Title

Development of a high-throughput technology capable of detecting in situ DNA-RNA interactions and transcriptomes in single cells

Author

Wan, Xueyi

Publication Date

2022

UNIVERSITY OF CALIFORNIA SAN DIEGO

Development of a high-throughput technology capable of detecting *in situ* DNA-RNA interactions and transcriptomes in single cells

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Biology

by

Xueyi Wan

Committee in charge:

Professor Sheng Zhong, Chair
Professor Enfu Hui, Co-Chair
Professor Cornelis Murre

2022

The Thesis of Xueyi Wan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

University of California San Diego

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

sciMARGI	single cell *in situ* Mapping of RNA-Genome Interactome
SPRITE	Split-pool Recognition of Interactions by Tag Extension
scSPRITE	single-cell Split-pool Recognition of Interactions by Tag Extension
iMARGI	*in situ* Mapping of RNA-Genome Interactome
GRID-seq	Global RNA Interactions with DNA by deep sequencing
G&T-seq	Genome and Transcriptome sequencing
sci-RNA-seq	single cell Combinatorial Indexing RNA sequencing
SPLiT-seq	Split Pool Ligation-based Transcriptome sequencing
Smart-seq	mRNA-Seq with SMART™ template switching technology
DroNC-seq	massively parallel sNuc-seq with droplet technology
FREE	Filled/truncated Right End Edit
lncRNA	long non-coding RNA
eRNA	enhancer RNA
sisRNA	stable intronic sequence RNA
miRNA	microRNA
mRNA	messenger RNA
TAD	topologically associating domain
GWAS	genome-wide association study
SNP	single nucleotide polymorphisms
FACS	flow-activated cell sorting
DSG	disuccinimidyl glutarate

FA	formaldehyde
EthD-1	Ethidium Homodimer-1
SDS	sodium dodecyl sulfate
PCR	polymerase chain reaction
NLS	sodium dodecyl sulfate
DOC	sodium deoxycholate
UMI	unique molecular identifier
TS-RT	template-switching reverse transcription
TSO	template-switching oligo
RIN	RNA integrity number
SP	sequencing primer
DF	duty factor
PIP	peak incident power
HEK	human embryonic kidney 293
H1	human embryonic stem cell line H1
mESC	mouse embryonic stem cell
OH	hydroxyl group
App	adenylated group
Pho	phosphate group
wt/vol	weight/volume
bp	base pair
nt	nucleotide

LIST OF SUPPLEMENTAL FILES

wan_sciMARGI_protocol_cell_lines.docx

wan_sciMARGI_protocol_tissue.docx

wan_sciMARGI_oligos_list.xlsx

wan_sciMARGI_final_library_configuration.pdf

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Sheng Zhong for his support as the chair of my committee. His guidance and encouragement have been invaluable in my two years of master's study.

I would also like to acknowledge Dr. Zhifei Luo, Dr. Tri C Nguyen, and Xingzhao Wen, without whom this project would not have progressed this far in less than two years. Having the opportunity to collaborate with them has been a rewarding experience for me.

This thesis, especially Sections 3, 4, and 5, contains unpublished material coauthored with Dr. Zhifei Luo and Xingzhao Wen. The thesis author was the primary author of these chapters.

ABSTRACT OF THE THESIS

Development of a high-throughput technology capable of detecting *in situ* DNA-RNA interactions and transcriptomes in single cells

by

Xueyi Wan

Master of Science in Biology

University of California San Diego, 2022

Professor Sheng Zhong, Chair
Professor Enfu Hui, Co-Chair

While methods for detecting single cell transcriptome and genome architecture have been established, there are currently no technologies for detecting single-cell DNA-RNA interactions. Thus, while a regulatory role for chromatin and RNA organization has been hypothesized, we still lack a comprehensive understanding of where and how chromatin and RNA interact, as well as the consequences for gene regulation.

In this thesis, I describe a breakthrough technology called single cell *in situ* Mapping of RNA-Genome Interactome (sciMARGI). Through integrating two promising technologies, combinatorial indexing and microdroplet methods, we can decode single-cell single-complex information by labeling interacting DNA and RNA molecules with unique cell and cluster barcodes. Currently, a DNA library based on human brain tissue has been successfully generated and shown to have a proper single cell cluster distribution and DNA-DNA interaction map. While the RNA library still needs additional adjustments, the libraries as a whole demonstrated promising results.

When fully developed, the sciMARGI technology will not only detect three-way *in situ* interactions between DNA and DNA, DNA and RNA, and RNA and RNA, but also quantify RNA within the single nucleus, establishing a direct link between chromatin-RNA organization and gene expression. Due to its exceptional versatility, sciMARGI can be applied to not only cell lines but also complex tissues, allowing us to resolve some of the most critical obstacles regarding gene expression regulatory processes.

1. Introduction and Significance

The RNA expression profile has long been regarded as a fundamental criterion for evaluating biological processes. It is capable of reflecting not just beneficial or detrimental cell conditions, but also initiating cell fate decisions (Hwang et al., 2018). Due to the discovery that gene expression is variable even among similar cell types, single-cell RNA-seq technologies have grown in popularity and prominence. To more precisely decipher the relationship between gene expression and phenotypic outcomes, it is necessary to thoroughly explore the mechanisms governing the regulation and modification of transcriptome patterns. The effects of chromatin structure and RNA-chromatin interactions on gene expression patterns are two distinct types of regulatory mechanisms that have emerged only in the last two decades and remain mostly unknown.

The genome in the eukaryotic nucleus has a three-dimensional architecture, which includes stable chromosome territories, TAD structure, and transient DNA-DNA interactions such as enhancer-promoter loops. It has been discovered that the three-dimensional genome organization within a single nucleus regulates a variety of nuclear processes, including DNA replication, RNA processing, and gene expression (Arrastia et al., 2021). Because of their transient nature, the hypothesized mechanism for fine-tuning transcription through chromatin interacting with distal regulatory components such as enhancers and promoters remains unclear. Considering the heterogeneity of cell types and even individual cells, it is desirable to understand precisely how changes in chromatin structure could affect the gene expression profile at the single-cell level.

Additionally, RNA and chromatin could interact to influence nuclear structure and gene expression. It has been discovered that long non-coding RNAs (lncRNAs), enhancer RNAs (eRNAs), stable intronic sequence RNAs (sisRNAs), and a variety of other types of transcripts need to work together to ensure the nucleus's chromatin is tightly regulated (Khelifi and Hussein, 2020). While developing research has identified numerous examples of short- and long-range RNA-chromatin interactions (for example, LUNAR1 and XIST), we still lack a comprehensive picture of the distribution of RNA-chromatin interactions and the significance of the hypothesized processes. As a result, it's been a long-standing goal to figure out where and how RNA interacts with chromatin, as well as the consequences for gene regulation.

To address the issues mentioned above and obtain a more comprehensive understanding of how chromatin structure and RNA-chromatin interactions affect biological functions, we are developing a novel technique termed single cell *in situ* Mapping of RNA-Genome Interactome (sciMARGI). This method allows the detection of three-way interactions between DNA and DNA, DNA and RNA, and RNA and RNA. Additionally, it quantifies RNA within the nucleus at the same time, establishing a relationship between chromatin and RNA organization and gene expression. SciMARGI's capacity to show DNA RNA interactomes and gene expression profiles enables it to address significant challenges in gene expression regulatory mechanisms. More precisely, our proposed method is extremely versatile and can be used to investigate a wide variety of biological questions, including identifying the target genes of enhancers harboring a GWAS SNP, examining chromatin associated RNAs, searching for miRNA and its mRNA targets, and identifying regulatory elements that contribute to risk gene expression in human diseases such as cancer, Parkinson's disease, and Alzheimer's disease.

2. Technology Reviews and Inspirations

2.1 Chromatin-structure detection technology

To create a breakthrough technology that can connect genome structure and RNA-chromatin interaction to gene expression, we must draw inspiration and directions from existing technologies in the relevant fields.

At the moment, there are two primary technologies for identifying genome architecture at the single-cell level: single-cell Hi-C and single-cell SPRITE (scSPRITE). Single-cell Hi-C provides a genomic perspective of chromatin interactions in single cells and has lately become one of the most frequently utilized tools for detecting genome architecture. However, because of its proximity-ligation-based detection method, single-cell Hi-C is limited in its capacity to detect long-range DNA interactions (Nagano et al., 2013). In comparison, single-cell SPRITE is based on combinatorial indexing rather than proximity ligation, allowing for the identification of multiple DNA interactions. Apart from pairwise interactions, single-cell SPRITE uncovers higher-order interactions, such as those found in clusters surrounding nuclear bodies and nucleoli (Arrastia et al., 2021).

The technologies discussed above, particularly scSPRITE, have enabled the investigation of chromatin architecture in complex tissues such as the brain and tumors. However, it remains challenging to establish a direct link between those heterogeneous genomic architectures and the gene expression profiles of individual cells. For instance, we are still unsure if interaction with a particular regulatory element results in an increase or decrease in the expression of target genes, particularly for regulatory elements containing mutations.

2.2 RNA-chromatin detection technology

Due to the fact that RNA-chromatin interaction is a relatively new area, few tools for detecting RNA-DNA interactions at the single-cell level have been developed. Currently, only bulk-level interactomics technologies such as iMARGI and GRID-seq are capable of detecting genome-wide DNA-RNA interactions, and their proximity-ligation-based approaches provide a significant hurdle to their advancement to the single-cell level (Li et al., 2017). Typically, those technologies employ a linker that binds to DNA on one end and RNA on the other end, which not only limited their detection of higher-order interactions in complexes, but also exposed them to a major risk of losing critical information as chromosome coverage decreased dramatically in single-cell technologies (Wu et al., 2019). Therefore, present RNA-DNA interactomics techniques are not only incapable of providing precise information about how RNA-chromatin interactions affect gene regulation in single cells, but also are not suited for advancement to the single-cell level due to technological constraints.

To gain a better understanding of the interaction between DNA and RNA at the single-cell level, scientists also developed single-cell multiomics technologies. For instance, Macaulay et al. introduced G&T-seq, a technology that combines single-cell genome and transcriptome sequencing, allowing for the simultaneous reading of DNA and RNA expression profiles from individual cells (Macaulay et al., 2015). However, this type of technology does not provide information on which RNA and DNA molecules interact, and we still do not understand where and how RNA molecules interact with chromatin.

Despite the fact that the field of DNA-RNA interaction currently faces significant technological challenges, we were surprised that the SPRITE combinatorial indexing method provided useful ideas. Apart from utilizing SPRITE to identify DNA-DNA interactions, Quinodoz et al. ligated an adaptor to RNA molecules, allowing for the mapping of interacting DNA and RNA molecules inside each crosslinked cluster (Quinodoz et al., 2018). We may also simultaneously tag the interacting DNA and RNA molecules in single cells using the flexibility offered by such a combinatorial indexing strategy, thus providing a partial image for building single-cell RNA/chromatin interactome detection technologies.

2.3 Single-cell transcriptomes detection technology

Many effective single-cell RNA-seq (scRNA-seq) technologies have been well-established and even commercialized in recent decades. scRNA-seq approaches can be classified into three broad categories based on their isolation methods: limited dilution, flow-activated cell sorting (FACS), and microfluidic techniques (Hwang et al., 2018).

One of the most traditional and widely utilized approaches is limited dilution. Although this is not the most efficient method, it does not require the use of a microscope and simply requires pipettes to isolate individual cells. Sci-RNA-seq (Cao et al., 2017) and SPLiT-seq (Rosenberg et al., 2018) are two current RNA-seq technologies that use limited dilutions; and SPRITE's combinatorial indexing method employs a limited dilution strategy (Quinodoz et al., 2018).

FACS is the most commonly used approach for isolating highly pure single cells. It is frequently employed in a variety of single-cell RNA sequencing technologies, including Smart-seq (Picelli et al., 2014) and Microwell-seq (Han et al., 2018). However, it necessitates the use of monoclonal antibodies to target proteins of interest, as well as a high initial volume of cells.

Microfluidic and microdroplet-based isolation techniques are the most recent advancements in single-cell isolation technologies. In comparison to FACS approaches, it requires fewer samples and has a lower analytic cost for single-cell RNA captures, as demonstrated by Drop-seq (Macosko et al., 2015) and DroNC-seq (Habib et al., 2017). Furthermore, commercial platforms with great consistency and high-throughput, such as the Chromium built by 10X Genomics, are already available. We have been inspired by such a novel and efficient method in the design of our technology. Because the Chromium platform is capable of isolating not just single cells or nuclei, but also single molecules such as nucleic acids. Thus, by repurposing the Chromium platform, we can isolate DNA-RNA-protein complexes and selectively capture target molecules via the addition of particular adaptors. This new insight contributes to the development of our single-cell single-complex multi-way interactome detection technology.

2.4 Inspirations for sciMARGI technology

After evaluating the existing technologies in the fields of chromatin structure, DNA-RNA interactomes, and transcriptomes, we can now identify the challenge of revealing single-cell RNA-chromatin interactome information, as well as the direction to uncover the impact of genome structure and interactomes on biological processes. Because of this, the goal of inventing a technology that can reveal the information in interacting complexes is affirmed to be substantial and extremely valuable.

In terms of general technological design, we integrated two promising technologies, scSPRITE and Chromium, to reveal single-cell single-complex information. Due to the flexibility of combinatorial indexing, we can modify the DNA and RNA ends in isolated nuclei to include a combinatorial barcode, providing each nucleus a unique "cell barcode." Because commercialized microdroplet technology is convenient and consistent, we can repurpose the 10X genomic platform to affix a unique "cluster barcode" to the molecules in complexes once they are released from the nuclei. By adding the cell and cluster barcodes together, we can determine which DNA and RNA are interacting and trace them back to the original cell. Additionally, after ligating to the cell barcodes and being collected in 10X microdroplets, the RNA molecules can provide the single-cell gene expression information on their own.

This way, we may provide three levels of information about a single cell via this cell barcode plus cluster barcode design. To begin with, we have information about the chromatin structure simply by examining the DNA readings. Second, we could obtain an RNA expression profile by analyzing the RNA readings. Lastly, we can reveal the interacting DNA and RNA molecules by sorting sequencing reads with identical cell and cluster barcodes. Thus, we will achieve our goal of inventing a technology capable of directly correlating changes in genome structure and RNA-chromatin structure with gene expression profiles in single cells.

3. Overview of the sciMARGI procedures

To begin, we start with single nuclei rather than single cells. Due to the difficulty of dissociating complex tissues, whether fresh or archived, and obtaining intact single cells, employing single nuclei rather than single cells may make it easier to apply our method to complex tissues. Furthermore, while there may be an enrichment of nascent transcripts in the nucleus relative to the rest of the cell, data already indicates that the average expression profile of single nuclei agrees with the rest of the cell (Habib et al., 2017). As a result, using nuclei rather than whole cells will not invalidate our results for gene expression sequencing.

Before adaptors or barcodes are added, the sciMARGI approach crosslinks the nuclei, ensuring the covalent bonding of the interacting DNA-RNA-protein complexes. We next permeabilize the nuclei to make them accessible for subsequent treatments. We have established and evaluated the processes for working with cell lines and complex tissues. Section 4.1 below discusses the design and considerations for nuclei preprocessing.

After acquiring the prepared nuclei, we must fragment the chromatin in order to secure the appropriate size of the DNA and RNA molecules. Section 4.2 discusses the logic and validation tests for selecting the ideal sizes.

Then, we will use the combinatorial indexing method to add cell barcodes to the nuclei, considering its ability to modify both DNA and RNA molecules. We begin by inserting adaptors for DNA and RNA ends in the crosslinked complexes into each nucleus. To differentiate single nuclei, those adaptors will be ligated to distinct cell barcodes in three rounds of split-pool ligation, yielding 96^3 (884,736) combinations capable of separating thousands of nuclei. Sections 4.3 and 4.4 detail the design and considerations for the adaptors and the cell barcodes, respectively.

Sonication is then used to rupture the nuclear membranes, releasing the covalently bound DNA-RNA-protein complexes into solution. Section 4.5 describes the optimization tests used to select the sonication parameters.

We choose the microdroplet technique for this stage not only because it is reliable and high-throughput, but also because it avoids the capture of DNA-RNA-protein complexes by NHS beads, hence lowering the possibility of DNA-RNA interaction clusters being lost when compared to the scSPRITE technology (Arrastia et al., 2021). To be more precise, complexes from thousands of nuclei are divided into numerous droplets, each of which contains one of the 3.5 million distinct barcodes. Because the ends of DNA and RNA are already barcoded, the 3.5 million cluster barcodes may be used to separate thousands of single complexes in each nucleus. Section 4.6 details the full architecture and mechanism for repurposing the 10X Genomics platform.

After adding both cell and cluster barcodes to DNA and RNA ends, we can uncrosslink the complexes and then produce cDNA from the RNA ends using reverse transcription and template switching mechanisms. Finally, we will add sequencing adaptors to the ends of the DNA and RNA libraries, amplify the barcoded DNA and RNA libraries, and sequence the final DNA and RNA libraries. Sections 4.7 and 4.8 discuss the strategies and considerations for those final library construction processes.

To better visualize the entire sciMARGI library preparation procedure, Figure 3.1.1 depicts the main flowchart for the processes discussed before.

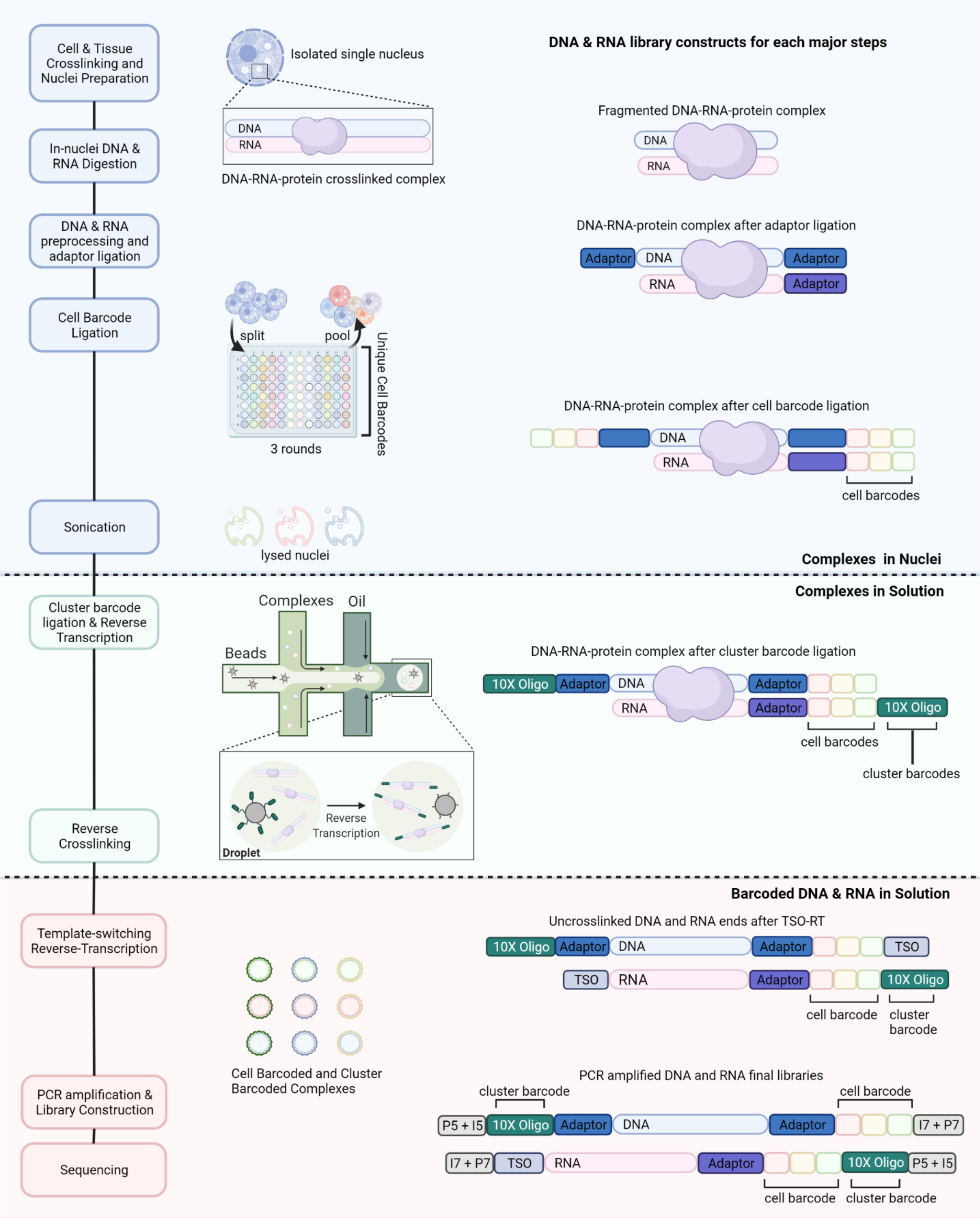

Cell & Tissue Crosslinking and Nuclei Preparation
In-nuclei DNA & RNA Digestion
DNA & RNA preprocessing and adaptor ligation
Cell Barcode Ligation
Sonication
Cluster barcode ligation & Reverse Transcription
Reverse Crosslinking
Template-switching Reverse-Transcription
PCR amplification & Library Construction
Sequencing
Isolated single nucleus
DNA
RNA
DNA-RNA-protein crosslinked complex
split
pool
Unique Cell Barcodes
3 rounds
lysed nuclei
DNA & RNA library constructs for each major steps
Fragmented DNA-RNA-protein complex
DNA
RNA
DNA-RNA-protein complex after adaptor ligation
Adaptor
DNA
Adaptor
RNA
Adaptor
DNA-RNA-protein complex after cell barcode ligation
DNA
RNA
cell barcodes
Complexes in Nuclei
Complexes in Solution
Complexes
Oil
Beads
Reverse Transcription
Droplet
DNA-RNA-protein complex after cluster barcode ligation
10X Oligo
Adaptor
DNA
Adaptor
RNA
Adaptor
10X Oligo
cell barcodes
cluster barcodes
Barcoded DNA & RNA in Solution
Uncrosslinked DNA and RNA ends after TSO-RT
10X Oligo
Adaptor
DNA
Adaptor
TSO
TSO
RNA
Adaptor
10X Oligo
cell barcode
cluster barcode
Cell Barcoded and Cluster Barcoded Complexes
PCR amplified DNA and RNA final libraries
cluster barcode
cell barcode
P5 + I5
10X Oligo
Adaptor
DNA
Adaptor
I7 + P7
I7 + P7
TSO
RNA
Adaptor
10X Oligo
P5 + I5
cell barcode
cluster barcode

Figure 3.1.1: High-level overview of the sciMARGI processes and designs

4. Experimental Designs and Considerations

4.1 Cell and tissue crosslinking and nuclei preparation

To maximize the detection of indirect RNA-DNA interactions generated by intermediate proteins, we adopted a dual crosslinking approach. We begin by crosslinking nucleic acids to proteins using formaldehyde, followed by protein-protein crosslinking with DSG (Nguyen, 2018). Although strong crosslinking conditions have been shown to aid in the stabilization of long-range interactions, they also tend to complicate nucleus isolation and reduce nuclei permeability, preventing linkers and enzymes from entering the nuclear membrane during subsequent phases (Zhou et al., 2019). To assess the ideal crosslinking conditions for sciMARGI, we evaluated both weak (2 mM DSG + 1% (wt/vol) formaldehyde) and strong (2 mM DSG + 3% (wt/vol) formaldehyde) conditions on HEK and H1 cell lines, as well as human brain tissues. We discovered that while the strong crosslinking condition can be employed in certain cell lines (such as HEK), it results in significantly larger DNA fragments in other cell lines (such as H1) [Figure 4.1.1 a]. However, only 1% formaldehyde can be utilized to achieve the desired DNA fragment size in human brain tissue [Figure 4.1.1 b]. Additionally, we investigated why elevated formaldehyde levels would result in such large DNA fragments in tissue samples. We observed that increased reverse crosslinking time under 100% permeabilization can result in smaller DNA fragments in mouse brain tissues following restriction enzyme digestion [Figure 4.1.1 c]. Apart from decreasing nucleus permeabilization, we speculate that 3% formaldehyde limits the accessibility of the chromatin structure to restriction enzymes, thus resulting in insufficient chromatin digestion. As a result, for consistency across samples, we recommend using the crosslinking condition with only 1% formaldehyde. When employing only certain cell lines, you can try crosslinking with the strong condition to obtain additional data.

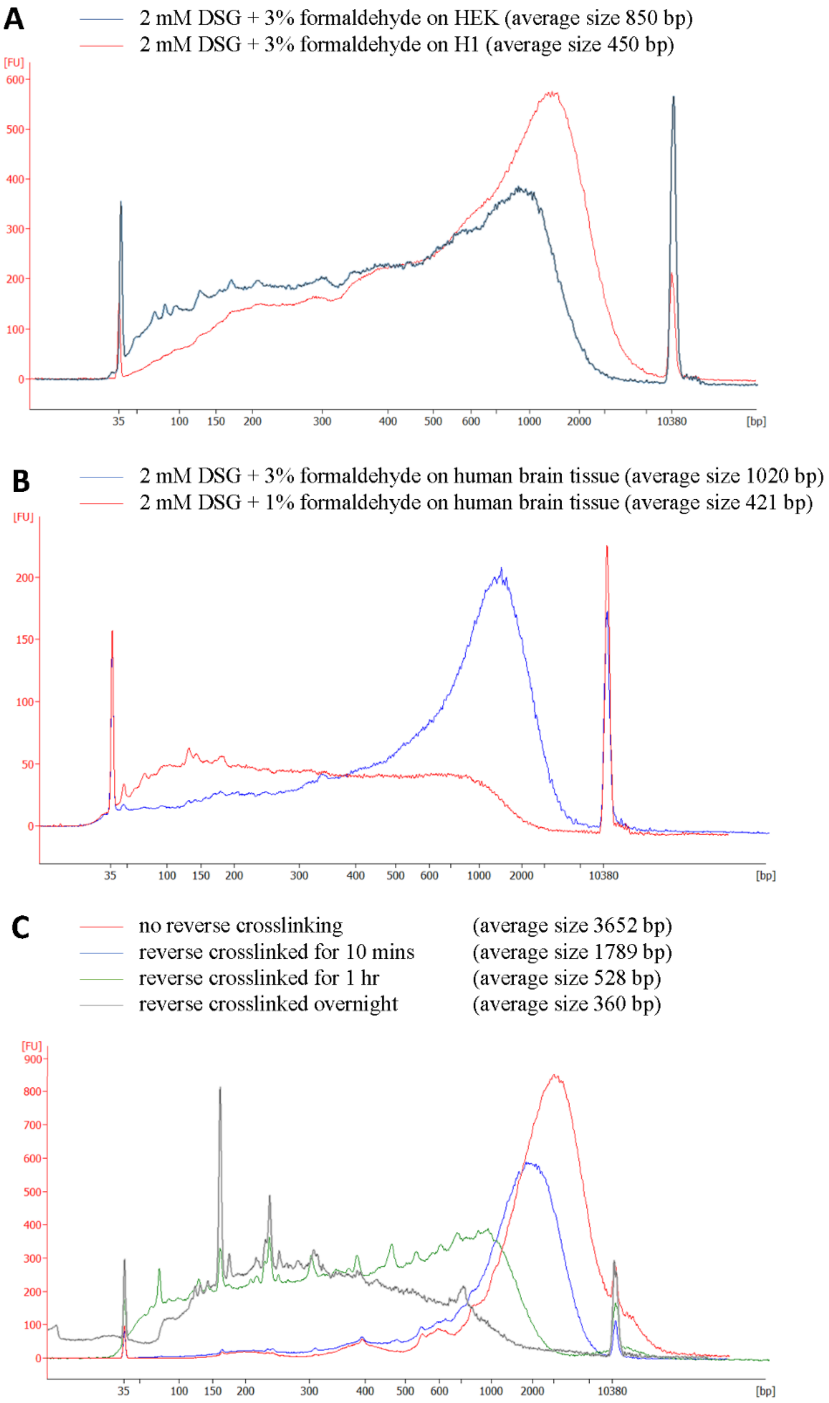


Figure 4.1.1: DNA fragmentation results under various crosslinking conditions
A) H1 and HEK cells were treated with the same crosslinking condition (2 mM DSG + 3% formaldehyde), fully permeabilized, digested with restriction enzymes, and then DNA distribution was measured using Bioanalyzer. B) human brain tissue was treated with different crosslinking conditions (3% and 1% formaldehyde, respectively), fully permeabilized, digested with restriction enzymes, and then DNA distribution was measured using Bioanalyzer. C) mouse brain tissue treated with the same crosslinking condition (with 3% formaldehyde), fully permeabilized, and then reverse crosslinked for various times. The DNA distributions after restriction enzyme digestion are measured by Bioanalyzer.

To efficiently and reliably isolate single nuclei from complex tissues, we adapted and modified the gradient centrifugation approach developed by Howard Yang's group (Corces et al., 2017). We collected high-quality single nuclei successfully from fresh mouse brain and frozen human brain tissues. Figure 4.1.2 shows an example of microscopic images of nuclei collected using this method. The ethidium homodimer stain demonstrates that all isolated nuclei are permeabilized. The supplementary files contain the complete approach and adjustments.

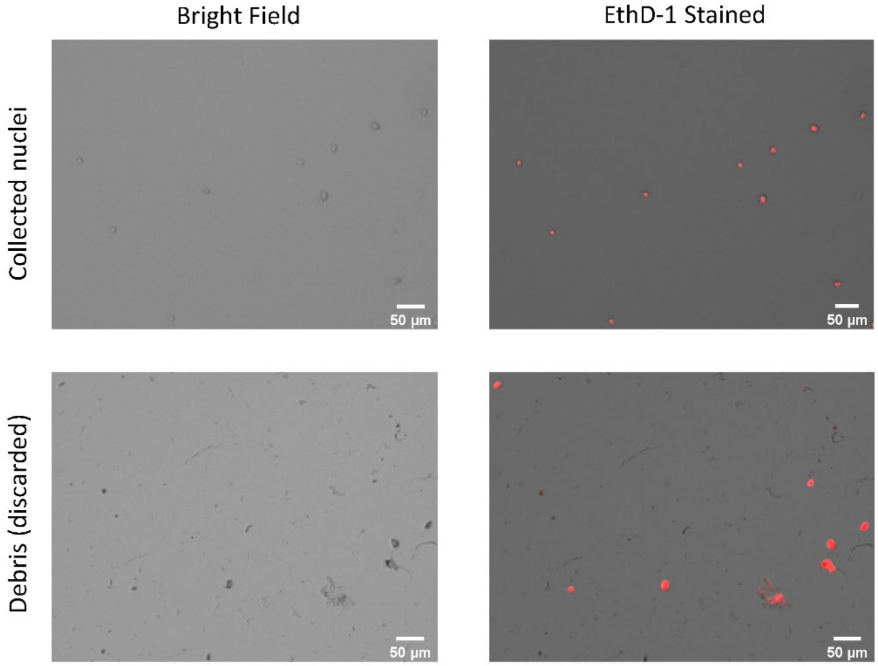


Figure 4.1.2: Isolated nuclei from frozen human brain tissues
The upper row shows the isolated nuclei obtained via density gradient centrifugation, whereas the lower row shows the debris that will be discarded. Permeabilized nuclei are red-stained by Ethidium Homodimer-1 (EthD-1).

Sufficient permeabilization of isolated nuclei is required to allow enzymes and oligos to enter the nucleus. To increase the nuclear permeabilization ratio without compromising nuclear integrity, we investigated numerous permeabilization conditions on a variety of cell lines and frozen tissues. For HEK, mESC, and H1, we determined that treating cells with 0.3 percent SDS at 65C for 10 minutes is sufficient to completely permeabilize the nuclei while keeping their shape. For human and mouse brain tissues, as our nuclei isolation process already completely permeabilized the nuclei, we noticed that treating with SDS does not help further permeabilize the nuclei, but may possibly damage the nuclei. Figure 4.1.3 shows an example of isolated mESC nuclei with a permeabilization ratio of 100% with an undamaged shape. Although these criteria may be applied to most cell lines and tissues, we still recommend performing a pre-experiment to examine the nuclear integrity under a microscope and the permeabilization ratio using EthD-1 stains to ensure that the nuclei are thoroughly permeabilized and undamaged.

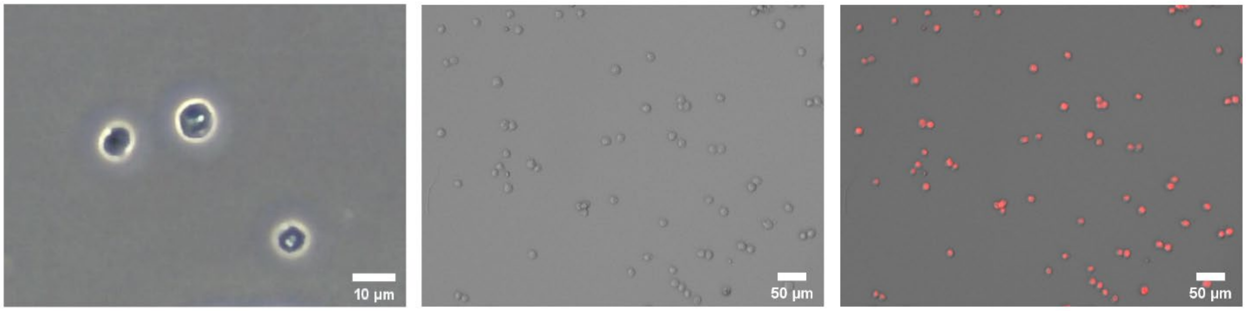


Figure 4.1.3: Permeabilized nuclei from mouse embryonic stem cells
After cell lysis, the nuclei are treated with 0.3% SDS at 65C for 10 minutes and then quenched with 3% Triton X-100 at 37C for 15 minutes. The left image was captured using a 40X microscope, whereas the middle image was captured with a 0.6X microscope. EthD-1 staining of permeabilized nuclei is shown on the right.

4.2 In nuclei digestion of DNA and RNA molecules

To maximize the efficiency of sequencing on the Illumina platform, the ideal range of DNA and RNA fragments should be 250-400 bp, with the majority of DNA and RNA fragments being between 200-800 bp (Quail et al., 2009). To ascertain which restriction enzyme could generate the requisite DNA fragment size, we found references suggesting that the HpyCH4V enzyme can make DNA fragments with an average size of 823 bp (Arrastia et al., 2021). However, the HpyCH4V enzyme alone cannot meet our requirement for the majority of DNA fragments to be between 200 and 800 bp in length. Hence, we evaluated additional 4-base pair restriction enzyme combinations [figure 4.2.1 a, b]. We ultimately chose to use HpyCH4V in combination with HaeIII, because this combination of enzymes not only achieves the appropriate DNA size in both cell lines and tissues, but also eliminates the necessity for post-digestion DNA end-repair. Additionally, we evaluated DNA products after 4 hours and overnight enzyme digestion to determine whether extending the enzyme digestion period could result in a smaller DNA fragment size. We found that prolonging the digesting period beyond 4 hours had no discernible effect on DNA size distribution [figure 4.2.1 c]. As a result, we determined that four hours is the best digesting time.

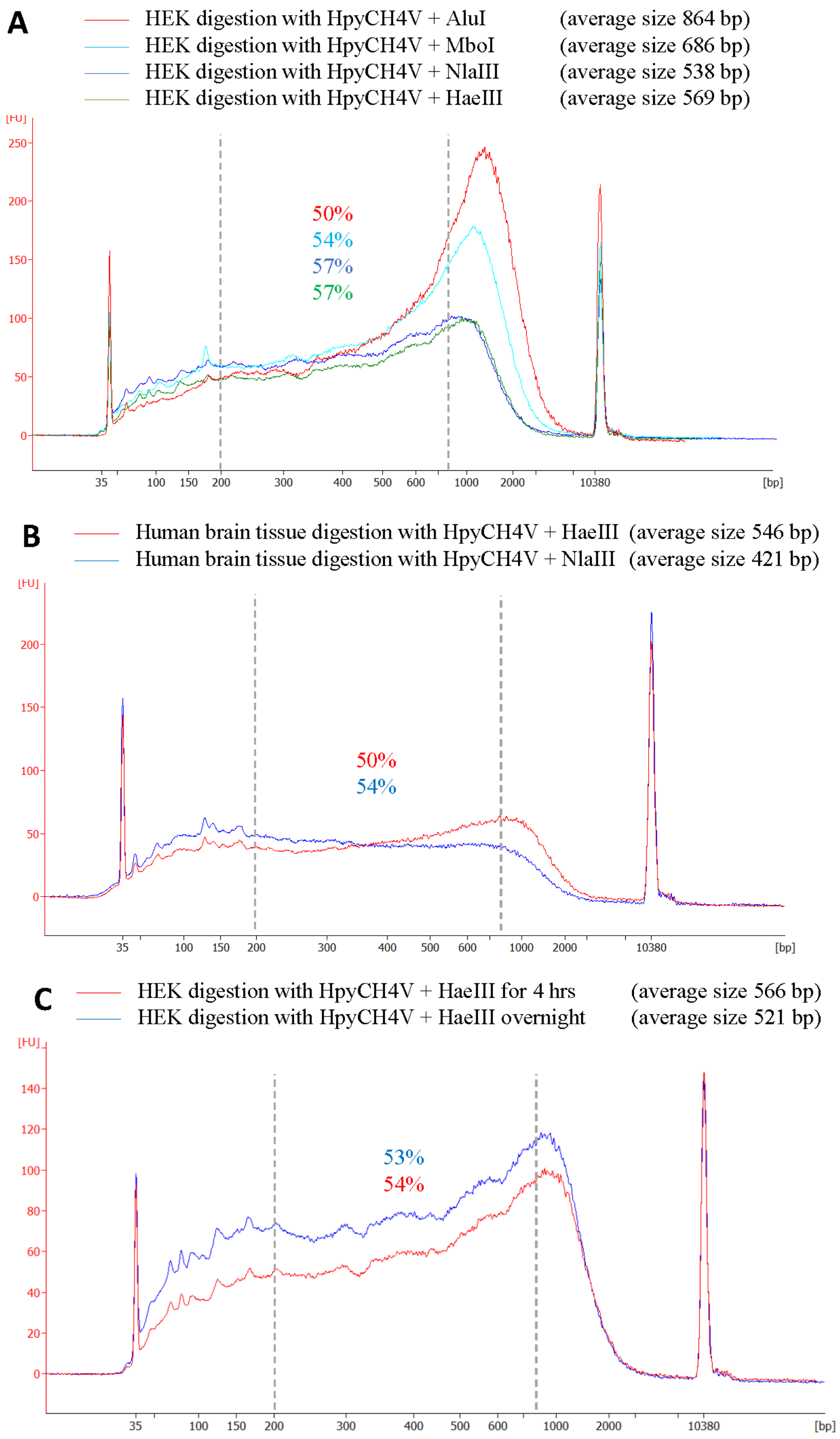


Figure 4.2.1: DNA fragmentation results under various digestion conditions
A) 3% FA crosslinked HEK cells are treated with different combinations of restriction enzymes for 4 hours. B) 1% FA crosslinked human brain tissues are treated with different combinations of restriction enzymes for 4 hours. C) 3% FA crosslinked HEK cells are treated with the same restriction enzymes for different time periods. The DNA distribution is measured by Bioanalyzer. The percentages of total DNA fragments with a length of 200 bp to 800 bp are labeled on the diagrams.

Using the same logic, we should ensure that the majority of the RNA fragments are between 200 and 800 nt in length prior to adding any adaptors. Given the ease with which RNA degrades, we compared situations in which we added RNase inhibitors to all preceding steps with situations in which we did not add any RNase inhibitors. Regardless of whether we added RNase inhibitors or not, the RNA was already substantially degraded after in-nuclei DNA digestion. Additionally, we noticed that further treatment of Rnase at this stage had no discernible effect on the RNA molecule sizes [figure 4.2.2 a]. For almost all of the frozen human brain samples, unfortunately, the RNA integrity level is already low at the beginning [figure 4.2.2 b]. Consequently, we eliminated the RNA fragmentation step from this process, although we still recommend adding RNase inhibitors and handling the material with care to avoid further RNA degradation.

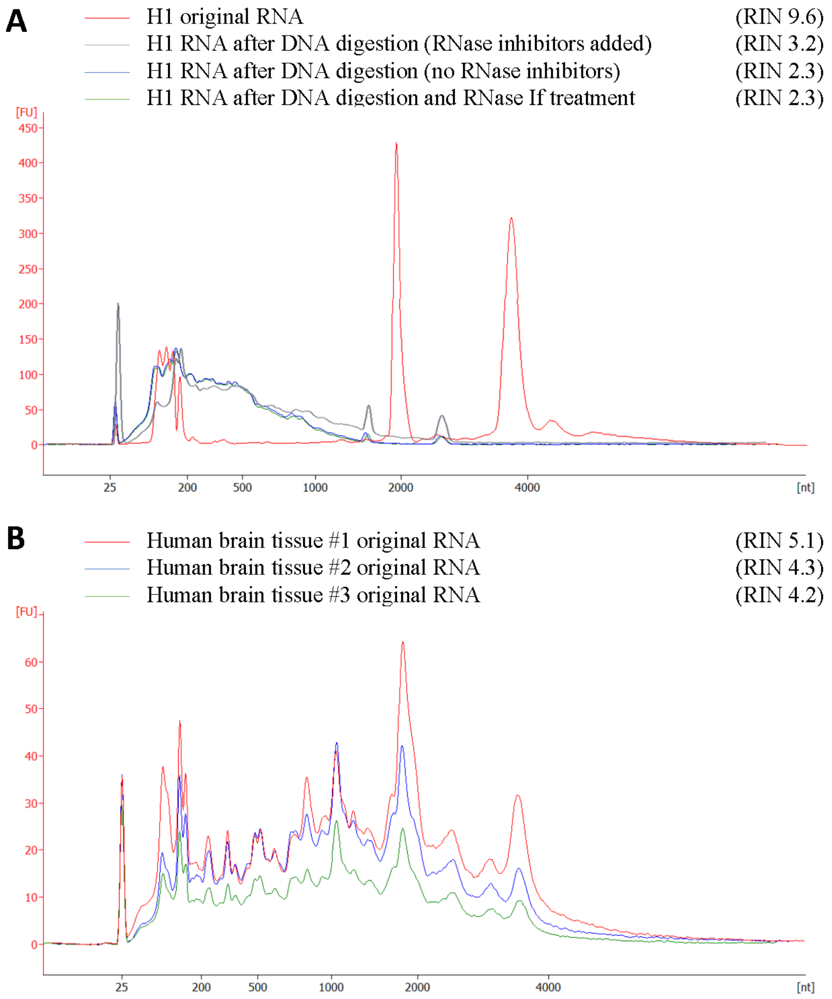


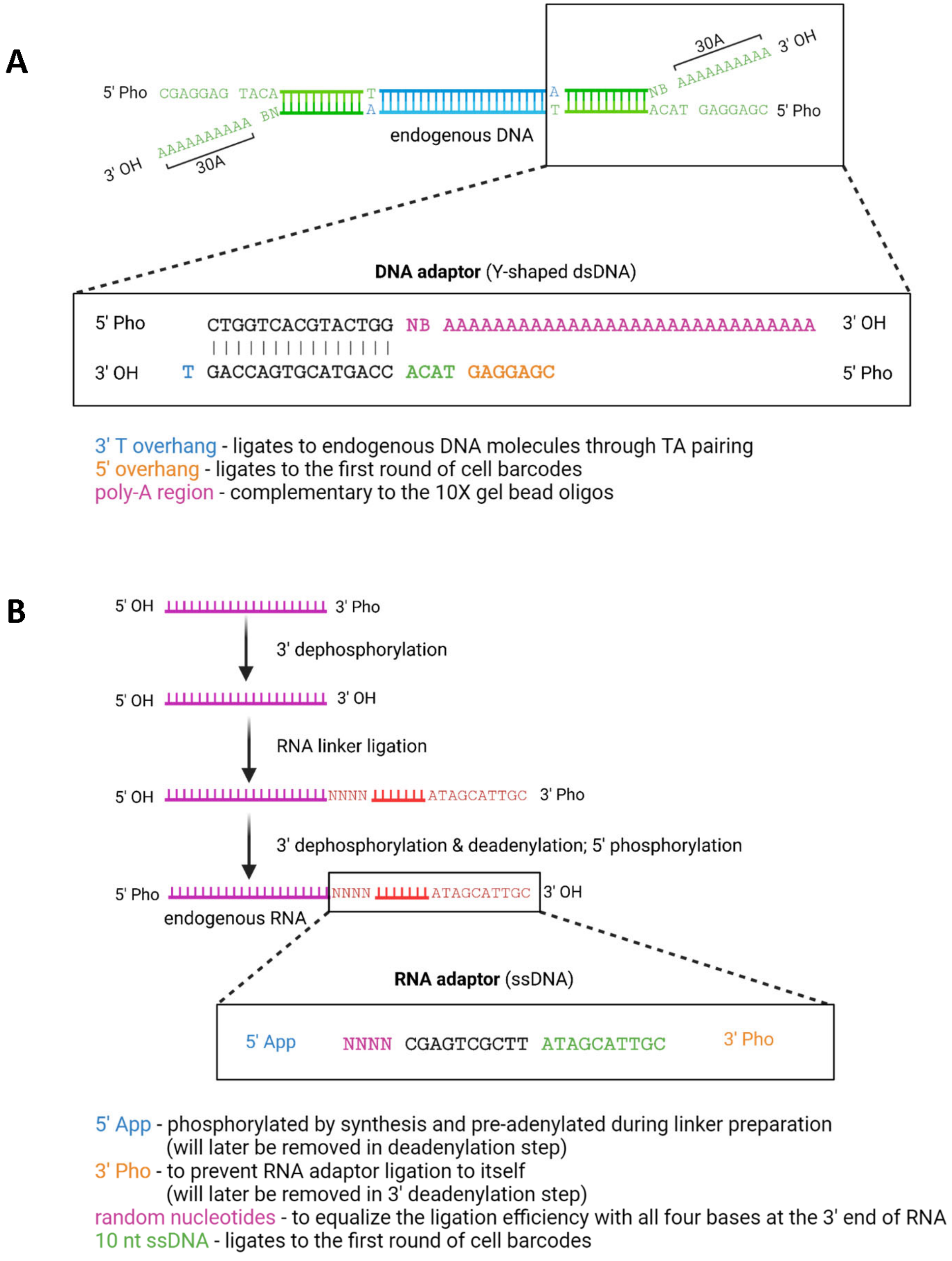
Figure 4.2.2: Comparison of RNA distribution patterns in original samples and under different treatment conditions

A) H1 cell lines are used to proceed until the in-nuclei DNA digestion step, with or without the addition of RNase inhibitors. The sample without the use of RNase inhibitors is further digested with 1 ul of 1:100 diluted RNase If enzyme. B) RNAs are extracted directly from three frozen human brain tissue samples chosen at random. Each RNA sample is extracted using Trizol and analyzed using Bioanalyzer. RIN: RNA integrity number measured by Bioanalyzer.

4.3 DNA and RNA preprocessing and adaptor ligation

To successfully add cell barcodes to both DNA and RNA ends in crosslinked complexes, DNA and RNA ends must first be ligated with specially designed adaptors.

T/A pairing is used to increase the efficacy of ligating DNA fragments with DNA adaptor molecules. Because we fragmented the DNA molecules with restriction enzymes that leave only blunt ends, we needed to add a 3' dA tail before adaptor ligation. The DNA linker was purpose-built to be a Y-shaped adaptor capable of being ligated to both ends of DNA molecules. The DNA adaptor's double-stranded portion would be ligated to fragmented DNA via TA pairing. With its poly-A sequence, the top single strand of the DNA adaptor is designed to be partially complementary to the 10X gel bead oligos, while the bottom single strand gets ligated to the first-round split-pool barcode. Using this Y-shape design, we can add both cell barcodes (split-pool barcodes) and cluster barcodes (10X gel bead oligos) to the DNA [figure 4.3.1 a]. Additionally, we verified that none of the sequences had potential restriction sites for the HpyCH4V and HaeIII enzymes.



3' T overhang - ligates to endogenous DNA molecules through TA pairing
5' overhang - ligates to the first round of cell barcodes
poly-A region - complementary to the 10X gel bead oligos

5' App - phosphorylated by synthesis and pre-adenylated during linker preparation
(will later be removed in deadenylation step)
3' Pho - to prevent RNA adaptor ligation to itself
(will later be removed in 3' deadenylation step)
random nucleotides - to equalize the ligation efficiency with all four bases at the 3' end of RNA
10 nt ssDNA - ligates to the first round of cell barcodes

Figure 4.3.1: Adaptor ligation steps for DNA and RNA
A) the structural and functional characteristics of the DNA adaptor molecule. B) the structure and function of the RNA adaptor molecule, as well as the procedures involved in preparing the RNA end.

Because cell and cluster barcodes cannot be added to DNA molecules without the adaptor, it is critical to maximize ligation efficiency at this stage. We evaluated a variety of ligation parameters and discovered that the ligation condition given in the SPRITE study (ligating with Instant Sticky-end Ligation Master Mix for 3 hours) produces the maximum yield, with approximately 30% efficiency. Notably, we observed that the DNA adaptors tend to self-ligate, resulting in a construct that can be ligated to barcodes but contains no endogenous DNA. In comparison to the Blunt/TA Master Mix, ligating with Instant Sticky-end Ligation Master Mix produces much fewer self-ligated adaptors (figure 4.3.2). Although they are artifacts, those self-ligated adaptors can be removed later via gel purification, so they will have little effect on the final library's quality.

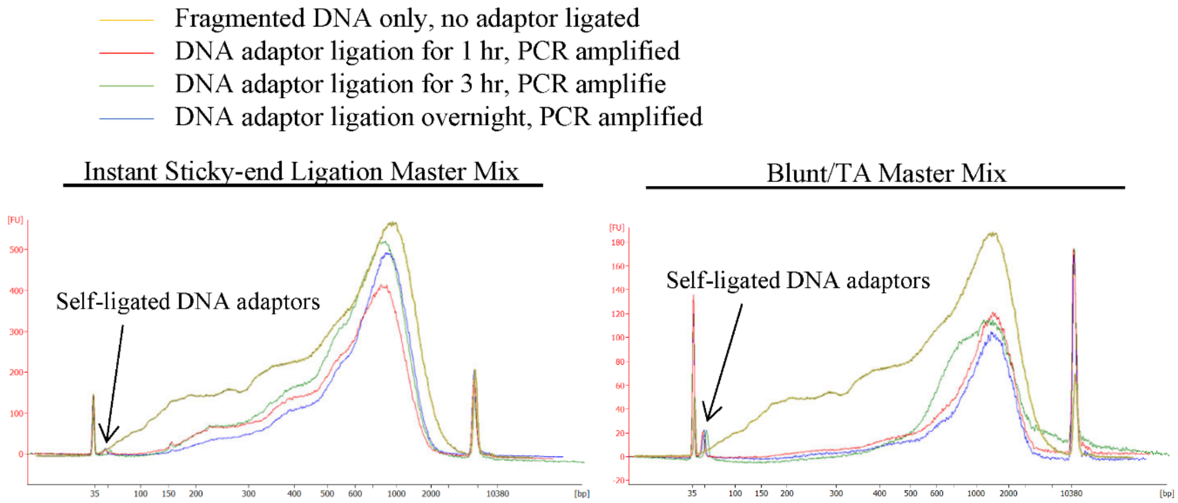


Figure 4.3.2: Comparison of DNA adaptor ligated products across different ligation enzymes and time durations
H1 cells were fragmented, and a portion of the DNA was analyzed without PCR amplification using Bioanalyzer After ligating the H1 cells with DNA adaptors under various conditions, they were amplified using primers complementary to the DNA adaptors. Bioanalyzer was used to evaluate the PCR amplified products. The self-ligated DNA adaptors (about 50 bp) are highlighted.

The RNA adaptor is a single-stranded DNA that ligates to the endogenous RNA on one end and to the first round of cell barcodes on the other [figure 4.3.1 b]. To increase the effectiveness of RNA adaptor ligation, we catalyzed this reaction with a mutant variant of T4 RNA Ligase 2 (T4 RNA Ligase 2, truncated KQ). Without the use of ATP, this enzyme ligates single-stranded RNA with 3' hydroxyl (3'-OH) to single-stranded RNA or DNA with a 5' adenylated group (5'-App), hence preventing ligation between endogenous RNAs. The 5' end of the RNA adaptor is phosphorylated during synthesis and pre-adenylated during the adaptor preparation stage to generate the 5'-App. Additionally, we added a protective 3' phosphate group during adaptor synthesis to prevent the adaptor from ligating to itself. Figure 4.3.1 b illustrates the exact design and layout of the RNA adapter.

Due to the fact that the RNA molecules have already been digested by endogenous RNase, leaving them with a 3' phosphate and a 5' hydroxyl group, we will begin with an RNA 3' dephosphorylation procedure. After ligating the RNA molecules with the ssDNA linker, we will use the T4 PNK enzyme to remove the protective 3' phosphate group. To avoid the unspecific ligation of excessive RNA adaptor oligos, we will additionally remove the 5' App group using the 5' deadenylase enzyme. Finally, because the RNase digested RNA molecules retain their 5' OH group, we will perform a 5' phosphorylation reaction to allow for oligos to ligate to the RNA 5' end. Figure 4.3.1 b illustrates the stages involved in RNA end preparation and linker ligation.

4.4 Split-pool cell barcode ligation

We used the combinatorial indexing method from the scSPRITE publication to uniquely barcode and differentiate single nuclei (Arrastia et al., 2021). To summarize, we distribute nuclei over a 96-well plate and ligate the complexes within the nuclei with a unique DNA and RNA barcode in each well. After pooling the nuclei and repeating the split-pool method with the second and third round barcodes, the number of possible barcode combinations ($96^3 = 884{,}736$) will surpass the nuclei count (5 to 10 thousand) by a factor of around 100. Assuming we use no more than 10,000 nuclei for downstream sequencing, the probability of only one nucleus being assigned to a single barcode combination is about 99 percent, hence minimizing the likelihood of two nuclei being ligated with the same cell barcode.

Using the scSPRITE barcode design as a starting point, we modified it to create our own cell barcodes. DNA and RNA cell barcodes are both constructed with a 7-nt ssDNA overhang and a 14-bp dsDNA random region. To prevent DNA and RNA barcodes from ligating to each other, DNA barcode duplexes have a 5' overhang, whereas RNA barcode duplexes have a 3' overhang. For more effective library creation, the bottom strand of the third round DNA barcode contains the Nextera Read 2 primer sequence, and the bottom strand of the third round RNA barcode contains a complementary sequence of the Customer Index 1 sequencing primer. Additionally, the top strand of the third round RNA barcode has a poly-A region that is complementary to the 10X gel bead oligos. The specific design rationales and layouts for the DNA and RNA cell barcodes are shown in figure 4.4.1.

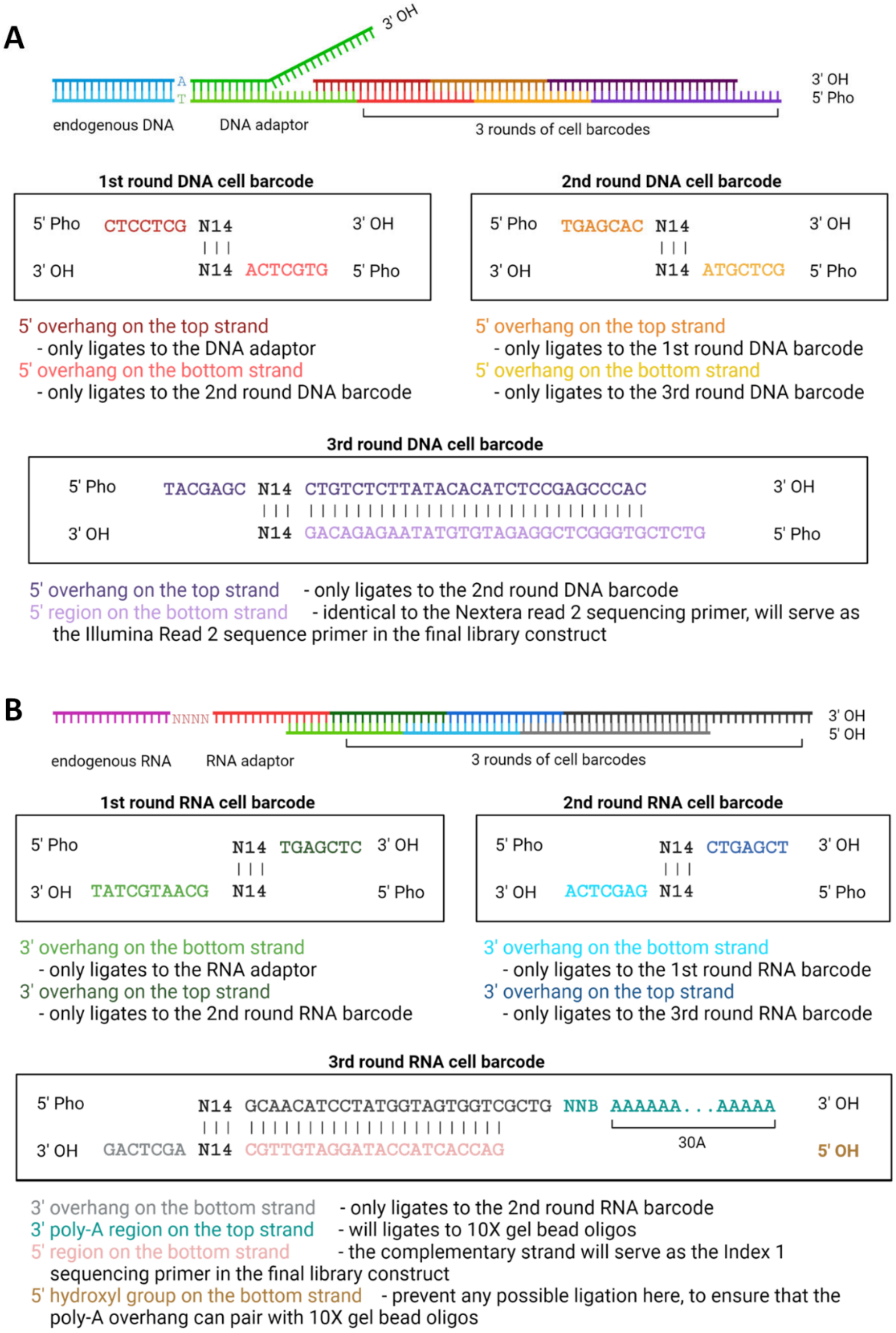

A
3' OH
A
T
3' OH
5' Pho
endogenous DNA
DNA adaptor
3 rounds of cell barcodes
1st round DNA cell barcode
5' Pho CTCCTCG N14 3' OH
3' OH N14 ACTCGTG 5' Pho
5' overhang on the top strand
- only ligates to the DNA adaptor
5' overhang on the bottom strand
- only ligates to the 2nd round DNA barcode
2nd round DNA cell barcode
5' Pho TGAGCAC N14 3' OH
3' OH N14 ATGCTCG 5' Pho
5' overhang on the top strand
- only ligates to the 1st round DNA barcode
5' overhang on the bottom strand
- only ligates to the 3rd round DNA barcode
3rd round DNA cell barcode
5' Pho TACGAGC N14 CTGTCTCTTATACACATCTCCGAGCCCAC 3' OH
3' OH N14 GACAGAGAATATGTGTAGAGGCTCGGGTGCTCTG 5' Pho
5' overhang on the top strand - only ligates to the 2nd round DNA barcode
5' region on the bottom strand - identical to the Nextera read 2 sequencing primer, will serve as the Illumina Read 2 sequence primer in the final library construct
B
NNNN
3' OH
5' OH
endogenous RNA
RNA adaptor
3 rounds of cell barcodes
1st round RNA cell barcode
5' Pho N14 TGAGCTC 3' OH
3' OH TATCGTAACG N14 5' Pho
3' overhang on the bottom strand
- only ligates to the RNA adaptor
3' overhang on the top strand
- only ligates to the 2nd round RNA barcode
2nd round RNA cell barcode
5' Pho N14 CTGAGCT 3' OH
3' OH ACTCGAG N14 5' Pho
3' overhang on the bottom strand
- only ligates to the 1st round RNA barcode
3' overhang on the top strand
- only ligates to the 3rd round RNA barcode
3rd round RNA cell barcode
5' Pho N14 GCAACATCCTATGGTAGTGGTCGCTG NNB AAAAAA...AAAAA 3' OH
30A
3' OH GACTCGA N14 CGTTGTAGGATACCATCACCAG 5' OH
3' overhang on the bottom strand - only ligates to the 2nd round RNA barcode
3' poly-A region on the top strand - will ligates to 10X gel bead oligos
5' region on the bottom strand - the complementary strand will serve as the Index 1 sequencing primer in the final library construct
5' hydroxyl group on the bottom strand - prevent any possible ligation here, to ensure that the poly-A overhang can pair with 10X gel bead oligos

Figure 4.4.1: Design of DNA and RNA cell barcodes

A) Illustration of one end of a DNA molecule ligated to a DNA adaptor and cell barcodes. B) Illustration of one end of an RNA molecule ligated to an RNA adaptor and cell barcodes. The boxes showcase the exact layout and function of each round of DNA and RNA cell barcodes.

Instead of spending time and resources generating and decoding barcode combinations, we used the 14-nucleotide barcode from the FREE barcode library provided by Hawkins (Hawkins et al., 2018). This 14-nt barcode library is capable of correcting a maximum of two substitution, insertion, and deletion errors. Additionally, they are built with a balanced GC content, low homopolymer runs, and a small potential for internal hairpins. After excluding barcodes that contained potential restriction sites, we selected 576 (96 x 6) of them for the double-stranded random section of our cell barcodes.

We examined multiple ligation enzymes and varying overhang lengths in barcode duplexes to decrease non-specific ligation products. We discovered that barcode duplexes with a 7 nt overhang produce the fewest non-specific ligation products *in vitro* under the enzyme condition described in the SPRITE study at 20C [figure 4.4.2 a]. To further optimize the ligation situation *in vivo*, we also investigated various ligation times and calculated the ligation efficiency in H1 cell lines. Across the three time periods tested, the ligated products looked comparable, and the ligation efficiency was nearly identical (about 55 percent) [figure 4.4.2 b]. Due to those findings, the optimal ligation time was determined to be one hour at 20C.

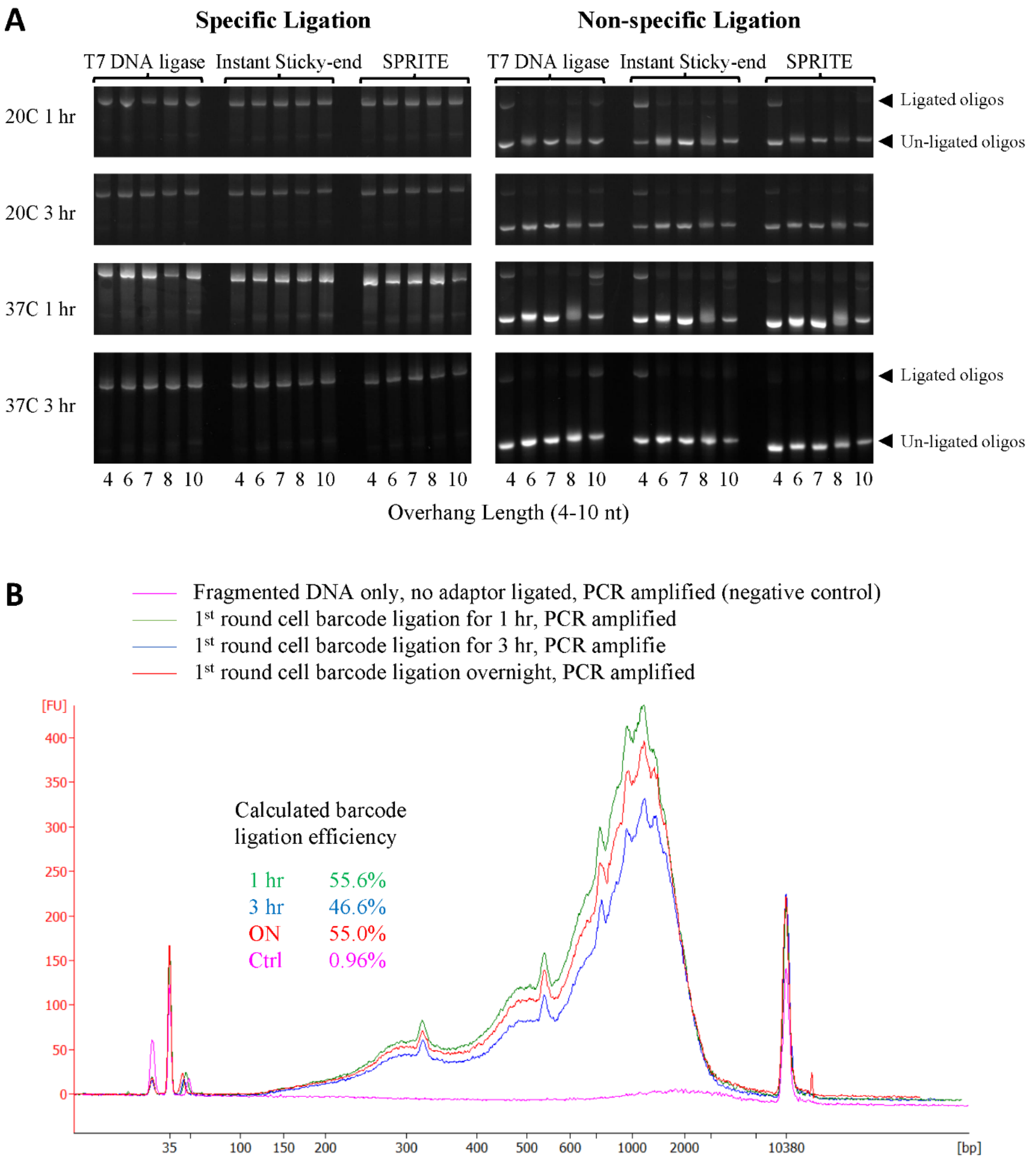


Figure 4.4.2: Optimization of cell barcode ligation conditions

A) *in vitro* ligation of the DNA adaptor and the first round of cell barcode under various ligation conditions. There are five different overhang lengths (4-10 nt) evaluated, three different ligation enzyme master mixes, and four different ligation time and temperature conditions. The left panel shows oligo ligation with matched oligos, whereas the right panel shows oligo ligation with unmatched oligos. The bands of ligated and un-ligated oligos are pointed out. B) H1 cells are fragmented and ligated to DNA adaptors, then ligated to only the first-round cell barcode for different time periods. The ligated product is amplified for eight cycles using primers specific to the cell barcode and then quantified using a bioanalyzer. The efficiency of barcode ligation is calculated.

4.5 Sonication

To remove probable nuclei clumps before adding the cluster barcode, we filtered nuclei to get isolated nuclei. Given that the total number of cell barcode combinations is around one million, we will take no more than ten thousand single nuclei to advance to the next phase.

Sonication will be used to break the nuclear membrane and release the crosslinked complexes from the nuclei. After testing various sonication strengths and times, we verified two conditions that can dissolve at least 80% of nuclei while avoiding fragmentation of barcoded nucleic acids [figure 4.5.1 a]. One condition is to use water as the sonication buffer and to sonicate for 5 minutes with a 5% duty factor. The other condition adds some ionic detergent to the buffer and sonicates for only 2 minutes at a decreased strength (3.3% Duty Factor). The recovery rate of adaptor ligated DNA molecules is almost identical between the two conditions and the positive control [figure 4.5.1 b]. As a result, we believe that adding detergent not only aids in the dissolution of the nuclear membrane, but also facilitates the fragmentation of the nucleic acids. Because the detergent in the sonication buffer will interfere with the emulsion formation process in the subsequent cluster barcoding stage, we recommend sonicating with only water. When using detergent to reduce sonication duration, it is required to do a detergent removal step immediately following sonication.

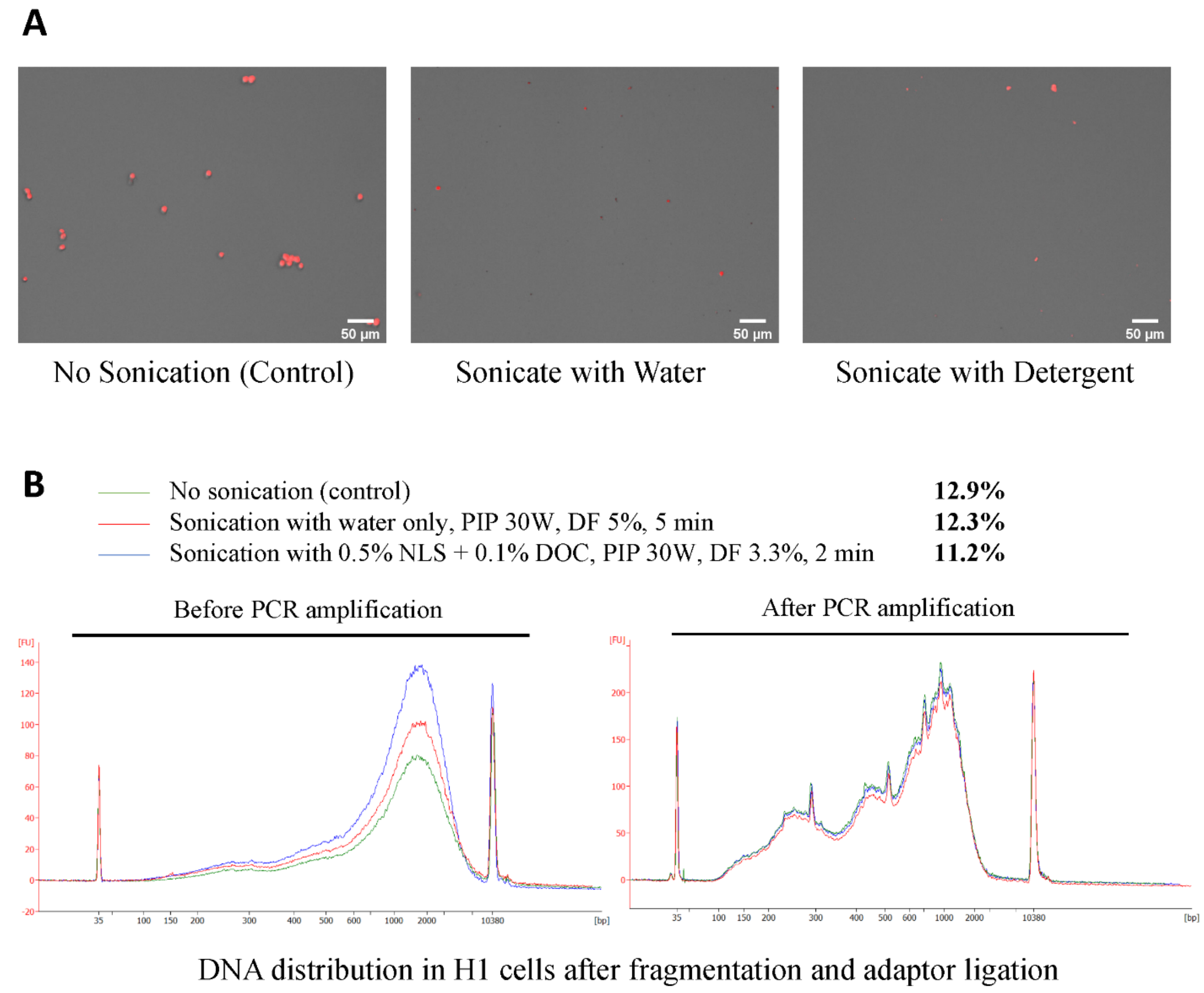


Figure 4.5.1: Two validated sonication conditions capable of dissolving at least 80% of nuclei without fragmenting barcoded nucleic acids

H1 cells are fragmented and ligated with DNA adaptors, followed by sonication under various conditions. A) image of residual nuclei following sonication treatment. B) the pattern of DNA size distribution following sonication treatment. The right panel shows the DNA products obtained after eight cycles of PCR amplification using primers complementary to the DNA adaptor. The percentage of total DNA that contains intact DNA adaptor constructs is determined and is shown by bolded numbers.

4.6 10X Genomics cluster barcode ligation and reverse transcription in microdroplets

Using the commercialized 10X Genomics platform, we adapted microdroplet technology to uniquely barcode and identify single complexes. Briefly, the complexes and 10X gel beads are delivered separately to capillary channels, generating millions of liquid-in-oil droplets, each containing a single gel bead. Following emulsion formation, the oligos contained within the gel bead will be released and hybridized to the barcoded DNA and RNA ends trapped within each droplet (Wang et al., 2020). Figure 4.6.1 illustrates the exact design of 10X gel bead oligos and the locations of their hybridization. The 10X barcode (N16) region will be used to differentiate single complexes, the unique molecular identifier (N14) region will be used to compensate for PCR bias, and the poly(dT)VN region will hybridize to the poly-A region in the DNA adaptor and the third round RNA barcode.

According to the 10X Genomics' protocol, the 10X barcode region is unique to each gel bead, and the Single Cell 3' v3.1 kit contains a pool of around 100,000 gel beads. Using the Poisson distribution model, the probability of only one complex being allocated to one unique 10X gel bead is over 99% if each single cell contains at most 1000 clusters (Arrastia et al., 2021). Therefore, the chance of DNA and RNA ends from two complexes being ligated with the same 10X barcode is negligible.

Following emulsion formation, reverse transcription will be carried out inside each microdroplet using 10X gel bead oligos as primers. After reverse transcription extension, the UMI and cluster barcodes from the 10X gel bead oligos will be integrated into the 3' end of the DNA adaptor [figure 4.6.1 b]. Additionally, the 10X gel beads hybridized to the RNA ends will integrate the cell barcode and the RNA adaptor. However, only part of the endogenous RNA can be integrated due to the steric hindrance created by covalent bindings in the DNA-RNA-protein complexes [figure 4.6.1 c]. After reverse transcription is complete, the microdroplets will be ruptured and the complexes extracted from the liquid-oil mixture. Excessive 10X gel bead oligos will be digested by Exo I and removed in later steps with size selection beads.

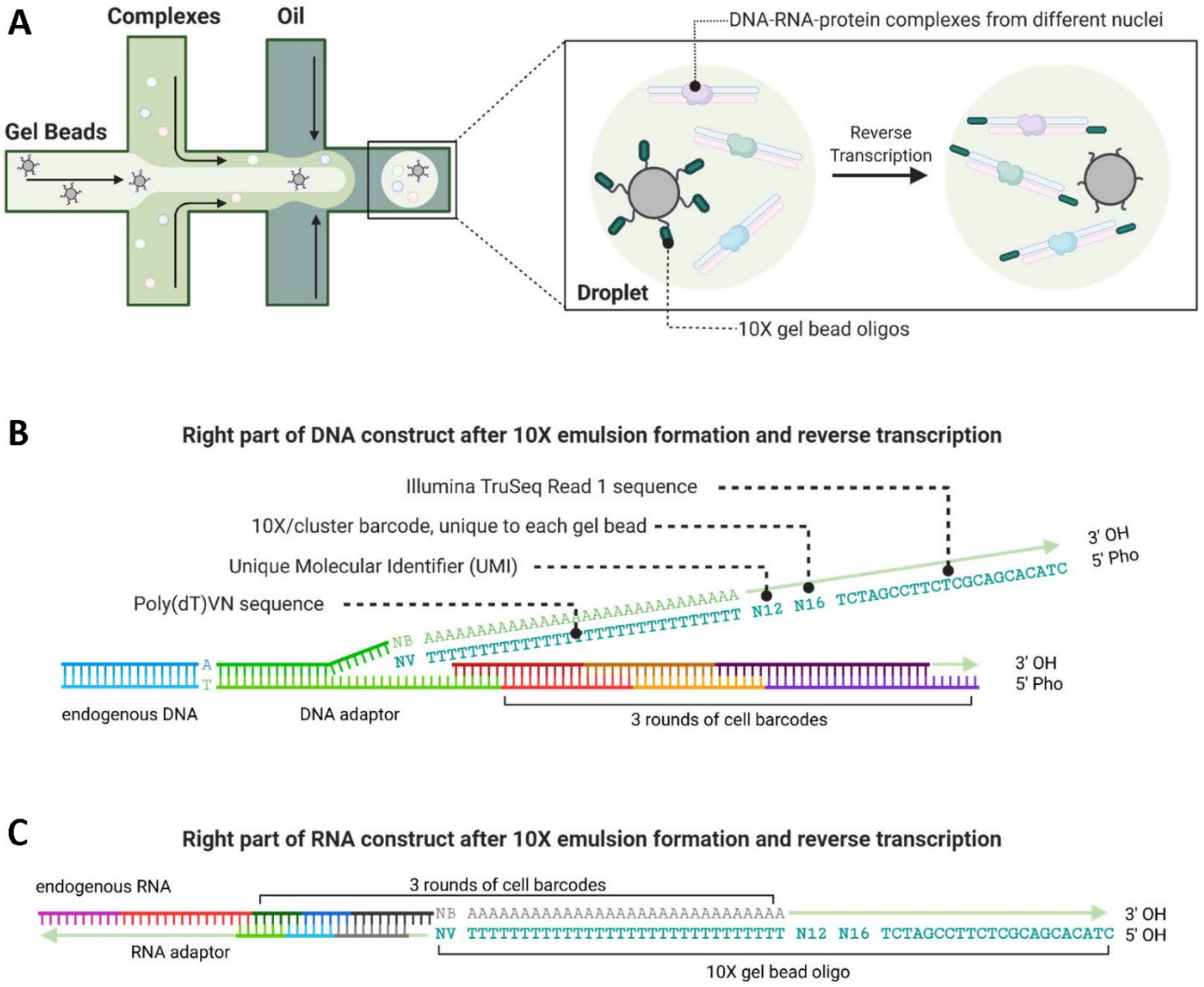

A
Complexes
Oil
Gel Beads
DNA-RNA-protein complexes from different nuclei
Reverse Transcription
Droplet
10X gel bead oligos
B
Right part of DNA construct after 10X emulsion formation and reverse transcription
Illumina TruSeq Read 1 sequence
10X/cluster barcode, unique to each gel bead
Unique Molecular Identifier (UMI)
Poly(dT)VN sequence
NB AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
NV TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT N12 N16 TCTAGCCTTCTCGCAGCACATC
3' OH
5' Pho
A
T
3' OH
5' Pho
endogenous DNA
DNA adaptor
3 rounds of cell barcodes
C
Right part of RNA construct after 10X emulsion formation and reverse transcription
endogenous RNA
3 rounds of cell barcodes
NB AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
NV TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT N12 N16 TCTAGCCTTCTCGCAGCACATC
3' OH
5' OH
RNA adaptor
10X gel bead oligo

Figure 4.6.1: Design of 10X gel bead oligos, hybridization locations, and reverse transcription procedures
A) overview of the microdroplet technique. B) DNA construct design after 10X gel bead oligo hybridization and reverse transcription. C) RNA construct design after 10X gel bead oligo hybridization and reverse transcription. Both the oligo hybridization and reverse transcription take place inside each microdroplet. The light green arrows in panels B and C indicate the reverse transcription positions.

To validate the efficiency of the 10X Genomics platform in barcoding complexes, we compared the lysate before and after the emulsion formation and reverse transcription procedures. The general DNA distribution pattern of the lysate remained unchanged, and a capture rate of 127 percent was calculated [figure 4.6.2]. This capture rate of greater than 100% could be explained by the production of cDNA during the reverse transcription reaction. Overall, those results imply that the 10X Genomics platform successfully captures and retrieves the majority of the complexes.

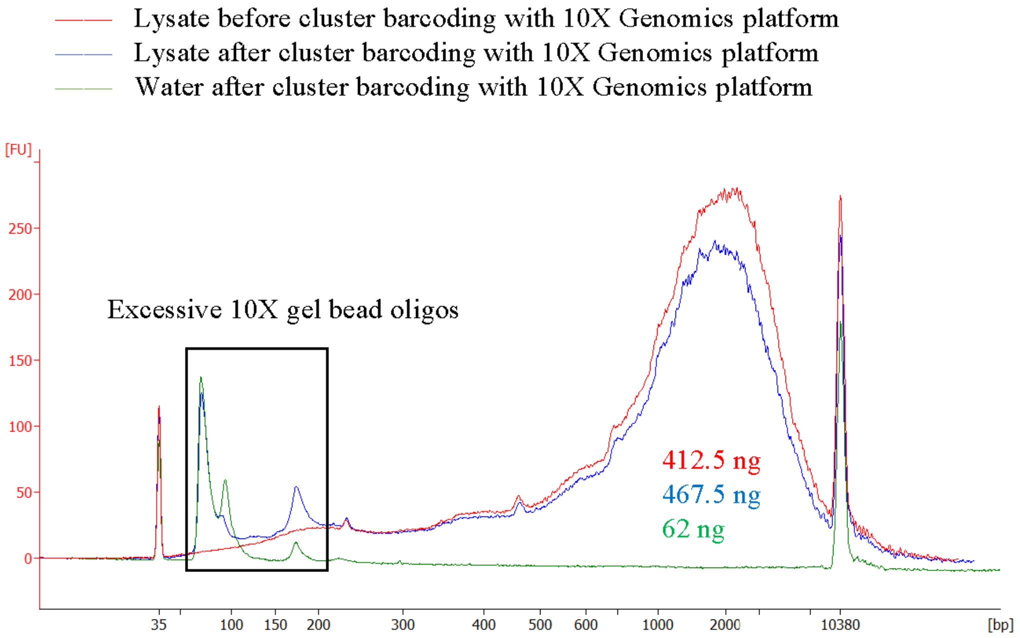


Figure 4.6.2: Validation of repurposing 10X Genomics to barcode complexes
Mouse embryonic stem cells were treated with sciMARGI procedures until the cluster barcoding step. The DNA distribution pattern in the lysate (which contains only barcoded complexes) was determined with Bioanalyzer before and after the cluster barcoding step. The negative control sample shows the size distribution of excessive 10X gel bead oligos (black box). The total amount of DNA in each sample is shown in colored numbers. The capture rate of complexes by 10X Genomics was calculated to be 127%.

4.7 Reverse crosslinking, template-switching reverse transcription, and PCR amplification

After adding the cell and cluster barcodes, we will break the covalent bonds used to crosslink the DNA-RNA-protein complexes, purify the nucleic acids, and complete the cDNA generation of RNA molecules. Finally, we will amplify the appropriate DNA and RNA constructions and incorporate sequencing adaptors into the libraries. Figures 4.7.1 and 4.7.2 depict the workflows for DNA and RNA library construction, respectively.

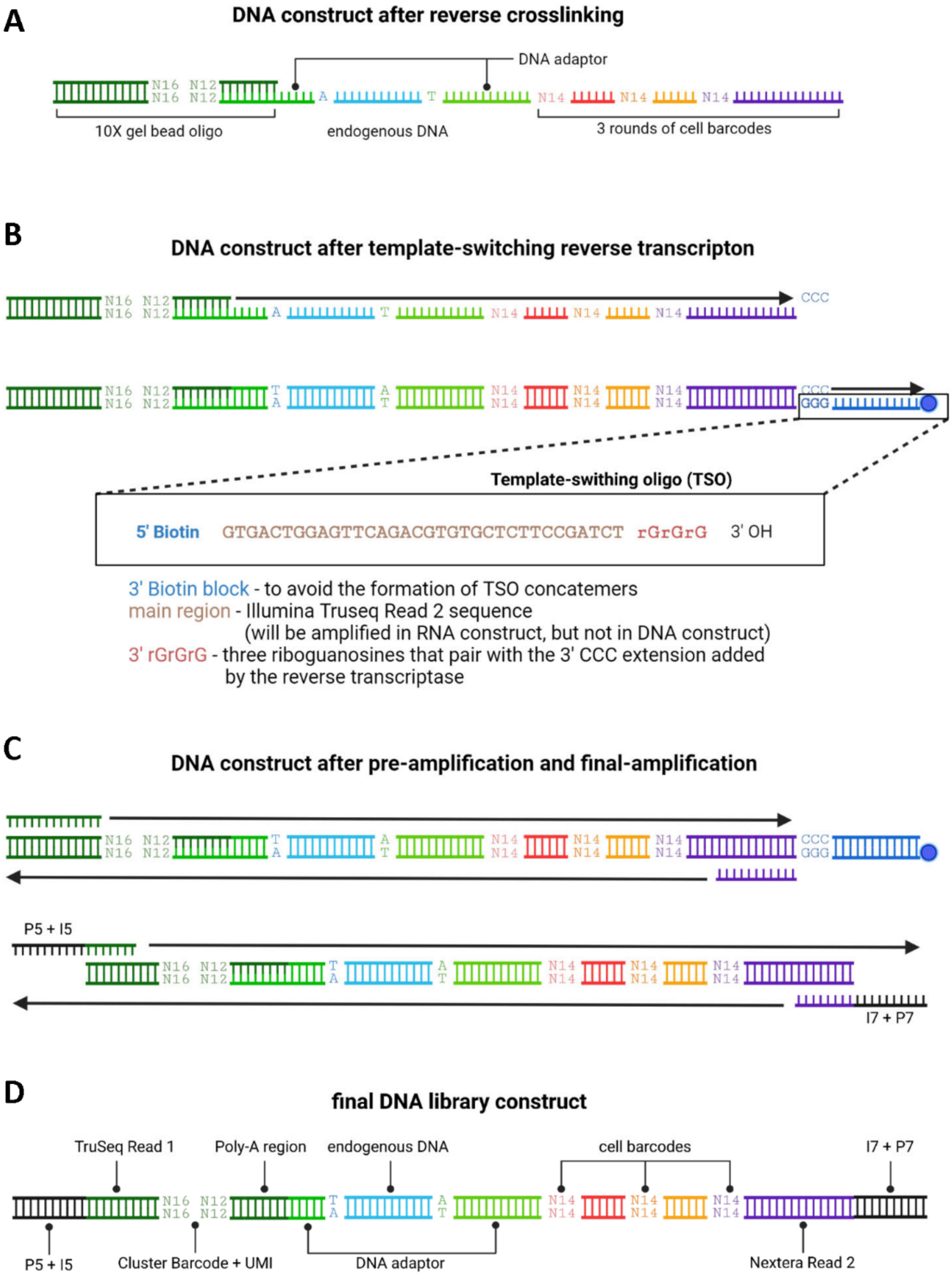

A
DNA construct after reverse crosslinking
DNA adaptor
N16 N12
N16 N12
A
T
N14
N14
N14
10X gel bead oligo
endogenous DNA
3 rounds of cell barcodes
B
DNA construct after template-switching reverse transcripton
CCC
GGG
Template-swithing oligo (TSO)
5' Biotin GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT rGrGrG 3' OH
3' Biotin block - to avoid the formation of TSO concatemers
main region - Illumina Truseq Read 2 sequence
(will be amplified in RNA construct, but not in DNA construct)
3' rGrGrG - three riboguanosines that pair with the 3' CCC extension added
by the reverse transcriptase
C
DNA construct after pre-amplification and final-amplification
P5 + I5
I7 + P7
D
final DNA library construct
TruSeq Read 1
Poly-A region
endogenous DNA
cell barcodes
I7 + P7
P5 + I5
Cluster Barcode + UMI
DNA adaptor
Nextera Read 2

Figure 4.7.1: DNA construct design of TS-RT, PCR amplification, and final library
A) Reverse crosslinked DNA construct. B) DNA construct after template-switching reverse transcription. The layout and functionality of template-switching oligo (TSO) are illustrated. C) DNA construct after pre-amplification and final-amplification. D) Final DNA library construct.

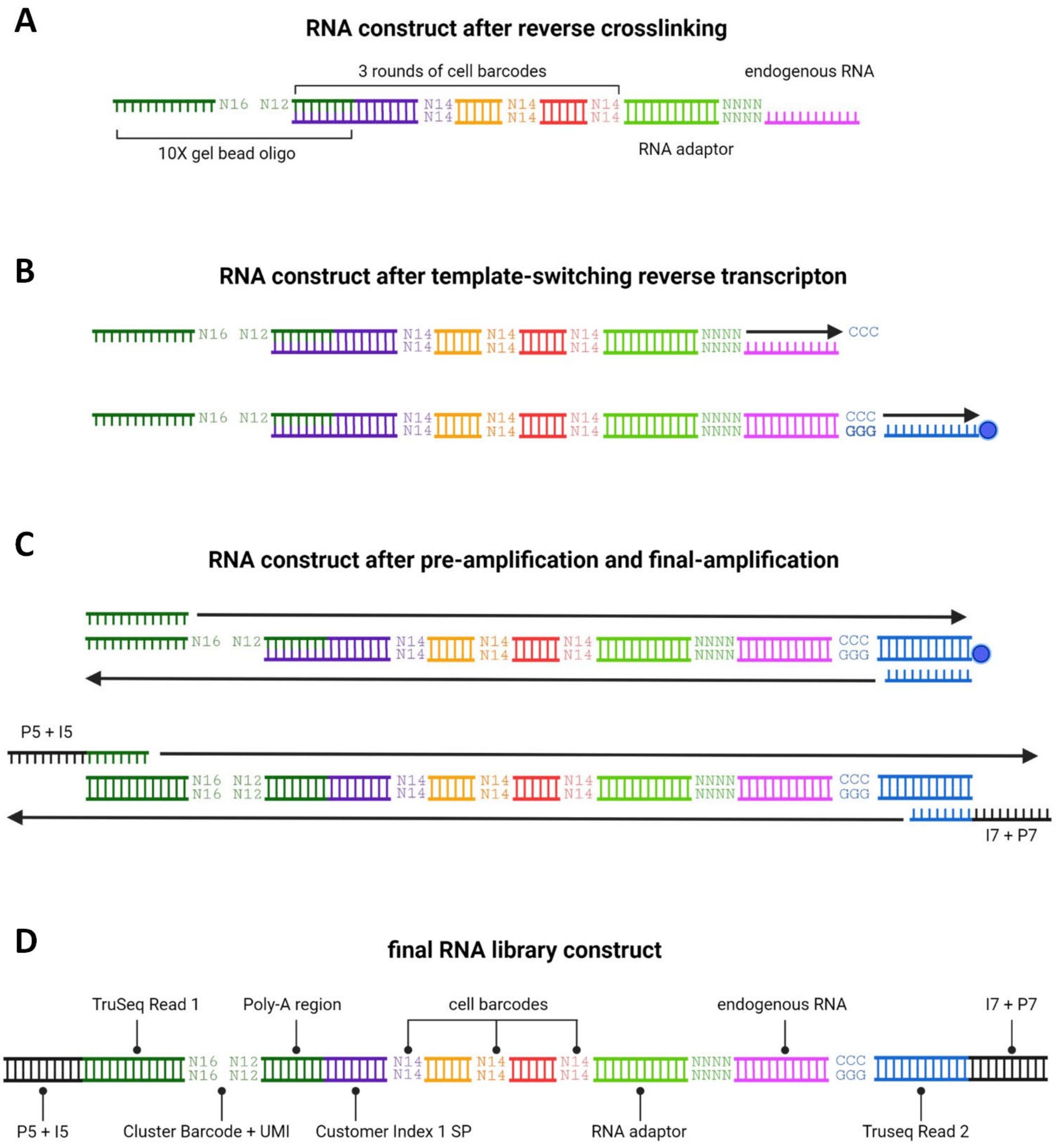


Figure 4.7.2: RNA construct design of TS-RT, PCR amplification, and final library
A) Reverse crosslinked RNA construct. B) RNA construct after template-switching reverse transcription. The layout and functionality of template-switching oligo (TSO) are identical to those shown in figure 4.7.1. C) RNA construct after pre-amplification and final-amplification. D) Final RNA library construct. SP: sequencing primer.

First, we reverse crosslink the complexes in order to remove the covalent bonds that impede the amplification process. Because RNA molecules are susceptible to degradation, we conducted tests to ensure that our reverse crosslinking condition does not result in a significant decrease in RNA quality [figure 4.7.3].

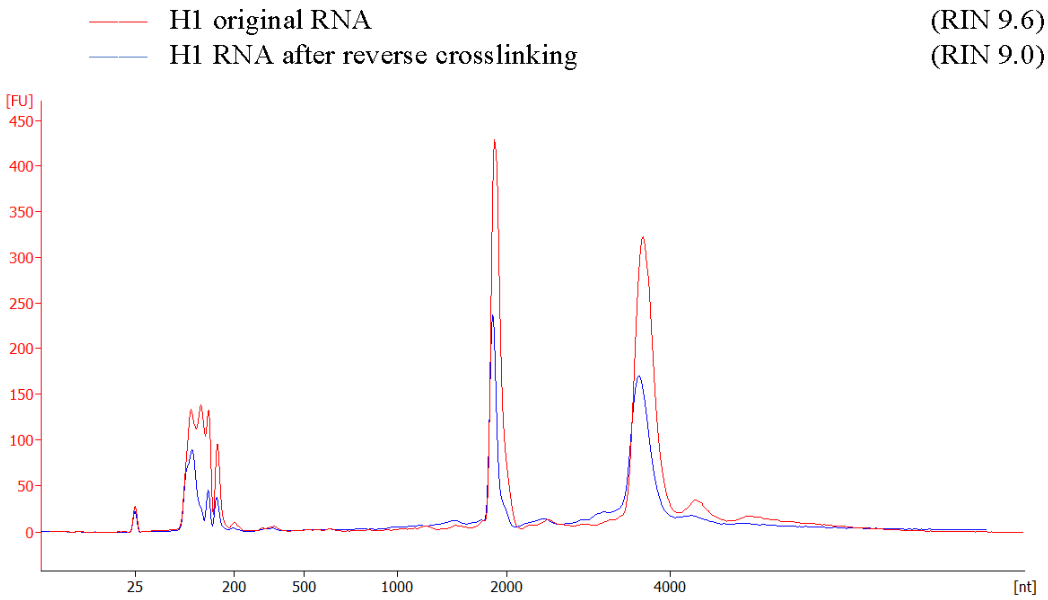


Figure 4.7.3: Reverse crosslinking has no significant effect on the integrity of RNA
Total RNA is isolated from H1 cells using Trizol and reverse crosslinked overnight. Bioanalyzer was used to assess the original RNA and reverse crosslinked RNA. RIN: RNA integrity number.

To complete the cDNA synthesis of barcoded RNA molecules, we employed a template-switching mechanism to complete the first and second strand synthesis procedures. Importantly, biotin is used to block the 5' end of the template-switching oligo (TSO) during synthesis, preventing the formation of oligo concatemers during reverse transcription [figure 4.7.4 a]. The complementary strands of barcoded DNA and RNA molecules will be synthesized using Maxima-H reverse transcriptase and the TSO appended to the 5'-end of the template strands [figure 4.7.1 b and figure 4.7.2 b].

Following the generation of the entire cDNA construct for the RNA library, we will amplify the appropriate DNA and RNA constructs. To begin with, we divide our library into two aliquots: one for DNA and one for RNA. Then we perform pre-amplification for approximately seven to ten PCR cycles for DNA and RNA libraries, respectively. The TSO region of the RNA construct is used as one of the amplification primers, whereas the TSO region of the DNA construct is left unamplified. Finally, we will add the P5 and P7 sequencing adaptors to the libraries using a 10- to 15-cycle final amplification [figure 4.7.1 c, figure 4.7.2 c].

To assure the quality of the final library constructs, we strongly recommend performing a Bioanalyzer check on the library size distribution prior to sequencing. Figure 4.7.4 b provides an example of the libraries before and after amplification. Notably, the final DNA usually contains a peak of approximately 250 bp, which is an artifact created by the self-ligated DNA adaptors. Before sequencing, we recommend adding a gel purification step to select only libraries between the 350 bp and 1500 bp range to increase the portion of meaningful constructs in the final libraries.

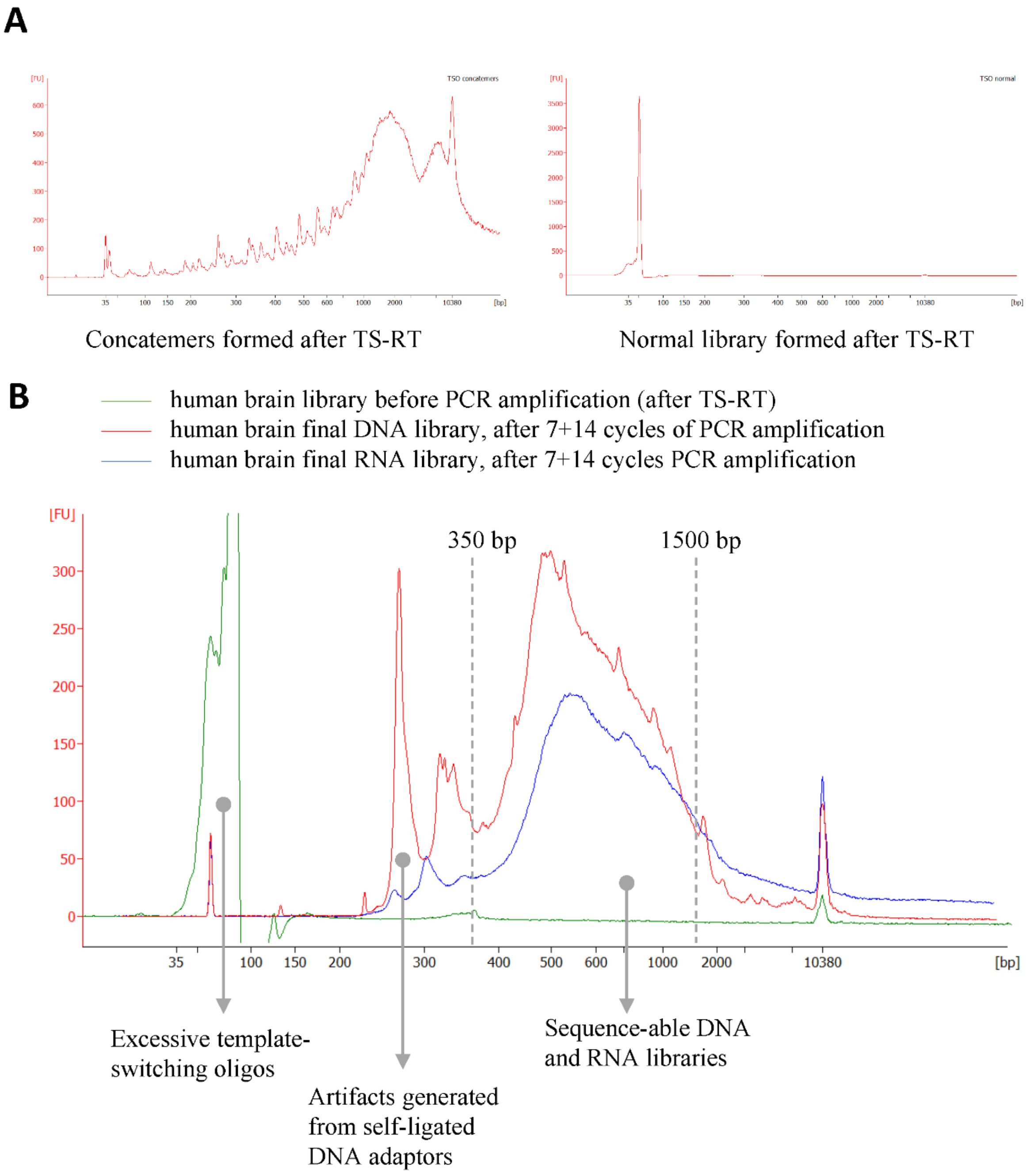
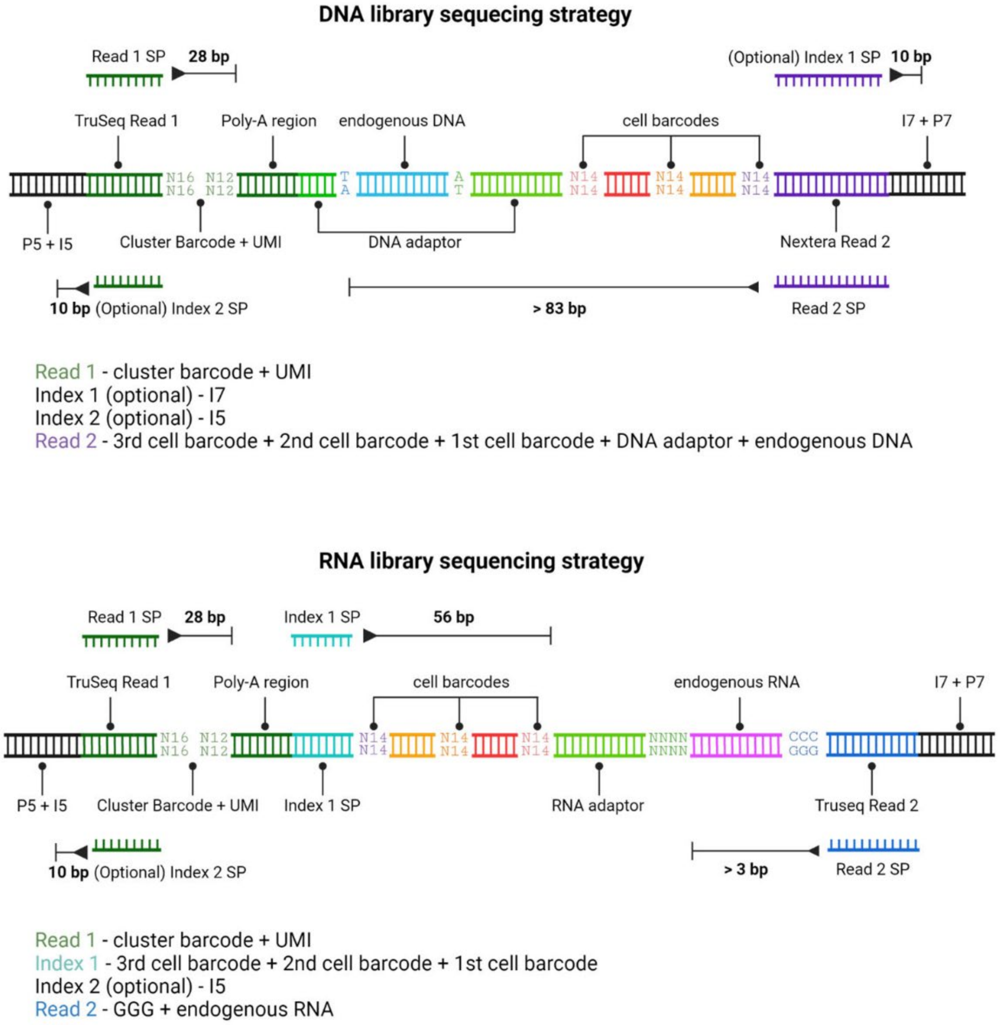


Figure 4.7.4: Examples of sciMARGI library products before and after amplification
A) comparison of products after the TS-RT step using the oligos with and without the 5’ biotin block. SciMARGI protocols were used to treat HEK cells until the template-switching step, and then TS-RT was performed separately using oligos with and without the 5' biotin block. On the left are the library products generated using the oligo without the 5' biotin block, while on the right are the library products generated using the oligo with the 5' biotin block. B) An example of human brain libraries before and after amplification. The labeled artifacts can be removed by adding a gel purification step. Only libraries with a size range of 350 to 1500 bp can be effectively sequenced.

4.8 Library sequencing

Finally, through pair-end sequencing, we will read the DNA and RNA libraries separately. Due to the exceedingly low diversity of the poly-A/T region, reading through it would have a detrimental effect on the base call quality of the succeeding sequences, and hence we must avoid reading through it. To extract information from the cluster barcode, the cell barcode, and endogenous nucleic acids, we developed distinct sequencing strategies for DNA and RNA libraries, as illustrated in figure 4.8.1.


DNA library sequecing strategy

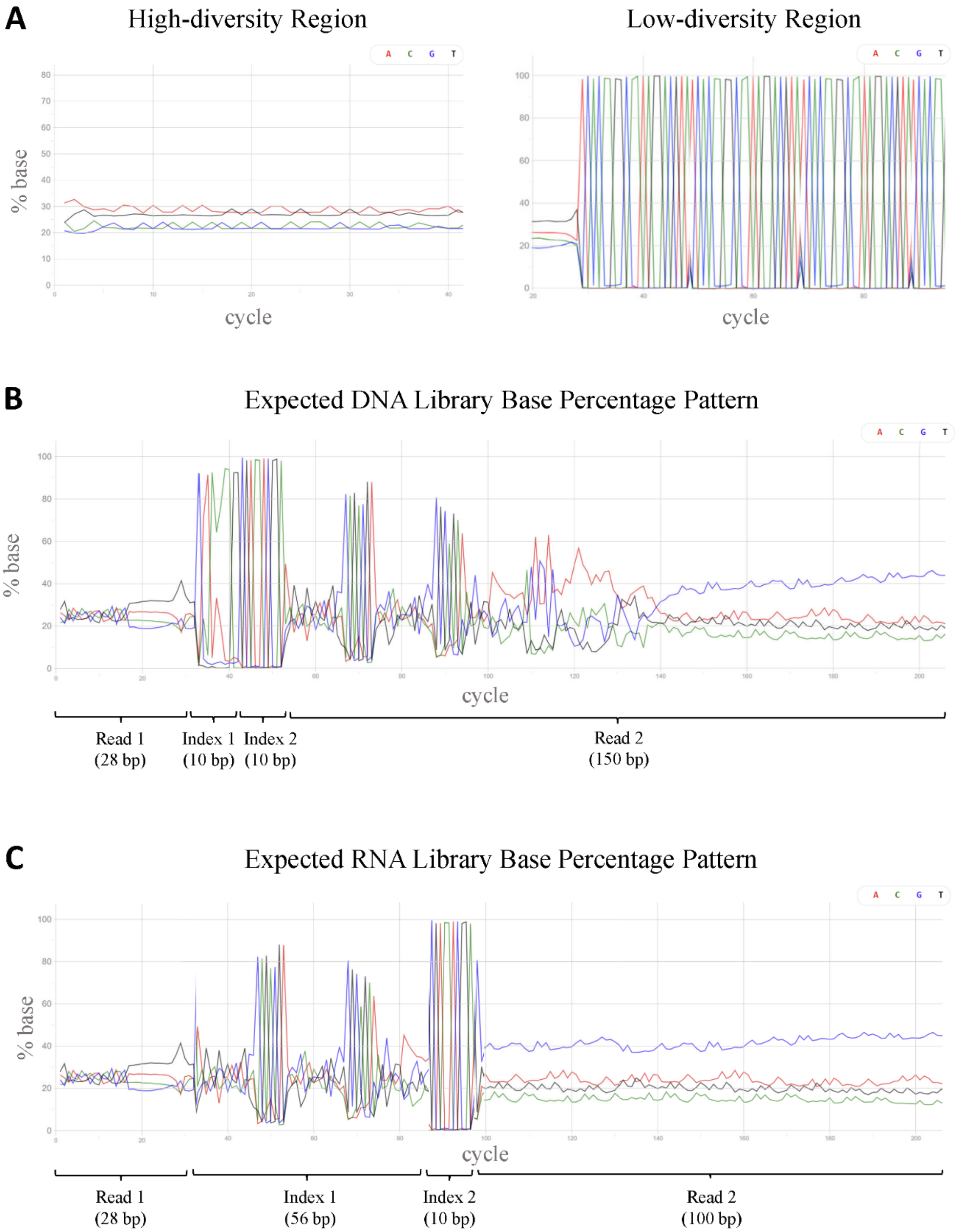
RNA library sequencing strategy

Figure 4.8.1: DNA and RNA library sequencing strategies
The lengths of each read are shown in bolded numbers (excluding the uncertain length of endogenous nucleic acids). SP: sequencing primer.

The read 1 sequencing primer for the DNA library would hybridize to the TruSeq Read 1 region added by the 10X Genomics platform and read through the cluster barcode and UMI. The read 2 sequencing primer would bind to the Nextera Read 2 region introduced during the third round of cell barcode ligation. It would then read through the three rounds of cell barcodes, the DNA adaptor, and lastly the endogenous DNA. Additionally, the read 2 sequence must be longer than 83 bp in order to read into endogenous DNA. The index 1 and index 2 sequencing primers are optional but may be included if several libraries need to be sequenced simultaneously.

The read 1 sequencing strategy for the RNA library is identical to that for the DNA library. The read 2 sequencing primer, on the other hand, would hybridize to the TruSeq Read 2 region introduced during the template switching step and read directly into the endogenous RNA sequence. Due to the varying length of endogenous RNA, we used a customer Index 1 sequencing primer to hybridize to the Nextera Read 2 area near the 3rd round RNA cell barcode. This index 1 read enables us to collect the information from all the cell barcodes. Again, the index 2 sequencing primer is optional and may be included for the purpose of multiplexing.

Due to the varied levels of diversity inside a DNA or RNA construct, we expect specific base percentage patterns to indicate that our library builds are valid. For example, comparing different DNA sequences, the 14-nt random cell barcodes are highly diverse, whereas the 7-nt linkers between the cell barcodes have a low degree of diversity. Thus, we can expect to see a significant difference between those two types of regions in the percent base graphs [figure 4.8.2 a]. According to this rationale, the predicted base percentage for the proper DNA construct should be as shown in figure 4.8.2 b, whereas that for the correct RNA construct should be as shown in figure 4.8.2 c. By comparing the sequencing results to the anticipated patterns, we could get a sense of whether our libraries mostly contain the correct structures.


A High-diversity Region

Low-diversity Region

B Expected DNA Library Base Percentage Pattern

C Expected RNA Library Base Percentage Pattern

Figure 4.8.2: Theoretical base percentage patterns under various circumstances
A) comparison of percent base patterns from high-diversity and low-diversity regions. B) the expected percent base pattern for a DNA library containing mostly correct constructs. The image was taken directly from our actual results. C) the estimated percent base pattern for an RNA library with mostly correct constructs. The image has been recreated and is not a true representation of the result. Neither the DNA nor the RNA libraries are multiplexed.

5. Results of the current sciMARGI design

5.1 Overview of the current results

Due to the extensive optimization work that has been conducted, we have only been able to build a complete DNA and RNA library for a human brain tissue sample using the current sciMARGI design. When we compared the sequencing data to the predicted base percentage patterns, we discovered that the DNA library was valid, but the RNA library had abnormalities. By evaluating the sequencing readouts, we were able to confirm the DNA library and identify ways to improve the RNA library design. We then evaluated the cell barcodes and cluster barcodes from the DNA side, demonstrating that our technique is capable of constructing a single-cell DNA-DNA interaction map.

Dr. Zhifei Luo and Xingzhao Wen handled all of the informatics processing in this part.

5.2 Positions of barcodes in read 2 sequences confirm the successful construction of DNA library

We evaluated the 150 bp of read 2 sequences to determine which proportion of our DNA library has the correct construction as designed. We searched for all DNA and RNA barcode (N14) sequences, as well as adaptor sequences, and plotted the read counts of the sequences that were 100 percent matched against their starting positions [figure 5.2.1].

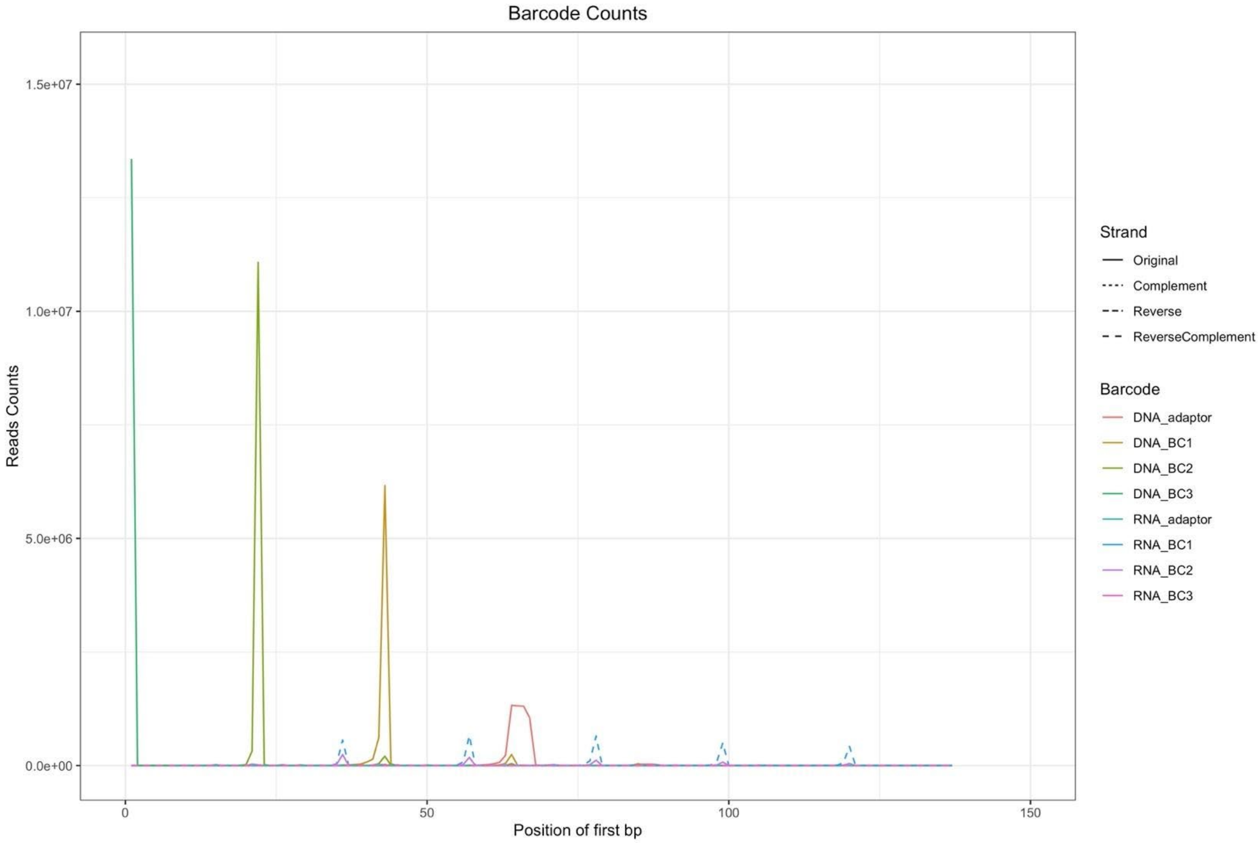


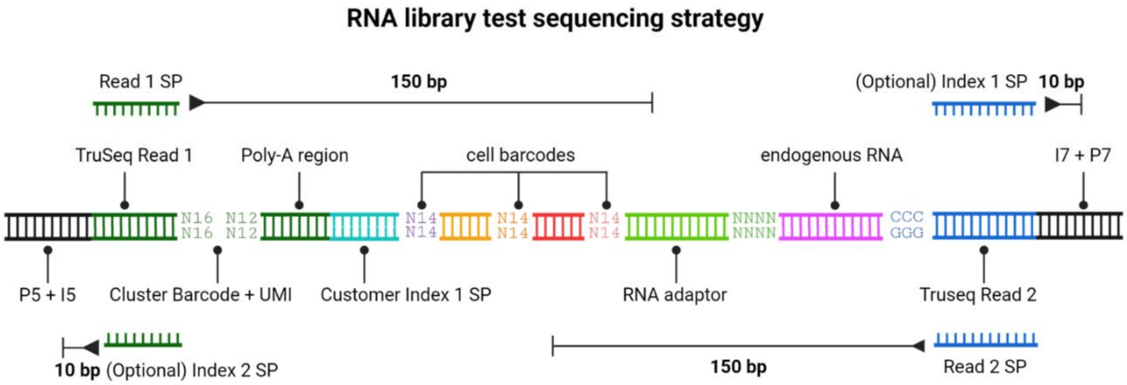
Figure 5.2.1: Counts and positions of all barcodes in the DNA library
All DNA and RNA barcode and adaptor sequences are searched, including the original sequence (the sequences that are supposed to be read), complementary sequences, reverse sequences, and reverse complementary sequences. The read counts are shown against the starting positions of the sequences that are 100 percent matched. The Y axis represents the number of read counts scaled to the total number of reads. The X axis represents the location of the initial base pair in a matched sequence.

Ideally, DNA Barcode 3 (DNA BC3), BC2, and BC1 should begin at 1, 22, and 43 bp, respectively, and the DNA adaptor should begin at 64 bp. As shown, the plotted barcode starting positions closely match the theoretical expectations, demonstrating that the majority of DNA libraries are valid and meaningful. There is an increase in the starting position dispersion for each matched sequence from the start to the end of the read, which could be attributable to phasing and pre-phasing errors that occur naturally during the sequence run. There is also a decrease in the read counts of the matched sequences, which is understandable given that the ligation efficiency is not 100%, thus the oligos that are ligated later in the protocol would have a greater chance of being included in the sequencing library. Some RNA barcode regions can also be identified in this DNA library. However, because they accounted for less than 5% of total reads, the RNA contamination would have little effect on the DNA libraries and could be eliminated later during informatics processing.

In general, the sequencing library is primarily composed of correct DNA constructs, signifying that we could successfully acquire information through the cluster barcodes from Read 1 and the cell barcodes from Read 2.

5.3 RNA library sequencing results reveal directions for improving the sciMARGI design

After comparing the base percentage graphs of the RNA sequencing results to the expected pattern, we found substantial abnormalities. To determine the structures of the major artifacts, we sequenced the RNA library again using a different test strategy [figure 5.3.1]. Because we are unsure whether the Nextera Read 2 region is in the correct position, we removed the customer index 1 sequencing primer and increased the read lengths of both read 1 and read 2 to 150 bp. To force the read 1 sequence through the poly-A/T region, we spiked in 80% of the PhiX control. Although read 1 successfully sequenced the cluster barcode and UMI, the base call quality declines dramatically after the poly-A/T region, thus we used read 2 instead of read 1 to extract information about the artifact structures in the following analysis.



Read 1 - cluster barcode + UMI + poly-A + customer index 1 SP + cell barcode + RNA adaptor
Read 2 - GGG + endogenous RNA + RNA adaptor + cell barcode

Figure 5.3.1: Sequencing strategy to troubleshoot the RNA library
SP: sequencing primer

To determine the structures of valid constructs and artifacts, we searched all DNA and RNA barcodes and adaptors and plotted the read counts of the sequences that were 100 percent matched against their starting positions, just as we did to the DNA sequencing library.

Additionally, we searched the last 14 bases of the Customer Index 1 sequencing primer and the final 14 bases of the Nextera Read 2 sequencing primer [figure 5.3.2]. Ideally, read 2 should begin with 3Gs and be followed by RNA inserts of varying lengths; it will then read the RNA adaptor followed by RNA barcodes in reverse complementary orientation. Apart from those, none of the sequences, including the Nextera Read 2 sequencing primer, are predicted to exist. Surprisingly, there are a significant number of barcodes that are enriched at specific loci, which implies that the final RNA library contains a substantial proportion of artifacts that lack the correct customer primer binding sites.

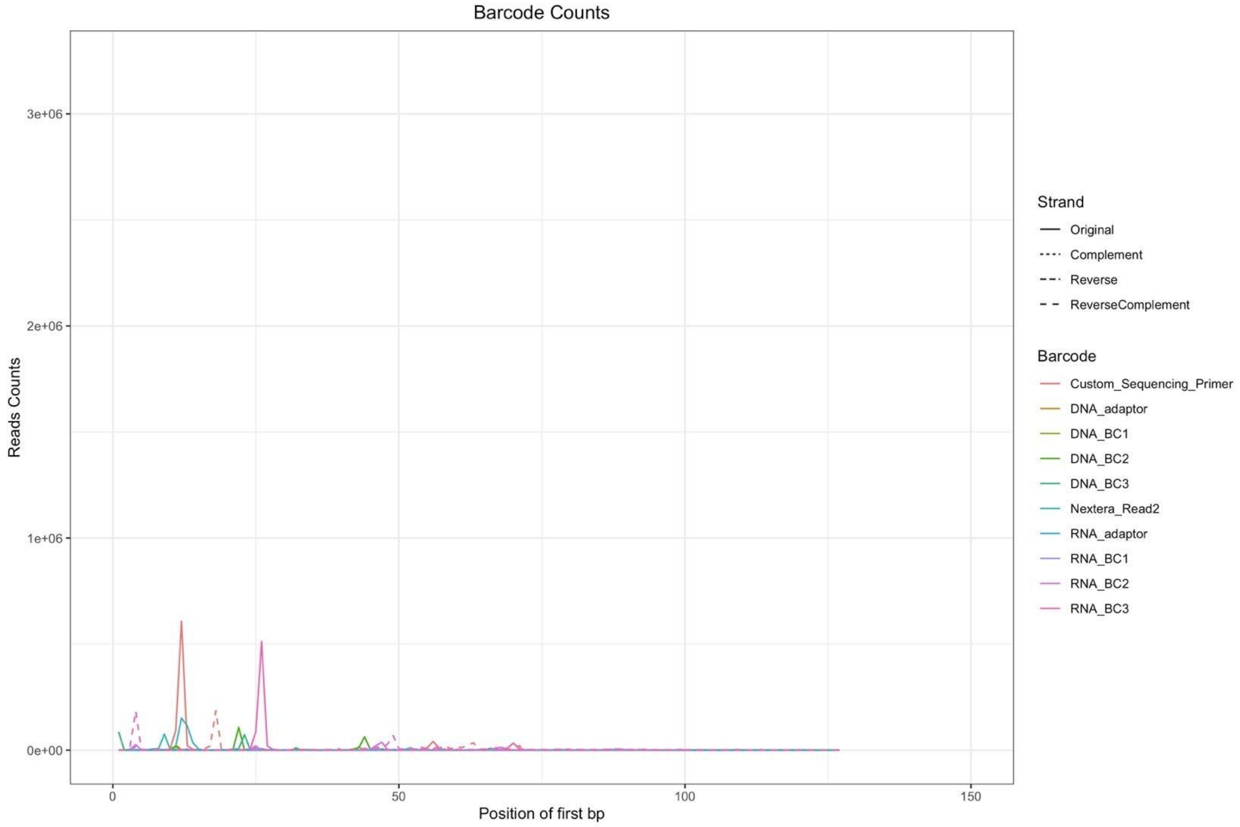


Figure 5.3.2: Counts and positions of all barcodes in the RNA library
All DNA and RNA barcode sequences are searched, as well as adaptor sequences, customer index 1 sequencing primer, and Nextera read 2 sequencing primer, including the original sequence (the sequences that are supposed to be read), complementary sequences, reverse sequences, and reverse complementary sequences. The read counts are shown against the starting positions of the sequences that are 100 percent matched. The Y axis represents the number of read counts scaled to the total number of reads. The X axis represents the location of the initial base pair in a matched sequence. Notably, this library comprises 80% of PhiX, which does not contain any matching sequences.

To assess the artifacts, we first investigated the enrichment pattern for RNA library-specific oligos and discovered that there are primarily two types of artifacts [figure 5.3.3]. One type of artifact begins with the reverse complementary strand of the 3rd round RNA cell barcode at 4 or 47 bp, followed by the reverse complementary strand of the customer index 1 sequence primer. We hypothesize that the artifact structures resemble those depicted in Figure 5.3.4 a and b, in which the template-switching oligos are introduced to the third round of RNA cell barcoding and get amplified. The other form of artifact begins with the original strand of the customer index 1 sequence primer at 11 or 56 bp, followed by three rounds of RNA cell barcodes. Figure 5.3.4 c and d illustrate the possible artifact structures, where the template-switching oligo is introduced to the other end of the 3rd round RNA cell barcode.

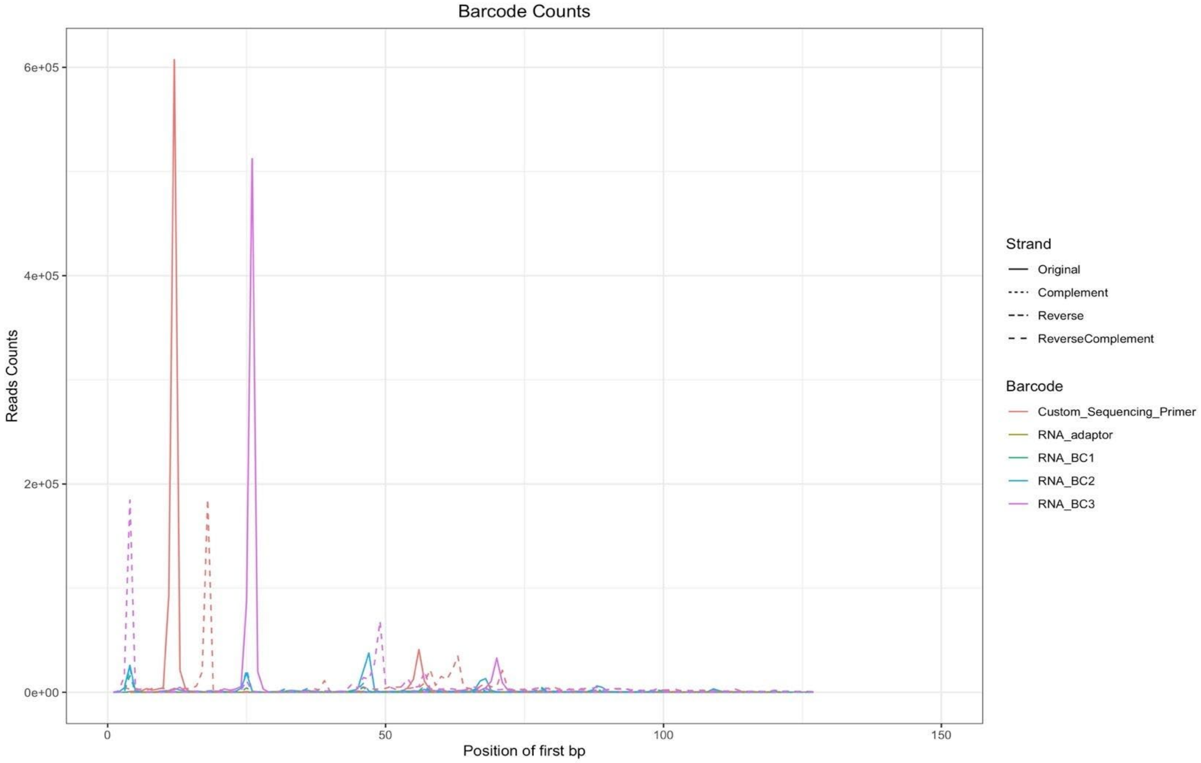
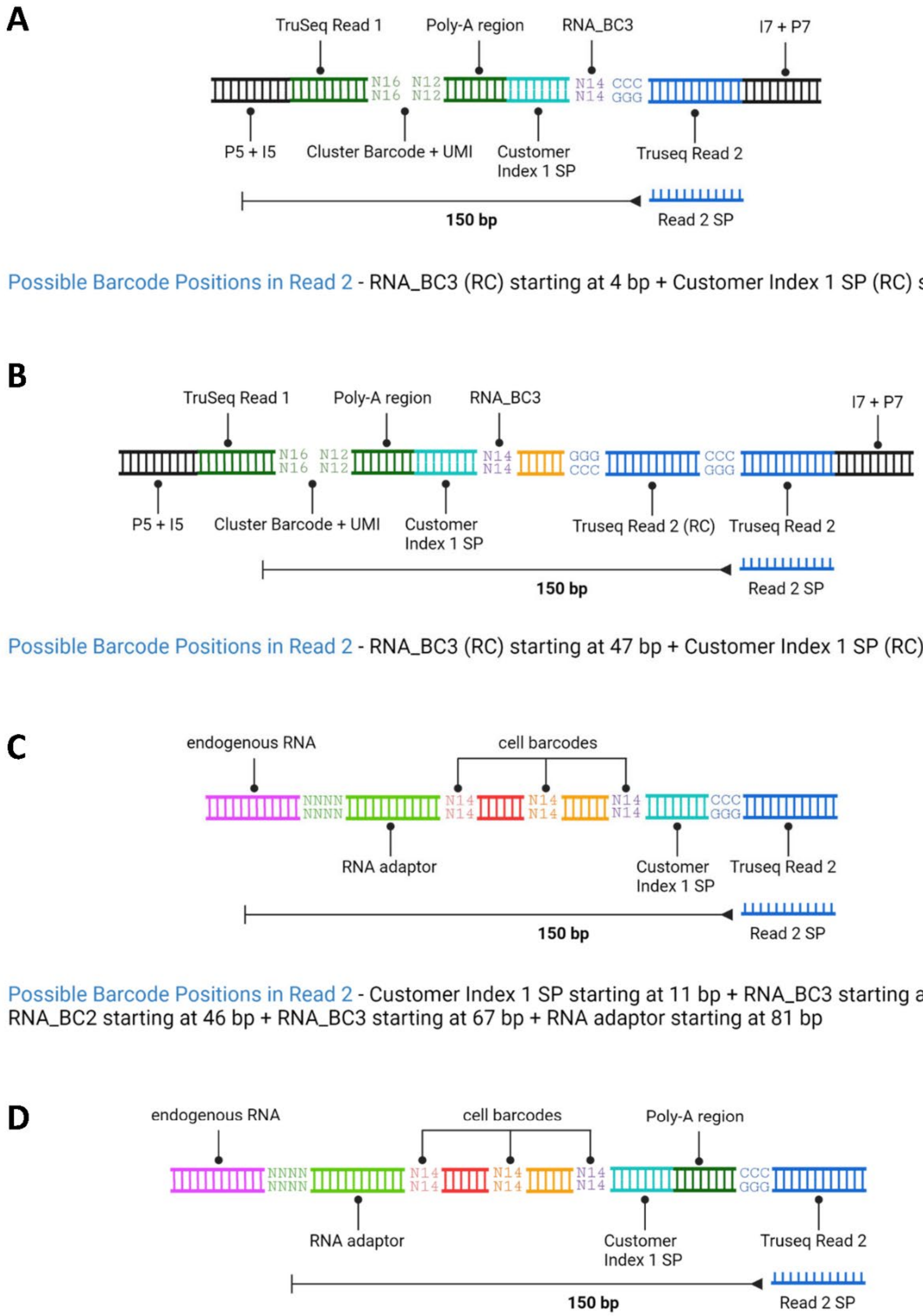


Figure 5.3.3: RNA-specific barcode positions in the RNA library sequencing result
Only oligos specific to the RNA library are searched, including the RNA cell barcodes, RNA adapter, and customer index 1 sequencing primer. The read counts are shown against the starting positions of the sequences that are 100 percent matched. The Y axis shows the number of read counts but is not scaled to the total reads. The X axis represents the location of the initial base pair in a matched sequence.



Possible Barcode Positions in Read 2 - RNA_BC3 (RC) starting at 4 bp + Customer Index 1 SP (RC) starting at 18 bp

Possible Barcode Positions in Read 2 - RNA_BC3 (RC) starting at 47 bp + Customer Index 1 SP (RC) starting at 61 bp

Possible Barcode Positions in Read 2 - Customer Index 1 SP starting at 11 bp + RNA_BC3 starting at 25 bp + RNA_BC2 starting at 46 bp + RNA_BC3 starting at 67 bp + RNA adaptor starting at 81 bp

Possible Barcode Positions in Read 2 - Customer Index 1 SP starting at 56 bp + RNA_BC3 starting at 70 bp + RNA_BC2 starting at 91 bp + RNA_BC3 starting at 112 bp + RNA adaptor starting at 126 bp

Figure 5.3.4: Hypothesized structures of the RNA artifacts exist in the RNA final library
RC: reverse complementary. SP: sequencing primer.

Due to the high number of matched DNA barcodes in the RNA library, we investigated the pattern for DNA library-specific oligos as well. In figure 5.3.5, we can detect an evident pattern for two groups of DNA cell barcodes, with the third-round barcode beginning at 1 bp and 23 bp, respectively. This indicates that the ends of some DNA library products, whether complete or incomplete, were integrated with template-switching oligos, which serve as RNA pre-amplification primers and library read 2 sequencing primers in subsequent steps [figure 5.3.6].

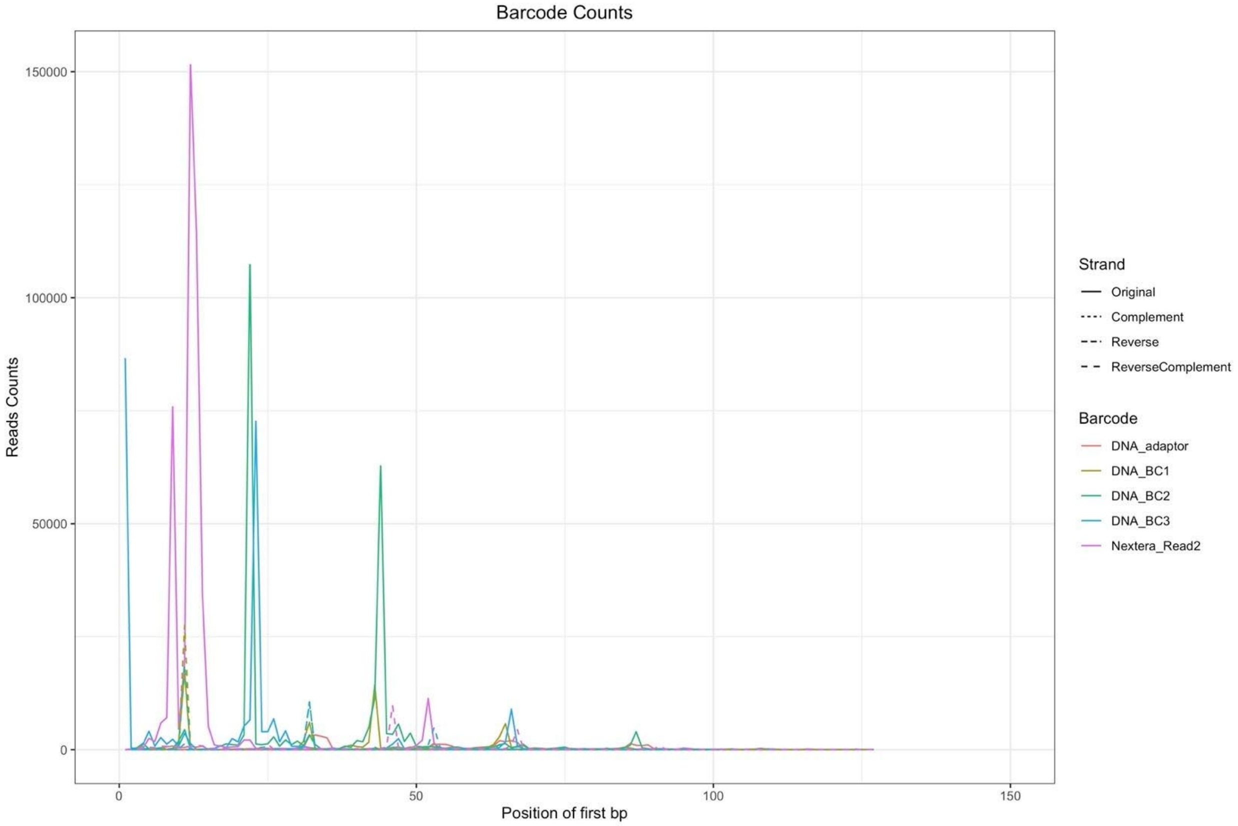
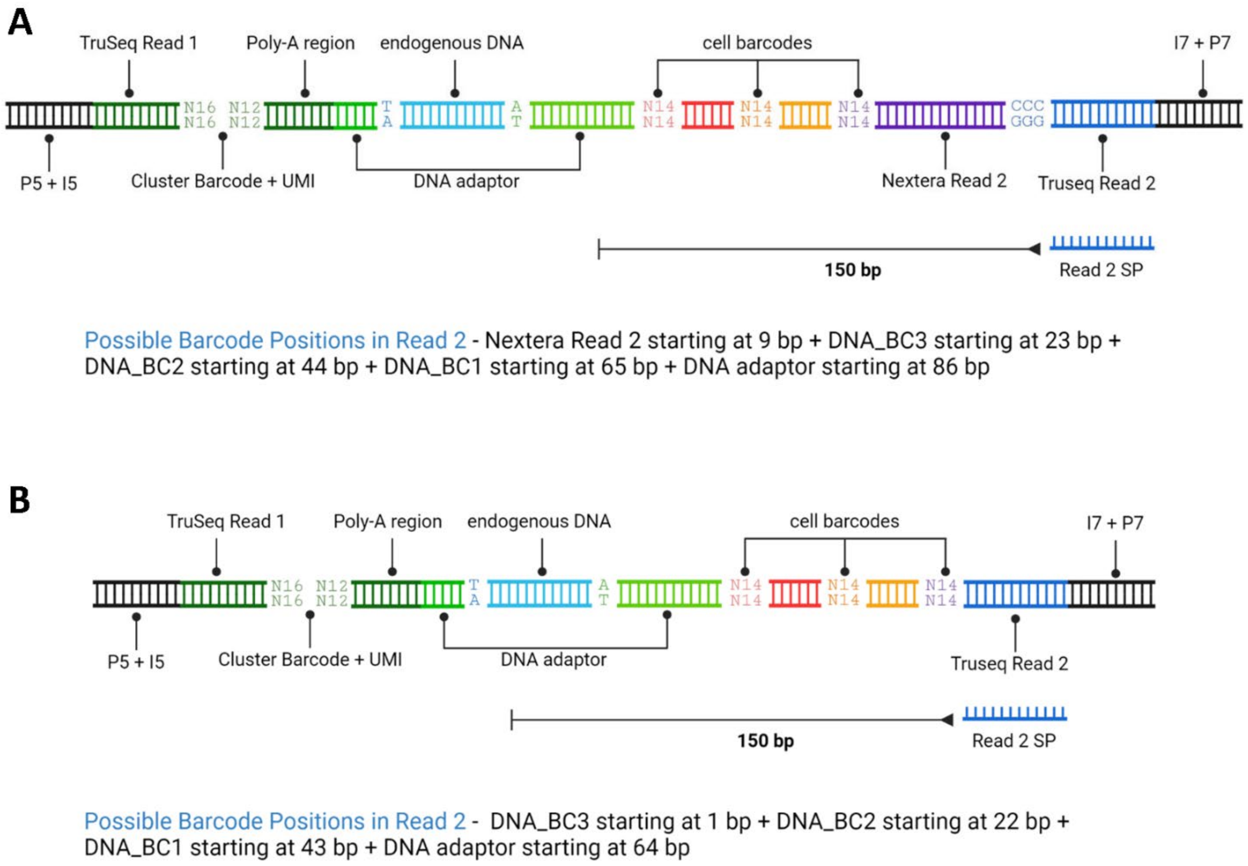


Figure 5.3.5: DNA-specific barcode positions in the RNA library sequencing result
Only oligos specific to the DNA library are searched, including the DNA cell barcodes, DNA adaptor, and Nextera Read 2 sequencing primer. The read counts are plotted against the starting positions of the sequences that are 100 percent matched. The Y axis represents the number of read counts but is not scaled to the total number of reads. The X axis represents the location of the initial base pair in a matched sequence.



Possible Barcode Positions in Read 2 - Nextera Read 2 starting at 9 bp + DNA_BC3 starting at 23 bp + DNA_BC2 starting at 44 bp + DNA_BC1 starting at 65 bp + DNA adaptor starting at 86 bp

Possible Barcode Positions in Read 2 - DNA_BC3 starting at 1 bp + DNA_BC2 starting at 22 bp + DNA_BC1 starting at 43 bp + DNA adaptor starting at 64 bp

Figure 5.3.6: Hypothesized structures of the DNA artifacts exist in the RNA final library
SP: sequencing primer.

Together, those artifacts comprised the majority of the RNA library, with RNA artifacts accounting for around 45% and DNA artifacts accounting for approximately 20% of the reads. Despite the fact that the RNA sequencing library was not successfully created, the sequencing results provided us with valuable information and guidance for improving our sciMARGI procedure. To minimize the artifacts, we will redesign our RNA first and second strand synthesis procedures, ideally by avoiding the template-switching step, which un-specifically adds oligos to the ends of all constructs.

5.4 Cluster distribution in cells based on DNA reads

Now that the DNA sequencing library has been successfully generated, we first evaluated the cell and cluster barcoding results to determine the distribution of clusters within cells and the interacting DNA sequences inside clusters. For the three rounds of 14 bp cell barcodes, no more than two mismatches in all the 42 base pairs are deemed correct. With the 16 bp cluster barcode, only a perfect match is considered correct. Using such standards, around 30.1% of all the DNA raw reads have both the correct cell and cluster barcodes. Furthermore, 65% of all reads with valid barcodes were uniquely mapped to the hg38 reference genome. Once again, this data demonstrates that the DNA components in sciMARGI libraries were successfully constructed.

After deduplication, the alignments result in a total of 11036 molecular clusters. The majority of clusters (n = 9431) contain only two reads, whereas 202 clusters contain more than ten reads [figure 5.4.1 a]. Additionally, we detected a total of 268 distinct cell barcodes. The cluster and read counts per cell are depicted in Figure 5.4.1 b and c, respectively.

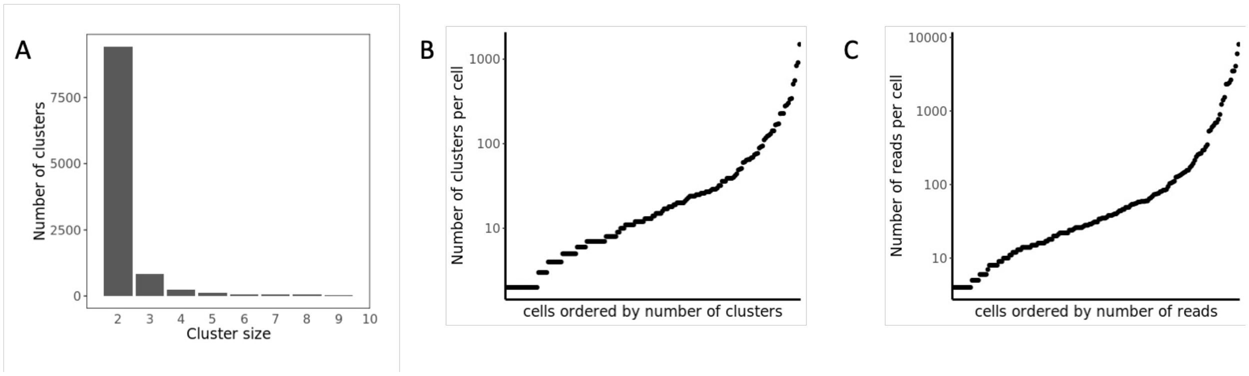


Figure 5.4.1: Distributions of reads per cluster, clusters per cell, and reads per cell
The raw reads from the DNA library sequencing results are filtered to ensure that they contain the correct cell barcodes (no more than two mismatches), cluster barcodes (no mismatches), and adaptor sequences (no mismatches). The reads with valid constructs were aligned to hg38 and the mapping quality (mapg > 20) was used to filter them. Finally, the mapped reads are deduplicated and organized by cluster and cell barcodes.

Although the number of detected cells is considerably less than the number of input nuclei (around 3 thousand), we obtained comparable cluster counts per cell to the scSPRITE study, which uses a combinatorial indexing method rather than the microdroplet technique to barcode the single complexes (Arrastia et al., 2021). This demonstrates that we successfully repurposed the 10X Genomics platform to barcode single complexes, which is innovative and may serve as a model for future technological advancements.

5.5 Cluster barcodes reveal genomic DNA interaction map

To determine if our DNA cluster barcodes truly represented the genome structure, we created a DNA-DNA contact matrix using all of the clusters from the human brain DNA library results [figure 5.5.1]. The pairwise contact matrix for chromosomes 1 to 15 was displayed. While the data are sparse, some higher-order structure and concentrated local interactions along the diagonal are detectable. These findings provide a proof of concept for the capability of our sciMARGI technology to map genomic DNA-DNA interactions from single complexes.

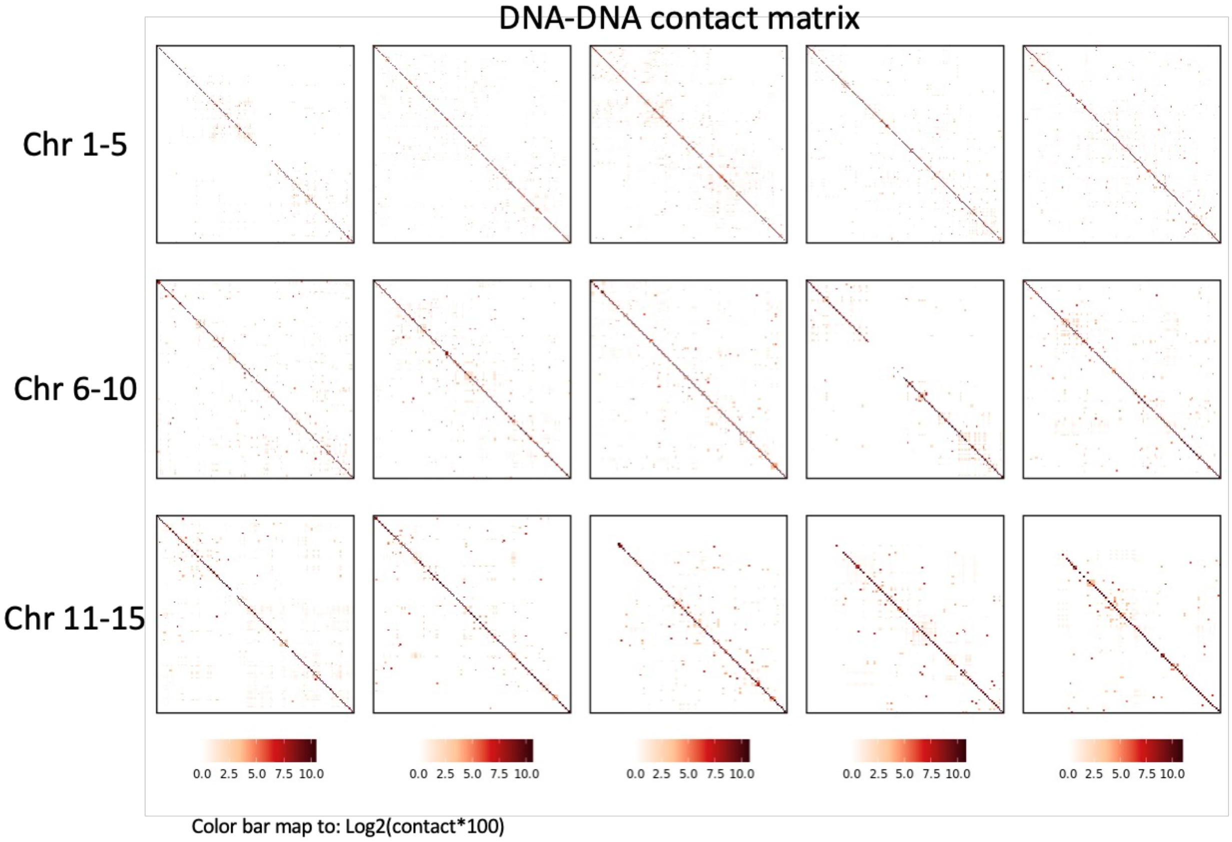
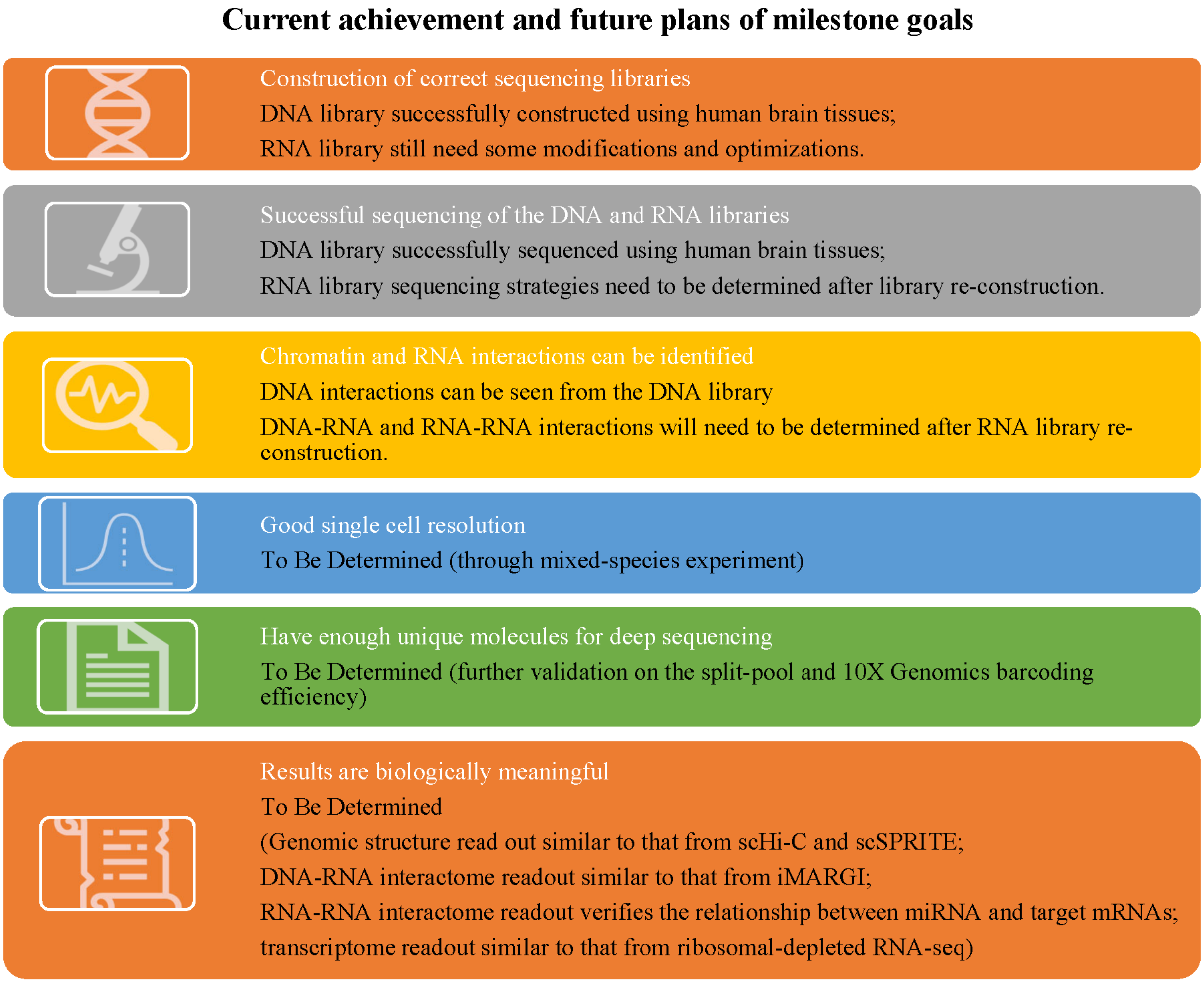


Figure 5.5.1: DNA-DNA contact matrix of chromosome 1 to 15 for human brain sciMARGI library
The pairwise interactions among all 11036 clusters are generated using the mapped and deduplicated reads. After that, each pair is normalized by $\frac{2}{n \times (n-1)}$, where n denotes the number of reads in each cluster.

6. Conclusion and future work

In this thesis, I’ve discussed the detailed design and optimization testing for the sciMARGI technology under development, as well as preliminary sequencing results from a human brain tissue library. Through adopting the combinatorial indexing and microdroplet techniques, we appended unique cell and cluster barcodes to the interacting DNA and RNA molecules in single cells. Notably, our sciMARGI technology is capable of not only detecting DNA-DNA, DNA-RNA, and RNA-RNA interactions, but also of correlating this interactomics information to single-cell gene expression profiles.

There are a total of six milestone goals that we have established in order to finish the whole development procedure of this new technology. The first goal is to successfully generate DNA and RNA libraries that contain the proper constructs. This is the most significant challenge because our library preparation techniques incorporate both the split-pool method and the 10X Genomics platform. We have validated that the DNA library has been created effectively, yet the RNA library requires more tweaking and optimization. The second objective is to successfully sequence the libraries that have been constructed. Because current next-generation deep sequencing techniques have certain requirements in read length, sequence content, and sequence primers, we need to specifically design the sequencing strategies. At the moment, the DNA library sequencing results are showing great promise, while the RNA sequencing strategy might need to be redesigned. The third milestone is to identify the interactions between chromatin and RNA. Based on the current DNA-DNA interaction results from the human brain tissue library, the 10X Genomics platform appears to be capable of reliably barcoding single complexes. Furthermore, to effectively disclose DNA-RNA interactomics data, optimal crosslinking strength is essential as well. The next two milestones will be to confirm that we have sufficient single-cell resolution and that we have detected enough single-complexes for deep sequencing, respectively. To determine whether the cell barcodes effectively distinguish single nuclei, we will conduct a mixed species experiment and look for clusters that contain both human and mouse genomic information. To obtain a sufficient number of single-complexes, we need to ensure that both cell barcodes and cluster barcodes are ligated efficiently to the target sequences. The final milestone is to validate that the sciMARGI results are biologically meaningful. We will compare our results to established approaches such as scHi-C and scSPRITE for the DNA-DNA interaction. Due to the lack of a single-cell DNA-RNA interactome detection method at the moment, we will compute a pseudo-bulk library and compare its resemblance to the iMARGI results. For RNA-RNA interactions, we will look for enrichment of the miRNA and its target mRNA. In terms of transcriptome profile, we will compare our results with those from ribosomal-depleted RNA-seq experiments. Figure 6.1 illustrates an overview of the milestones and current status described above.



Current achievement and future plans of milestone goals

Construction of correct sequencing libraries

DNA library successfully constructed using human brain tissues;

RNA library still need some modifications and optimizations.

Successful sequencing of the DNA and RNA libraries

DNA library successfully sequenced using human brain tissues;

RNA library sequencing strategies need to be determined after library re-construction.

Chromatin and RNA interactions can be identified

DNA interactions can be seen from the DNA library

DNA-RNA and RNA-RNA interactions will need to be determined after RNA library re-construction.

Good single cell resolution

To Be Determined (through mixed-species experiment)

Have enough unique molecules for deep sequencing

To Be Determined (further validation on the split-pool and 10X Genomics barcoding efficiency)

Results are biologically meaningful

To Be Determined

(Genomic structure read out similar to that from scHi-C and scSPRITE;

DNA-RNA interactome readout similar to that from iMARGI;

RNA-RNA interactome readout verifies the relationship between miRNA and target mRNAs;

transcriptome readout similar to that from ribosomal-depleted RNA-seq)

Figure 6.1: A summary of the milestone goals achieved and future plans for sciMARGI's development

Acknowledgements

This thesis, especially Sections 3, 4, and 5, contains unpublished material coauthored with Dr. Zhifei Luo and Xingzhao Wen. The thesis author was the primary author of these chapters.

REFERENCES

Arrastia, M. v., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F., and Guttman, M. (2021). Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nature Biotechnology 2021 1–10. https://doi.org/10.1038/s41587-021-00998-1.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., Adey, A., Waterston, R.H., Trapnell, C., and Shendure, J. (2017). Comprehensive single cell transcriptional profiling of a multicellular organism. Science *357*, 661. https://doi.org/10.1126/SCIENCE.AAM8940.

Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., Satpathy, A.T., Rubin, A.J., Montine, K.S., Wu, B., Kathiria, A., Cho, S.W., Mumbach, M.R., Carter, A.C., Kasowski, M., Orloff, L.A., Risca, V.I., Kundaje, A., Khavari, P.A., Montine, T.J., Greenleaf, W., and Chang, H.Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods *14*, 959–962. https://doi.org/10.1038/NMETH.4396.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D.A., Rozenblatt-Rosen, O., Zhang, F., and Regev, A. (2017). Massively-parallel single nucleus RNA-seq with DroNc-seq. Nat Methods *14*, 955. https://doi.org/10.1038/NMETH.4407.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., Wang, Y., Yue, R., Li, T., Huang, H., Orkin, S.H., Yuan, G., Chen, M., and Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091-1107.e17. https://doi.org/10.1016/J.CELL.2018.02.001.

Hawkins, J.A., Jones, S.K., Finkelstein, I.J., and Press, W.H. (2018). Indel-correcting DNA barcodes for high-throughput sequencing. Proc Natl Acad Sci U S A *115*, E6217–E6226. https://doi.org/10.1073/PNAS.1802640115/SUPPL_FILE/PNAS.1802640115.SD01.CSV.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Experimental & Molecular Medicine 2018 50:8 *50*, 1–14. https://doi.org/10.1038/s12276-018-0071-8.

Khelifi, G., and Hussein, S.M.I. (2020). A New View of Genome Organization Through RNA Directed Interactions. Frontiers in Cell and Developmental Biology *8*, 517. https://doi.org/10.3389/FCELL.2020.00517/BIBTEX.

Li, X., Zhou, B., Chen, L., Gou, L.T., Li, H., and Fu, X.D. (2017). GRID-seq reveals the global RNA-chromatin interactome. Nature Biotechnology *35*, 940–950. https://doi.org/10.1038/nbt.3968.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P.D., Quail, M.A., Swerdlow, H.P., Zernicka-Goetz, M., Livesey, F.J., Ponting, C.P., Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature Methods *12*, 519–522. https://doi.org/10.1038/nmeth.3370.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A., McCarroll, S.A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202. https://doi.org/10.1016/J.CELL.2015.05.002.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 2013 502:7469 *502*, 59–64. https://doi.org/10.1038/nature12593.

Nguyen, T.C. (2018). Development of high-throughput technologies to map RNA structures and interactions.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols 2013 9:1 *9*, 171–181. https://doi.org/10.1038/nprot.2014.006.

Quail, M.A., Swerdlow, H., and Turner, D.J. (2009). Improved Protocols for Illumina Sequencing. Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.] *0 18*. https://doi.org/10.1002/0471142905.HG1802S62.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., Guttman, M. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744-757.e24. https://doi.org/10.1016/j.cell.2018.05.024.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., Pun, S.H., Sellers, D.L., Tasic, B., Seelig, G. (2018). SPLiT-seq reveals cell types and lineages in the developing brain and spinal cord. Science *360*, 176. https://doi.org/10.1126/SCIENCE.AAM8999.

Wang, Y., Cao, T., Ko, J., Shen, Y., Zong, W., Sheng, K., Cao, W., Sun, S., Cai, L., Zhou, Y.L., Zhang, X.X., Zong, C., Weissleder, R., Weitz, D. (2020). Dissolvable Polyacrylamide Beads for High-Throughput Droplet DNA Barcoding. Advanced Science *7*, 1903463. https://doi.org/10.1002/ADVS.201903463.

Wu, W., Yan, Z., Nguyen, T.C., Bouman Chen, Z., Chien, S., and Zhong, S. (2019). Mapping RNA–chromatin interactions by sequencing with iMARGI. Nature Protocols *14*, 3243–3272. https://doi.org/10.1038/s41596-019-0229-4.

Zhou, B., Li, X., Luo, D., Lim, D.H., Zhou, Y., and Fu, X.D. (2019). GRID-seq for comprehensive analysis of global RNA–chromatin interactions. Nature Protocols *14*, 2036–2068. https://doi.org/10.1038/s41596-019-0172-4.