zachary S Zumsteg, Felix Y Feng, and Scott A Tomlins. 2016. "Translational and Clinical Implications of the Genetic Landscape of Prostate Cancer." *Nature Reviews Clinical Oncology* 13: 597–610.

Suzuki, Yutaka, Daisuke Ishihara, Masahide Sasaki, Haruhito Nakagawa, Hiroko Hata, Takeshi Tsunoda, Manabu Watanabe, Takami Komatsu, Toshio Ota, Takao Isogai, and others. 2000. "Statistical Analysis of the 5′ Untranslated Region of Human MRNA Using 'Oligo-Capped' CDNA Libraries." *Genomics* 64 (3): 286–97.

"The Nobel Prize in Physiology or Medicine 1987." n.d. Accessed June 3, 2018. https://www.nobelprize.org/nobel_prizes/medicine/laureates/1987/.

Viovy, Jean-Louis. 2000. "Electrophoresis of DNA and Other Polyelectrolytes: Physical Mechanisms." *Reviews of Modern Physics* 72 (3): 813.

Vitak, Sarah A, Kristof A Torkenczy, Jimi L Rosenkrantz, Andrew J Fields, Lena Christiansen, Melissa H Wong, Lucia Carbone, Frank J Steemers, and Andrew Adey. 2017. "Sequencing Thousands of Single-Cell Genomes with Combinatorial Indexing." *Nature Methods* 14 (3): 302–8. https://doi.org/10.1038/nmeth.4154.

Westra, Jurjen W., Richard R. Rivera, Diane M. Bushman, Yun C. Yung, Suzanne E. Peterson, Serena Barral, and Jerold Chun. 2010. "Neuronal DNA Content Variation (DCV) with Regional and Individual Differences in the Human Brain." *The Journal of Comparative Neurology* 518 (19): 3981–4000. https://doi.org/10.1002/cne.22436.

Xi, Larry, Alexander Belyaev, Sandra Spurgeon, Xiaohui Wang, Haibiao Gong, Robert Aboukhalil, and Richard Fekete. 2017. "New Library Construction Method for Single-Cell Genomes." Edited by Ruslan Kalendar. *PLOS ONE* 12 (7): e0181163. https://doi.org/10.1371/journal.pone.0181163.

Zangle, Thomas A, Ali Mani, and Juan G Santiago. 2010. "Theory and Experiments of Concentration Polarization and Ion Focusing at Microchannel and Nanochannel Interfaces." *Chemical Society Reviews* 39 (3): 1014–35.

Zhan, Hans, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. 2017. "Scalable Whole-Genome Single-Cell Library Preparation without Preamplification." *Nature Methods* 14: 167–73.