Enhancing Single-Cell RNA Sequencing to Elucidate Host Cellular Response to Ebola Virus Infection

Enhancing Single-Cell RNA Sequencing to Elucidate Host Cellular Response to Ebola Virus Infection

A thesis presented by

Nadine Marie Khoury

to

The Faculty of the

Harvard John A. Paulson School of Engineering and Applied Sciences

In partial fulfillment of the requirements for

The Bachelor of Arts degree with honors in

Biomedical Engineering

Faculty Advisor: Prof. Pardis Sabeti

Harvard University

Cambridge, MA

April 3, 2020

Honor Code

In submitting this thesis to the Harvard John A. Paulson School of Engineering and Applied Sciences in partial fulfillment of the requirements for the degree with honors of Bachelor of Arts, I affirm my awareness of the standards of the Harvard College Honor Code.

Name: Nadine Marie Khoury

Signature:

Nadine Khoury

Acknowledgements

Firstly, I would like to thank my Principal Investigator and mentor Dr. Pardis Sabeti for welcoming me into her lab my freshman year and advising me through my undergraduate years. This project would not have been possible without my project mentor Dr. Aaron Lin, who was immensely supportive in guiding me through the world of scientific research, every step of the way. I would further like to thank my thesis co-readers and advisors, Dr. Alain Viel and Dr. Sean Eddy, for their support and guidance in the making of this thesis, and Harvard's Office of Undergraduate Research and Fellowships for their support through the Herchel Smith and PRISE (Program for Research in Science and Engineering) Fellowships. I send my gratitude to my previous research mentors over the years: Dr. Alain Viel, who introduced me to the fascinating field of genetics and DNA sequencing during his freshman seminar "Building a Living Cell One Brick at a Time," my high school research mentor Dr. Beata Jarosiewicz at Brown University's BrainGate Lab, and Dr. Michael Young at Rockefeller University. Lastly, I would like to thank my parents, sister, and family for their lifelong support and commitment to my success and wellbeing.

Statement of Research

This thesis marks the culmination of my three-year-long efforts in the Sabeti Lab, at the Broad Institute of MIT and Harvard. This project was undertaken in collaboration with the National Institutes of Health (NIH), funded by the United States Food and Drug Administration. Scientists at the NIH used their Biosafety Level 4 (BSL4) facility to infect nonhuman primates (rhesus macaques) with Ebola virus, and collected samples from the primates over the course of infection. They began the relevant single-cell RNA-Seq (scRNA-Seq) preparation in their facilities, and sent the samples to our lab, where Drs. Aaron Lin, Kayla Barnes, and others, completed the scRNA-seq library preparation. My work primarily experimented on these libraries. Under the training and guidance of Dr. Lin, I developed and completed the assays exploring CRISPR-Cas9 depletion, cell specific PCR, and LCMV transcriptomics. These efforts have culminated in co-authorship in a paper to be submitted in the next few days:

- Single-Cell profiling defines viral and host transcriptional dynamics in Ebola infection. Dylan Kotliar, Aaron E. Lin, Kayla G. Barnes, James Logue, David McIlwain, Travis Hughes, **Nadine M. Khoury**, Marc Wadsworth, Garry Nolan, Richard S. Bennett, Alex K. Shalek, Lisa E. Hensley, Pardis C. Sabeti. *Manuscript to be submitted in coming days.*

Aside from the work in this thesis, I have researched two other projects at the Sabeti Lab:

- Using DNA-RNA hybrid selection to extract and amplify transcriptomes containing Ebola virus cDNA, also under the mentorship of Dr. Aaron Lin.
- Using CRISPR-Cas9 to deplete abundant host background for sequencing infectious pathogens, under the mentorship of Mr. Simon Ye.

Abstract

Recent and current outbreaks of Ebola Virus Disease in West Africa and the Democratic Republic of Congo highlight the need to understand the disease and its pathophysiology. Single-Cell RNA sequencing (scRNA-seq) is a new development that can deeply characterize how the host's different cell-types modulate their gene expression in response to infection. Recently, our lab used scRNA-seq to study how Ebola virus disease attacks host cell machinery, the first such study for a Risk Group 4 (RG-4) pathogen. ScRNA-seq was tested on blood samples from nonhuman primates infected with Ebola virus, but an adapter artifact from the scRNA library preparation over-amplified in many libraries, and reduced the quantity and quality of meaningful reads. In this study, I use CRISPR-Cas9 to target and degrade this artifact, leaving cDNAs from host genes intact. The CRISPR-Cas9 assays reduced the presence of adapter multimers 10-fold on average, with reductions reaching 37-fold*,* allowing us to sequence over 15 libraries that failed previously and generate thousands of additional transcriptomes. I then developed a "cell specific PCR," a method aimed at deeply sequencing the transcriptome of a specific single cell. This method selectively amplifies transcriptomes of interest by employing PCR primers corresponding to a cell's DNA barcode assigned during scRNA-seq preparation. Finally, I began developing models to study other viral hemorrhagic fevers in a lower containment setting. In all, this study plays a direct role in improving scRNA-seq data to study RG-4 pathogens. By recovering scRNA-libraries and amplifying transcriptomes of interest, we help uncover mechanisms used by different cell-types to respond to viral infection, paving the way for more effective vaccines, diagnostics, and antiviral medication to prevent and treat tragic outbreaks around the world.

Table of Contents

1. Introduction

1.1 Ebola Virus Overview

Ebola virus is a negative-sense, single stranded RNA virus in the filovirus family that causes Ebola virus disease (EVD), a severe hemorrhagic fever in humans and non-human primates (Basler 2017). During 2013–2016, Ebola spread through West Africa, causing over 11,000 deaths and 28,000 reported cases (WHO, 2015). Though the virus was known to circulate in human populations as early as 1976, in December 2013, the virus spread from Guinea to Liberia, Sierra Leone, and greater West Africa at an unforeseen rate via sustained human to human transmission (Gire et al. 2014).

Since the epidemic, scientists have sought to discover the mechanisms of EVD from both the viral and host standpoint. While scientists have studied the components of the virus' structure (Nathan et al. 2020) and mutations (Diehl et al. 2016) that contributed to its lethality, equally crucial to understanding Ebola's mechanism of attack is to explore the host cell's reaction to infection. For example, Reynard et al. used RT-qPCR, ELISA, and Magnetic Bead analysis of Plasma taken from Guinean patients to demonstrate how EBOV's ability to induce overexpression of cytokines can cause an intense immune response characteristic of fatal cases. By contrast, patients lacking this "cytokine storm" had a well-balanced immune response and survived (Basler 2017; Reynard et al. 2019). Understanding the regulatory mechanisms our immune systems employ in response to Ebola infection can ultimately help indicate which systems to target when developing effective treatments and vaccines.

1.2 RNA Sequencing Overview

A comprehensive method to analyze cellular activity in response to infection is to examine a cell's *transcriptome*, which comprises all of the mRNA inside a cell. These mRNAs code for proteins that the cell uses to conduct its various functions. The transcriptome reveals which genes are transcribed in a cell and to what degree, and can therefore indicate a cell's state or activity.

Measuring mRNA levels by sequencing (RNA-Seq) provides a snapshot of which genes are up- or down-regulated under specific conditions. In the context of infection, activation of the immune response can affect the expression of hundreds of antiviral genes. Since each cell-type can express its own response to infection, scRNA-seq can provide information on how each antiviral gene is differentially expressed within different cell-types over the course of infection at an unprecedented level of precision (Cristinelli and Ciuffi 2018).

Performing RNA sequencing can also serve to quantify the copies of viral RNA within a transcriptome, a powerful tool in tracking the presence and progression of infection (Zanini et al. 2018). RNA sequencing can be accomplished in two ways: Bulk Sequencing and Single-Cell Sequencing:

1.2.1 Bulk RNA Sequencing

Up until recently, to measure mRNA, scientists have relied solely on *bulk RNA sequencing,* in which thousands of cells are lysed together, and the total the mRNA of all sampled cells is sequenced in bulk. This results in average gene expression values for all tested cells. Once sequences are obtained, scientists can then map these sequences to known genes and pathways in the organism, thus indicating which genes are actively being expressed within the cells. While bulk mRNA sequencing is employed in a variety of different contexts not limited to infectious disease, Cross et al. applied bulk sequencing to analyze Ebola infection in rhesus macaques, finding that proinflammatory and prothrombotic signaling processes were induced upon Ebola infection (Cross et al. 2018).

Bulk sequencing provides deep sequences of an entire gene, enabling scientists to distinguish between multiple transcript variants of a given gene. Furthermore, bulk sequencing allows for higher starting input mRNA, which improves the quality of libraries, decreasing the risk of RNA degradation (Chen, Ning, and Shi 2019). As a result, bulk sequencing's advantages lie in the ability to analyze variants within a gene for novel splicing patterns and transcript variants. However, despite these benefits, bulk sequencing prevents scientists from efficiently differentiating between the expression levels within different cell-types. For example, bulk sequencing may not provide information on a gene's expression in B cells versus T cells, because all cells are lysed together and a pool of all mRNAs are sequenced simultaneously, providing general average expression values; in other words, *heterogeneity* is not preserved. This lack of heterogeneity is a significant drawback, as different cell-types have different functions within the body, express different genes to carry out those functions, and can express vastly different regulatory responses to infection. As a result, this lack of heterogeneity could result in dulling potentially significant regulation signals, preventing scientists from isolating specific targets within the body susceptible to infection.

1.2.2 Single-Cell RNA Sequencing

Single-Cell RNA Sequencing (scRNA-seq) is a novel development that has provided an opportunity to characterize *each individual cell's* transcriptome, rather than pooling all cells to produce average values. In doing so, scRNA-seq preserves the heterogeneity in gene expression across cells (Figure 1). For example, unlike bulk sequencing, scRNA-seq can separate B cells and T cells, thus helping scientists explore biological differences between the cells after infection (Waickman et al. 2019; Choi and Kim 2019). Furthermore, by counting the copies of viral transcripts within a host cell's transcriptome, scRNA-seq can allow scientists to detect which cells are infected with a virus and to what extent. Unlike bulk seq, scRNA-seq only generates a portion of the transcript, since its main aim is to identify and count genes, a process that does not require the full sequence of the transcript. Specifically, scRNA-seq provides low-depth sequences of transcripts' 3'end, which in turn can only confirm the presence and identity of transcripts corresponding to a given gene (Hwang, Lee, and Bang 2018).

However, scRNA-seq preparation is typically harder and less cost-effective than bulk RNAseq. While bulk sequencing is cost-effective and can be easily transported in and out of Biosafety Level 4 facilities, scRNA-seq poses logistical challenges when applied to RG-4 pathogens such as Ebola virus. These challenges stem from the specific devices typically required to complete scRNA-seq preparation methods. For example, existing effective scRNA-seq methods such as 10X, Drop-Seq, and inDrops all require microfluidic devices and droplet generators, posing challenges for manufacturing and transporting samples, since the devices themselves cannot leave the given BSL4 facility (Macosko et al. 2015) (Klein and Macosko 2017) (Zilionis et al. 2017). Furthermore, since the cells must remain intact during transport, scientists cannot use harsh inactivating chemicals when transporting samples to and from BSL4 facilities. To address these challenges, Gierahn et al. recently presented SeqWell, a novel method to sequence single cells in a portable and cost-efficient manner (Gierahn et al. 2017). By sealing single cells within a microarray of microscopic wells with semipermeable membranes (with each well containing one cell), Gierahn et al. increase the efficiency and range of scRNA-seq capabilities, opening the door for unforeseen applications within high-risk or under-resourced environments.

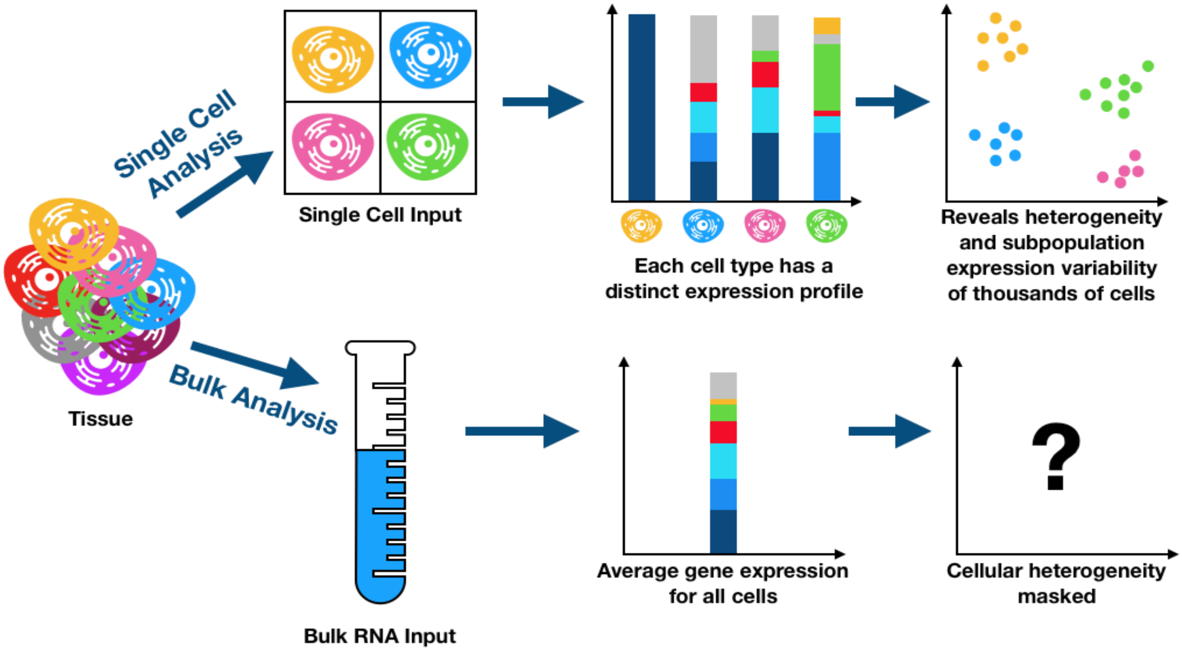

Single Cell Analysis
Single Cell Input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Tissue
Bulk Analysis
Bulk RNA Input
Average gene expression for all cells
?
Cellular heterogeneity masked

Figure 1: Bulk versus Single-Cell RNA Sequencing

1.3 Single-Cell RNA Sequencing Workflow

Since the Ebola virus is a lethal pathogen that requires Biosafety Level 4 containment, the first portion of the single-cell sequencing process is completed at the National Institutes of Health (NIH).

The process commences with a glass slide containing a small microarray of wells. Special spherical polymer beads containing unique DNA barcodes are washed over the microarray. The beads are sized such that one bead fits in each well (Gierahn et al. 2017). When harvesting blood from the infected nonhuman primates, researchers purify the immune cells by centrifugation. After scientists extract the immune cells they use a microscope to count the number of cells present and then dilute them such that the number of cells is approximately one tenth the number of wells in the microarray. The user then loads the cells onto the array. The dilution helps ensure that each well does not contain more than one cell (doublets). Doublets distort the results of single-cell sequencing; if one well contains two cells, scRNA-seq analysis would report this data as a single cell's transcriptome, when the well actually reflects the expression profiles of two separate cells. Conflating two transcriptomes for one during analysis can lead to contradictory or inaccurate results (Hwang, Lee, and Bang 2018; McGinnis, Murrow, and Gartner 2019).

Each bead is connected to a large array of primers. Each of these primers is comprised of a universal adapter named SeqB (that serves as a handle for PCR), a cell barcode, a transcript barcode, and an oligo polyT sequence (which helps anneal the polyA tail on mRNAs to the bead) (Figure 2). The cell barcode is common for all primers on a given bead, but differs across different beads. This approach allows us to identify each individual cell, since each well ideally contains one bead and one cell (Gierahn et al. 2017). The transcript barcode is different for each primer on a given bead, thus identifying each individual transcript (to prevent double counting of transcripts, which would inflate expression values) (Gierahn et al. 2017).

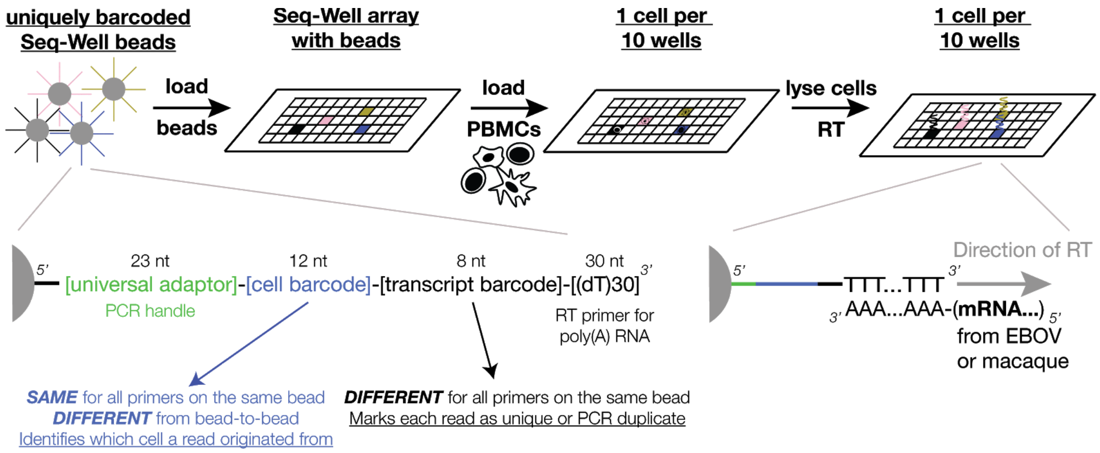

uniquely barcoded Seq-Well beads
Seq-Well array with beads
1 cell per 10 wells
1 cell per 10 wells
load beads
load PBMCs
lyse cells RT
5'
23 nt
12 nt
8 nt
30 nt
3'
[universal adaptor]-[cell barcode]-[transcript barcode]-[(dT)30]
PCR handle
RT primer for poly(A) RNA
Direction of RT
5'
3'
TTT...TTT
3' AAA...AAA-(mRNA...) 5'
from EBOV or macaque
SAME for all primers on the same bead
DIFFERENT from bead-to-bead
Identifies which cell a read originated from
DIFFERENT for all primers on the same bead
Marks each read as unique or PCR duplicate

Figure 2: Visual Summary of the extraction of single cells' transcriptomes (Lin, AE)

The user then washes the array with a high concentration of guanidinium, which denatures proteins, lyses the single cells, and releases the mRNA (the transcriptome). This released mRNA can then bind its polyA tail with the oligo dT sequence on the end of the bead's primer (Gierahn et al. 2017). Ideally, one mRNA sequence attaches to one primer, such that the bead is populated with all the mRNAs of a transcriptome.

Then, Reverse Transcription (RT) elongates along the transcript to produce a double stranded sequence. Once RT reaches the end of the template, it generates an oligo dC sequence. At the end of this RT step, a phenomenon known as "template switching" occurs. Scientists add an additional oligo (3' GGG-SeqB 5'), which then anneals onto the oligo dC generated at the end of RT. At this point, RT then elongates the *transcript* in accordance with the GGG-SeqB sequence, which now serves as the template, hence the term "template switching" (because the RT template starts as the mRNA and then switches to GGG-SeqB). The SeqB sequence also serves as the PCR handle attached to the bead; as a result, the SeqB oligo ends up on both sides of the cDNA attached to the bead (see Figure 3).

This product (at the end of Step 2 in Figure 3) is sent to our lab, where we finish the sequencing preparation. The RNA on the transcriptome hybrid and the original GGG-SeqB template are then degraded, leaving the reverse-transcript cDNA attached to the replicated GGG-SeqB sequence (Gierahn et al. 2017). Now all the material is single stranded cDNA. We then perform second strand synthesis to generate a complementary DNA strand, resulting in double stranded cDNA, with the 5' end attached to the bead (see Figure 3).

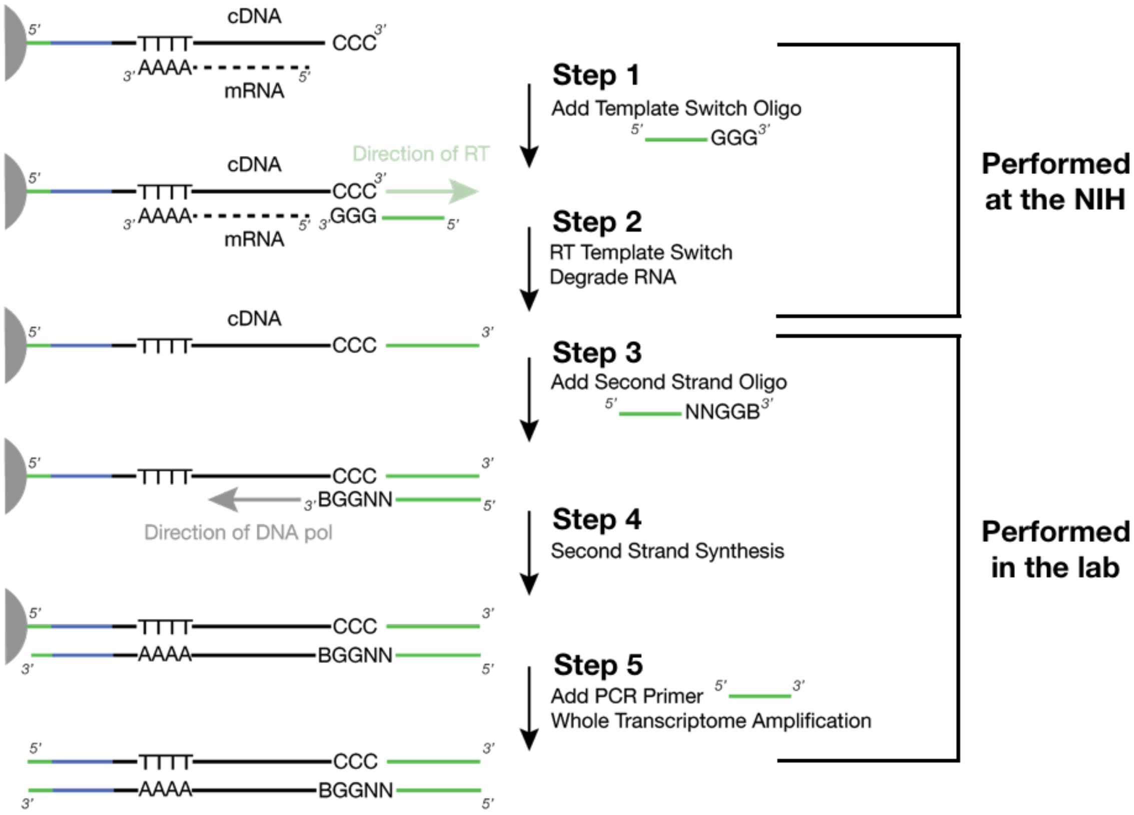


Figure 3: Visual representation of Template Switch and transcriptome amplification process (Lin, AE)

Following the template switching step, the cDNA is now flanked by the SeqB sequence on either side. As a result, scientists can amplify the pooled transcriptome using one copy of the SeqB primer, which anneals onto both ends of the cDNA. This is a benefit to the template switching step during RT; template switching allows for a single common primer to be used during amplification rather than two (Wulf et al. 2019). The results of this step consist of an amplified number of double stranded cDNAs, each containing a PCR handle (SeqB), barcodes, transcript, and the GGG-SeqB oligo.

The DNA product after amplification spans approximately one to two kilobases, which is too long for short-read sequencers such as Illumina. The Nextera kit's Tn5 Transposase enzyme solves this problem by making a single cut at a random site in each dsDNA and adding its specific transposon sequence (Ran et al. 2015).

During Nextera PCR, a primer containing the P5 end attached to a SeqB oligo anneals to the SeqB sequence on the 3' end of the dsDNAs (see Figure 4). This primer does not attach to the 5' SeqB sequence, as the P5 ends in an 'AC' sequence, which will not anneal with the 5' end of the cDNA. The Nextera P7 reverse primer then anneals to the transposon sequence that was attached earlier. The resulting segments will span the PCR handle and barcodes as well as a random portion of the transcriptome. We only need a portion of the transcript (not all) since we merely seek to identify the gene (by its geneID) and count it, which does not require a comprehensive sequence. We use this portion of the transcript to identify the corresponding gene using mapping software. Then, we calculate the frequency of each gene per cell and infer which amplified genes produce a unique expression profile and are differentially expressed in infected cells.

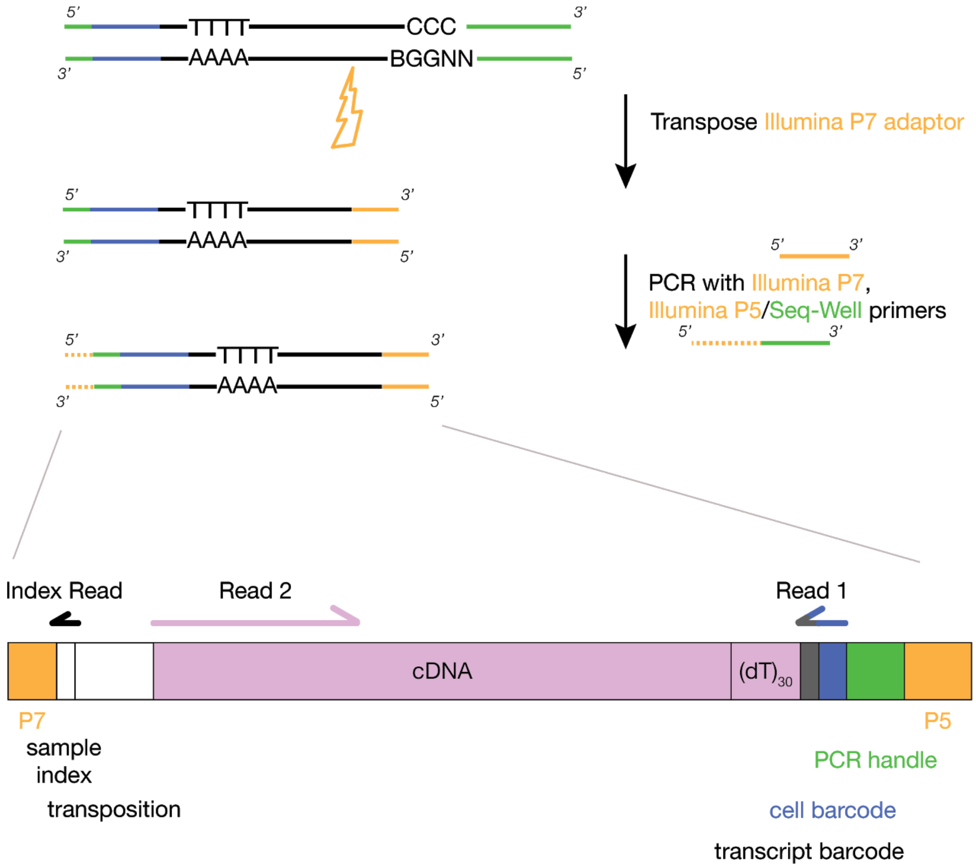


Figure 4: Visual representation of Nextera library preparation after isolating the dsDNA transcriptome (Lin, AE)

1.4 Challenges and Areas of Improvement in scRNA-Seq

1.4.1 'SeqB' Over Amplification

While a powerful technology, scRNA-seq comes with a number of challenges, particularly in the Biosafety Level 4 setting required for this study. In particular, when blood samples from nonhuman primates infected with Ebola underwent scRNA-seq, Illumina sequencing generated a mere 20% of the expected 3.7 billion reads. Upon further examination, our lab discovered a short 17-base long artifact of the scRNA-seq preparation (SeqB) had over amplified, producing multimers, and affecting the quality of the entire sequencing run. This obstacle could have stemmed from either low quantity of the original RNA, and/or RNA degradation due to transport of samples and RNA instability. Regardless, this apparent amplification generated excessive amounts of SeqB multimers, interfering with our lab's ability to extract meaningful information from the valuable transcriptomes collected during this high-risk study. I address this problem of SeqB over amplification in Chapter 2 of this thesis, where I detail our efforts to deplete the SeqB multimer using CRISPR-Cas9.

1.4.2 Deep Sequencing of a Specific Transcriptome of Interest

Another area of growth within the field of scRNA-seq lies in amplifying the transcriptome of a specific cell of interest. For example, if a specific cell exhibits an interesting or abnormal genetic profile, scientists hoping to further examine this cell with scRNA-seq would undertake a large-scale NovaSeq run, which requires large amounts of input DNA and is time consuming. Scientists could potentially obtain the same valuable information from a single cell from a smaller-scale MiSeq run that does not necessarily consume the resources of the NovaSeq (the machine currently employed to sequence most scRNA-seq assays). As a result, transcriptomics studies would benefit greatly from the ability to isolate and deeply sequence the transcriptome of a single cell of interest. In Chapter 3, I address this area of growth, detailing our efforts to develop a "cell specific PCR" to selectively amplify and sequence a single transcriptome of interest within a pool of cell transcriptomes.

1.5 Problem Statement and Research Objectives

In this project, researchers at the National Institutes of Health (NIH) infected rhesus macaques with Ebola virus, and extracted from blood and tissues at various time points. Blood samples from these infected nonhuman primates then underwent single-cell RNA sequencing in our lab.

In collaboration with the researchers at NIH, who conducted the relevant BSL4 work in their Bethesda, Maryland facility, this project aims to elucidate the ways different cell-types modulate their gene expression levels in response to Ebola virus infection over time. The specific aims of my thesis are to design and implement original molecular and quantitative methods to:

1) Deplete SeqB multimer contamination using CRISPR-Cas9 (Chapter 2)
2) Selectively amplify and sequence the transcriptome of a single cell of interest by designing a cell specific PCR (Chapter 3)

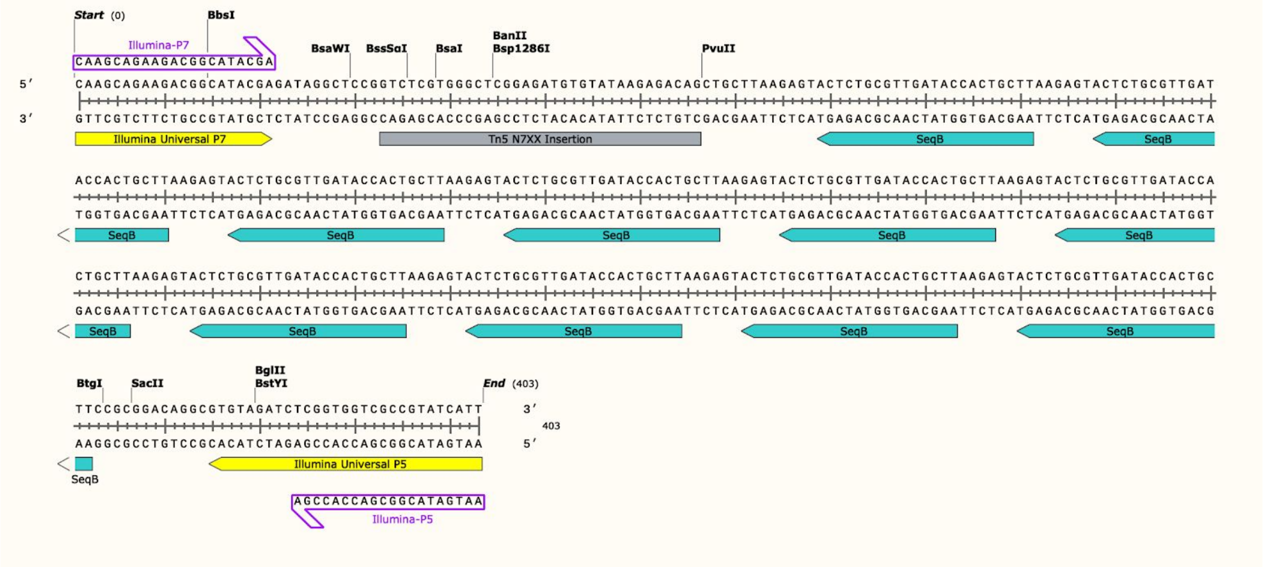
I then use cell-culture to explore new applications of scRNA-seq through a study of lymphocytic choriomeningitis virus (LCMV), an arenavirus containing four genes that can be easily isolated and co-expressed within the confines of a BSL2 facility (Chapter 4).

2. Using CRISPR-Cas9 to Deplete Abundant Multimers for scRNA-Seq

<u>2.1 Introduction</u>

2.1.1 Multimer 'SeqB' Contamination

Sequencing the initial samples prepared by the standard scRNA-seq protocol resulted in low-yield sequencing runs, generating a mere 750 million reads, 20% of the expected 3.7 billion reads normally generated. This yield is not enough to interpret the results of over 60 samples, each with thousands of cells containing its own transcriptome. Upon further examination, we discovered that this low yield was due to an overamplification of the adapter sequence "SeqB." We observed overwhelming amounts of SeqB multimers in a number of reads, one of which is depicted in Figure 5.



Start (0)
BbsI
Illumina-P7
CAAGCAGAAGACGGCATACGA
BsaWI
BssSαI
BsaI
BanII
Bsp1286I
PvuII
5′
CAAGCAGAAGACGGCATACGAGATAGGCTCCGGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGAT
3′
GTTCGTCTTCTGCCGTATGCTCTATCCGAGGCCAGAGCACCCGAGCCTCTACACATATTCTCTGTCGACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTA
Illumina Universal P7
Tn5 N7XX Insertion
SeqB
SeqB
ACCACTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGATACCA
TGGTGACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTATGGT
SeqB
SeqB
SeqB
SeqB
SeqB
CTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGATACCACTGCTTAAGAGTACTCTGCGTTGATACCACTGC
GACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTATGGTGACGAATTCTCATGAGACGCAACTATGGTGACG
SeqB
SeqB
SeqB
SeqB
SeqB
BtgI
SacII
BglII
BstYI
End (403)
TTCCGCGGACAGGCGTGTAGATCTCGGTGGTCGCCGTATCATT
3′
403
AAGGCGCCTGTCCGCACATCTAGAGCCACCAGCGGCATAGTAA
5′
Illumina Universal P5
SeqB
AGCCACCAGCGGCATAGTAA
Illumina-P5

Figure 5: Schematic of a read that is contaminated with the SeqB multimer
As seen above, the entire read contains SeqB copies (blue) and not much other genetic information; it does not provide any transcriptome information and can interfere with our ability to obtain useful reads. This is an example of the type of sequence we aim to deplete with CRISPR-Cas9.

This SeqB multimer issue places a heavy burden on our sequencing efforts. Firstly, the amplified multimers overwhelm the sample, leading to the sequencer generating many reads that reflect the multimer, rather than meaningful transcriptome sequences. Secondly, multimers can cause malfunctions in the sequencer, which can disrupt the quality of useful reads (containing the transcriptome) during the sequencing run, and in many cases, lead to machine crashes. The machines experience these complications in the presence of multimers because Illumina sequencing relies on a primer adding one base to each read and taking photographs of fluorescence across the flow cell (to which DNAs are attached) (Wulf et al. 2019; "What Is the Illumina Method of DNA Sequencing?" 2014). If a given read contains many copies of this SeqB sequence, and the sequencing primer anneals to the same SeqB sequence, the primer will attach onto the same DNA at multiple points and add many bases instead of one. Consequently, adding multiple bases at a single site causes a drastic increase in light intensity, resulting in a large spot on the resulting sequencing image taken by the sequencing machine. The camera tries to tune its settings, normalizing to this increased intensity, which then severely decreases the intensity of non-SeqB-multimer reads (transcriptome sequences) where the sequencing primer only annealed once. This severely decreases the quality of the run, yielding little useful information.

The appearance of SeqB multimers could be due to a number of issues detailed below:

1. <u>Low Input or Low Quality RNA:</u> If the number of cells is too low or the RNA quality is not reliable from the start, genetic noise within the sample can amplify during PCR. Specifically, low input or low quality RNA can cause a relative excess of the SeqB sequence within the sample. This excess SeqB can subsequently overamplify. This continuous amplification could lead to the

generation of excessive SeqB, the sequence that anneals to a PCR primer used in the library preparation method.

2. Hairpin Formation and Excess SeqB: After RT and template switching, when we add the second SeqB sequence onto the 5' end, it is possible that the SeqB on each end could fold into a hairpin. It would not anneal perfectly, allowing a part of the 5' SeqB to be exposed (Figure 6). During PCR, the first SeqB could be amplified. Since the primers are also composed of SeqB, SeqB could be priming on itself during the PCR annealing phase and elongating on itself to the multimers that we have observed.

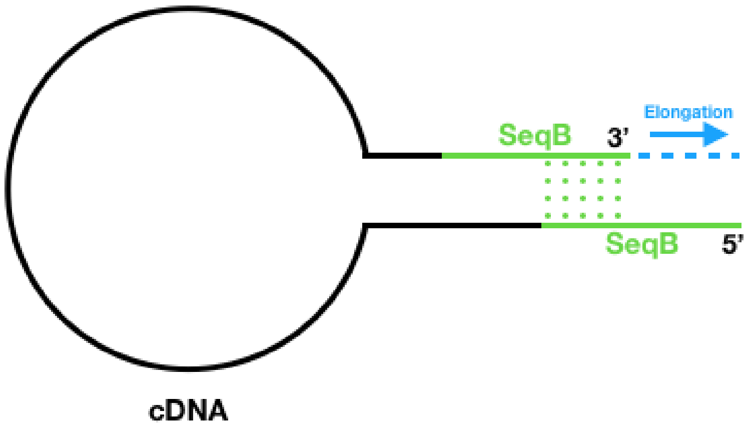


Figure 6: Potential hairpin formation in a cDNA that could give rise to the observed SeqB multimers

2.1.2 CRISPR-Cas9

A possible approach to solving the multimer contamination issue is to ensure that no multimers enter the sequencer. We implemented a CRISPR-Cas9 assay, or DASH (Depletion of Abundant Sequences by Hybridization) to address this issue.

The CRISPR-Cas9 complex consists of a guide RNA (sgRNA) attached to a Cas9 cutting protein. The sgRNA contains a spacer region, which anneals to the DNA target, as well as a scaffold, which attaches to the Cas9 protein. A protospacer adjacent motif (PAM) site in the template DNA is also required. The Cas9 recognizes the PAM site and starts scanning the 5' end to see if it is complementary to the sgRNA (Gu et al. 2016). If it is, the cutting of the Cas9 is activated, which allows the Cas9 to cleave the template at the desired location (Figure 7).

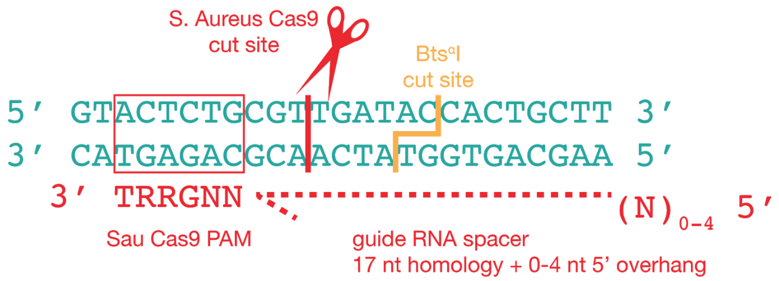

S. Aureus Cas9
cut site
Bts^αI
cut site
5′ GTACTCTGCGTTGATACCACTGCTT 3′
3′ CATGAGACGCAACTATGGTGACGAA 5′
3′ TRRGNN (N)0-4 5′
Sau Cas9 PAM
guide RNA spacer
17 nt homology + 0-4 nt 5' overhang

Figure 7: Visual representation mapping template DNA with the Cas9 cut site, PAM, and restriction enzyme cut site (Lin, AE)

The 23 base-long SeqB sequence contains a PAM site recognized by *Staphylococcus aureus* Cas9 (SaCAS9, PAM: 5' NNGRRT 3') (Tang et al., 2018). By designing custom guide RNAs annealing to the SeqB sequence, we can design an assay in which the CRISPR-Cas9 selectively cleaves dsDNAs containing the SeqB multimers at their many locations. Since this method would digest the multimers into small pieces, a simple cleanup would filter out the multimers, leaving the high-quality longer strands intact. Since the Cas9 cleaves the 3' end containing SeqB and required P5 end for sequencing, we subsequently reattach a new P5 end to restore the required sequencing end.

2.2 Materials and Methods

2.2.1 Synthesizing SeqB sgRNAs and DNA test substrates for Depletion of Abundant Sequences by Hybridization (DASH)

The 23-base SeqB adapter lacks the requisite 5'-NGG-3' protospacer adjacent motif (PAM) of the commonly used *Streptococcus pyogenes* Cas9 (SpyCas9), but contains the 5'-NNGRRT-3' PAM corresponding to *Staphylococcus aureus* Cas9 (SauCas9) (Tang et al., 2018). We designed a single guide RNA (sgRNA) targeting all 17 bases of SeqB adjacent to the six-base SauCas9 PAM. We prepended the sgRNA with four random bases and the first 'G' transcribed by T7 RNA polymerase during *in vitro* transcription (IVT), since short SauCas9 spacers have decreased cutting efficiency (Tycko et al. 2018). For our negative controls, we designed three sgRNA spacers containing either a scrambled sequence (based on GenScript's Sequence Scramble tool), a negative control sequence (based on IDT's Negative Control DsiRNA), or an enhanced green fluorescent protein sequence (based on IDT's EGFP-S1 DsiRNA).

To produce these sgRNAs, we ordered custom Gene Fragments (gBlocks) from IDT containing the spacer and SauCas9 scaffold sequences, cloned these gBlocks into plasmids and performed Sanger sequencing (Genewiz, South Plainfield, NJ) to ensure the absence of mutations. We generated linear IVT template DNA by PCR, using a forward primer that included the T7 RNA polymerase promoter, 0–4 random bases, the spacer sequence, and a reverse primer that annealed to the SaCas9 scaffold. We gel purified IVT template DNA, followed by 2.0X SPRI cleanup to remove residual guanidinium.

We then performed IVT using the MEGAshortscript T7 Transcription kit (Thermo Fisher Scientific, Waltham, MA) following the manufacturer's protocol, with the addition of 10U SUPERase-In RNase Inhibitor, at 37 °C for 16 hours. We purified sgRNAs using the RNA Clean & Concentrate Kit (Zymo, Irvine, CA) and verified the correct sgRNA length on a 15% TBE-urea gel (Thermo Fisher Scientific, Waltham, MA) stained with 1X SYBR Gold (Thermo Fisher Scientific, Waltham, MA).

As positive and negative controls, we cloned individual reads from library RA1639.D005.fresh.a1.std (~30% multimer content) into plasmids grown in *recA-* Endura Chemically Competent cells (Lucigen, Middleton, WI) to avoid rejection of SeqB multimers. We verified each sequence by Sanger sequencing (Genewiz, South Plainfield, NJ) and selected one pure cDNA read and one contaminated SeqB multimer read for testing. See Table 1 for the candidate control reads and their respective multimer content. We generated linear DNA by PCR with the Illumina P7 and P5 primers, gel purified, then SPRI purified each test DNA.

Read (#)	Length (bp)	SeqB count (#)
1	435	10
2	**524**	**12**
3	826	1
4	**634**	**1**
5	223	2
6	357	1
7	516	2
8	477	11
9	**372**	**1**
10	371	1
11	**403**	**10**
12	**350**	**1**

Table 1: Candidate template reads for the depletion experiment and their respective number of SeqB multimer repeats
In our preliminary testing, we used Templates 4, 9 and 11 for the depletion. Templates 4 and 9 represent a pure and uncontaminated template, with only one SeqB copy, and Template 11 represents a problematic adapter multimer we hope to deplete, containing 10 copies in a 400-base-long sequence. The restriction digest used Templates 2, 4, 9, 11, 12, and 12 no Bts$^{\alpha}$I cut site (not listed in table, discussed below).

2.2.2 Restriction Enzyme Validation

Before testing the Cas9 depletion, we tested the cutting of SeqB multimers using Restriction Enzyme (RE) digestion as a proof-of-concept. REs are single-subunit enzymes that serve as the gold standard for highly efficient and sequence-specific digestion. Therefore, we used RE digest as a benchmark for our new method.

The SeqB sequence contains a RE cut site for the enzyme Bts$^{\alpha}$I (Figure 6). Like most REs, Bts$^{\alpha}$I recognizes a six-base pair sequence, which occurs at a probability of 0.024% (1/4096) in random sequence. Indeed, by chance, one of the cDNA templates we used (Template 12, Table 1) also contained an internal Bts$^{\alpha}$I RE site within its cDNA sequence. To verify that Bts$^{\alpha}$I was cutting specifically, we performed site-directed mutagenesis to create an alternate Template 12 that did not contain the internal Bts$^{\alpha}$I RE site (called '12 noBts$^{\alpha}$I). Testing the digest on the template without the internal Bts$^{\alpha}$I cut site would allow us to directly compare cutting efficiency between two templates with and without the cut site (Figure 8).

We digested 100ng of each template with 10U of Bts$^{\alpha}$I enzyme (1µL 10,000U/mL) in 1X CutSmart buffer at 55˚C for 2 hours. We then added 10µL EDTA followed by a 0.8X SPRI with a 10µL elution. Finally, we performed qPCR to measure the rate at which templates reamplified. Since the enzyme cuts a part of the SeqB fragment, we ran the qPCR using Illumina P7 primer with a P5 primer attached to the remaining SeqB fragment with thermocycling conditions: 98˚C 30 seconds; and 12 cycles of 98˚C 10 seconds, 50˚C 30 seconds, and 72˚C 15 seconds.

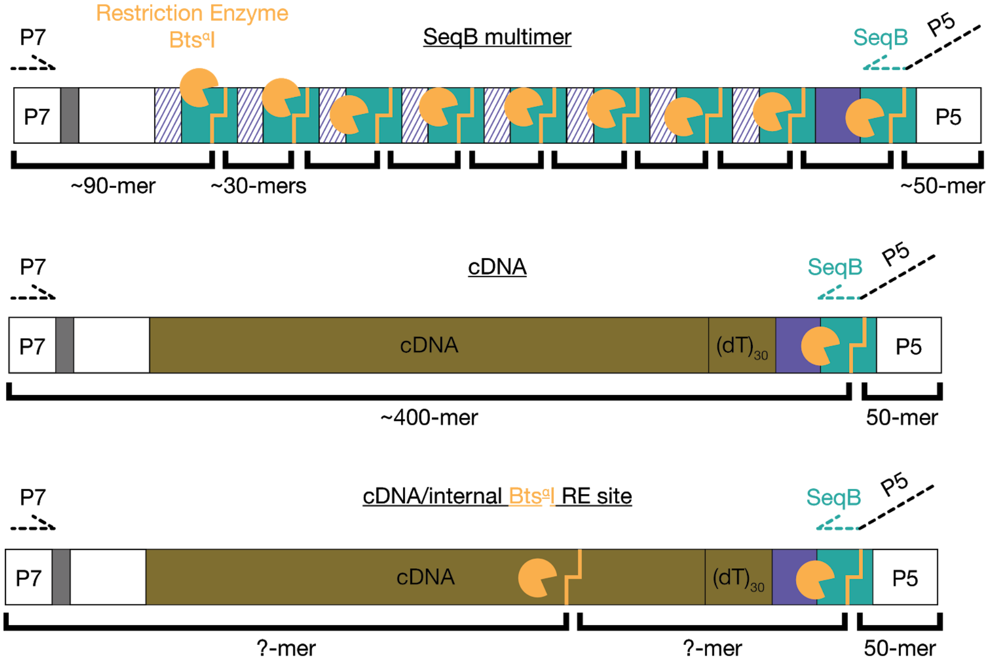


Figure 8: Visual representation of expected restriction enzyme cutting for various templates
The bottom template represents Template 12, which contains a secondary Bts$^{\alpha}$I cut site outside of the SeqB sequence at the end. The middle template represents a useful transcriptome with only one copy of SeqB. The top template represents an adapter multimer. Because the top template contains many copies of SeqB, the RE is expected to cut at each site, digesting the template into small pieces (Lin, AE).

2.2.3 DASH optimized for S. aureus Cas9

We followed the SauCas9 digestion conditions listed in (Kaur, n.d.; Biolabs and New England Biolabs, n.d.), except that we used an increased sgRNA:SauCas9:DNA ratio because each SeqB DNA multimer contained multiple copies of the SeqB target sequence. We first incubated 5 pmol SauCas9 and 10 pmol sgRNA in 1.05X NEBuffer 3.1 at 25 °C for 10 min. To start the digestion reaction, we then added 5 fmol of DNA library (molar ratio of 2000:1000:1 sgRNA:SauCas9:DNA) for a final volume of 20 µL and incubated the mixture at 37 ˚C for 2 hours. We quenched the reaction by adding a final concentration of 50 µM EDTA, 1% SDS, and

0.1 U/µL Proteinase K. To remove short, digested fragments, we performed two 0.8X SPRI cleanups using AmpureXP beads (Beckman Coulter, Brea, CA).

Digesting SeqB also removes the Illumina P5 sequence from non-multimer, proper cDNAs; therefore, we reattached P5 by PCR using a primer that anneals to the remaining SeqB fragment and contains an overhang with the P5 sequence: We performed PCR using NEB Next Ultra II Q5 Master Mix (New England Biolabs, Ipswich, MA) and thermocycling conditions: 98˚C 30 sec; 12 cycles of 98˚C 10 sec, 50˚C 30 sec, 72˚C 15 sec, followed by a 0.8X SPRI purification.

During initial testing, we visualized the results of this depletion using quantitative PCR (qPCR) as well as tapestation, a method that separates dsDNA bound to dye by size through electrophoresis within small capillaries stored in a "tape." When performing the depletion on experimental samples (coming directly from contaminated nonhuman primate libraries), we quantified and pooled the libraries for subsequent sequencing.

2.2.4 Computational Analysis Methods

To interpret the sequencing results, I used jupyter Python notebook and Matlab. Upon gathering the sequencing data, I wrote methods to extract sequences from fastq files and determine the proportion of reads with multimer contamination. I then refined my code to account for single nucleotide polymorphisms that would not directly match the SeqB multimer sequence but were nonetheless contaminated. To visualize our results, I then created a program that could visually label the SeqB multimer, mutated SeqB, polyA tail, and terminating sequence in each read by replacing the sequence with a user-friendly label. This labeling method allows us to interpret our results in a sequence specific manner, allowing the naked eye to visualize multimer placement with respect to the predicted structure of the read. The full code can be found in the appendix at the end of this report.

2.3 Results

2.3.1 Preliminary Depletion Testing

When generating our DNA control libraries for testing, we selected three control templates to test from the reads listed in Table 1: Template 4, Template 9, and Template 11. After Sanger sequencing, we verified that Templates 4 and 9 contained only one copy of the SeqB sequence, and thus represented the "pure" control samples in our depletion tests. Template 11, on the other hand, consisted entirely of SeqB multimers (10 copies, as seen in Table 1), and thus represented the fully contaminated control. Consequently, a successful depletion would allow Template 9 to remain intact and appear similarly before and after depletion, while fully digesting Template 11 to show negligible content after depletion.

The preliminary CRISPR-Cas9 depletion tests were successful in that they reduced the concentration of the multimer template without severely affecting the template lacking the multimer (Figure 9). Figures 9A and B, showing the qPCR after the reactions on Template 11 (containing many SeqB multimers) shows a clear difference in quantity between samples that underwent the depletion and samples that did not. Meanwhile, Template 9 (which only had one SeqB copy) showed little difference in the amplification plots of depleted and undepleted samples. This shows that the CRISPR-Cas9 reaction does, in fact, selectively degrade the multimers to some extent. Furthermore, we sought to explore whether the efficiency of Cas9 is dependent on the length of sgRNA. As a result, we added zero to four extra bases to the SeqB sgRNA spacer sequence and tested the various lengths in our depletion reaction. The data shows little difference between the zero and four-base-long random add-ons to the sgRNA (named 0N and 4N). In Figure 9C, we can confirm once again that the depletion can effectively eradicate solutions with 100% multimer. Furthermore, by creating various mixes of Template 11 and Template 9, we generated test libraries containing an intermediate percentage of multimer (25% and 50%), and tested the depletion (Figure 9C). As the percent multimer increases, the bands become thinner, as expected. 4N sgRNA seems to produce slightly thicker bands than 0N, which may imply the ability to preserve more of the pure sample.

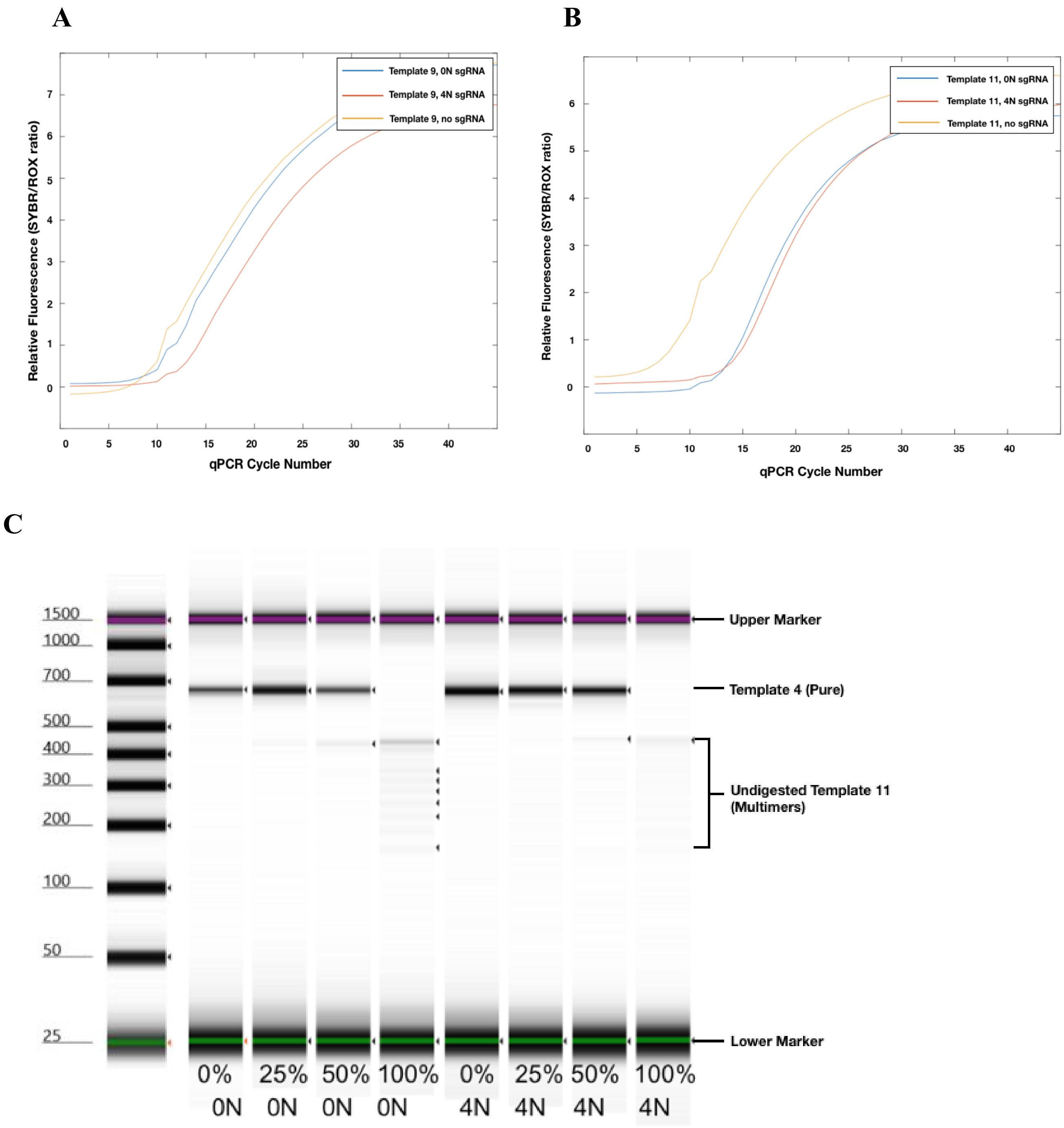


Figure 9: Comparing the CRISPR-Cas9 depletion on pure versus contaminated control templates using various sgRNA lengths

(A) Depletion on Template 9 (uncontaminated control). (B) Depletion on Template 11 (contaminated control) (C) Using tapestation to test additional varied concentrations of multimer template (Template 11, 403 bp and lower; Template 4, 634 bp) alongside varying sgRNA lengths (0N and 4N add-ons).

We then tested specific sgRNA controls to ensure that our assay was sequence-specific. We generated three different sgRNA controls, each with a specific purpose:

1. “Scrambled” sgRNA: This sequence consists of the normal SeqB sgRNA sequence but in randomized order. This control preserves nucleotide content (same percent GC) and ensures that the depletion does not target SeqB. However, there is no guarantee that it does not target any other sequence in our assay.
2. “GFP” sgRNA: This sequence targets GFP. Unlike the scrambled sgRNA, we can ensure that this sgRNA sequence does not target any sequences in rhesus macaque.
3. “Negative Control” sgRNA (NC): This sgRNA serves as a targeting control; it is designed by IDT not to target any sequence in human, mouse, or rat.

As expected, the SeqB sgRNA left the pure sample (0% lane) fully intact with a thick band while depleting the other solutions containing multimers. The control guides also performed as expected (Figure 10). For Scrambled, GFP, and Negative Control sgRNA sequences, the 0% multimer remained fully intact, and the thickness of the pure product’s band decreased as percent multimer increased. Simultaneously, as SeqB multimer content increased, the thin undigested bands began to appear for all control guides, while appearing in very low quantities within the product containing experimental sgRNA targeting SeqB, as expected.

We then optimized the depletion assay, testing various run-times and SeqB sgRNA concentrations. The tapestation in Figure 11 suggests that all assays were able to effectively deplete the multimer samples, though the long and high concentration run proved most efficient in clearing Template 11’s multimers, and only exhibited bands corresponding to Template 9.

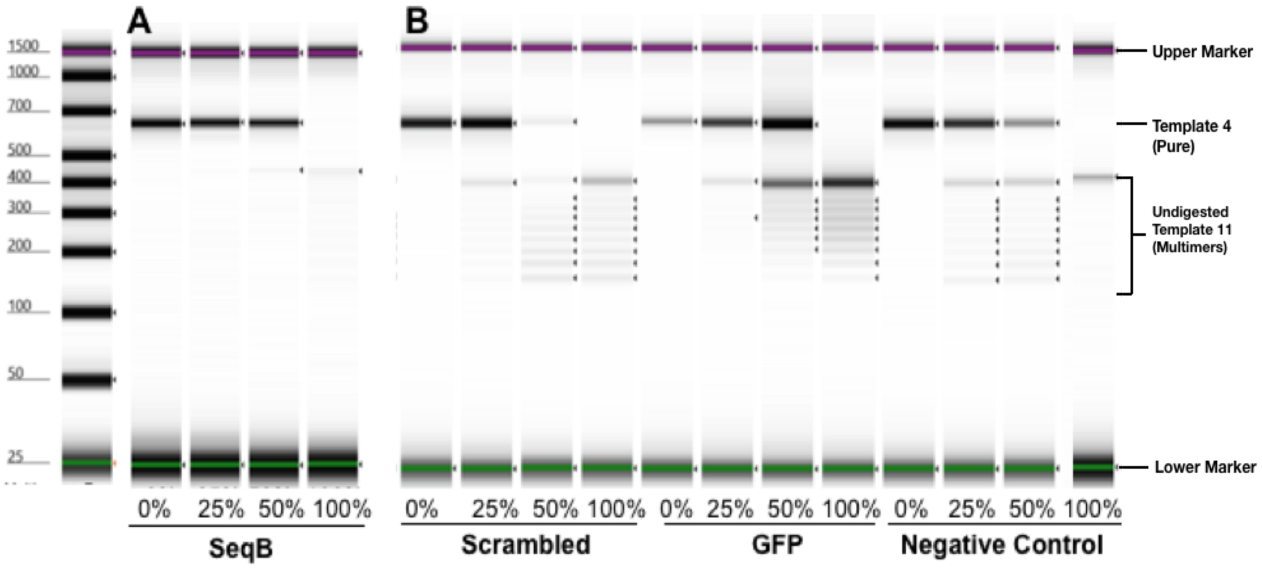


Figure 10: Testing Depletion with sgRNA controls
Tapestation (electrophoresis through capillaries that separate dsDNA by size) results from the depletion test against control sgRNAs at varying multimer concentrations (percent). As expected, Template 4 is preserved at 634 bp in all assays, while Template 11 is absent in depleted samples, and undigested in control sgRNAs (403bp and lower).

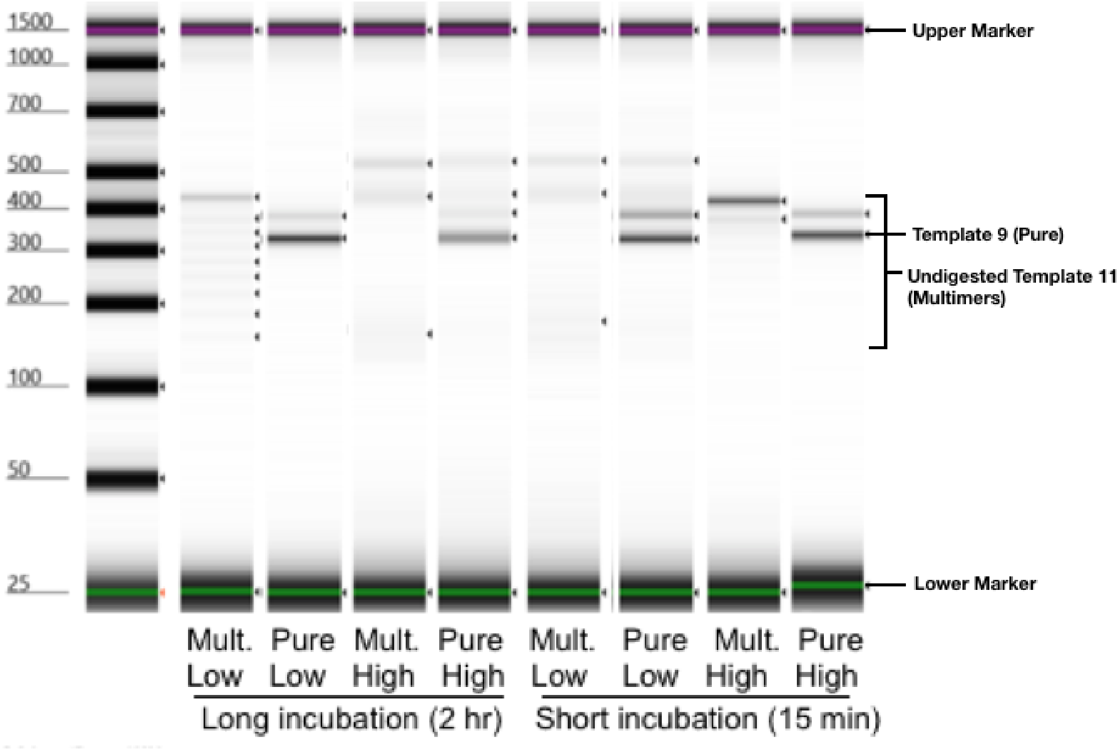


Figure 11: Depletion Reaction optimization results
"Mult." refers to a contaminated sample (Template 11) containing excess SeqB copies (403 bp and lower), "Pure" refers to an uncontaminated library (Template 9) containing only one SeqB copy (372 bp). "Low" and "High" refer to sgRNA concentrations (High: 2000:1000:1 sgRNA:SauCas9:DNA, Low: 100-fold RNA and Cas9 dilution relative to DNA). Incubation time refers to the duration that the Cas9 was left to deplete the sample.

2.3.2 Restriction Enzyme (RE) Validation

The RE digest was successful, selectively cutting regions containing Bts$^{\alpha}$I, while leaving other regions intact, as seen in Figure 12. Templates that only contained one copy of the SeqB sequence displayed similar amplification plots with and without the restriction enzyme (Templates 4, 12 noBts$^{\alpha}$I, and 9 – Figures 12A, 12B, and 12E). Templates with many SeqB multimers exhibited sharp differences when exposed to the REs as compared to uncut (Templates 2 and 11 – Figures 12D and 12F). When comparing Template 12's digest with the digest of Template 12 without the second Bts$^{\alpha}$I cut site (Figures 12B and C), the RE made the additional cut to the template containing the Bts$^{\alpha}$I cut site, resulting in the shifted curve (representing a smaller quantity). These results verify the ability to accomplish specific cutting and digestion targeted at sites that match a region of the SeqB sequence, and serve as a point of reference when evaluating the results of the test CRISPR-Cas9 depletion reactions. By comparing the depletion results to the RE digest, we can gain insight into the efficiency and effectiveness of the CRISPR-Cas9 cutting, relative to our verified RE digest method.

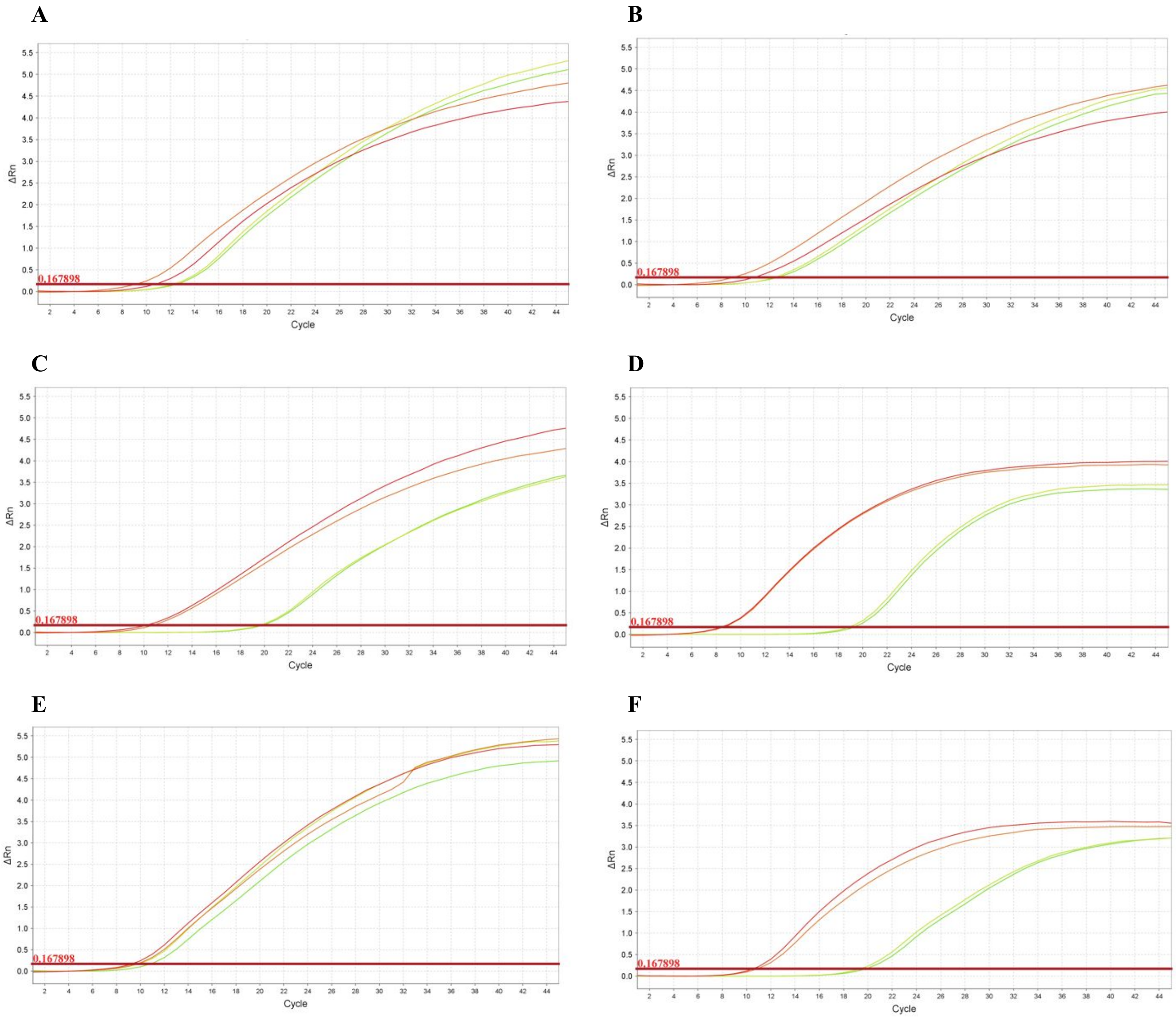


Figure 12: Restriction Enzyme Validation Assays

(A) qPCR Amplification plot for Template 4 with and without restriction digestion. Red curves represent the uncut samples and green curves represent the cut samples. There are two curves per color because duplicates were tested. (B) qPCR Amplification plot for Template 12 not containing the second Bts$^{\alpha}$I site, with and without restriction digestion. Red curves represent the uncut samples and green represent the cut samples. (C) qPCR Amplification plot for Template 12 containing the second cut site with and without restriction digestion. Red curves represent the uncut samples and green curves represent the cut samples. (D) qPCR Amplification plot for Template 11 with and without restriction digestion. Red curves represent the uncut samples and green curves represent the cut samples. (E) qPCR Amplification plot for Template 9 with and without restriction digestion. Red curves represent the uncut samples and green curves represent the cut samples. (F) qPCR Amplification plot for Template 2 with and without restriction digestion. Red curves represent the uncut samples and green curves represent the cut samples.

2.3.3 Depletion on Experimental Samples

Having confirmed the depletion viability on control libraries, we ran the depletion on 48 samples coming directly from nonhuman primates infected with Ebola. The samples we chose had previously displayed high amounts of SeqB multimer contamination, similar to that of contaminated test libraries. The sequencing results of the depletion show that for the majority of contaminated experimental libraries, the multimer was effectively eliminated. In fact, on average, multimer quantity decreased 10-fold on average, with one sample surpassing a 36-fold reduction. However, five samples (bolded in Table 2) showed very little reduction in their multimer content. Further analysis showed that samples with a higher amount of initial multimer content, exhibit less efficient depletion (Figure 13A). This suggests that very high multimer content could be the result of low levels of input RNA, leading to low amounts of cDNA template undergoing scRNA-seq library preparation. To explore why a select few libraries did not deplete as effectively, I computationally examined a contaminated sample that did not exhibit a promising level of depletion and compared it to a pure, uncontaminated experimental sample. This comparison served to provide a more detailed break-down of the reads from each sample, which could provide insight in determining the cause of the depletion inefficiencies. I first performed initial calculations of percent multimer present in each sample, and validated them with the corresponding values provided by the sequencing software. However, I also designed a program that would detect the presence of sequences that matched SeqB allowing for 1 SNP, allowing us to account for underestimates of multiple percentages due to slight deviations from a perfect SeqB match. This recalculation increased the percent multimer present from 85 to 93% (Figure 13B). Consequently, this high percentage could be indicative of a sample with little to no initial template to begin with, allowing the multimer to completely encompass the sample, leading to the unexpectedly low depletion rates observed.

A

Percent Multimer Pre Depletion	Percent Multimer Post Depletion	Fold Reduction (Pre/Post)
25.419	2.443	10.402
22.979	3.445	6.700
13.659	1.454	9.391
27.232	**0.740**	**36.780**
28.557	7.529	3.793
45.916	**46.498**	**0.987**
46.035	**22.728**	**2.025**
21.527	1.871	11.504
34.909	3.459	10.092
18.389	1.281	14.356
10.100	0.319	31.624
33.599	1.952	17.209
15.746	1.001	15.728
17.668	1.100	16.124
32.256	2.995	10.770
19.314	1.761	10.967
45.916	**43.135**	**1.064**
40.808	**26.096**	**1.564**
45.378	**30.340**	**1.496**
38.616	8.171	4.726
46.035	12.840	3.585
44.089	10.137	4.349
39.433	8.075	4.883
	MEAN FOLD REDUCTION:	**10.004**

Table 2: Fold reduction table for each depleted sample
Bolded rows represent samples that did not deplete as effectively (two-fold reduction and below). Underlined rows represent samples that depleted exceptionally effectively.

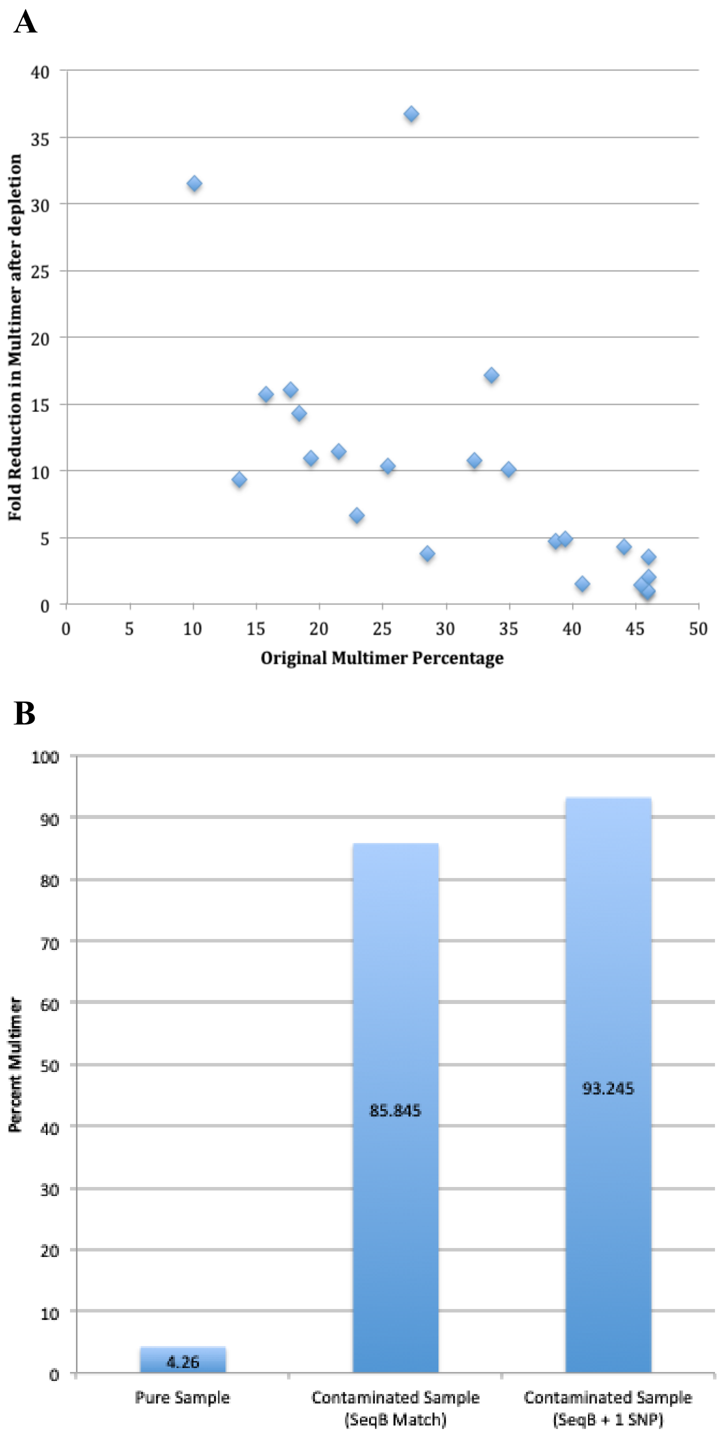


Figure 13: Examining reductions in Multimer contamination within experimental scRNA-Seq Libraries before and after CRISPR depletion

(A) Fold Reduction in Multimer After Depletion versus Original Multimer Percentage. (B) Calculating multimer concentrations accounting for 1 SNP.

2.4 Discussion

The results above exhibit optimized testing conditions, detailed control testing, and a successful attempt to deplete abundant multimers in contaminated libraries. The results show that on average, the CRISPR-Cas9 depletion reduced multimer concentration by over 10-fold. However, while the majority of contaminated libraries were effectively purified, five libraries showed little to no decrease in multimer percentage after depletion. This poses a unique conundrum: Why would most of the libraries deplete as expected, and yet a small subset shows little depletion? The libraries that showed less depletion tended to have a very high percentage of multimer in the original library. As the original multimer percentage increased, the depletion efficiency decreased. One hypothesis to explain this result is that the library had low quantity and quality DNA to begin with. It could be that a very low quantity and quality sample made it into the SeqWell preparation; as a result, since there is very little DNA template entering the scRNA-seq library preparation, SeqB could multimerize and subsequently amplify upon itself. The specific mechanism might be at the template switch phase (described in background), which could involve the addition of many units of SeqB onto the end of the DNA. To explore this further, we seek to examine specific sequences and their structural makeup in the future, which could potentially inform how these multimers arose in the first place.

Our success in depleting multimer contamination has broad implications that extend beyond the specific SeqWell preparation method. Users could potentially apply this method to a wide range of experiments in which over-amplified multimers have contaminated cDNA libraries. Furthermore, the results of our work could potentially inform scientists using droplet-based approaches to scRNA-seq. For example, a common issue is the “empty droplet” problem, in which it may be difficult to differentiate between a noisy or abnormal transcriptome due to an “empty droplet” (no input), and excessive contamination of a droplet that did contain input RNA (Lun et al. 2019). Our results suggest that depletion of contaminant multimers could help solve this issue; The depletion may preserve transcriptomes that might have been contaminated, while droplets with no input would maintain noise after depletion.

Overall, this experimental success marks a newfound ability to preserve valuable sequencing libraries that may have experienced contamination during the rigorous scRNA-seq library preparation, and may expand past the realm of the specific SeqWell workflow. Furthermore, in preserving over 15 libraries that had failed previously, we employed this depletion to sequence over 1000 additional transcriptomes. We are now using this newly generated data to produce a deeper, and more comprehensive analysis of cell specific responses to Ebola virus infection, an exploration that has grand implications in future outbreak prevention and response.

3. Designing a Cell Specific PCR to Amplify the Transcriptome of a Single Cell

3.1 Introduction

scRNA-seq has become a commonplace effort in profiling the transcriptomes of thousands of single cells. As described in Section 1.3, the scRNA-seq methodology rests on pooling many libraries together, sequencing this pool, and separating cells and transcripts by their individual barcodes. This places a challenge for scientists who may seek to profile a few select cells within this massive pool of libraries. To address this challenge, we design and implement a "cell specific PCR," a novel method that scientists can perform before sequencing to amplify the transcriptome of specific cells of interest within a greater pool of transcriptome libraries. Specifically, a cell specific PCR would selectively amplify transcriptomes containing a unique cell barcode corresponding to a specific cell of interest within a pool of cell transcriptomes, with the hope of gaining more information on expression levels within the specific isolated cell.

Previous research by Ranu et al. exhibits success using a cell specific PCR within the FACS single-cell sequencing methodology; the study targeted ultra-rare cell-types, and, in doing so, reduced the "required sequencing effort to profile single cells by 100-fold" (Ranu et al. 2019).

Cell specific PCR can be accomplished by taking advantage of the fact that every cell transcript is given a unique cell barcode, during library preparation. By designing unique primers that anneal to the barcode of a specific cell's libraries, we can selectively amplify the transcriptome of a single cell. We can then ligate the original Illumina sequencing adapters back onto the amplified product, and subsequently sequence the product to gain deeper insight on the amplified transcriptome of interest.

3.2 Materials and Methods

3.2.1 Primer Design

We then designed and ordered a 20-nt primer with 5'-CAGAG-/U/[6 base sticky end]-AC-[cell barcode]. The CAGAG anneals to the end of the SeqB sequence of a cDNA, and is followed by a 'U.' This U is ultimately be cleaved by the USER enzyme (which specifically cleaves U bases) to generate sticky ends during sticky end ligation. The cell barcode of the primer anneals to the corresponding cell barcode on the cDNA sample during cell specific PCR. We employ the 6-base ligation sequence and 'AC' in the ligation protocol, detailed in Section 3.2.4.

3.2.2 NexteraXT Tagmentation

Before transposing the P7 and P5 adapters to the cDNA samples, we combined 10uL of each sample at 180pg/uL with 10uL 4X Amplicon Tagment Mix Enzyme and 20uL 2X Tagment DNA buffer. The solution was then incubated at 55ºC for five minutes. We then immediately added 10uL of Neutralization Buffer, mixed by pipetting, and centrifuged the solution. We then incubated the solution at room temperature for 5 minutes and subsequently moved the solution to ice.

3.2.3 Cell Specific Indexing PCR

After the Nextera reaction, instead of proceeding to the standard PCR to attach Illumina's P5/P7 ends for sequencing, we performed the cell specific PCR reaction in order to only amplify the cDNAs with the cell barcode that matches our cell of interest.

For each reaction, we combined the 50uL of the cDNA that had undergone tagmentation with 14uL water, 30uL Nextera PCR Mix, 2uL of the forward primer containing the cell barcode (designed in Section 3.2.1) at 10uM, and 4uL of the standard Nextera indexing reverse primer at 5uM. For the sake of testing, we sometimes ran this PCR as a qPCR, which involved adding 1.67uL SYBR green (with 12.33uL water instead of 14uL).

To isolate the sample at the point at which only the transcriptome of interest had amplified, we then aliquoted each 100uL reaction, and ran a 10uL subset as a qPCR. If the PCR had gone to completion (i.e. ~35 cycles), all the material would have amplified, including primer dimers, which would have flooded our sample with unnecessary contamination and undesired amplified sequences. We recorded the cycle number at which the sample began to exponentially rise in quantity (Ct threshold), and ran the remaining 90uL in a PCR that amplified up to this recorded cycle number. The PCR conditions were as follows: 95°C, 30 sec, X cycles of: 95°C, 10 sec; 55°C, 30 sec; 72°C, 30 sec (where X is the recorded cycles needed to achieve the Ct threshold obtained from the qPCR), 72°C, 5 min, 4°C, hold

We then performed a 0.8X SPRI DNA cleanup using (Ampure XP beads). We eluted the wash in 45uL of water, and saved 44uL.

3.2.4 P5 Adapter Ligation

At this point, the product of the PCR contains the Illumina P7 end but lacks the P5 end required for sequencing. We explored two ways to ligate the P5 end back onto the product: blunt end and sticky end ligation. To attach the P5 adapter by blunt end ligation, we diluted the adapter to 1.6uM. We then prepared and added 32uL of ligation master mix to 10uL of each cDNA sample. This master mix contained 22uL water, 8uL 5X Quick Ligation buffer, 5uL Sticky End Master Mix, 3uL propanediol, 2uL 1.6uM adapter. We then incubated the reaction at 20ºC for 15 minutes, and quenched with 5uL 0.5M EDTA. The reaction then underwent a 0.8X SPRI cleanup with a 20uL elution.

We tested sticky end ligation by first generating a sticky end in the cell PCR product. Specifically, we generated this sticky end by incubating our samples with an enzyme (USER) that cleaves at a specific base (U) that was incorporated into the DNA through the primers of the cell specific PCR (as described in Section 3.2.1). Specifically, we incubated the 44uL PCR product with 1uL USER II Enzyme (to cleave at U), and 4uL CutSmart Buffer for 15 minutes at 37ºC, followed by a heat inactivation at 65ºC for 10 minutes and a 0.8X SPRI cleanup (10uL elution). To subsequently attach the adapter, the sticky end ligation then proceeds identically to the blunt end ligation, adding 32uL of the ligation master mix, followed by a 15 minute incubation at 20ºC, quenching with 5uL 0.5M EDTA, and a 0.8X SPRI cleanup with a 20uL elution.

3.2.5 P7-P5 PCR

We amplified fully-ligated products by first running a qPCR on a subset of the mixture, recording the amplification threshold (Ct), and then running a PCR on the remaining samples with a cycle number equivalent to the recorded Ct value (as completed in the initial cell specific PCR). To perform this PCR, we combined 18uL of the ligation product with 42uL of PCR master mix containing: 6.2uL water, 30uL 2X Ultra Q5 MasterMix, 1uL SYBR Green I, and 2.4uL 10uM Illumina P7 and P5 primers. 10uL of this mix underwent a qPCR, and once we determined the Ct, the remaining 50uL underwent a PCR with a cycle number equivalent to the Ct value. We purified the PCR product with a 0.8X SPRI and 15uL elution. We then quantified this product using the TapeStation High Sensitivity D1000 assay.

3.2.6 Prepare for Sequencing

Using the quantification values obtained after the PCR, we pooled the cell libraries at an equal molarity and performed a 0.8X SPRI cleanup followed by a 14uL elution. We then quantified the pool using the Qubit High Sensitivity DNA assay as well as the TapeStation High Sensitivity D1000 assay, and loaded it onto the Illumina MiSeq sequencing machine.

3.3 Results

3.3.1 Using Control Templates to Test and Optimize the Cell Specific PCR

To test the cell specific PCR, we controlled the template DNA by utilizing the same cDNA control reads used in the CRISPR-Cas9 depletion experiments (detailed in Chapter 2).

Specifically, we selected Templates 4 and 12 for preliminary testing. The first step in completing the cell specific PCR is to design custom primers; these primers must only amplify the transcriptome of the specific cell of interest. Thus, one must design the primers to match the cell barcode of the cell of interest. We designed primers corresponding to the cell barcodes of Template 4 and Template 12, and ran the PCR to test for selective amplification as well as detect non-specific amplification (i.e. a primer amplifying an incorrect template). Figure 14 shows the qPCR data for this initial testing. When Template 4 underwent the reaction with Template 4's primers, the qPCR exhibited the highest amplification. Furthermore, when we used Template 4's primers against a mixture of cDNAs containing Template 4 and Template 12, the qPCR also displayed selective amplification. To test non-specific amplification, we reacted the cDNA of Template 12 with the primers of Template 4, expecting to see no amplification (since the primer does not match the template cell barcode). However, although this sample amplified less than those with matching templates and primers, the qPCR shows that there was non-negligible non-specific amplification (Figure 14). We kept this in mind as we designed the experimental protocol, in which we stop the amplification at a given Ct value before non-specific amplification can take effect (Section 3.2.3).

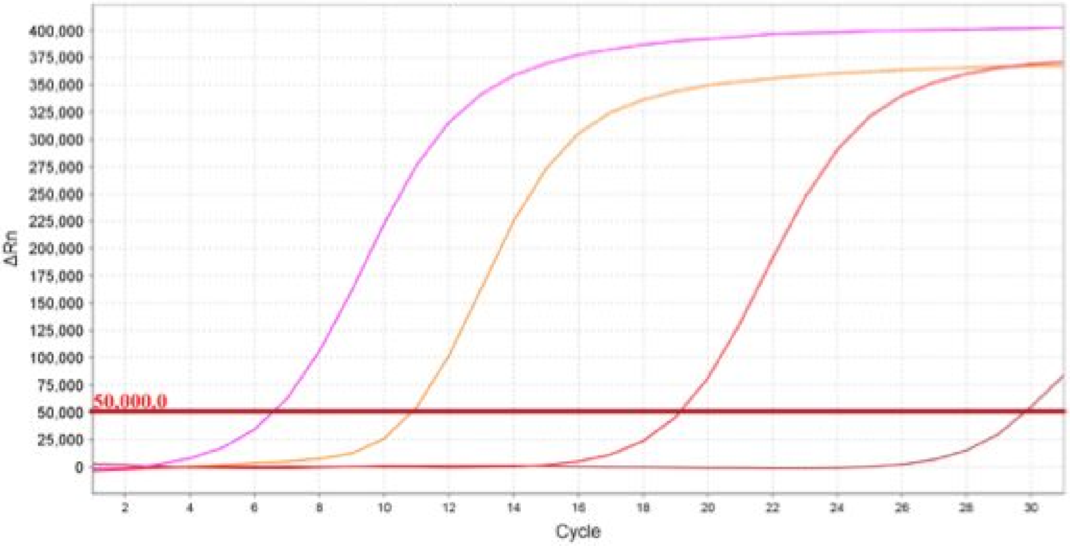


Figure 14: Cell Specific qPCR on Test Libraries
Here, we selectively amplify the transcript libraries of Template 4. As you can see, when using the primers that match the barcode of Template 4, the amplification is greatest (pink). When amplifying a mix of a small concentration of Template 4 with another Template 12, the qPCR still exhibits amplification of Template 4 (orange). However, when running the qPCR using mismatched primers with a different cell barcode (Template 4 primers against Template 12 cDNA), we see significantly less amplification (light red, over 5 cycles difference). The dark red line represents the water control.

After successfully testing the cell specific PCR on our control libraries, we proceeded to the ligation step, in which we ligate the P7 and P5 ends necessary for sequencing. After sticky end ligation, the sample, now containing the cell specific PCR product with a sticky end, can then ligate to a custom-ordered adapter sequence that contains the corresponding sticky end followed by the sequence matching the Illumina P5 primer. Ideally, once the adapter ligates, the resulting DNA will contain both sequencing ends, allowing for Illumina primers to anneal during Nextera PCR, leading to effective Illumina sequencing. Blunt end ligation is similar, but does not involve the generation of a sticky end and as a result does not require a cleavage enzyme. The custom blunt end adapter would contain the Illumina P5 sequence required for Illumina sequencing, and would ligate directly onto the end of cDNA. As a result, the P5 and P7 primers would anneal in PCR leading to effective sequencing.

We tested both sticky and blunt end ligation, and ultimately decided to use the blunt end ligation in our further assays. This was due to inconsistency exhibited in the sticky end ligation assay. Specifically, when testing a negative control consisting of template DNA that lacked the USER II cutting enzyme, we expected no amplification in the subsequent P7 P5 qPCR, since there would be no enzyme to generate the sticky end, thus preventing the P5 adapter from ligating (without a ligated P5 end, the cDNA should not amplify during a PCR using P7/P5 primers). However, as seen in Figure 15, Template 4 without the USER II enzyme also amplified. Although there is a 5-cycle difference between this negative control and the positive controls, the strand without a sticky end should not have been able to ligate to P5 and amplify at all, which indicates that either some strains were able to ligate without a sticky end, or there was cross-contamination from full Template 4 libraries containing the sequencing ends. Due to this inconsistency and the potentially unreliable sticky end generation, we opted for the blunt end ligation for the assays amplifying experimental samples from nonhuman primates.

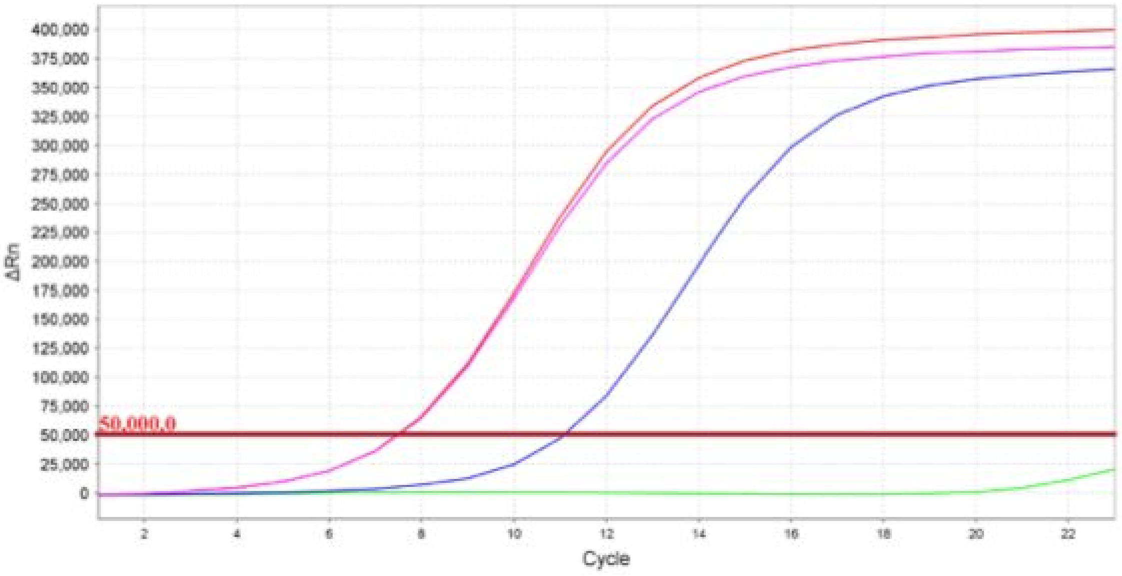


Figure 15: Ligation testing with and without sticky ends
Here, we test the ligation of the adapter containing the Illumina P5 primer through qPCR. The red and pink lines represent incubation with the USER II enzyme (that generates the sticky end required for ligation). Specifically, the red curve contains pure DNA from Template 4, and the pink line contains a mixture of 10% Template 4 DNA. The blue line, which represents Template 4 without the sticky end enzyme, also amplified 5-cycles ahead of the USER-treated samples. The green line represents the water control.

3.3.2 Testing the Cell Specific PCR on Experimental Libraries

After optimizing the cell specific PCR assay on test libraries, we tested the assay on an experimental library from nonhuman primates infected with Ebola virus. This library had previously undergone scRNA-seq, providing us with a source of comparison (comparing the sequencing results with versus without cell specific PCR). We selected five cell candidates within this library whose transcriptomes we sought to amplify; we isolated the barcode of these five cells, designed custom primers to amplify transcripts corresponding to these barcodes, and named the samples CELL01, CELL02, CELL03, CELL04, and CELL05. Specifically, two cells, CELL01 and CELL03, each exhibited abnormally high levels of EBOV genes in their transcriptomes (49.8% and 43.9% respectively) after an initial total scRNA-seq run, a finding we sought to explore further through the cell specific PCR. The remaining three cells were uninfected. We picked these five candidates to contain between 1,000 and 10,000 total transcript counts after initial scRNA-seq.

Although the assay amplified successfully in control conditions (Figure 14), upon testing the cell specific PCR on experimental samples, sequencing yielded lower counts of total transcripts and genes as compared to the same library that had not undergone cell specific PCR. This result could be due to the fact that our cell specific PCR started with less input DNA than the non-cell specific PCR samples; we split an existing library five ways (for each of the five cell specific PCR reactions), a division we did not make when initially sequencing the library without cell PCR. Another reason that could account for the decrease in gene and transcript counts is the sheer amount of cells within an experimental sample; the cell of interest could be in such a small proportion relative to the other cells in the sample that any amplification would eventually plateau. In addition, the expression profiles of a single cell were not significantly different with versus without cell specific PCR. This is evidenced by the strong correlation charts and Pearson coefficients shown in Table 3, which suggest that a cell specific PCR preserves a cell’s general expression pattern and does not necessarily bring about a different result.

		Number Genes	**Total Counts**	**Correlation Coefficient**
CELL01	None	496	1370	0.996
	Cell PCR	559	1204	
CELL02	None	2301	9710	0.971
	Cell PCR	1711	4815	
CELL03	None	623	1710	0.996
	Cell PCR	329	800	
CELL04	None	1914	5290	0.852
	Cell PCR	1292	2215	
CELL05	None	617	1020	0.775
	Cell PCR	522	691	

Table 3: Total transcript counts and number of detected genes after sequencing with and without cell specific PCR.

3.4 Discussion

The results above mark a promising step in scRNA-seq analysis, in that we demonstrate an ability to amplify and sequence the transcriptome of a specific cell of interest within a pool of cell transcriptomes.

Still partially unanswered is why gene counts are lower in samples that have undergone cell specific PCR. This could be due to the depth of our NovaSeq sequencing machine. By design, NovaSeq provides roughly 300 million reads per sample even without cell specific PCR, many of which are not necessarily useful for analysis if a user is only interested in a few specific transcriptomes. The fact that we pursue this deep sequencing method for each sample without cell specific PCR means that NovaSeq likely already provides a deep sequence of all the mRNAs in a sample, even without selective amplification, which could potentially explain our results. By consequence, our data suggests that a cell specific PCR may be best employed on a lower depth sequencer, or in an environment that may not have the resources to obtain or repeatedly use a high-depth, high-cost NovaSeq machine. As a next step to test this hypothesis, we plan to run a cell specific PCR, sequence the result on a lower-depth MiSeq machine, and determine whether the MiSeq can provide the whole single-cell transcriptome as effectively as the NovaSeq. If valid, this hypothesis makes an argument for the resourcefulness of a cell specific PCR: If scientists would like the transcriptomes of a few rare cells, our results suggest that these researchers would not need to dispense excess resources to obtain and run a full NovaSeq, as the NovaSeq would generate extraneous information on cells that are not of interest; the lower depth MiSeq may provide the same depth when sequencing a few specific cells of interest after cell specific PCR. In all, these findings can potentially usher in a more resourceful analysis of specific cells of interest, providing more information on expression levels within a specific isolated cell, information that could be valuable when analyzing cell specific regulatory changes during infection.

4. Current Efforts and Future Work: Application to LCMV

4.1 Introduction

Since scientists can only directly study Ebola virus in a BSL4 facility, I have begun to expand my study to explore Lymphocytic Choriomeningitis Virus (LCMV), a virus that resembles Lassa Virus, a fatal viral hemorrhagic fever that affects approximately 100,000 people each year (Günther and Lenz 2004). Containing only four genes, LCMV provides a new and unique opportunity to perform single-cell RNA sequencing in the lab, through BSL2 facilities. This past semester I began a study to express all permutations of the four LCMV genes in cell-culture with the hope of uncovering how each gene and its interactions shapes host gene expression, through western blots, immunofluorescence, co-immunoprecipitation, and ultimately scRNA-seq.

4.2 LCMV Background

LCMV is an arenavirus that bears resemblance to Lassa virus, and consists of two ssRNA strands (de la Torre 2009). The virus contains a total of four genes: L, which encodes the RNA dependent RNA polymerase; NP, which encodes the viral nucleoprotein; Z, which encodes the matrix protein that incorporates glycoproteins into viral particles, and GPC which encodes the glycoproteins (Cornu and de la Torre 2001). LCMV studies allow researchers to break a virus down into its individual components, which is important for in-depth viral studies. Since studies of Ebola virus infection are only accessible within BSL4 facilities, scientists cannot effectively study the impacts of Ebola virus infection in cell-culture within a standard laboratory. Thus, by expanding to LCMV, we hope to pursue a more in-depth, cell-culture based methodology to correlate viral gene expression with the host gene expression.

More fundamentally, the key question of this LCMV study is: how do viral proteins interact to modulate the regulation of host genes involved with immune response? Our strategy is to break the virus down into its individual components (or proteins), rebuild it by combining these various components in multiple permutations, and exploring which combinations produce interactions known to confer viral infection; this approach is relatively convenient since LCMV only contains four genes. Moreover, by mutating genes that encode proteins known to interact, one can disturb protein-protein interactions and subsequently elucidate the functions of these interactions.

4.3 Materials and Methods:

4.3.1 Obtaining Plasmids

To selectively express specific genes and mutations of interest, we generated plasmids, each containing a short peptide tag fused to one of LCMV's four genes. Mutations predicted to disturb key protein-protein interactions were generated using Site Directed Mutagenesis. All plasmids were then cloned into bacteria and grown for 16-18 hours at 37ºC. We then performed a Qiagen "Miniprep" to extract and purify the plasmids from the bacterial cells. We subsequently verified the desired sequence (including the plasmid template, antibody tag sequence, and gene sequence) by Sanger sequencing.

4.3.2 Transfection

After creating and validating plasmids, we transfected them into HEK 293FT cells in cell-culture to express the desired LCMV genes. Specifically, for each well in a 12-well plate (scaled up for larger plate sizes), we added 50uL OptiMem, to 3uL 220ng/uL plasmid DNA and 1uL water. We then added a mixture containing 50uL of OptiMem and 4uL L2K Lipofectamine. We vortexed the mixture and incubated it at room temperature for 30 minutes. We subsequently added 100uL of this mixture to a well within a 12-well plate (If performing immunofluorescence, we placed a glass coverslip in each well before adding the plasmid for microscopy purposes). Meanwhile, for each reaction, we mixed 800uL of media with 160uL of cells (either HEK 293FT if performing a western blot, or Vero cells if performing immunofluorescence), and the 960uL was added to each well. We then placed the plate in a 37ºC rocker for approximately 2-3 minutes, and then incubated the transfected cells at 37ºC for 2-3 days.

4.3.3 Immunofluorescence (Under Development)

To conduct immunofluorescence (IF) assays, we first prepared 2% Paraformaldehyde (PFA) by combining 7mL PBS with 1mL Paraformaldehyde. We then obtained our plate containing transfected cells (detailed in Section 4.3.2), aspirated the old media, and washed with 1mL PBS per well. We aspirated the PBS and added 1mL 2% PFA to each well to fix the cells and rocked the plate for 15 minutes. We subsequently prepared glycine quench buffer by combining 20mL PBS with 150.14g Glycine (which disturbs the crosslinking incurred by PFA by providing a Nitrogen nucleophile), and Triton Perm Buffer (9.9mL PBS and 100uL Triton X-100). We washed the cells three times with 1mL Glycine quench buffer per well, then added

1mL Triton Perm buffer per well, and rocked the plate for 15 minutes at room temperature to permeabilize the cells. We then washed the cells three times with 1mL Tween Wash Buffer (containing 79.84mL PBS and 160uL Tween 20), rocking the plate for 5 minutes between washes. We subsequently blocked cells with 1mL BSA blocking buffer per well (containing 27.93mL PBS, 7mL BSA, and 70uL Tween 20), and left the cells to rock at room temperature. After one hour, we added 1mL of our primary antibody (diluted in the BSA buffer), rocked the plate for 30 minutes (see Section 4.4.4 for explanation of antibody choice), and performed three Tween washes. We then added 1mL of diluted secondary antibody to the wells, which were then rocked in the dark (aluminum foil cover) for one hour, and washed three times with Tween. We diluted Hoechst dye to 100ug/mL in BSA blocker; this dye binds to the nucleus of the cell and fluoresces blue, allowing us to locate each cell's nucleus under the microscope. We added 1mL Hoechst to each well, and rocked the plate for another 15 minutes, followed by 3 subsequent Tween washes. We then loaded the glass slips within each well onto glass microscope slides and visualized the slides on our lab's microscope. Table 4 below details the tags and antibodies we used in this protocol.

Antibody	Manufacturer
Primary Antibodies:	
Rabbit anti-vinculin	clone E1E9V, Cell Signaling Technologies
Mouse anti-V5	clone SV5-Pk1, Bio-Rad
Mouse anti-FLAG	clone M2, Sigma-Aldrich
Mouse anti-HA	clone HA-7, Sigma-Aldrich
Mouse anti-Myc	clone 9B11, Cell Signaling Technologies
Mouse anti-Actinin	clone 69758S, Cell Signaling Technologies
Rabbit anti-HA	clone C29F4, Cell Signaling Technologies
Rabbit anti-V5	clone D3H8Q, Cell Signaling Technologies
Secondary Antibodies:	
Goat anti-mouse, HRP-conjugated	Jackson ImmunoResearch #115-035-174
Goat anti-rabbit, HRP-conjugated	Jackson ImmunoResearch #111-035-144
Goat anti-mouse, Alexa Fluor 488 Plus-conjugated	Thermo Fisher Scientific #A32723
Goat anti-rabbit, Alexa Fluor 555 Plus-conjugated	Thermo Fisher Scientific #A32732

Table 5: Antibodies and tags used in immunofluorescence and western blot protocols

4.4 Initial Results:

4.4.1 Selecting Mutants

Table 5 below reports the mutants we tested along with their functions (Casabona et al. 2009; Ortiz-Riaño et al. 2012; Hastie et al. 2016).

Gene	Mutation	Function
NP	D471A	Disrupts NP-NP interactions and NP-Z interactions
NP	D471G	Disrupts NP-NP interactions, and transcription/replication
Z	G2A	Highly conserved region, mutant disturbs viral budding and infectivity
Z	L57R	Does not allow the Z protein to form an oligomer (only monomeric Z)
Z	K69A	Does not allow the Z protein to form an oligomer (only monomeric Z)
Z	L72A	Disturbs NP-Z interaction, decreases infectivity
Z	P73A	Does not allow the Z protein to form an monomer (only oligomeric Z)

Table 5: Z and NP mutants tested along with their functions

4.4.2 Western blot/Co-Immunoprecipitation Assays (Under Development)

After transfecting and growing cells, we have begun to test whether or not two LCMV proteins interact. We co-transfected two plasmids into the same population of cells: One plasmid contained a protein A fused to tag 1 and another contained a protein B fused to tag 2 (These proteins could be NP and a Z mutant, for example, fused to the V5 and HA tags respectively). While we are currently in the process of completing these tests, one would then harvest the cells, conduct a co-immunoprecipitation assay followed by a western blot that would involve an antibody binding *one* of these tags, say tag A, for example. If the proteins were not interacting or binding, protein A-tagA would bind antibody A, and one protein band would be seen on the western blot. However, if proteins A and B were interacting, protein A-tagA would be attached to protein B. Consequently, antibody A would pull out both proteins A and B, yielding a western blot with two bands, representing proteins A and B. While we have not implemented this entire assay yet, we have begun to transfect and express individual proteins and mutants into cells, and observe resultant western blots. Figure 16 below displays one such western blot.

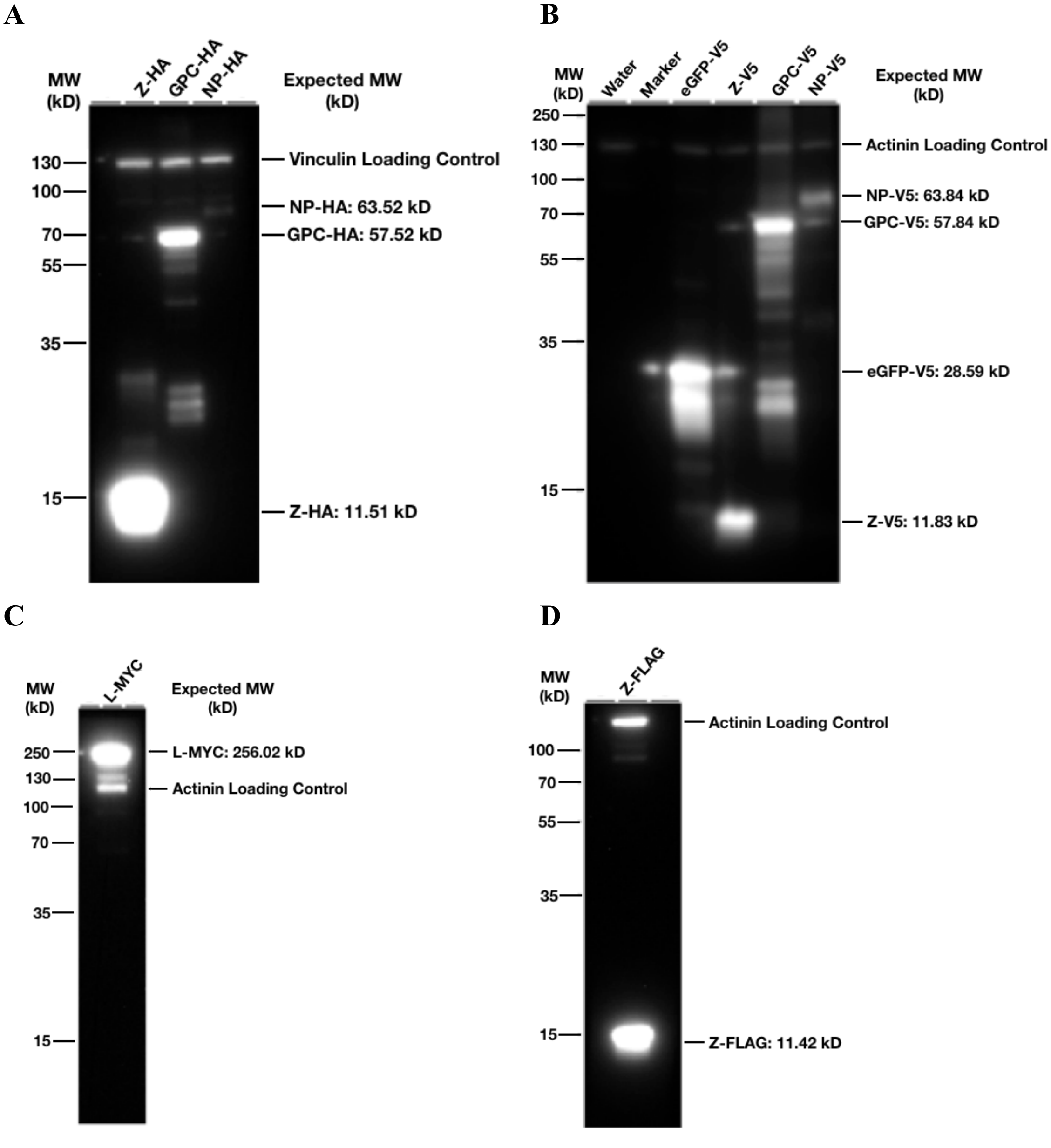


Figure 16: Initial western blot results of cells containing each LCMV gene (NP, Z, L, GPC)
Loading Control: Vinculin/Actinin (common cell proteins); Positive Control: eGFP; Negative Control: water. We used the following antibodies: A) Mouse anti-HA; Goat anti-mouse B) Mouse anti-V5; Goat anti-mouse, C) Mouse anti-MYC; Goat anti-mouse D) Mouse anti-FLAG; Goat anti-mouse. Loading controls: Actinin - Mouse anti-Actinin; Goat anti-mouse. Vinculin: Rabbit anti-Vinculin; Goat anti-mouse.

4.4.3 Immunofluorescence

After transfecting and growing cells, we began to pursue immunofluorescence assays. If two proteins interact, we would expect the immunofluorescence to exhibit two proteins visually close to one another, colocalized within the cell. We began preliminary testing by individually expressing the LCMV genes and performing immunofluorescence. Our initial results (Figures 17-18) provide images of the expressed LCMV proteins, but exhibit a lack of resolution that could be due to limitations of the focal plane. Furthermore, the wide-ranging fluorescence in Figures 17-18 implies that transfected cells may be overexpressing the proteins, leading these proteins to accumulate throughout the cytoplasm. Section 4.5.1 interprets this finding further and offers alternative experimental approaches to address these challenges.

4.4.4 Antibody Choice for Immunofluorescence

In tests co-expressing two proteins with different tags, we used two primary antibodies from different animals, with each primary antibody targeting a different tag (one tag corresponding to each protein). The purpose of using two species is to produce different colors in the immunofluorescence, enabling us to distinguish the proteins from one another under the microscope. The secondary antibody is specific for each primary antibody, and targets each species. For example, in an assay expressing Z-FLAG and NP-V5, our primary antibodies included a Rabbit anti-V5 and a Mouse anti-FLAG. The secondary antibody solution contained two antibodies, Goat anti-Rabbit and Goat anti-Mouse. These secondary antibodies are each conjugated to distinct fluorescent molecules, allowing us to distinguish between the two proteins by color under the microscope. Initial immunofluorescence results for various transfected LCMV genes are shown in Figures 17 and 18 below. Specifically, Figure 18C depicts an initial attempt at co-expressing two LCMV proteins (Nucleoprotein NP and Z). The overlapping green and orange indicates a site where both Z (green) and NP (orange) were present.

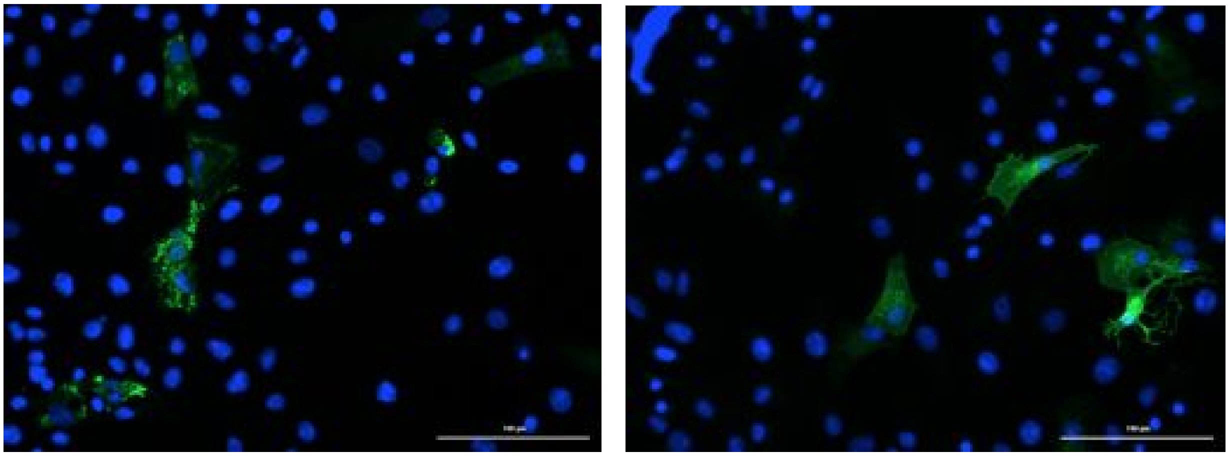


Figure 17: Initial immunofluorescence of cells containing two LCMV genes
This figure shows a preliminary immunofluorescence assay, detecting Vero cells transfected with plasmids containing LCMV's viral Nucleoprotein gene (left), and Z gene (right). The blue represents the nuclei of the cells (stained with Hoechst dye that binds dsDNA within the nucleus).

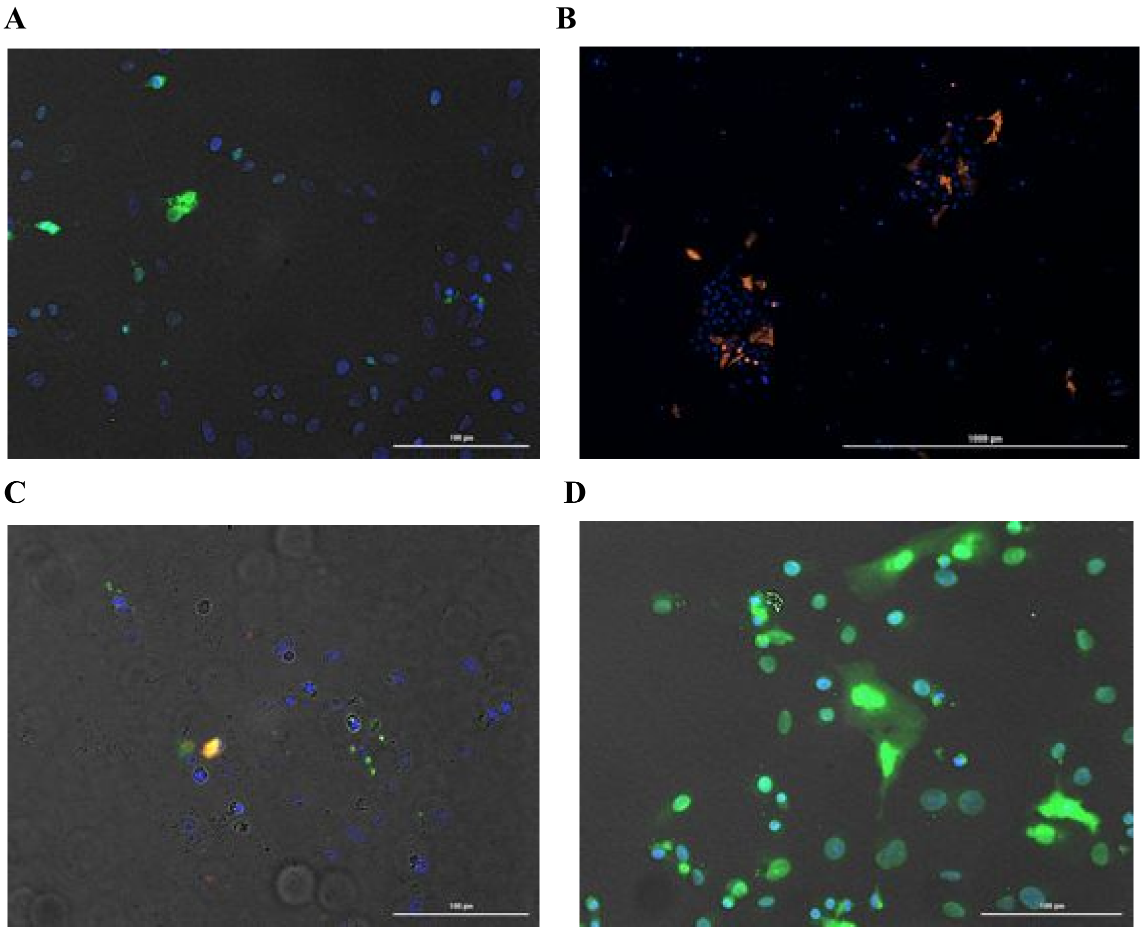


Figure 18: Immunofluorescence of Z (A), NP (B), NP and Z co-expressed (C), and eGFP (D) as a positive control.

4.5 Discussion

The approach and initial results described in this chapter lay the groundwork for elucidating how each LCMV gene and its interactions shape host gene expression. In the near future, we plan to further explore the results of our immunofluorescence and co-immunoprecipitation assays as detailed in the following sections.

4.5.1 Immunofluorescence and Interaction Studies

As stated in Section 4.4.4, Figure 18 suggests that the resolution is not clear enough to definitively suggest colocalization. This lack of resolution could be due to the limitations of the focal plane on microscopic resolution; a wide focal plane can show two proteins in close two-dimensional proximity, when, in reality, they might be far apart on the Z (depth) axis.

Now that we have ensured reasonable initial results, we hope to further analyze our immunofluorescence assays using confocal microscopy, which may narrow the focal plane, allowing us to more accurately determine whether two proteins in close microscopic proximity are truly colocalized within a cell.

Aside from confocal microscopy, another approach that can provide high resolution imaging is Direct Stochastic Optical Reconstruction Microscopy (dSTORM). When coupled with data analysis software, dSTORM can quantify the degree of interaction between a given pair of proteins within a cell. Specifically, dSTORM visualizes cellular structures by using light of different wavelengths to irradiate photoswitchable fluorescent dyes (Heilemann et al. 2008). For example, Schmider et al. integrate dSTORM with clustering algorithms to determine the localization of single molecules within a cell, and explore the organization of leukotriene biosynthesis (Schmider et al. 2019). We hope to incorporate this method into our future interaction studies to obtain a more accurate indication of protein proximity and ultimately elucidate colocalization and interaction.

Furthermore, the wide-ranging fluorescence in Figures 17-18 implies that cells may be over-expressing the transfected proteins, causing these proteins to accumulate throughout the cytoplasm, and preventing immunofluorescence from providing useful information on localization and interaction. We hope to address this over-expression in our future studies by using live LCMV to explore natural expression levels. Studying live LCMV would allow us to validate colocalization and ensure that any observed interactions during initial testing were not a result of our transfection causing over-expression within the cell.

4.5.2 Co-Immunoprecipitation Studies

We plan to continue our co-immunoprecipitation studies by testing additional permutations of genes and tags. For example, one planned assay will test self-association for individual LCMV proteins to elucidate whether a given protein can interact with itself to produce a significant effect.

Another planned co-immunoprecipitation test involves reversing the tags on two different genes. For example, in a test exploring interactions between protein A and protein B, a co-immunoprecipitation assay would involve both: one test with A-Tag1 and B-Tag2, and another test swapping the tags (A-Tag2 and B-Tag1). We expect the same results from both assays, since the interaction (or lack thereof) between A and B should be the same regardless of the tag. However, if the two tests yield different interactions, the results would imply that the assay produced a false negative or a false positive. A false positive could be the result of an experimental artifact giving rise to an observed interaction, rather than genuine protein activity. Furthermore, a false negative could occur if the addition of a specific tag interferes with the assay and abrogates an existing protein-protein interaction.

We hope to further validate our results by testing a co-immunoprecipitation assay in which both tags are fused to the N-terminus instead of the C-terminus of the proteins. We would expect both N-terminus and C-terminus tags to reveal the same result. If we obtain different results based on where the tag is fused, the data would suggest that the tag's presence is potentially interfering with protein folding and affecting protein interactions.

Moreover, in the future we plan to explore the effects of bicistronic and multicistronic expression vectors on our assay. On one hand, the quality of transcription might be more efficient in systems that simultaneously transfect multiple plasmids (with each plasmid expressing a given protein). On the other hand, by placing multiple genes within a single vector, the probability of successful transfection increases relative to our current assays, since these current assays rely on multiple simultaneous successful transfections into a given set of cells.

On a larger scale, we hope to apply our work to other BSL2 viruses, and potentially individual proteins from BSL4 viruses, a novel approach to study the genes and proteins of high risk-group pathogens within the confines of a BSL2 level facility. In all, this study begins the effort to use cell-culture techniques to uncover how proteins and their interactions can causally affect virus-like particle (VLP) production and host gene expression.

5. Conclusion

The key to addressing global epidemics is to harness emerging scientific technologies. By integrating novel sequencing and gene editing methods, we take strides in efficiently sequencing the transcriptomes of single cells taken from nonhuman primates infected with Ebola virus.

First, we successfully demonstrated the newfound ability to preserve valuable sequencing libraries that may have been contaminated during the rigorous scRNA-seq preparation method. We decreased contamination by over 10-fold on average, with some reductions as high as 37-fold. In analyzing the few cells that did not deplete as effectively, we offered potential explanations that have broad implications in the "empty droplet" problem faced by many scRNA-seq users. In preserving over 15 libraries that had failed previously, we employed this depletion to sequence over 1000 additional transcriptomes. Our work now lies in analyzing this additional data, searching for transcriptional changes in specific cell-types over the course of infection.

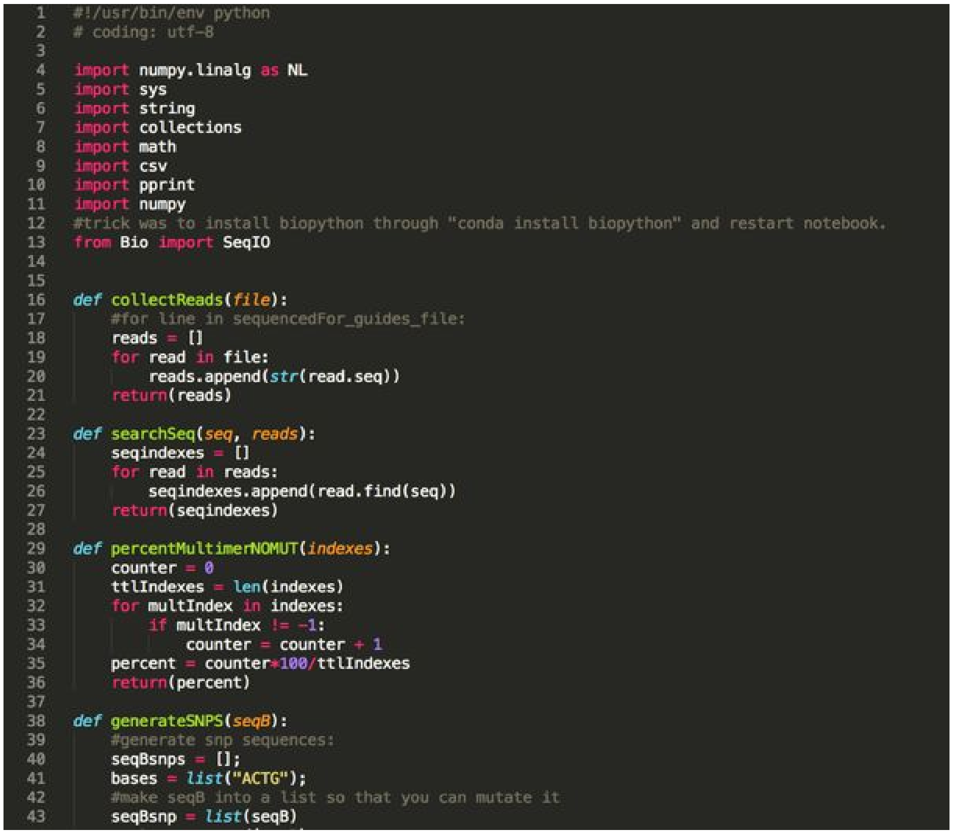
We further enhanced scRNA-seq methodologies by designing and implementing a "cell specific PCR," a development that can serve as a more resourceful means of gaining a deep insight on gene regulation within specific cells of interest. This insight can lead to a more informed conclusion on the effects of Ebola virus on specific host cells. In the future, we hope to test this cell specific PCR on a lower depth sequencer such as a MiSeq. This test would allow us to determine whether a lower depth machine could provide similar comprehensive profiles of transcriptomes from specific single cells, a potentially resourceful alternative to the current NovaSeq approach.

Moreover, we will continue our efforts to study LCMV in cell-culture, implementing additional co-immunoprecipitation and immunofluorescence assays. This work will allow us to further elucidate the role of each viral protein in VLP formation and infectivity.

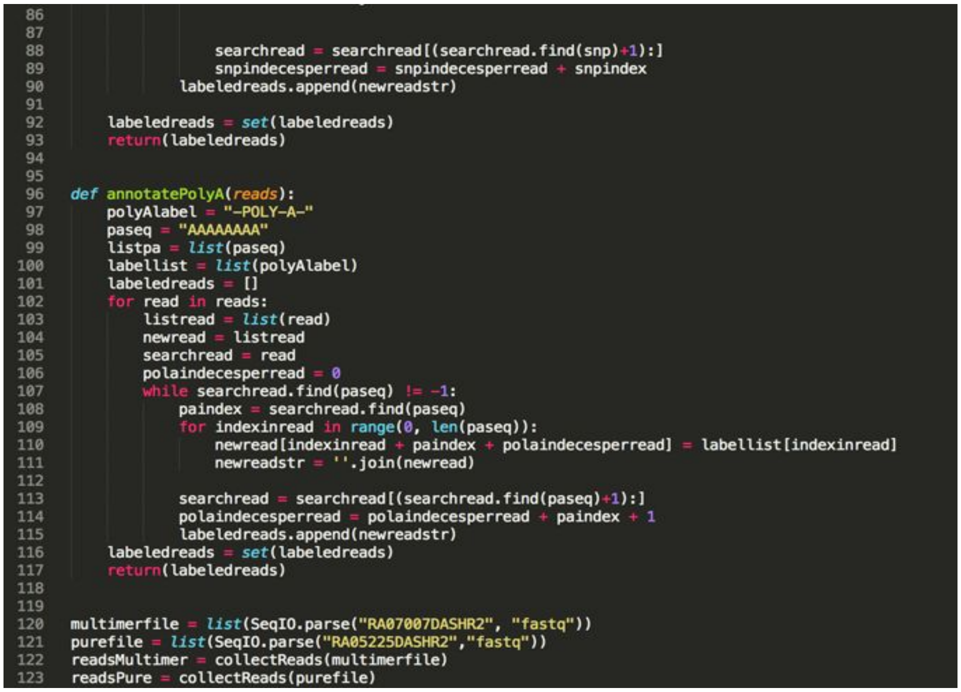
While there still remains unanswered questions in the field of transcriptomics, this research sets a foundation for future studies in high-risk BSL4 settings. This work can lead to a better understanding of how lethal pathogens such as the Ebola virus affect the immune response over time, which brings us one step closer to improving the prevention and response for tragic outbreaks around the world.

6. Coding Appendix

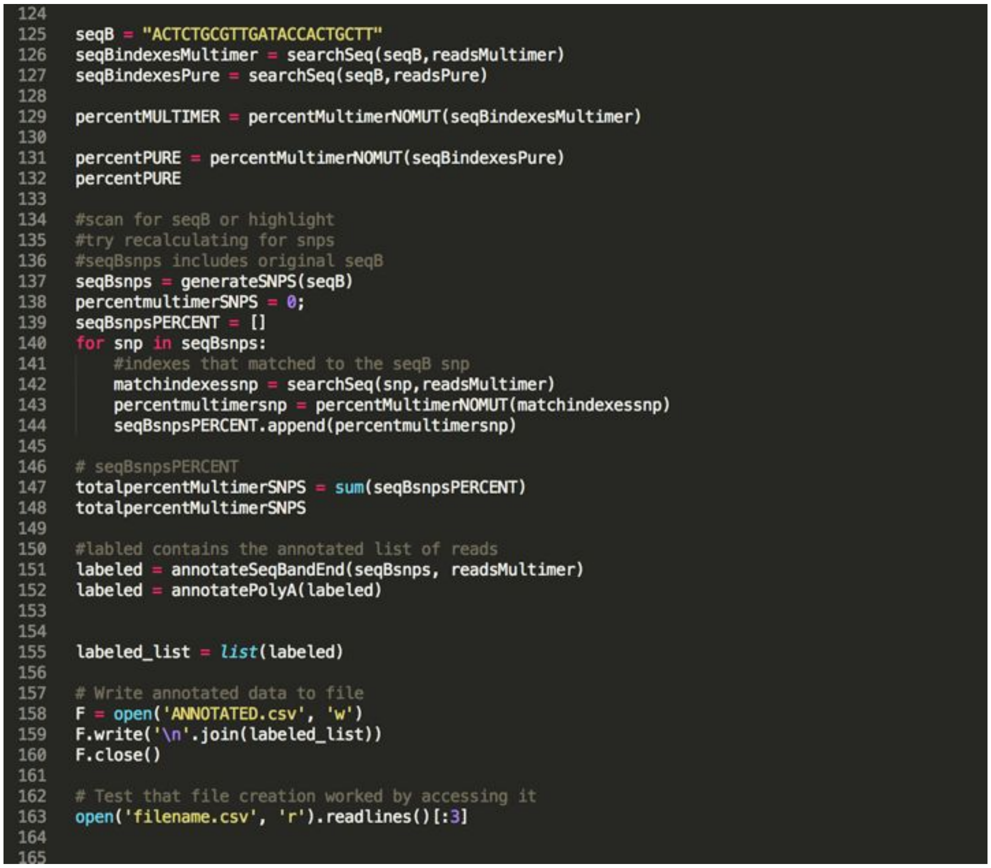
DASH Analysis on Experimental Samples



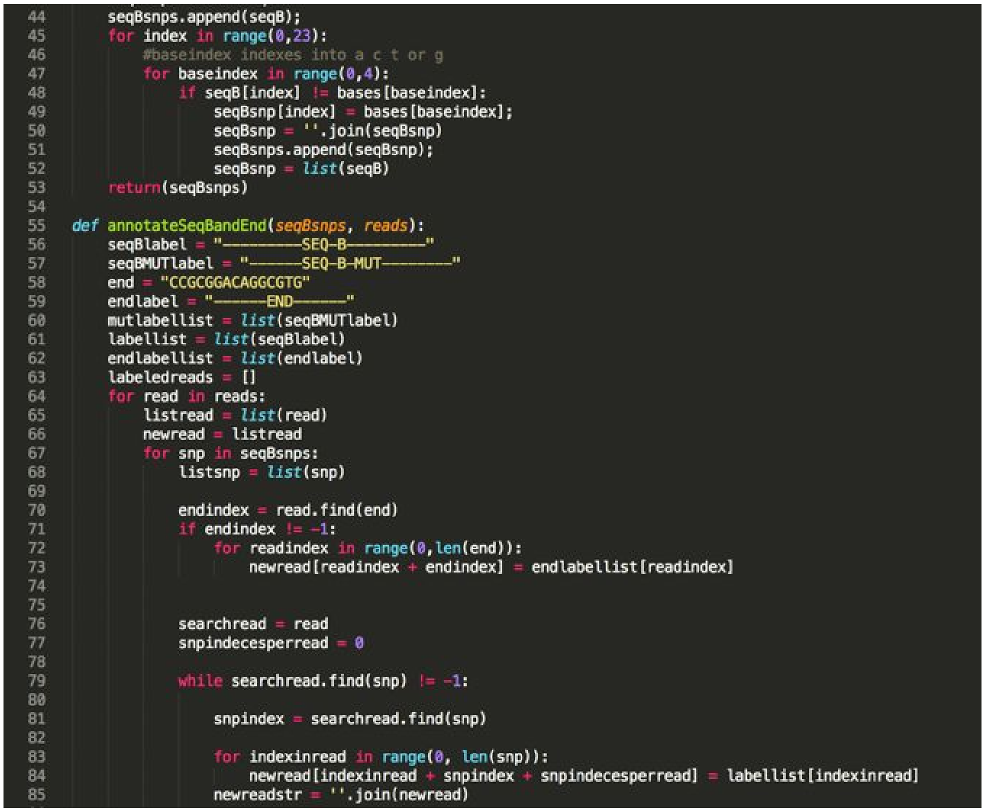
```python
#!/usr/bin/env python
# coding: utf-8

import numpy.linalg as NL
import sys
import string
import collections
import math
import csv
import pprint
import numpy
#trick was to install biopython through "conda install biopython" and restart notebook.
from Bio import SeqIO


def collectReads(file):
    #for line in sequencedFor_guides_file:
    reads = []
    for read in file:
        reads.append(str(read.seq))
    return(reads)

def searchSeq(seq, reads):
    seqindexes = []
    for read in reads:
        seqindexes.append(read.find(seq))
    return(seqindexes)

def percentMultimerNOMUT(indexes):
    counter = 0
    ttlIndexes = len(indexes)
    for multIndex in indexes:
        if multIndex != -1:
            counter = counter + 1
    percent = counter*100/ttlIndexes
    return(percent)

def generateSNPS(seqB):
    #generate snp sequences:
    seqBsnps = [];
    bases = list("ACTG");
    #make seqB into a list so that you can mutate it
    seqBsnp = list(seqB)
```



```python


                searchread = searchread[(searchread.find(snp)+1):]
                snpindecesperread = snpindecesperread + snpindex
            labeledreads.append(newreadstr)

    labeledreads = set(labeledreads)
    return(labeledreads)


def annotatePolyA(reads):
    polyAlabel = "-POLY-A-"
    paseq = "AAAAAAAA"
    listpa = list(paseq)
    labellist = list(polyAlabel)
    labeledreads = []
    for read in reads:
        listread = list(read)
        newread = listread
        searchread = read
        polaindecesperread = 0
        while searchread.find(paseq) != -1:
            paindex = searchread.find(paseq)
            for indexinread in range(0, len(paseq)):
                newread[indexinread + paindex + polaindecesperread] = labellist[indexinread]
                newreadstr = ''.join(newread)

            searchread = searchread[(searchread.find(paseq)+1):]
            polaindecesperread = polaindecesperread + paindex + 1
            labeledreads.append(newreadstr)
    labeledreads = set(labeledreads)
    return(labeledreads)


multimerfile = list(SeqIO.parse("RA07007DASHR2", "fastq"))
purefile = list(SeqIO.parse("RA05225DASHR2","fastq"))
readsMultimer = collectReads(multimerfile)
readsPure = collectReads(purefile)
```

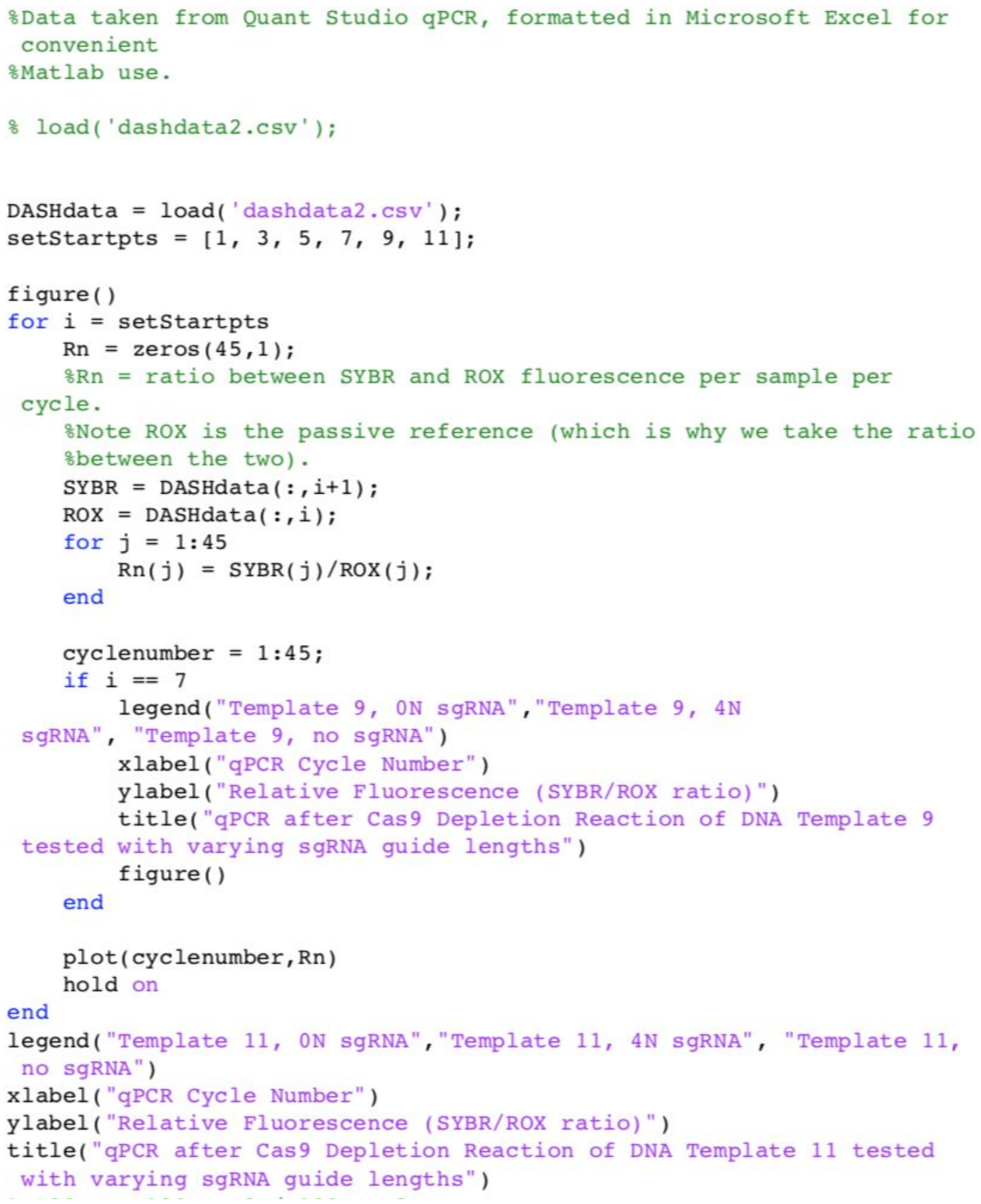


```
seqB = "ACTCTGCGTTGATACCACTGCTT"
seqBindexesMultimer = searchSeq(seqB,readsMultimer)
seqBindexesPure = searchSeq(seqB,readsPure)

percentMULTIMER = percentMultimerNOMUT(seqBindexesMultimer)

percentPURE = percentMultimerNOMUT(seqBindexesPure)
percentPURE

#scan for seqB or highlight
#try recalculating for snps
#seqBsnps includes original seqB
seqBsnps = generateSNPS(seqB)
percentmultimerSNPS = 0;
seqBsnpsPERCENT = []
for snp in seqBsnps:
    #indexes that matched to the seqB snp
    matchindexessnp = searchSeq(snp,readsMultimer)
    percentmultimersnp = percentMultimerNOMUT(matchindexessnp)
    seqBsnpsPERCENT.append(percentmultimersnp)

# seqBsnpsPERCENT
totalpercentMultimerSNPS = sum(seqBsnpsPERCENT)
totalpercentMultimerSNPS

#labled contains the annotated list of reads
labeled = annotateSeqBandEnd(seqBsnps, readsMultimer)
labeled = annotatePolyA(labeled)


labeled_list = list(labeled)

# Write annotated data to file
F = open('ANNOTATED.csv', 'w')
F.write('\n'.join(labeled_list))
F.close()

# Test that file creation worked by accessing it
open('filename.csv', 'r').readlines()[:3]

```



```
    seqBsnps.append(seqB);
    for index in range(0,23):
        #baseindex indexes into a c t or g
        for baseindex in range(0,4):
            if seqB[index] != bases[baseindex]:
                seqBsnp[index] = bases[baseindex];
                seqBsnp = ''.join(seqBsnp)
                seqBsnps.append(seqBsnp);
                seqBsnp = list(seqB)
    return(seqBsnps)

def annotateSeqBandEnd(seqBsnps, reads):
    seqBlabel = "---------SEQ-B---------"
    seqBMUTlabel = "------SEQ-B-MUT--------"
    end = "CCGCGGACAGGCGTG"
    endlabel = "------END------"
    mutlabellist = list(seqBMUTlabel)
    labellist = list(seqBlabel)
    endlabellist = list(endlabel)
    labeledreads = []
    for read in reads:
        listread = list(read)
        newread = listread
        for snp in seqBsnps:
            listsnp = list(snp)

            endindex = read.find(end)
            if endindex != -1:
                for readindex in range(0,len(end)):
                    newread[readindex + endindex] = endlabellist[readindex]


            searchread = read
            snpindecesperread = 0

            while searchread.find(snp) != -1:

                snpindex = searchread.find(snp)

                for indexinread in range(0, len(snp)):
                    newread[indexinread + snpindex + snpindecesperread] = labellist[indexinread]
                newreadstr = ''.join(newread)
```

DASH qPCR Analysis on Control Samples



```
%Data taken from Quant Studio qPCR, formatted in Microsoft Excel for
 convenient
%Matlab use.

% load('dashdata2.csv');


DASHdata = load('dashdata2.csv');
setStartpts = [1, 3, 5, 7, 9, 11];

figure()
for i = setStartpts
    Rn = zeros(45,1);
    %Rn = ratio between SYBR and ROX fluorescence per sample per
 cycle.
    %Note ROX is the passive reference (which is why we take the ratio
    %between the two).
    SYBR = DASHdata(:,i+1);
    ROX = DASHdata(:,i);
    for j = 1:45
        Rn(j) = SYBR(j)/ROX(j);
    end

    cyclenumber = 1:45;
    if i == 7
        legend("Template 9, 0N sgRNA","Template 9, 4N
 sgRNA", "Template 9, no sgRNA")
        xlabel("qPCR Cycle Number")
        ylabel("Relative Fluorescence (SYBR/ROX ratio)")
        title("qPCR after Cas9 Depletion Reaction of DNA Template 9
 tested with varying sgRNA guide lengths")
        figure()
    end

    plot(cyclenumber,Rn)
    hold on
end
legend("Template 11, 0N sgRNA","Template 11, 4N sgRNA", "Template 11,
 no sgRNA")
xlabel("qPCR Cycle Number")
ylabel("Relative Fluorescence (SYBR/ROX ratio)")
title("qPCR after Cas9 Depletion Reaction of DNA Template 11 tested
 with varying sgRNA guide lengths")
```

7. References

Basler, Christopher F. 2017. “Molecular Pathogenesis of Viral Hemorrhagic Fever.” *Seminars in Immunopathology*. https://doi.org/10.1007/s00281-017-0637-x.

Biolabs, New England, and New England Biolabs. n.d. “In Vitro Digestion of DNA with Cas9 Nuclease, S. Pyogenes (M0386) v1.” *Protocols.io*. https://doi.org/10.17504/protocols.io.ch2t8d.

Casabona, J. C., J. M. Levingston Macleod, M. E. Loureiro, G. A. Gomez, and N. Lopez. 2009. “The RING Domain and the L79 Residue of Z Protein Are Involved in Both the Rescue of Nucleocapsids and the Incorporation of Glycoproteins into Infectious Chimeric Arenavirus-Like Particles.” *Journal of Virology*. https://doi.org/10.1128/jvi.00329-09.

Chen, Geng, Baitang Ning, and Tieliu Shi. 2019. “Single-Cell RNA-Seq Technologies and Related Computational Data Analysis.” *Frontiers in Genetics* 10 (April): 317.

Choi, Yoon Ha, and Jong Kyoung Kim. 2019. “Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing.” *Molecules and Cells* 42 (3): 189–99.

Cornu, T. I., and J. C. de la Torre. 2001. “RING Finger Z Protein of Lymphocytic Choriomeningitis Virus (LCMV) Inhibits Transcription and RNA Replication of an LCMV S-Segment Minigenome.” *Journal of Virology*. https://doi.org/10.1128/jvi.75.19.9415-9426.2001.

Cristinelli, Sara, and Angela Ciuffi. 2018. “The Use of Single-Cell RNA-Seq to Understand Virus–host Interactions.” *Current Opinion in Virology*. https://doi.org/10.1016/j.coviro.2018.03.001.

Cross, Robert W., Emily Speranza, Viktoriya Borisevich, Steven G. Widen, Thomas G. Wood, Rebecca S. Shim, Ricky D. Adams, et al. 2018. “Comparative Transcriptomics in Ebola Makona-Infected Ferrets, Nonhuman Primates, and Humans.” *The Journal of Infectious Diseases* 218 (Suppl 5): S486.

Diehl, William E., Aaron E. Lin, Nathan D. Grubaugh, Luiz Max Carvalho, Kyusik Kim, Pyae Phyo Kyawe, Sean M. McCauley, et al. 2016. “Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic.” *Cell*. https://doi.org/10.1016/j.cell.2016.10.014.

Gierahn, Todd M., Marc H. Wadsworth 2nd, Travis K. Hughes, Bryan D. Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J. Christopher Love, and Alex K. Shalek. 2017. “Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput.” *Nature Methods* 14 (4): 395–98.

Gire, Stephen K., Augustine Goba, Kristian G. Andersen, Rachel S. G. Sealfon, Daniel J. Park, Lansana Kanneh, Simbirie Jalloh, et al. 2014. "Genomic Surveillance Elucidates Ebola Virus Origin and Transmission during the 2014 Outbreak." *Science* 345 (6202): 1369–72.

Günther, Stephan, and Oliver Lenz. 2004. "Lassa Virus." *Critical Reviews in Clinical Laboratory Sciences* 41 (4): 339–90.

Gu, W., E. D. Crawford, B. D. O'Donovan, M. R. Wilson, E. D. Chow, H. Retallack, and J. L. DeRisi. 2016. "Depletion of Abundant Sequences by Hybridization (DASH): Using Cas9 to Remove Unwanted High-Abundance Species in Sequencing Libraries and Molecular Counting Applications." *Genome Biology* 17 (March): 41.

Hastie, Kathryn M., Michelle Zandonatti, Tong Liu, Sheng Li, Virgil L. Woods Jr, and Erica Ollmann Saphire. 2016. "Crystal Structure of the Oligomeric Form of Lassa Virus Matrix Protein Z." *Journal of Virology* 90 (9): 4556–62.

Heilemann, Mike, Sebastian van de Linde, Mark Schüttpelz, Robert Kasper, Britta Seefeldt, Anindita Mukherjee, Philip Tinnefeld, and Markus Sauer. 2008. "Subdiffraction-Resolution Fluorescence Imaging with Conventional Fluorescent Probes." *Angewandte Chemie* 47 (33): 6172–76.

Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang. 2018. "Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines." *Experimental & Molecular Medicine* 50 (8): 96.

Kaur, Binnypreet. n.d. "In Vitro Digestion of DNA with Cas9 Nuclease, S. Pyogenes (M0386) v1 (protocols.io.rmud46w)." *Protocols.io*. https://doi.org/10.17504/protocols.io.rmud46w.

Klein, Allon M., and Evan Macosko. 2017. "InDrops and Drop-Seq Technologies for Single-Cell Sequencing." *Lab on a Chip* 17 (15): 2540–41.

Lin, AE. Courtesy of Dr. Aaron Lin. 2019.

Lun, Aaron T. L., Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C. Marioni. 2019. "EmptyDrops: Distinguishing Cells from Empty Droplets in Droplet-Based Single-Cell RNA Sequencing Data." *Genome Biology* 20 (1): 63.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14.

McGinnis, Christopher S., Lyndsay M. Murrow, and Zev J. Gartner. 2019. "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors." *Cell Systems* 8 (4): 329–37.e4.

Nathan, Lakshmi, Alex L. Lai, Jean Kaoru Millet, Marco R. Straus, Jack H. Freed, Gary R. Whittaker, and Susan Daniel. 2020. "Calcium Ions Directly Interact with the Ebola Virus Fusion Peptide To Promote Structure-Function Changes That Enhance Infection." *ACS Infectious Diseases* 6 (2): 250–60.

Ortiz-Riaño, Emilio, Benson Cheng, Juan Torre, and Luis Martínez-Sobrido. 2012. "D471G Mutation in LCMV-NP Affects Its Ability to Self-Associate and Results in a Dominant Negative Effect in Viral RNA Synthesis." *Viruses*. https://doi.org/10.3390/v4102137.

Picelli, S. et al. "Tn5 Transposase and tagmentation procedures for massively scaled sequencing projects." Genome Res. 2014 Dec; 24(12): 2033-2040. doi:10.1101/gr.177881.114. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248319/

Ran, F. Ann, F. Ann Ran, Le Cong, Winston X. Yan, David A. Scott, Jonathan S. Gootenberg, Andrea J. Kriz, et al. 2015. "In Vivo Genome Editing Using Staphylococcus Aureus Cas9." *Nature*. https://doi.org/10.1038/nature14299.

Ranu, Navpreet, Alexandra-Chloé Villani, Nir Hacohen, and Paul C. Blainey. 2019. "Targeting Individual Cells by Barcode in Pooled Sequence Libraries." *Nucleic Acids Research* 47 (1): e4.

Reynard, Stéphanie, Alexandra Journeaux, Emilie Gloaguen, Justine Schaeffer, Hugo Varet, Natalia Pietrosemoli, Mathieu Mateo, et al. 2019. "Immune Parameters and Outcomes during Ebola Virus Disease." *JCI Insight* 4 (1). https://doi.org/10.1172/jci.insight.125106.

Schmider, Angela B., Melissa Vaught, Nicholas C. Bauer, Hunter L. Elliott, Matthew D. Godin, Giorgianna E. Ellis, Peter A. Nigrovic, and Roy J. Soberman. 2019. "The Organization of Leukotriene Biosynthesis on the Nuclear Envelope Revealed by Single Molecule Localization Microscopy and Computational Analyses." *PloS One* 14 (2): e0211943.

Tang L, He X, Liu X, Zhou C, Liu J, Ge X, Li J, Liu C, Zhao J, Qu J, Song Z, Gu F. 2018. "SaCas9 Requires 5'-NNGRRT-3' PAM for Sufficient Cleavage and Possesses Higher Cleavage Activity than SpCas9 or FnCpf1 in Human Cells. - PubMed - NCBI." Accessed July 29, 2019. https://www.ncbi.nlm.nih.gov/pubmed/29247600.

Torre, Juan C. de la. 2009. "Molecular and Cell Biology of the Prototypic Arenavirus LCMV: Implications for Understanding and Combating Hemorrhagic Fever Arenaviruses." *Annals of the New York Academy of Sciences* 1171 Suppl 1 (September): E57–64.

Tycko, Josh, Luis A. Barrera, Nicholas C. Huston, Ari E. Friedland, Xuebing Wu, Jonathan S. Gootenberg, Omar O. Abudayyeh, Vic E. Myer, Christopher J. Wilson, and Patrick D. Hsu. 2018. "Publisher Correction: Pairwise Library Screen Systematically Interrogates Staphylococcus Aureus Cas9 Specificity in Human Cells." *Nature Communications* 9 (1): 3542.

Waickman, Adam T., Kaitlin Victor, Tao Li, Kristin Hatch, Wiriya Rutvisuttinunt, Carey Medin, Benjamin Gabriel, Richard G. Jarman, Heather Friberg, and Jeffrey R. Currier. 2019. “Dissecting the Heterogeneity of DENV Vaccine-Elicited Cellular Immunity Using Single-Cell RNA Sequencing and Metabolic Profiling.” *Nature Communications* 10 (1): 3666.

“What Is the Illumina Method of DNA Sequencing?” 2014. Yourgenome. September 16, 2014. https://www.yourgenome.org/facts/what-is-the-illumina-method-of-dna-sequencing.

World Health Organization. (2015). Ebola Situation Reports. http://apps.who.int/ebola/en/ebola-situation-reports. Accessed February 26, 2020.

Wulf, Madalee G., Sean Maguire, Paul Humbert, Nan Dai, Yanxia Bei, Nicole M. Nichols, Ivan R. Corrêa Jr, and Shengxi Guan. 2019. “Non-Templated Addition and Template Switching by Moloney Murine Leukemia Virus (MMLV)-Based Reverse Transcriptases Co-Occur and Compete with Each Other.” *The Journal of Biological Chemistry* 294 (48): 18220–31.

Yourik Y., Fuchs R., et al. “Staphylococcus aureus Cas9 is a multiple-turnover enzyme.” RNA: A publication of the RNA society. 2018. Doi: 10.1261/rna.067355.118 https://www.biorxiv.org/content/early/2018/06/07/340042

Zanini, Fabio, Makeda L. Robinson, Derek Croote, Malaya Kumar Sahoo, Ana Maria Sanz, Eliana Ortiz-Lasso, Ludwig Luis Albornoz, et al. 2018. “Virus-Inclusive Single-Cell RNA Sequencing Reveals the Molecular Signature of Progression to Severe Dengue.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (52): E12363–69.

Zilionis, Rapolas, Juozas Nainys, Adrian Veres, Virginia Savova, David Zemmour, Allon M. Klein, and Linas Mazutis. 2017. “Single-Cell Barcoding and Sequencing Using Droplet Microfluidics.” *Nature Protocols* 12 (1): 44–73.