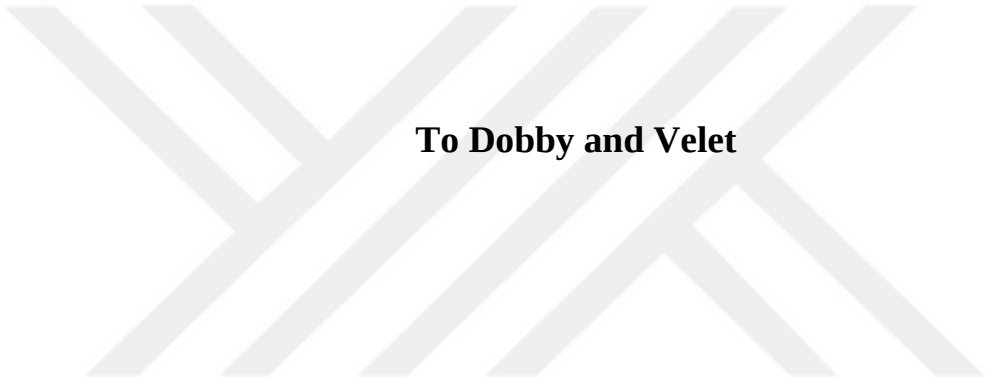




To Dobby and Velet

**SYNTHETIC REGULATORY SEQUENCE DESIGNS FOR mRNA
VACCINES**

**A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
MATERIALS SCIENCE AND NANOTECHNOLOGY**

**By
AHMET HINÇER
AUGUST 2023**

SYNTHETIC REGULATORY SEQUENCE DESIGNS FOR mRNA VACCINES

By Ahmet Hınçer

August 2023

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Urartu Özgür Şafak Şeker (Advisor)

Ayşe Elif Erson Bensan

İlyas Chachoua

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

**SYNTHETIC REGULATORY SEQUENCE DESIGNS FOR mRNA
VACCINES**

Ahmet Hınçer

M.Sc in Materials Science and Nanotechnology

Advisor: Urartu Özgür Şafak Şeker

August, 2023

mRNA-based therapeutics have demonstrated significant potential for enhancing human health across various applications. Notably, mRNA vaccines, a prominent subset of this therapeutic approach, showcased their efficacy during the COVID-19 pandemic by safeguarding billions of lives. Despite their success, the full scope of mRNA vaccines in addressing diverse health concerns, such as cancer, remains constrained by existing limitations in tunability and targetability. A deeper exploration of mRNA vaccine regulation is inevitable to harness their complete capabilities.

This thesis centers on comprehending and manipulating two pivotal regulatory domains within the mRNA molecule itself: the coding sequence and the 5' untranslated region (UTR). Regarding the coding sequence, we engineered an mRNA vaccine candidate

featuring a combined antigen coding region for SARS-CoV-2 to elicit a dual immune response against the virus. Our findings underscore that the resultant antigen exhibited interactions with distinct antibodies generated throughout the natural course of infection. This interaction profile potentially signifies a dual immune activity for enduring protection. Therefore, we practiced the essential stages of mRNA molecule manipulation requisite for an effective vaccine candidate.

In parallel, we devised an innovative methodology for constructing synthetic 5' UTR libraries tailored for selective expression within cancer cells. Collectively, this thesis advances our grasp of mRNA vaccine regulation and design. Considering the needs of the current state of mRNA vaccines, this heightened control over mRNA molecules promises novel avenues for addressing a spectrum of diseases.

Keywords: mRNA Vaccines, 5' UTR, Synthetic Biology, Bioinformatics, Breast Cancer

ÖZET

mRNA AŞILARI İÇİN SENTETİK DÜZENLEYİCİ DİZİN TASARIMLARI

Ahmet Hınçer

Malzeme Bilimi ve Nanoteknoloji, Yüksek Lisans

Tez Danışmanı: Urartu Özgür Şafak Şeker

Ağustos, 2023

mRNA tabanlı terapötikler, çeşitli uygulamalarda insan sağlığını yükseltme konusunda önemli potansiyel göstermiştir. Özellikle mRNA aşıları, bu terapötik yaklaşımın öne çıkan bir alt kümesi olarak, COVID-19 pandemisi sırasında milyarlarca yaşamı koruyarak etkinliklerini sergilemişlerdir. Başarılarına rağmen, kanser gibi farklı sağlık sorunlarına yanıt vermede mRNA aşılarının tam kapsamı, mevcut düzenleyebilirlik ve hedeflenebilirlik sınırlamaları tarafından kısıtlanmaktadır. mRNA aşısı düzenlemesinin daha derinlemesine bir keşfi, tam kapasitelerini kullanabilmek için kaçınılmazdır.

Bu tez, mRNA molekülü içindeki iki temel düzenleyici bölgeyi anlama ve manipüle etme üzerine odaklanmaktadır: kodlama dizisi ve 5' çevrilmeyen bölge (UTR). Kodlama dizisi ile ilgili olarak, SARS-CoV-2 için çift immün yanıt uyandırmak amacıyla birleştirilmiş antijen kodlama bölgesine sahip bir mRNA aşısı adayı üzerine

mühendislik çalışmaları yaptık. Bulgularımız, elde edilen antijenin enfeksiyonun doğal seyri boyunca oluşturulan farklı antikorlarla etkileşimler sergilediğini ortaya koymaktadır. Bu etkileşim profili, potansiyel olarak uzun vadeli koruma için ikili bağışıklık aktivitesini işaret edebilir. Bu nedenle, etkili bir aşı adayı için gereken mRNA molekül manipülasyonunun temel aşamalarını uyguladık.

Paralel olarak, kanser hücrelerinde seçici bir ifade sağlamaya uyarlanmış sentetik 5' UTR kütüphaneleri oluşturmak için yenilikçi bir metodoloji geliştirdik. Toplamda, bu tez, mRNA aşılarının düzenlemesi ve tasarımı konusundaki anlayışımızı iletirmektedir. Mevcut mRNA aşılarının ihtiyaçları göz önünde bulundurulduğunda, bu mRNA moleküllerinin üzerindeki bu artan kontrol, çeşitli hastalıklara yanıt vermek için yeni olanaklar vaat etmektedir.

Anahtar Kelimeler: mRNA Aşıları, 5' UTR, Sentetik Biyoloji, Biyoinformatik, Meme Kanseri

ACKNOWLEDGEMENTS

First of all, I would like to express my gratefulness to Asst. Prof. Urartu Özgür Şafak Şeker. I have literally grown up next to him. As an undergraduate student, he accepted me to his lab and gave me an opportunity to meet with the concepts that broadened my perspective on molecular biology. During my nearly six years in his laboratory, he supported me in every aspect of my life, including my private life. I am happy to meet him in my early career stage. It would be impossible to complete my work without his mentorship. Thank you for everything, Urartu Hocam.

I would also like to thank my valuable jury members, Prof. Ayşe Elif Erson Bensan and Asst. Prof. İlyas Chachoua for spending time on my study to review and improve.

I want to thank Dr. Elif Duman for sharing her knowledge about the mRNA concept and experimental procedures. I am very grateful to have a bright intern, Aslı Doğruer. She helped me a lot in nearly all steps of this thesis.

I would like to share my thanks with the former and present Synthetic Biosystems Laboratory group members; Dr. Ebru Şahin Kehribar, Dr. Elif Duman, Recep Erdem Ahan, Gizem Makas, Merve Erden, Merve Yavuz, Cemile Elif Özçelik, Suat Tüçer,

Senem Şen, Aslı Semerci, Melis Karaca, Gozeel Binte Shahid, Doğuř Akboęa, Mehmet Emin Bakar, Ece Avcı, Cem Dirse Polat, Nazlıcan Tunç, Abdurrahman Atilla, Damla Albayrak, Derin Akman, Gökçe Özkul, Nedim Kurt and Ilkay Çisil Köksaldı. It was a great journey for me with the environment you created. With all its upsides and downsides, it creates a “real and unique” experience resembling life itself, not an imitation of it.

I would like to thank especially Cemre Zekiye Araz and Sümeyye Bahar Demirkaya for their uplifting laughter and entertaining companionship; Ece Avcı for her heart-warming conversations, trustworthy friendship and her support during the thesis period; Cemile Elif Özçelik for her real sisterhood in every cases of my life, she is like a family member for me, there is no way to repay her support. I also appreciated to her cat, Patisu, for being so cute during my thesis writing process.

I would like to thank my comrade, Fulya Koç, for being the source of the happiness in my life. I felt, am feeling, and I hope I will feel her support and love for all times. Behind every great man is a great woman.

Lastly, I would like to express my greatest gratitude and deepest thanks to my family members, especially my sister Aynur Doęaner, my mother Habibe Hınçer and my father Osman Hınçer, for their endless emotional support and love. They are the ones that make me who I am right now. Thank you my cats, Dobby and Velet, for making me happy at every bad situation in my life.

Table of Contents

CHAPTER I: INTRODUCTION	1
1.1 The Significance of mRNA Molecule	1
1.2 The Regulatory Regions of mRNA Molecule	3
1.3 The History of mRNA-based Therapeutics.....	4
1.4 The Working Principle of mRNA Vaccines	6
1.5 The Importance of 5' Untranslated Regions of mRNA Vaccines.....	7
1.6 The Aim of Study.....	8
CHAPTER II: MATERIALS AND METHODS	12
2.1 Bacterial Cell Strains, Cell Maintenance and Growth Conditions	12
2.2 Construction of Plasmid Vectors.....	12
2.3 Chemical Transformation of Bacterial Cells.....	14
2.4 Plasmid Isolation, Next-Generation Sequencing and Its Alignments	15
2.5 <i>In Vitro</i> Transcription	16
2.6 Cap Structure Alteration Reaction	18
2.7 dsRNA Removal via Cellulose.....	19
2.8 dsRNA Removal via Column Chromatography.....	20

2.9	Phosphatase Treatment of mRNA	21
2.10	TBE Agarose Gel Electrophoresis for Quality Check of IVT mRNA.....	22
2.11	Mammalian Cell Line Maintenance	23
2.12	Mammalian Cell Transfection with mRNA and DNA.....	24
2.13	Protein Expression and Characterization of Transfected Cells	25
2.14	Analysis of 5' UTR Sequences for Differential Expression	27
2.15	Creation of 5' UTR Libraries by Random Assembly PCR	29
2.16	Flow Cytometry for Mean Fluorescence Value.....	31
CHAPTER III: RESULTS AND DISCUSSION.....		33
3.1	Combinatory Antigen Production via mRNA for SARS CoV2	33
3.1.1	mRNA Production via <i>In Vitro</i> Transcription from Plasmids	34
3.1.2	Removal of dsRNA.....	37
3.1.3	Transfection of HEK293T Cells by IVT mRNAs.....	41
3.2	Synthetic 5' UTR Library Creation for Differential Expression in Healthy and Cancer Breast Cells.....	45
3.2.1	Initial Analysis of Human 5' UTRs	47
3.2.2	Cloning of Chosen UTRs	48
3.2.3	Expression Profiles at HEK293T Cells	51
3.2.4	Shadowing Effect Analysis to Shift mRNA Usage to DNA Usage.....	56
3.2.5	Final Analysis of Human 5' UTRs and Consensus Production	58

3.2.6 Random Assembly PCR and Library Creation.....	61
CHAPTER IV: CONCLUSION	66
BIBLIOGRAPHY	68
APPENDIX A.....	79
DNA SEQUENCES OF CONSTRUCTS.....	79
APPENDIX B.....	88
LIST OF PRIMERS	88
APPENDIX C.....	93
PLASMID MAPS OF CONSTRUCTS USED IN THIS THESIS.....	93
APPENDIX D.....	103
SEQUENCING RESULTS.....	103
APPENDIX E	106
ADDITIONAL RECIPES	106
APPENDIX F	108
R SCRIPTS & API LINKS FOR BIOINFORMATIC ANALYSIS.....	108
Combination Script	108
Counting_statistics Script.....	119
Consensus Finder Function	121
Consensus Counter Function	122

LIST OF TABLES

Table 1: The scoring of chosen human 5' UTR with their corresponding plasmid names. The total number of the genes scored was 7791.....	48
Table 2: The accuracy of the expression profiles of chosen 5' UTRs.....	56
Table 3: The differential expression scoring of the human 5' UTRs between MCF7 and MCF10A cells. The total gene amount obtained was 4496.	59
Table 4: Consensus sequence analysis of the scored human 5' UTRs. Total number of found patterns was 433.	61
Table E. 1: LB Medium and Agar Plate Recipe	106
Table E. 2: PCR reaction with Q5	106
Table E. 3: Double Digestion Reaction	106
Table E. 4: Gibson Assembly Mixture (1.33X).....	106
Table E. 5: TSS Buffer Composition.....	107
Table E. 6: T4 Ligation Reaction.....	107
Table E. 7: 10X TBE Buffer Recipe	107
Table F. 1: API links for the proteomic data obtainment	123

LIST OF FIGURES

Figure 1: The Aims of this thesis. a) The creation of the dual-antigen expressing mRNA vaccine candidate for SARS-CoV2. b) Synthetic 5' UTR design for differential expression in healthy (MCF10A) and cancer (MCF7) breast cell lines....	10
Figure 2: Random assembly PCR process to create various length amplicons.	30
Figure 3: a) The structure of SARS CoV2. b) The illustrations of the antigen producing IVT mRNAs. The 5' and 3' UTRs (white regions) were obtained from human β globulin gene.....	34
Figure 4: The digestions of 974 (lane 1), 976s (lane 2) and 977 (lane 3). Expected bands are 6738 bps, 3804 bps and 5067 bps respectively.....	35
Figure 5: The mRNA bands after every step of manipulations described. In each section, mRNA after IVT and DNase I treatment first elution (lane 1), second elution (lane 2), mRNA after capping reaction first elution (lane 3), second elution (lane 4), mRNA after column purification (lane 5) and mRNA after phosphatase treatment (lane 6) were shown. a) 974 mRNA. Expected bands are 3974 bps. b) 976s mRNA. Expected bands are 1040 bps. c) 977 mRNA. Expected bands are 2033 bps.....	36
Figure 6: Removal of dsRNA via cellulose. Lanes are ssRNA ladder, dsRNA ladder, IVT reaction, first unbound fraction, second unbound fraction and elution of first bound fraction, respectively. Left marker shows the ssRNA ladder and right marker demonstrates the dsRNA ladder.....	38

Figure 7: The column chromatograms of the mRNAs of 974, 976s and 977. The UV280 nm absorbance values (blue line), conductivity (red line) and buffer mixing percentage (pink line) were shown. dsRNA peaks were obtained at sample loading (flow through). ssRNA regions were acquired at elution step.....40

Figure 8: The transfection of HEK293T cells with 922 IVT mRNA. In every section, left image is the bright field visualization of the cells, right image is the fluorescence version of the same location at the left image. a) Negative control. b) 200 ng of mRNA with 0.25 µl of Lipofectamine3000. c) 200 ng of mRNA with 0.5 µl of Lipofectamine3000. d) 400 ng of mRNA with 0.25 µl of Lipofectamine3000. e) 400 ng of mRNA with 0.5 µl of Lipofectamine3000. f) 800 ng of mRNA with 0.25 µl of Lipofectamine3000. g) 800 ng of mRNA with 0.5 µl of Lipofectamine3000.....42

Figure 9: Western blot image of the proteins produced via mRNAs. Both sections were blotted via anti-RBD antibody. a) Lane 1 is RBD positive control and lane 2 is 976s mRNA transfection lysate. Expected bands are 25 kDa. b) Lane 1 is Spike positive control and lane 2 is 974 mRNA transfection lysate. Expected bands are 134 kDa. 43

Figure 10: Western blot image of the proteins produced via 977 mRNA. a) This section was blotted via anti-RBD antibody. Lane 1 is Spike positive control for verification of antibody interaction and lane 2 is 977 mRNA transfection lysate. Expected bands are 134 kDa and 70 kDa, respectively. b) This section was blotted via anti-N antibody. Lane 1 is N positive control for verification of antibody interaction and lane 2 is 977 mRNA transfection lysate. Expected bands are 45 kDa and 70 kDa, respectively.....44

Figure 11: The dot blot image for IVT transfected cell lysates. a) Anti-N antibody was used. Circle 1 is N positive control (200 ng) and circle 2 is 977 mRNA transfection lysate. b) Anti-RBD antibody was employed. Circle 1, 2, 3 and 4 are RBD positive control (200 ng), 974 mRNA transfection lysate, 976s mRNA transfection lysate and 977 mRNA transfection lysate, respectively..	45
Figure 12: The illustration of scoring idea for human 5' UTRs.	46
Figure 13: The agarose gel image of 922 digestion. Expected band was 3519 bps. ...	49
Figure 14: The second PCR products to clone the chosen 5' UTRs. Lane 1, 2, 3, 4 and 5 are the amplicons for AH1, AH3, AH5, AH7 and AH9, respectively. Expected bands were approximately 1100 bps for all reactions.....	50
Figure 15: The colony PCR results for AH1, AH3, AH5, AH7 and AH9 constructs. 5 colonies were chosen for each construct (Lanes 1-5 AH1, lanes 6-10 AH3, lanes 11-15 AH5, lanes 16-20 AH7 and lanes 21-25 AH9). Expected bands were nearly 1100 bps.....	51
Figure 16: IVT reaction results for 922 (lane 1), AH1 (lane 2), AH3 (lane 3), AH5 (lane 4), AH7 (lane 5) and AH9 (lane 6). Expected bands are approximately 1050 bps.	52
Figure 17: The transfection of HEK293T cells with IVT mRNAs with chosen UTRs. In every section, left image is the bright field visualization of the cells, right image is the fluorescence version of the same location at the left image. a) 922. b) AH1. c) AH3. d) AH5. e) AH7. f) AH9.	53

Figure 18: The expression percentage of sfGFP obtained from flow cytometry for chosen UTRs. The mean fluorescence values are normalized between 0% (negative control) to 100% (922).....	55
Figure 19: The expression percentage of sfGFP obtained from flow cytometry for mRNA and DNA transfections. The mean fluorescence values are normalized between 0% (negative control) to 100% (400 ng 922 mRNA).....	57
Figure 20: The illustration of the consensus region analysis according to the differential expression scores of human 5' UTRs.....	60
Figure 21: The agarose gel image of random assembly PCR. Lane 1 is the result of first 10 cycles whereas lane 2, 3 and 4 are the end reactions of different pooling attempts. The expected bands were smear due to various lengths of amplicons.....	62
Figure 22: The agarose gel image of the CMV7 digestion. BamHI-EcoRI digestion (lane 1), only BamHI digestion (lane 2), EcoRI-PstI digestion (lane 3) and BamHI-PstI digestion (lane 4) were executed. The expected bands were 6139 (lane 1), 6195 (lane 2), 3959 and 2236 (lane 3) and 4015 and 2180 (lane 4) bps.....	63
Figure 23: The colony PCR results for library creation with ligation. Lanes numbered with white color show the presence of sfGFP whereas black colored ones demonstrate the presence of human β globulin 5' UTR. Lane 1 is CMV7 positive control. Lanes 2-7 are the colonies from negative control plate. Lanes 8-15 are the colonies from expected plate. Expected bands for both conditions were approximately 1020 bps (small increase for black colored ones).	63
Figure 24: The colony PCR results for library creation with Gibson assembly. Lanes numbered with green color show the presence of sfGFP whereas white colored ones	

demonstrate the presence of human β globulin 5' UTR. Lane 1 is CMV7 positive control. Lanes 2-7 are the colonies from negative control plate of ligation. Lanes 8-15 are the colonies from expected plate. Expected bands for both conditions were approximately 1020 bps (small increase for white colored ones).....	64
Figure D. 1: Sequencing result of AH1	103
Figure D. 2: Sequencing result of AH3	103
Figure D. 3: Sequencing result of AH5	104
Figure D. 4: Sequencing result of AH7	104
Figure D. 5: Sequencing result of AH9	104
Figure D. 6: Sequencing result of CMV7	105

CHAPTER I: INTRODUCTION

1.1 The Significance of mRNA Molecule

All living organisms operate based on a fundamental principle known as the Central Dogma of Life[1]. This process entails the transfer of information from DNA to protein. An essential player in this information transfer is Messenger RNA (mRNA), which plays a pivotal role in facilitating this process. While the genetic material of most organisms encodes information within DNA, the execution of these instructions necessitates the involvement of two distinct mechanisms: transcription and translation[2].

Transcription can be described as the process through which RNA is synthesized from DNA, facilitated by accessible promoter regions within the DNA-based genetic material and a complex of multi-megadalton proteins[3]. This intricate process results in the formation of various types of RNAs. While mRNA is primarily studied due to its role as a messenger, other RNA types such as microRNAs, transfer RNAs, small nucleolar RNAs, small nuclear RNAs, ribozymes, ribosomal RNAs, and long non-coding RNAs can also emerge during transcription[4]. These diverse RNA types, apart from mRNAs, significantly influence cellular regulation, playing a crucial role in the overall functioning of living organisms.

On the other hand, translation is the pivotal process responsible for synthesizing proteins based on the genetic code provided by mRNA. Ribosomes, which intriguingly contain an RNA-based compartment, initiate the reading of mRNAs. Proteins,

comprised of amino acids as their fundamental building blocks, are synthesized through the assistance of transfer RNAs (tRNAs) that interact with the codons present on the mRNA[5].

mRNAs stand at the intersection of transcription and translation, serving as a bridge between these two fundamental processes. This intersection, along with the presence of enzymatically active ribozymes, has given rise to the "RNA world" hypothesis, an evolutionary concept that remains a topic of debate among scientists[6].

From a chemical perspective, mRNA is a nucleic acid composed of four distinct ribonucleotides: Adenine, Cytosine, Uracil, and Guanine. It takes the form of a single-stranded molecule, featuring a backbone composed of sugar ribose and phosphate groups[7]. The classical mechanism of mRNA production within cells was elucidated by Watson and Crick[8]. This involves the transcription of the negative strand of DNA into the positive strand of mRNA through the action of RNA polymerases. In eukaryotic cells, this process takes place within the nucleus. Different types of RNAs are transcribed by various RNA polymerases, with RNA polymerase II specifically responsible for mRNA transcription[9]. This process is tightly regulated through the participation of transcription activators and repressors. These proteins can interact with other proteins to recruit them or remodel the chromatin structure of the genetic material[10]. Enhancers and promoters located upstream (5') of the transcription start site are influenced by these proteins, with the mediator complex connecting them to the basal transcription machinery, forming the preinitiation complex (PIC)[11]. Assembly of the PIC occurs onto the TATA box region according to the presence of this region. Phosphorylation of the RNA polymerase II carboxy-terminal domain by cyclin-dependent kinase 7 leads to polymerase release from PIC and elongation of the mRNA

strand[12]. The transcribed mRNA is referred to as pre-messenger RNA, containing intronic regions that are not translated during protein production. These introns are processed through splicing to yield fully functional, 5'-capped mRNAs[13].

1.2 The Regulatory Regions of mRNA Molecule

Eukaryotic mRNA molecules consist of four distinct regions that play crucial roles in regulating their stability and translational activity: the 5' cap, the coding region, poly(A) tails, and untranslated regions (UTRs) at both the 5' and 3' ends[14,15].

Every active eukaryotic mRNA molecule features a 5' cap, consisting of N7-methylated guanosine linked to the first nucleotide of the RNA. The regulatory functions of this structure commence within the nucleus, where the cap-binding complex (CBC) engages with it. This interaction orchestrates spliceosome assembly, RNA export, and nonsense-mediated decay. Upon export to the cytoplasm, the CBC-Cap interaction recruits the initial eIF4G protein, followed by other translational initiation factors, initiating protein synthesis. Substituting CBC with eIF4E results in stationary translation rounds. The presence of eIF4G leads to a pseudo-circularization of mRNA, facilitated by its interaction with the poly(A) binding protein, thus enhancing ribosome productivity. Moreover, this cap structure shields the mRNA from 5' to 3' exonuclease activity. In certain instances (Cap 1), it also signifies self-produced mRNA to the organism[16,17].

The coding region contains the instructions necessary for protein synthesis. The translational efficiency of the mRNA is primarily influenced by codon usage within this sequence. Frequent synonymous codons can enhance protein yield by recruiting the same tRNAs close to the ribosome[18–20]. Recent research suggests that ribosome

loading improvements attributed to coding region sequences, particularly related to codon usage, can augment mRNA stability[21].

Poly(A) tails, located at the 3' end, consist of repetitive adenosine nucleotides. They contribute to stability by interacting with the cap structure, circularizing mRNAs and expediting translation. This translational boost hinders mRNA degradation as the translational machinery obstructs nucleases[22].

Untranslated regions (UTRs) at both ends of the mRNA sequence do not affect the protein sequence itself but significantly regulate mRNA functionality. The 3' UTR plays roles in mRNA localization, stability, and translation[23]. Conversely, the 5' UTR establishes a ribosome docking environment and critically influences the control of translation initiation. Secondary structures within this region, such as IRES and G-quadruplexes, can activate or repress translation[24,25].

In essence, eukaryotic mRNA molecules are intricately structured entities where the 5' cap, coding region, poly(A) tail, and UTRs collaboratively govern stability, translational activity, and regulatory functions. This comprehensive orchestration ensures the precise and efficient production of proteins, vital to cellular operations.

1.3 The History of mRNA-based Therapeutics

The remarkable role of mRNA in living organisms has sparked considerable interest in its use as a therapeutic agent over the past three decades. In 1989, Robert W. Malone's work demonstrated that mRNA-loaded lipids could be translated in cultured eukaryotic cells, igniting inquiry into the potential applications of mRNA as a drug[26]. This exploration delved into the advantages of mRNA compared to DNA and protein-based approaches. Subsequent strides involved *in vitro* transcribed mRNA, unlocking diverse

applications like protein replacements and vaccinations, catalyzing a rapid acceleration of research in this domain.

From translating naked mRNA in mice to cancer antigen vaccinations, the fine-tuning of mRNA's biochemical properties, including nucleoside modifications to mitigate immunogenicity, coupled with enhanced delivery agents for precise targeting, birthed a new avenue of scientific investigation[27–30]. These efforts to refine therapeutic strategies deepened our understanding of mRNA's regulatory regions and manipulation techniques, a knowledge reservoir that proved invaluable, notably during the Covid-19 pandemic. mRNA-based technology used as a potent tool for developing a vaccine against SARS-CoV-2, galvanizing a surge of research and inspiring a new generation of scientists[31,32].

The foundation of most mRNA-based therapeutics hinges on the molecule's translational prowess within cells. Protein replacement therapies for conditions like diabetes, anemia, and myocardial infarction harness mRNA to express lacking proteins at their required locations. Expanding the frontier, genome engineering employs mRNA to induce site-specific modifications via the expression of nucleases like ZFN, TALEN, and Cas9, while genetic reprogramming leverages mRNA to induce pluripotency in cells, diverging from the exclusive focus on protein expression[33–39]. Notably, while not mRNA-based, the efficacy of antisense oligonucleotides (ASOs) and RNA interference (RNAi) underscore the potential of distinct RNA species as therapeutic agents[40].

Conversely, mRNA-based vaccination platforms for cancer immunotherapies and viral infections have surged ahead, propelled by the inherent advantages of mRNA over other vaccine platforms. Crucially, mounting an immunogenic response to cancer or viruses

necessitates transient antigen expression. The advantageous attributes of mRNA - facile scalability of production, cytoplasmic expression without nuclear entry, transient presence within cells, and the absence of genomic integration - render it a potent candidate for such cases[41].

1.4 The Working Principle of mRNA Vaccines

The successful application of mRNA vaccines hinges on the effective delivery of mRNA into cells. Among delivery platforms, lipid nanoparticles (LNPs) have emerged as the preferred choice due to their close resemblance to cellular lipid membranes. Comprising ionizable cationic lipids that interact with negatively charged mRNA molecules for encapsulation, along with phospholipids, cholesterol, and polyethylene glycol (PEG) to mimic cellular lipid structures, LNPs stand as a promising solution. Achieving targeted delivery is crucial for eliciting desired responses post-administration, prompting ongoing investigations into lipid mixture tropism and LNP surface decoration strategies[42,43].

Central to the operational mechanism of mRNA vaccines is mRNA translation within cells, aligning with the central dogma. Upon cell entry, mRNA features including the 5' cap, UTRs, and poly(A) tail provide mRNA to be treated as a self-transcribed molecule, while evading recognition by Toll-like receptors or the RIG1 pathway[44,45]. Various capping strategies during *in vitro* transcription, such as anti-reverse cap analog (ARCA) or CleanCap, facilitate this process, aligning with the canonical cap-mediated initiation mechanism[46,47].

Key to translation initiation is the recruitment of eukaryotic initiation factors (eIFs) to the mRNA's 5' cap. Among these, eIF4F, a trimeric protein complex encompassing

eIF4E, eIF4G, and eIF4A, plays a pivotal role. Specifically, eIF4E binds to the mRNA's 5' cap, while the polyA binding protein (PABP) on the poly(A) tail interacts with eIF4G, inducing a pseudo-circularization of the mRNA. eIF4G, in turn, recruits the 43S pre-initiation complex (PIC) through its eIF3 subunit, which includes the 40S ribosomal subunit. This complex traverses the mRNA's 5' UTR region in an ATP-dependent manner until it encounters the initiation codon, often accompanied by the Kozak sequence. Subsequent release of eIFs and association with the 60S ribosomal subunit shifts the mechanism towards translation elongation, ultimately leading to the expression of the encoded protein[24,48].

In the context of mRNA vaccine production, UTR regions are often selected from housekeeping genes, such as human globulin genes, to maximize production. Adjustments to poly(A) length and UTR structures can yield specific expression profiles within cells. The 5' UTR region assumes a pivotal role in translation control, as errors during initiation can result in premature translation termination or mRNA degradation, acting as safeguards against erroneous protein production[49].

1.5 The Importance of 5' Untranslated Regions of mRNA Vaccines

As previously discussed, the 5' untranslated region (UTR) serves as the initiation dock for mRNA translation, playing a crucial role in regulating the expression of the coding region. This control is orchestrated through a combination of cis-acting structures, including secondary and tertiary structures within the 5' UTR, and the involvement of trans-acting factors, collectively influencing the mRNA's translational fate[50].

Diverse structural elements within the 5' UTR, such as pseudoknots, hairpins, RNA G-quadruplexes, upstream open reading frames (uORFs), upstream start codons (uAUGs),

and internal ribosomal entry sites (IRESs), can intricately modulate translation. These elements contribute to the dynamic interplay between mRNA and the translational machinery, ultimately governing protein synthesis[24].

While a consensus sequence defining the 5' UTR remains elusive, research suggests that the mean ribosome load is altered in a sequence-specific manner[51]. An insightful study by Sample et al. harnessed machine learning to showcase that the mean ribosome loading and translational capacity can be modulated based on the mRNA sequence itself[52]. This emphasizes the sequence-dependent nature of translation initiation.

Further investigations into the regulatory mechanisms of the 5' UTR have revealed fascinating insights. For instance, the presence of uORFs has been implicated in mRNA decay, particularly in a manner dependent on UPF1, a key factor involved in nonsense-mediated mRNA decay (NMD). This reveals an intricate connection between translation initiation and mRNA surveillance mechanisms[53].

Moreover, the concept of a eukaryotic Toehold mechanism (eToehold) has emerged as a tool for activating translation through triggering. This innovative approach allows for the controlled initiation of translation in response to a specific trigger, offering a dynamic means of modulating gene expression[54].

1.6 The Aim of Study

The ongoing exploration of mRNA-based vaccine regulations is largely focused on delivery agents, mRNA biochemical modifications, and advanced computational methods for predicting ribosome loading capacities based on sequence characteristics. However, the need for novel approaches to enhance our understanding of mRNA translation and trafficking within cells remains paramount. In this thesis, our objectives

were two-fold: first, to develop an mRNA vaccine candidate for SARS-CoV2 by strategically modulating the coding region to generate a combined antigen capable of eliciting a robust immune response against various virus variants; and second, to construct a synthetic 5' UTR library aimed at inducing distinct translation patterns between healthy and cancerous breast tissue cells (Figure 1).



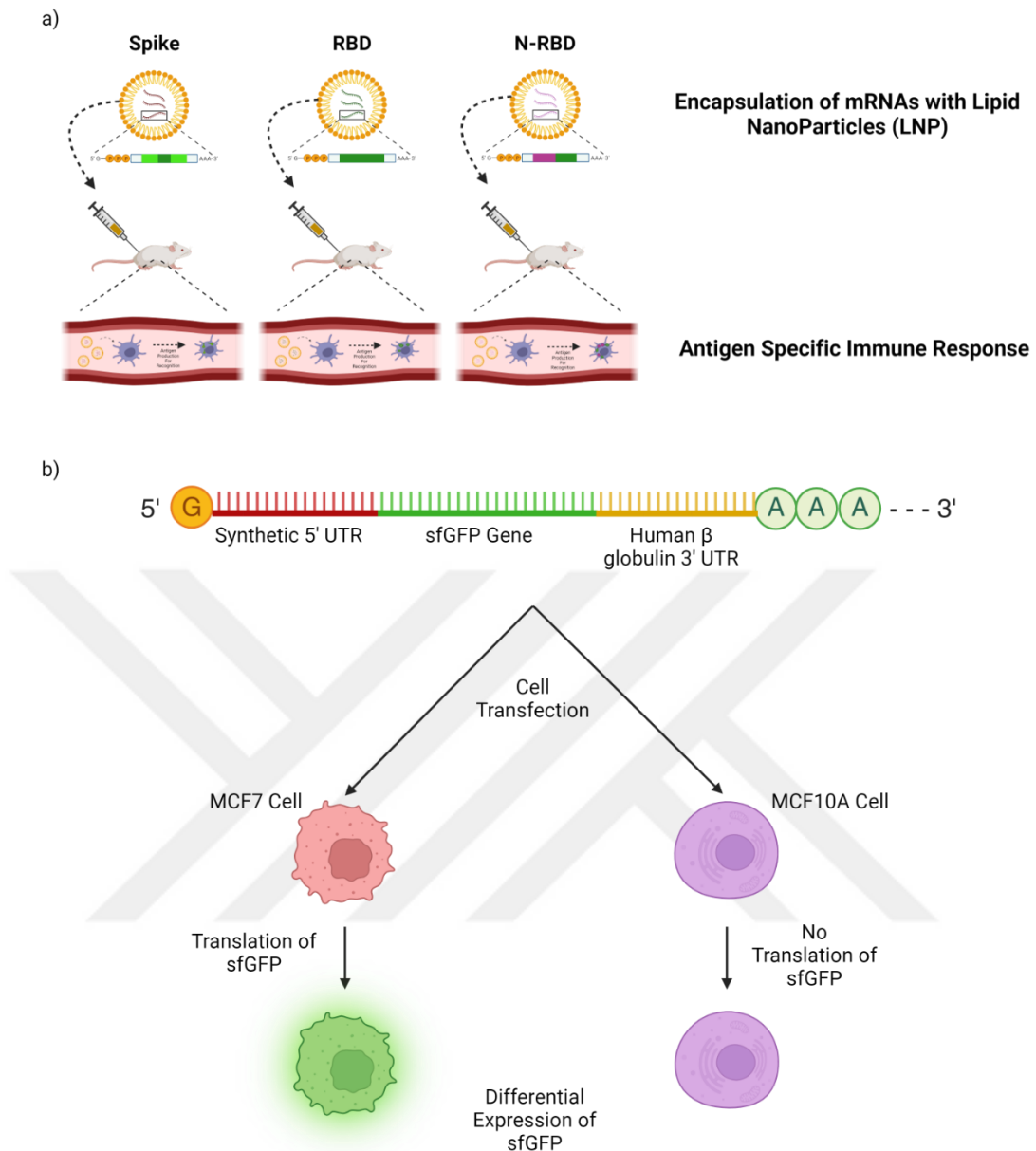


Figure 1: The Aims of this thesis. a) The creation of the dual-antigen expressing mRNA vaccine candidate for SARS-CoV2. b) Synthetic 5' UTR design for differential expression in healthy (MCF10A) and cancer (MCF7) breast cell lines.

The initial segment of this thesis encompassed an investigation into established techniques for creating vaccine candidates. However, the subsequent portion ventured into uncharted territory, introducing a bioinformatic workflow designed to assess the

translational and transcriptional properties of human cells' 5' UTRs. This innovative approach sought to uncover potential consensus sequences that could lead to selective protein expression in target cells, achieved through the random assembly of these sequences. The underlying premise is grounded in the notion that genes exhibiting substantial protein levels, even with comparatively lower transcript abundance, likely harbor intrinsic properties within their mRNA regulatory regions. Since translation is intimately linked to the 5' UTR, our methodology involved a comprehensive analysis of 5' UTRs of all available human genes matched with the protein-to-transcript ratios in diverse cell types. The resulting statistical significance guided the creation of consensus regions, representing potential targets for achieving selective expression in desired cell populations.

This workflow, centered around identifying consensus sequences for tailored translation patterns, can be readily adapted for various differentiation purposes targeting distinct diseases or conditions. The approach's versatility holds promise not only for vaccine development but also for precision medicine applications and therapeutic interventions. By unraveling the intricacies of mRNA translation through the lens of the 5' UTR, this thesis provides a novel avenue for advancing our understanding of cellular regulatory mechanisms and offers a robust framework for enhancing the efficacy of mRNA-based therapies across diverse medical contexts.

CHAPTER II: MATERIALS AND METHODS

2.1 Bacterial Cell Strains, Cell Maintenance and Growth Conditions

The cloning experiments utilized the *E. coli* DH5 α PRO strain, selected for its *recA1* and *endA* gene mutations, ensuring heightened insert stability and yield. To maintain bacterial cultures transformed with the plasmid were stored at -80°C in 1 ml aliquots, preserved with a 25% glycerol solution blended with Lysogeny Broth (LB) (Appendix E, Table E.1).

For plasmid retrieval, a requisite volume of LB medium was measured in accordance with the upcoming sequential experiment. This medium was then supplemented with the appropriate antibiotic, ampicillin in this instance. Subsequently, the stored bacterial culture was introduced into this liquid medium, with the container capacity set at five times the volume of the medium employed (e.g., 3 ml of medium in a 15 ml Falcon tube).

The inoculated medium was then placed in a shaking incubator set at 37°C, and incubated overnight. This process facilitated the proliferation and growth of the bacterial culture, enabling the subsequent stages of experimentation.

2.2 Construction of Plasmid Vectors

Plasmids containing sfGFP, Spike, RBD, and RBD-Nucleocapsid combination, namely PUC57-pIVT-sfGFP (922), PUC57-pIVT-Spike (974), PUC57-pIVT-RBD (976-S), and PUC57-pIVT-RBD_Nucleocapsid (977), were obtained from our lab. These plasmids were equipped with the T7 promoter region, the 5' and 3' UTRs of the human β globulin gene, and a poly(A) tail consisting of 118 adenosine repeats. Plasmid 922

was utilized as the starting point for introducing various 5' UTR regions, aligning with the initial concept of creating synthetic 5' UTRs. To replace the 5' UTR, PCR was performed using 2 distinct forward primers with overhangs, along with a shared reverse primer (Appendix E, Table E.2). Subsequently, plasmid 922 underwent digestion using KpnI-HF and XboI enzymes at 37°C for 3 hours (Appendix E, Table E.3).

Following PCR and digestion, the resulting PCR products and digested backbone were subjected to electrophoresis on a 1% agarose gel at 135 V for 35 minutes. Desired fragments were cut, and purification was carried out using the MN PCR Clean-up kit. The backbone, at 50 ng, was combined with PCR products at a 3-fold molar excess, and 1.33x Gibson Assembly reactions prepared in our lab were performed (Appendix E, Table E.4). Incubation of this mixture at 50°C for 1 hour facilitated the circularization of plasmids. These constructs were chemically transformed into the specified *E. coli* strain, and the cloned vectors were named AH1, AH3, AH5 etc., to minimize potential experimental biases.

For the final iteration of the synthetic 5' UTR creation workflow, pCDNA3-eGFP from our lab was employed as the backbone. Digestion of 2000 ng of pCDNA3-eGFP using SacI resulted in the creation of an insert-free vector. Through bioinformatic analysis of the CMV promoter using Element, the 5' UTR region for insertion was identified. Amplification of the sfGFP-3' UTR region from plasmid 922, using primers with overhangs, yielded an insert featuring EcoRI-5' UTR-BamHI-sfGFP-3' UTR segments, with homology regions at both ends for Gibson assembly. Following the same protocol as previously described, the plasmid CMV7 was successfully generated.

2.3 Chemical Transformation of Bacterial Cells

To procure chemically transformable *E. coli* DH5 α PRO cells, an overnight culture of bacteria from this strain was diluted at a 1:100 ratio in LB medium. This diluted mixture was then incubated in a shaking incubator (RPM 200) at 37°C for approximately 2 hours. Upon reaching an OD600 value of 0.3-0.5, the culture was evenly divided into two ice-chilled falcon tubes, which were subsequently stored on ice or at +4°C for 15 minutes. Following this cooling step, the tubes were centrifuged at +4°C for 10 minutes at 3000 rpm. The resulting supernatants were carefully aspirated, and the bacterial pellets were gently resuspended in ice-chilled TSS buffer at a 1:10 ratio (Appendix E, Table E.5). For instance, if the pellet originated from 30 ml of culture, 3 ml of TSS was utilized. The resuspended pellets in TSS were then aliquoted into 1.5 ml or 2 ml Eppendorf tubes in 100 μ l volumes, maintaining a cold environment.

Subsequently, one portion of these competent cells was employed to inoculate a selected antibiotic-containing medium to ascertain the absence of contamination. Another portion was reserved for plasmid transformation to assess its competence. The remaining competent cells were preserved at -80°C.

For the chemical transformation of plasmids, Gibson Assembly, or T4 ligation (Appendix E, Table E.6) reactions, the prepared competent cells were thawed at +4°C for around 15 minutes. Following thawing, the plasmid or reaction mixture was introduced to these cells and allowed to interact for an additional 20 minutes at the same temperature. The cell-plasmid mixture was then subjected to a brief heat shock at 42°C for 45 seconds, followed by immediate placement on ice for 2-5 minutes. After this cooling period, 1 ml of LB medium without antibiotics was introduced to the cells,

which were then incubated at 37°C for 45 minutes either in a shaking or stationary incubator to facilitate recovery and proliferation.

Subsequent to the 45-minute incubation, the cells were centrifuged at 5800 rpm for 5 minutes, and the supernatant (approximately 900 µl) was discarded. The remaining 100 µl was used to gently resuspend the cell pellets. This resuspended liquid was spread evenly onto LB-agar plates supplemented with the appropriate antibiotic based on the antibiotic resistance gene present in the transformed DNA. The plates were subsequently placed in a stationary incubator set at 37°C for overnight growth.

2.4 Plasmid Isolation, Next-Generation Sequencing and Its Alignments

Plasmid isolation was carried out in two distinct scales, namely miniprep and midiprep, in order to obtain the requisite quantity of DNA for consecutive experimental stages. Miniprep sufficed for all cloning, sequencing, mammalian DNA transfection, and library experiments. Plasmids acquired through midiprep were exclusively utilized as the source for *in vitro* transcription (IVT) reactions, owing to the demand for a substantial amount of mRNA transcribed from a correspondingly high DNA quantity. The GeneJET Plasmid Miniprep Kit from ThermoScientific was employed for miniprep, while the NucleoBond Xtra Midi Kit was employed for midiprep. For initiating the miniprep, 3 ml of overnight-cultured bacteria were centrifuged at 8000 rpm for 5 minutes, whereas 50 ml of bacterial growth was sufficient for the midiprep protocol. The established protocols provided by both kits were followed to carefully.

To facilitate next-generation sequencing (NGS), the plasmids obtained through miniprep were submitted to the Intergen company located in Ankara. A final volume of

20 µl, containing 1000-2000 ng of DNA, was dispensed into PCR tubes. The ensuing NGS results were disseminated via email. For a comprehensive alignment of the NGS outcomes with the anticipated plasmid map, the map was extracted in .gb format from Benchling.com. The Geneious application was instrumental in effecting map-to-reference alignments. Substantial pairwise identity scores were observed for pivotal regions on the map, exceeding 50%. The consensus sequence emerging from this alignment was subsequently uploaded onto Benchling, augmenting clarity regarding the presence of any mutations.

2.5 *In Vitro* Transcription

For *in vitro* transcription (IVT) of the desired plasmids, the HiScribe T7 ARCA mRNA Kit was employed. In suitable vectors designated for IVT, the T7 promoter was strategically positioned preceding the target mRNA sequence. The plasmids were sourced from bacterial stocks through midiprep, ensuring a substantial DNA yield. To prevent aberrantly elongated mRNA transcripts, plasmids were subjected to digestion by HindIII and SapI restriction enzymes to linearize the templates. The recognition sites of these enzymes were strategically positioned at the terminal region of the Poly(A) tail, facilitating dissociation of the T7 RNA polymerase, a component of the kit, for transcription of the sequence encompassing the 5' UTR, coding region, 3' UTR, and poly(A) tail.

A quantity of 20 µg of the plasmid of interest underwent digestion with HindIII and SapI endonucleases, incubated at 37°C for 4 hours. Following the incubation, a volume equivalent to 500 ng of the plasmid was loaded onto a 1% agarose gel, electrophoresed at 135 V for 35 minutes. The remaining digestion mixture was maintained at 37°C until

the completion of gel electrophoresis. Once plasmid linearization was verified, the Monarch Nucleic Acid Purification Kit (PCR Clean-Up) was utilized to isolate the linear plasmid remaining aside from the amount resolved on the gel.

For the production of IVT mRNA from the linearized plasmid, a mixture containing 10 µg of DNA, 100 µl of 2X ARCA/NTP Mix, 10 µl of 100 mM DTT, 20 µl of T7 RNA polymerase, and nuclease-free water up to a total volume of 200 µl was prepared in a dedicated RNA-based experiment fume hood. The expected amount of mRNA from 10 µg of DNA was approximately 250 µg. The reaction was thoroughly mixed and incubated in a stationary incubator at 37°C for 30 minutes. To eliminate template DNA, 20 µl of DNase I was introduced into the mixture, necessitating an additional 15 minutes at 37°C. The Monarch RNA Cleanup Kit was employed to isolate the IVT mRNA from this mixture. Initially, 400 µl of RNA cleanup binding buffer was added to the IVT reaction. Subsequently, 600 µl of ethanol (>95%) was meticulously combined with the binding buffer-IVT mixture. This composite solution was loaded onto a kit column up to a volume of 900 µl, followed by centrifugation at 14000 rpm for 1 minute. The flow-through was discarded, and the process was reiterated until all solution containing IVT mRNA was effectively loaded onto the column. For the subsequent washing step, 500 µl of RNA cleanup wash buffer was added to the column and centrifuged under the previous conditions, performed twice. Post-washing, the empty column was spun to thoroughly remove residual materials. A volume of 100 µl of nuclease-free water was introduced to the column and allowed to interact for 3 minutes, facilitating contact with the retained IVT mRNA. Subsequently, the column was placed over a 1.5 ml eppendorf tube and centrifuged at 14000 rpm for 3 minutes. To prevent undesired RNase degradation, 2.5 µl of Human placenta RNase inhibitor was incorporated into the eluted mRNA. A subsequent elution step was conducted with

50 µl of nuclease-free water to ascertain the presence of any residual mRNA on the column. For this secondary elution, 1.25 µl of RNase inhibitor was added.

From both the initial and secondary elution steps, 3 µl of the samples were transferred into separate PCR tubes for concentration analysis and gel loading. The resulting IVT mRNA solutions were stored at -80°C until subsequent manipulations or experiments.

2.6 Cap Structure Alteration Reaction

To enhance the efficiency of transfection experiments, the creation of the Cap-1 structure involved the addition of a methyl group to the 2'-O position of the first nucleotide, adjacent to the existing cap structure (Cap-0 or ARCA)[16]. To achieve this modification, the mRNA Cap 2'-O-Methyltransferase enzyme from NEB was employed. The reaction scale for Cap-1 modification can be adjusted based on the total mass of IVT mRNA (µg).

For the alteration of Cap-1 on 250 µg of mRNA, the mRNA volume was adjusted to 400 µl using nuclease-free water. The resulting mixture was initially heated to 65°C for 5 minutes, followed by cooling on ice for 5 minutes. Subsequently, a sequential addition of 50 µl of 10x capping buffer, 25 µl of freshly prepared 4 mM SAM (S-adenosyl methionine), and 25 µl of methyltransferase enzyme was carried out. The complete reaction was then incubated in a stationary incubator at 37°C for a duration of 60 minutes.

Following the modification process, mRNA purification and analysis were performed using the previously described kit and protocol from the IVT section. The resulting Cap-1 capped IVT mRNA was subsequently stored at -80°C for future use.

2.7 dsRNA Removal via Cellulose

To prevent the potential issues arising from dsRNA contamination, which can lead to undesirable effects such as inhibition of protein production, cytokine secretion, and compromised immune responses to mRNA-encoded antigens, Baierdörfer et al. introduced a purification technique that effectively separates ssRNA from dsRNA based on an ethanol-dependent approach[55].

To create a customized microcentrifuge column with cellulose, the silica portion of a miniprep column from the previously mentioned miniprep kit was carefully removed using tweezers. The column end was then blocked with autoclaved cotton to prevent cellulose from escaping into the collection tube. Subsequently, a chromatography buffer was prepared containing 10 mM HEPES at pH 7.2, 0.1 mM EDTA, 125 mM NaCl, and 16% (v/v) ethanol.

Microcrystalline cellulose (MCC) was mixed with the chromatography buffer to achieve a concentration of 0.2 g MCC per ml. This mixture was incubated on a shaker for 10 minutes. Next, 700 µl of the MCC-buffer solution was added to the prepared microcentrifuge column and centrifuged at 14000 rpm for 1 minute. The resulting flow-through was discarded, and the column was washed with 500 µl of chromatography buffer, followed by another centrifugation step.

Subsequently, 500 µl of chromatography buffer with IVT-mRNA (molarities at the end should be the same, 2X buffer was prepared), was loaded onto the column. The column was then incubated at room temperature with vigorous shaking for 30 minutes. Following centrifugation, dsRNA was anticipated to bind to the cellulose, while ssRNA would pass through as the flow-through. The dsRNA-bound cellulose was eluted using

nuclease-free water. This entire procedure was repeated for the flow-through solution using a newly prepared MCC-containing column to enhance separation efficiency.

For buffer exchange of the ssRNA, an additional step involved adding 0.1 volume of 3 M NaOAc (pH 5.5) and 1 volume of isopropanol. This mixture was incubated at -20°C overnight. Following centrifugation at 14000 rpm for 10 minutes, the resulting mRNA pellets were resuspended in nuclease-free water. The purified samples were stored at -20°C for short-term use and at -80°C for long-term storage.

2.8 dsRNA Removal via Column Chromatography

An alternative purification method for isolating ssRNA from IVT mRNA was established using a C4 HLD monolithic column obtained from Sartorius. This column was integrated into an ÄKTA Start Fast Protein Liquid Chromatography (FPLC) device, allowing for real-time UV monitoring during the purification process to differentiate peaks corresponding to dsRNA, proteins, residual DNA, and ssRNA. This UV tracking is the main reason for the reliability of this method.

To execute this protocol, five different RNase-free solutions were prepared using DEPC-treated water as a cost-effective approach. Buffer solutions were prepared according to the manufacturer's instructions in an RNase-free fume hood. The critical parameters of these buffers were conductivity and pH, as the method was based on the cationic interaction between mRNA and specific salts. To minimize prolonged mRNA exposure to these salts, inline dilution of IVT-mRNA was performed. After dilution, the recommended loading concentration of mRNA onto the column was achieved (0.15-0.3 µg per µl), and the FPLC device's UV sensor, which can measure 280 nm

absorbance, gave peaks for nucleic acids and proteins at 50% sensitivity compared with 260 nm absorbance, which was suitable for nucleic acids.

The protocol's outline was as follows: The FPLC device's tubing was thoroughly cleaned with RNase-free water. Following equilibration of the column and tubing with the recommended buffer, the IVT mRNA sample was diluted with a specific buffer to attain a desired salt molarity. The mRNA-loaded column was then eluted with a low-salt buffer to collect fractions. The UV280 absorbance values were used to identify and collect the ssRNA-containing fraction.

To remove excess salts from the ssRNA-containing fraction, the solution was diluted with the same buffer lacking salts, adjusting the final cation concentration suitable for isopropanol precipitation[56]. Subsequently, isopropanol (1 volume) was added, and the mixture was incubated at -20°C overnight. Centrifugation at +4°C and 14000 rpm for 10 minutes (adjustable if necessary) separated the RNA pellet, and the supernatant was carefully discarded. The pellet was washed twice with 500 µl of ice-cold 70% ethanol under similar centrifugation conditions. Finally, the pellet was resuspended in nuclease-free water for subsequent experiments. A 3 µl aliquot of the resuspended mRNA was taken for concentration analysis and gel electrophoresis.

2.9 Phosphatase Treatment of mRNA

To prevent RIG-I activation in cells by dephosphorylating the 5' and 3' ends of mRNAs, Quick Calf Intestinal Alkaline Phosphatase (CIP) from NEB was employed[57]. The required picomol amount of mRNA ends was determined based on

RNA length and mass using NEBioCalculator. This calculated amount served as the basis for scaling up the 1 pmol mRNA phosphatase treatment protocol.

For the dephosphorylation reaction involving 1 pmol of mRNA ends, the following components were mixed: 1 μ l of 10X CutSmart buffer, 0.5 μ l of Quick CIP, the mRNA solution, and nuclease-free water to make a final volume of 10 μ l. This mixture was then incubated at 37°C for 10 minutes. In cases where a larger amount of enzyme was required based on the pmol amount of mRNA, a reduced enzyme quantity could be utilized, with a corresponding increase in the incubation time. Following the incubation period, the Monarch RNA Clean-up kit was employed to purify the mRNA, separating it from any residual components of the reaction.

2.10 TBE Agarose Gel Electrophoresis for Quality Check of IVT

mRNA

For mRNA gel electrophoresis, a dedicated gel tank exclusively used for mRNA-based studies was employed. To prevent any potential RNase contamination that could affect the accuracy of mRNA assessment, all compartments of the gel tank were thoroughly cleaned with 70% ethanol and nuclease-free water prior to gel preparation.

To create the agarose gel, 1% agarose (mass-to-volume ratio) was mixed with 1X TBE buffer (Appendix E, Table E.7). TBE buffer was preferred over TAE thanks to its inhibitory effects due to borate on many enzymes, including RNases. In this study, 60 ml of 1X TBE buffer was combined with 600 mg of agarose. The mixture was then microwaved for 3 minutes to melt the agarose polysaccharides and form a homogeneous solution in the buffer. Subsequently, the solution was cooled under water to a temperature where the gel wouldn't solidify and the SyberSafe nucleic acid staining

agent wouldn't evaporate. This nucleic acid dye was utilized at a concentration of 0.001% (volume/volume). The final mixture was poured into the gel tank and left to solidify for 60 minutes. Once solidification was complete, the samples were loaded onto the gel.

To prepare the samples for loading, the mRNA samples and ssRNA ladder were mixed with 2X RNA loading dye in a 1:1 ratio. These combined samples were denatured at 70°C for 10 minutes and then immediately placed on ice for 2 minutes. Following this, the gel was run for 35 minutes at 125 V. The gel was visualized using a blue light transilluminator or an imaging system capable of agarose gel visualization, chemiluminescence, and Coomassie staining.

2.11 Mammalian Cell Line Maintenance

In this thesis, the HEK293T adherent cell line was chosen due to its advantages in terms of transfection efficiency, translational capacity, and ease of maintenance. For the preparation of a complete growth medium, 440 ml of high glucose Dulbecco's Modified Eagle's Medium (DMEM) was combined with 50 ml of heat-inactivated Fetal Bovine Serum (FBS), 5 ml of 100x L-Glutamine (200 mM), and 5 ml of 100x Penicillin/Streptomycin. This mixture was subsequently filtered using a Corning Disposable Filter Unit with a pore size of 0.45 µm. The resulting medium was stored at +4°C.

To initiate the HEK293T cell stock, the complete medium was warmed in a 37°C water bath for 15 minutes. The frozen cell stock from the liquid nitrogen tank was thawed under the same conditions until only a small ice piece remained in the stock. The thawed stock was then transferred to a 15 ml falcon tube containing 5 ml of complete growth

medium. After centrifugation at 1500 rpm for 5 minutes, the supernatant was discarded, and the cell pellet was resuspended using 4 ml of medium. The resuspended cells were placed in a 25 cm² tissue culture flask (T25), which was subsequently located in a 37°C humidified incubator supplied with 5% CO₂.

Regular subculturing of the cells was required approximately every 3-4 days, determined by observing the cells reaching around 90% confluency under a bright field microscope. During subculturing, the medium in the flask was discarded, and the cells were washed with 1X PBS. Following the removal of PBS, an appropriate volume of 0.05% Trypsin-0.53 mM EDTA, previously warmed at 37°C, was added to the cells. The flask was then placed in a 37°C incubator for 5 minutes. Once cell dissociation was achieved, an amount of complete growth medium three times that of the used trypsin-EDTA was added to inhibit the trypsin-EDTA activity. The cell mixture was centrifuged at 1500 rpm for 5 minutes, and the supernatant was discarded. The resulting pellet was resuspended in 1 ml of medium, and 100 µl of this resuspension was added to a T75 flask containing 12 ml of medium for subsequent experiments. If necessary, the remaining cells were counted for experimental setup.

2.12 Mammalian Cell Transfection with mRNA and DNA

During all transfection experiments, a total of 10,000 cells suspended in 100 µl of complete growth medium were evenly distributed into each well of a 96-well plate. Following an overnight incubation, the transfection process was carried out using the Lipofectamine 3000 reagent. For mRNA-based studies, the appropriate quantities of mRNA and Lipofectamine 3000 components were determined through experimentation (Figure 8), leading to the establishment of the final transfection protocol. All

transfection experiments were conducted in triplicates unless it indicated, resulting in the preparation of four separate mixtures.

The following steps outline the finalized transfection protocol as follows. Opti-MEM Medium was pre-warmed at 37°C in a water bath for 15 minutes. One PCR tube containing 25 µl of Opti-MEM was combined with 3.2 µl of P3000 reagent and 1600 ng of mRNA. A second PCR tube was used to dilute 2 µl of Lipofectamine 3000 with 25 µl of Opti-MEM. The mRNA-containing tube was gently transferred onto the Lipofectamine 3000 dilution, and this combined mixture was incubated at room temperature for 15 minutes. After the incubation period, one-fourth of the final mRNA-Lipofectamine mixture was added to each well of the 96-well plate, which already contained the overnight-cultured cells. Following a 24-hour incubation at 37°C, the cells were observed under a bright field or fluorescence microscope and subsequently utilized for further experimental analyses, such as protein analysis or flow cytometry.

For DNA transfections, a similar procedure was followed with adjustments to the amounts of DNA, P3000 agent, and Lipofectamine 3000. In this case, 2 µl of P3000 and 1 µl of Lipofectamine 3000 were used for each of the 4X mixtures. The subsequent steps of the protocol were the same as those for mRNA transfections.

2.13 Protein Expression and Characterization of Transfected Cells

To analyze protein expression and characterize mRNA-transfected cells, the following procedures were employed. The HEK293T cells were detached using a 0.05% Trypsin-0.53 mM EDTA solution preheated to 37°C, following the same steps used in the subculturing technique. For cells in 96-well plates, the medium was discarded, and 100 µl of trypsin-EDTA solution was added. After 5 minutes of incubation at 37°C,

dissociation was confirmed under a bright field microscope. Complete growth medium (300 μ l) was then added to inhibit trypsin.

The cell suspension was transferred to a 1.5 ml Eppendorf tube and centrifuged at 1500 rpm for 5 minutes. The supernatant was discarded based on pellet visibility. If the pellet was not visible, the supernatant was used to resuspend the invisible pellet location. In both cases, cells were lysed using M-PER mammalian protein extraction reagent. 50 μ l of the agent was added to cells and incubated on ice for 20 minutes. After incubation, the solution was centrifuged at 12000 rpm for 15 minutes to collect the supernatant containing expressed proteins.

Cell lysates were concentrated using Amicon Ultracentrifugal Filter units (cut-off 10 kDa) for Western blotting and dot blotting of Spike, RBD, and RBD-N proteins. Samples from the same mRNA transfection type were combined in a filter unit and centrifuged at 5000 g until the total volume reached 50 μ l. These 50 μ l concentrated samples were divided into 2 tubes, one containing 40 μ l for Western blotting and the other containing 10 μ l for dot blotting.

For Western blotting, two 10% SDS-PAGE gels were prepared. Samples were mixed with SDS loading dye (6x) and heated to 95°C for 8 minutes. After denaturation, 20 μ l of samples were loaded onto each gel and electrophoresed at 140 V for 80 minutes. Gels were transferred to PVDF membranes using the Trans-blot Turbo device. Membranes were blocked with 5% BSA in TBST for 90 minutes at room temperature to prevent non-specific interactions. Primary and secondary antibodies were prepared in ratios of 1:2500 and 1:5000 in 1% BSA-TBST solution, respectively. Primary antibodies included rabbit anti-RBD polyclonal antibodies (Sinobiologicals) for RBD recognition and human anti-N antibody (Genscript) for N recognition. Antibody

incubations were conducted at room temperature for 60 minutes. After primary antibody incubation, membranes were washed with TBST and secondary antibody incubation was performed, followed by repeated washing steps. Luminescence substrate kit was applied to membranes for visualization using a chemiluminescence imaging system.

For dot blotting, PVDF membranes were treated with methanol for 30 seconds followed by distilled water for 2 minutes. Samples (10 μ l) were dropped onto the membranes and allowed to dry at room temperature. Subsequent steps for blocking, primary antibody incubation, washing, secondary antibody incubation, washing, and visualization were similar to the Western blotting protocol. The blocking and antibody solutions were also consistent with the Western blotting process.

2.14 Analysis of 5' UTR Sequences for Differential Expression

For the initial analysis of human 5' UTRs, the following steps were undertaken. Gene median TPM values obtained from GTEx tissue-categorized transcriptomic data and proteome abundance data from Wang et al. were combined using Microsoft Excel[58,59]. The direct ratio of a gene's protein abundance level (iBAQ value) to its TPM value in the same tissue was calculated and formed the basis of the scoring matrix.

This matrix was then merged with the 5' UTR sequences and their lengths extracted from the HumanUTRDb[60]. The matrix was processed based on two criteria: first, the UTR length had to fall within the range of 30-600 nucleotides, and second, the UTRs were filtered based on an average score ± 2 * standard deviation.

The final matrix was categorized by scores from high to low according to the tissue of interest. Within each score partition, UTRs with similar lengths were selected to

minimize the potential impact of size differences and focus on the intended intrinsic properties.

For the final approach in human 5' UTR analysis, proteomic data for MCF7 and MCF10A cell lines were acquired through the ProteomicsDb Application Programming Interface (API) using Excel Power Query (Appendix F, Table F.1).

Transcriptomic datasets GSE208731 and GSE114332 were obtained from NCBI GEO[61,62]. Challenges arose from differences in gene identifiers between transcriptomic and proteomic data. To address this, Ensemble grch37 and grch38 versions were gathered for all genes from Biomart, including attributes such as Ensemble gene name, transcript length, gene id, and gene id versions[63,64]. UTR sequences were accessed from Ensemble grch38. Gene identifiers, thus harmonized, were used to link proteomic, transcriptomic, and UTR data, with preference given to the longest UTR sequence for genes with multiple options via an R script (Appendix F).

Upon merging the data for both cell lines, genes with TPM values higher than 0.5, calculated from raw counts for both MCF7 and MCF10A, were retained[65]. The UTR sequences of these genes underwent consensus region analysis. This involved identifying consensus patterns across a range of lengths within UTRs that scored higher in MCF7 cells compared to MCF10A cells (total ratio > 3). The presence of these consensus patterns was quantified for both UTR categories that scored higher and lower (Threshold = 1). This threshold demonstrated the UTRs which had tendency to expressed higher in MCF7 cells. Significance of the consensus sequences was determined using Fisher's Exact Test and Odds Ratio.

2.15 Creation of 5' UTR Libraries by Random Assembly PCR

To generate 5' UTR libraries, two distinct annealing sequence regions were devised through combinations of four selected patterns derived from differential expression analysis. These annealing sequences were employed as complementary segments during the assembly process[66,67]. A total of 22 chosen patterns were strategically positioned between these annealing sequences. To facilitate the assembly, primers denoted as AH_L0 to AH_L21 were synthesized in accordance with the specified arrangement. Additionally, four primers were designed to incorporate suitable recognition sites for restriction enzymes and homology regions. These primers were employed for the generation of randomly assembled amplicons, which were further used for ligation or Gibson assembly (Figure 2).

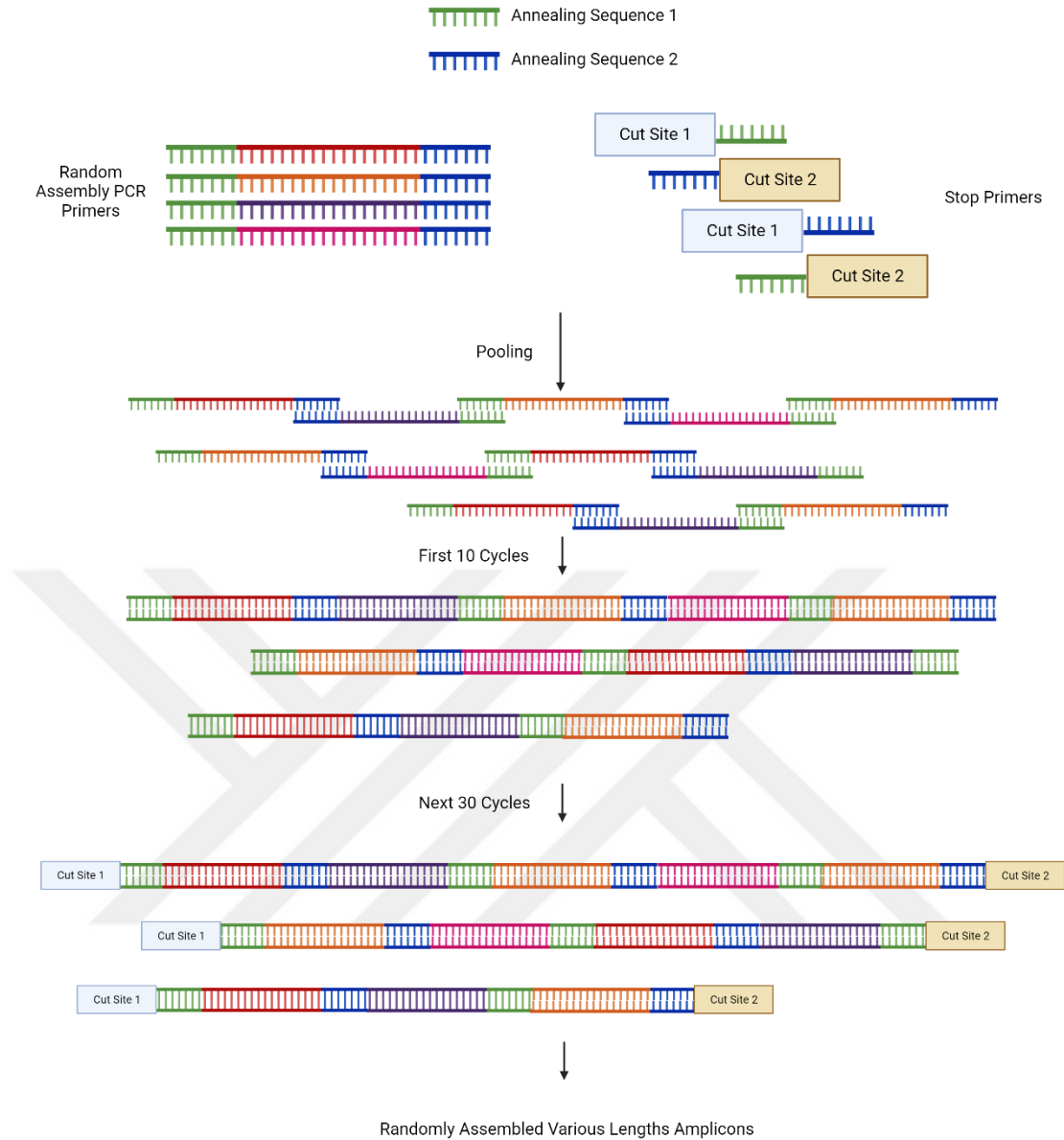


Figure 2: Random assembly PCR process to create various length amplicons.

For the random assembly PCR mixture, 2 μl of 100 μM primers containing consensus regions and 0.4 μl of initial end primers were combined. This mixture was then diluted in a 1:100 ratio. Subsequently, 10 μl of 5X Q5 reaction buffer, 1 μl of dNTPs, 0.5 μl of Q5 polymerase, and 38.5 μl of the diluted mixture were mixed. The PCR protocol consisted of 10 cycles with an appropriate annealing temperature based on the annealing sequences. Following this step, a final concentration of 25 μM stop primers

was added (1 μ l each). An additional 30 cycles of PCR were performed using the annealing temperature specific to the final stop primers. The resulting products, containing randomly assembled amplicons, were analyzed using agarose gel electrophoresis. The smear region indicative of these amplicons was extracted using a PCR clean-up kit.

Subsequent to the extraction, the extracted smear was employed for either ligation or Gibson assembly to generate the library elements. In both assembly methods, CMV7 plasmids underwent digestion using EcoRI and BamHI enzymes. For the ligation approach, the extracted smear was also subjected to enzymatic cleavage by these enzymes. Ultimately, a chemical transformation process was employed to obtain bacterial stocks of the library elements, facilitating their subsequent applications.

2.16 Flow Cytometry for Mean Fluorescence Value

Flow cytometry analysis was performed using the BD Accuri C6 flow cytometry device. The samples were prepared following the protocols outlined in sections 2.12 and 2.13, excluding the step involving cell lysis. Cell resuspension was accomplished using 100 μ l of PBS. Subsequently, each sample was loaded onto the flow cytometry device with a volume of 80 μ l at a medium speed.

To initiate the analysis, background signals arising from the PBS solution were assessed. This aided in the identification of the cell population by utilizing a forward scatter (FSC-A) vs side scatter (SSC-A) graph. Following the gating of this region, a new gated population was established, and this population was further examined using a forward scatter (FSC-A) vs forward scatter height (FSC-H) graph to discern singlets. Subsequently, these isolated singlets were subjected to analysis of their mean FL1-H

values. This analysis involved evaluating the fluorescence production of this population in comparison to the negative control.



CHAPTER III: RESULTS AND DISCUSSION

3.1 Combinatory Antigen Production via mRNA for SARS CoV2

To develop a comprehensive antigen strategy for SARS-CoV-2 and assess its efficacy, the protein mutational rates across different virus variants was searched to decide which antigens would be combined. Through this process, it was observed that the commonly targeted antigen, the Spike protein, exhibited reduced sensitivity to vaccine-induced neutralization for certain new variants. Conversely, the nucleocapsid (N) protein, responsible for genome packaging in the virus, demonstrated a lasting profile for memory T-cells. Notably, individuals possessing the memory T-cells for N protein from the SARS-CoV-1 outbreak in 2002-2003, exhibited cross-reactivity when encountering SARS-CoV-2. This cross-activity was attributed to the high similarity in amino acid sequences between the nucleocapsids of two viruses, suggesting a low mutation rate for the N protein[68].

In light of these findings, a novel approach was formulated: the creation of a single antigen sequence that combined the receptor binding domain (RBD) of the Spike protein with the N protein (Figure 3). This strategy aimed to establish a dual immunity against various SARS-CoV-2 variants. By incorporating both RBD and N protein sequences into the antigen, the potential risks associated with reduced recognition of new virus variants by the immune system could be mitigated. If the effectiveness of neutralizing antibodies declines for one component of the combination, the presence of antibodies against the other component could compensate, resulted in enhancing overall immunity and protection against evolving variants of the virus.

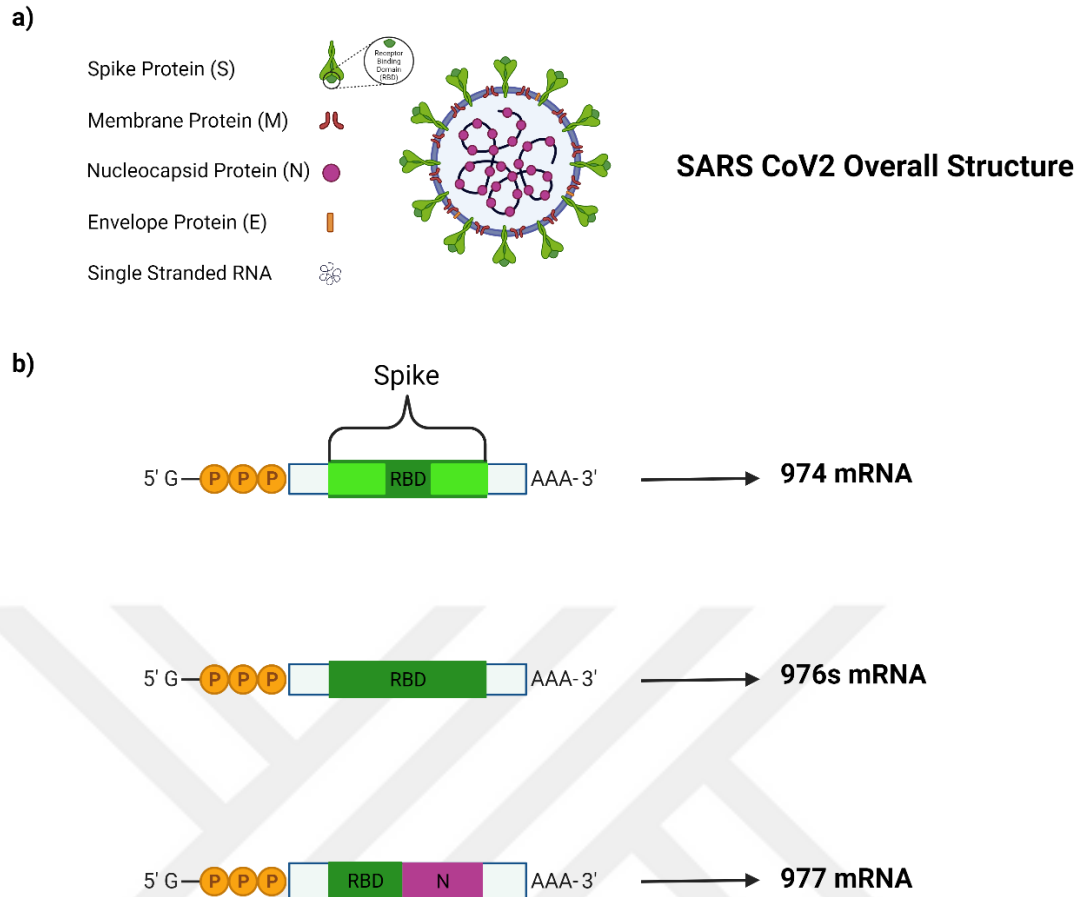


Figure 3: a) The structure of SARS CoV2. b) The illustrations of the antigen producing IVT mRNAs. The 5' and 3' UTRs (white regions) were obtained from human β globulin gene.

3.1.1 mRNA Production via *In Vitro* Transcription from Plasmids

To generate the mRNA constructs for conducting vaccination studies to investigate the immune responses induced by the RBD-N combination, a series of molecular manipulations were performed. Specifically, plasmids 974, 976s, and 977 were utilized, each containing common 5' and 3' UTRs of the human β globulin gene, as well as an identical 118-nucleotide-long poly(A) tail. However, these plasmids differed in terms of their coding regions, which respectively contained the sequences for the spike protein, RBD, and the RBD-N fusion.

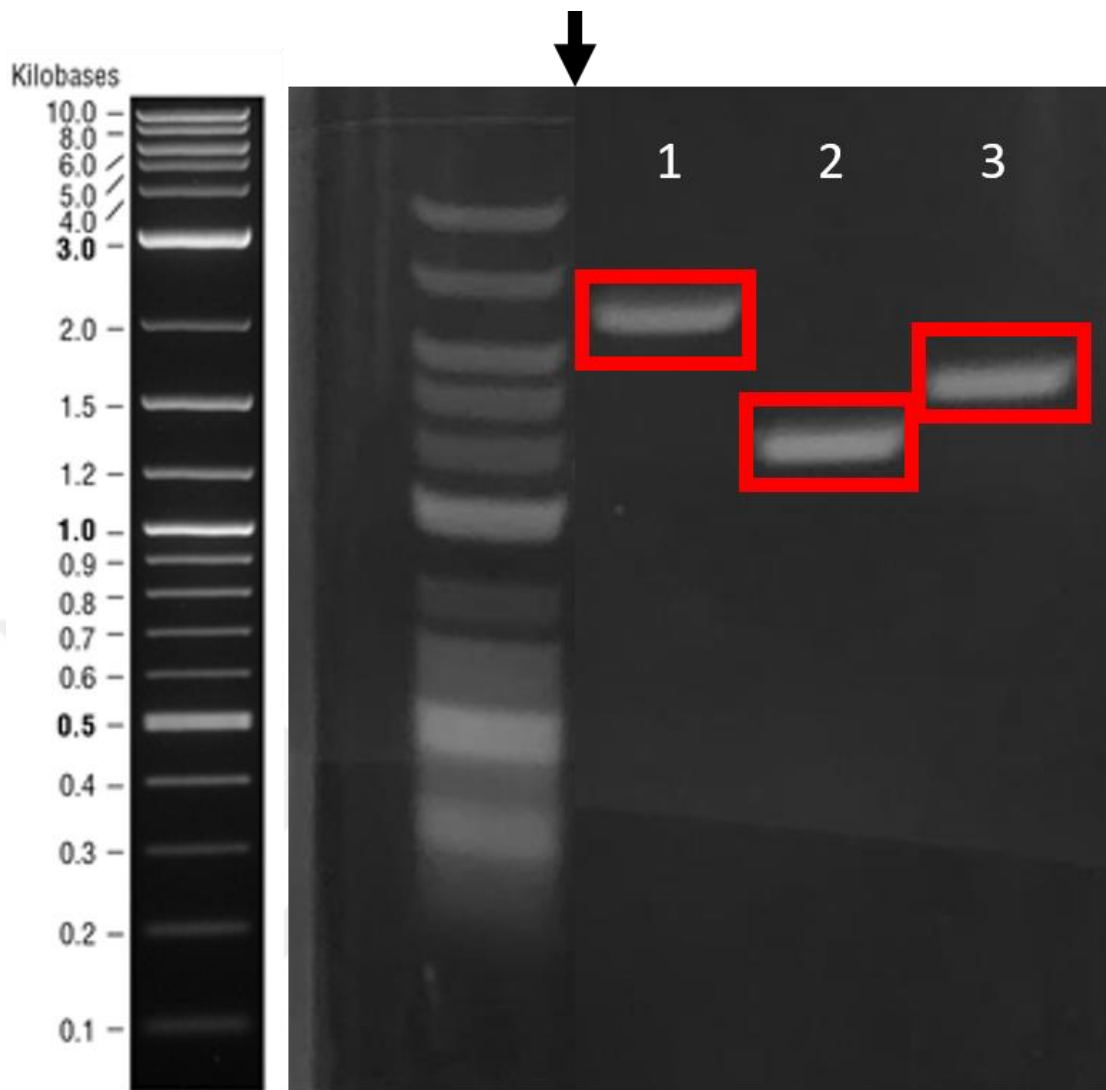


Figure 4: The digestions of 974 (lane 1), 976s (lane 2) and 977 (lane 3). Expected bands are 6738 bps, 3804 bps and 5067 bps respectively.

To begin the process, the plasmids were subjected to enzymatic digestion using HindIII and SapI restriction enzymes. The outcome of this enzymatic cleavage can be observed in Figure 4. Following the digestion, a cleanup procedure was carried out to purify the resulting digested plasmids. Subsequently, *in vitro* transcription reactions were conducted utilizing the ARCA capping kit. This sequence of steps was then followed by Cap1 transfer, removal of double-stranded RNA (dsRNA), and dephosphorylation treatment (as depicted in Figure 5).

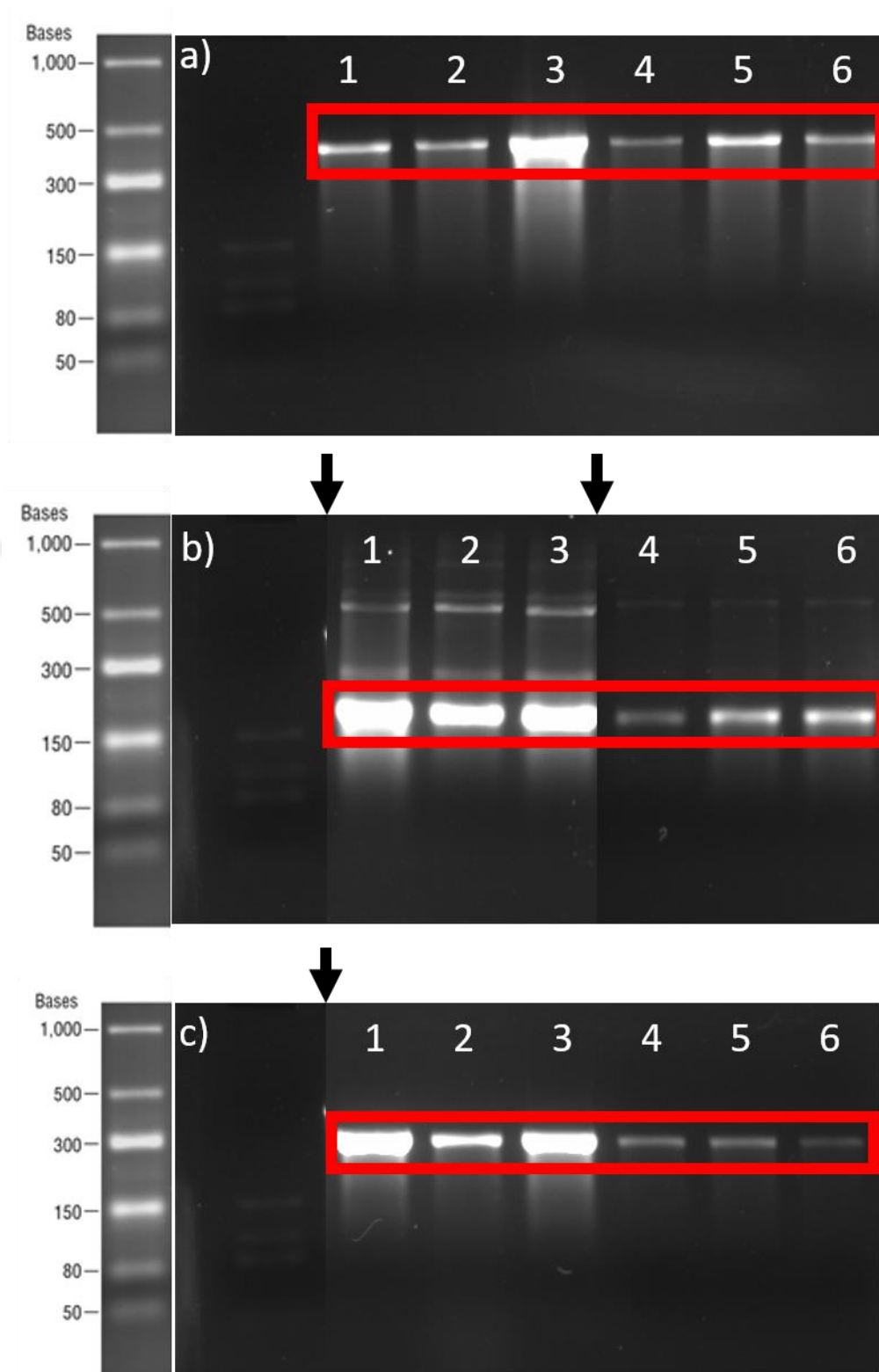


Figure 5: The mRNA bands after every step of manipulations described. In each section, mRNA after IVT and DNase I treatment first elution (lane 1), second elution (lane 2), mRNA after capping reaction first elution (lane 3), second elution (lane 4), mRNA after

column purification (lane 5) and mRNA after phosphatase treatment (lane 6) were shown. a) 974 mRNA. Expected bands are 3974 bps. b) 976s mRNA. Expected bands are 1040 bps. c) 977 mRNA. Expected bands are 2033 bps.

The purpose of these sequential manipulations was to generate the desired mRNA constructs that incorporate the RBD-N combination and possess the necessary modifications to optimize their translational and immunogenic properties. These mRNA constructs would then serve as critical components for the subsequent vaccination studies aimed at assessing the immune effects of the RBD-N antigen combination.

3.1.2 Removal of dsRNA

In this particular segment of the thesis, our primary objective was to develop an mRNA vaccine featuring a combinatorial antigen. Our primary focus revolved around identifying the essential criteria that needed to be met in order to effectively stimulate immunity towards specific antigens within the body. One of the most significant challenges encountered during this process was the prevention of antigen-specific antibody generation and memory cell formation due to inherent characteristics of *in vitro* transcribed (IVT) mRNA[69].

A notable complication arose from the intrinsic behavior of T7 RNA polymerase, leading to the production of double-stranded RNA (dsRNA). This occurrence stemmed from premature transcription termination and template-independent RNA polymerase activity. The resulting short RNA fragments synthesized at the initial stages of the IVT reaction had the potential to initiate the transcription of complementary RNA sequences. The presence of dsRNA within an mRNA vaccine could trigger cellular

defense mechanisms against viral infections, resulting in the inhibition of translation and the activation of pathways such as TLR3, RIG1, and MDA5. These pathways, while serving as protective measures against viral threats, posed obstacles to eliciting the desired immune responses to the intended antigen[70].

Given these challenges, the critical necessity emerged to eliminate dsRNA during the production of mRNA vaccines.

3.1.2.1 Cellulose Methodology

In our initial efforts, we sought to remove dsRNA within the *in vitro* transcription (IVT) reaction, a critical step in large-scale mRNA vaccine production. We pursued an economical approach involving cellulose purification, leveraging ethanol's unique property to influence nucleic acid binding to cellulose. Our choice of the IVT reaction, involving plasmid 922 which encoded the sfGFP antigen for easy fluorescence-based analysis, aided in subsequent experiments, particularly transfections in mammalian cell lines.

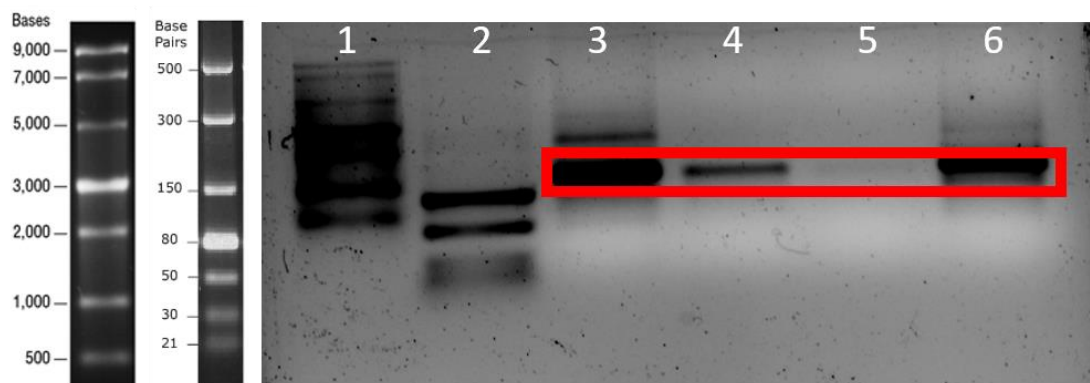


Figure 6: Removal of dsRNA via cellulose. Lanes are ssRNA ladder, dsRNA ladder, IVT reaction, first unbound fraction, second unbound fraction and elution of first bound fraction, respectively. Left marker shows the ssRNA ladder and right marker demonstrates the dsRNA ladder.

As depicted in Figure 6, the IVT reaction encompassed a mixture of both dsRNA and ssRNA. The presence of dsRNA resulted in a band shift due to its comparatively higher molecular weight in contrast to ssRNA. Initial cellulose purification experiments displayed promising potential, evident from the elimination of the higher band in the first unbound fraction. However, complications emerged during the elution phase, where both dsRNA and ssRNA were present. This phenomenon can be attributed to the challenge of precisely tuning the ethanol concentration to the optimal 16% (volume/volume) requirement. Even slight deviations in this percentage could influence the nucleic acid binding affinity.

Furthermore, the observable differences in band intensities between the unbound and bound fractions indicated that the loss of ssRNA after dsRNA removal fell beyond acceptable thresholds. Given that cost-effectiveness was a paramount consideration for both our combinatorial approach and mRNA vaccine production, this outcome was not within the desired range. Consequently, we made the strategic decision to shift our methodology towards column chromatography, aiming to overcome these challenges and achieve successful dsRNA elimination while maintaining the integrity of the ssRNA that was crucial for our vaccine production.

3.1.2.2 Column Chromatography

For the removal of dsRNA through column chromatography, we adhered to the manufacturer's well-established protocol. We processed IVT mRNAs derived from plasmids 974, 976s, and 977, incorporating steps to eliminate template DNAs and introduce the Cap1 structure. Roughly 200 µg of the mRNA samples were loaded onto the column, and the UV280 values were monitored to identify the fractions containing ssRNA during the elution process (Figure 7).

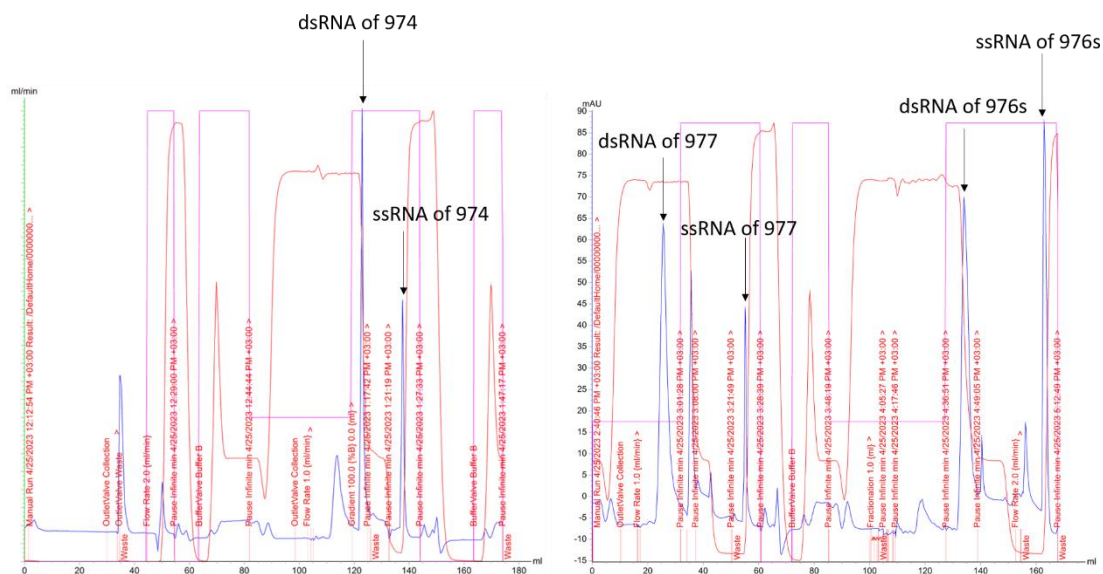


Figure 7: The column chromatograms of the mRNAs of 974, 976s and 977. The UV280 nm absorbance values (blue line), conductivity (red line) and buffer mixing percentage (pink line) were shown. dsRNA peaks were obtained at sample loading (flow through). ssRNA regions were acquired at elution step.

Due to concerns about potential interference from salts present in the elution buffer, we opted for an isopropanol precipitation method to exchange the buffer for nuclease-free water. Although the recommended approach for buffer exchange was tangential flow filtration (TFF), this equipment was unavailable in our laboratory at the time. Unfortunately, the isopropanol precipitation resulted in a reduction in initial mRNA amounts for all three versions (974, 976s, and 977) ranging from 65% to 77.5%. Consequently, the final resuspended mRNA quantities for these versions were 90 µg, 80 µg, and 45 µg, respectively.

It was noted that the precipitation process appeared to be more challenging for the 977 IVT mRNA compared to the other versions. This discrepancy suggests that intrinsic properties of the produced mRNA, such as secondary structures, might influence the yield of this particular process. The presence of smears in the gel wells could be

attributed to nuclease contamination during the column loading step (Figure 5). While efforts were made to prepare and clean the pipes, it's important to acknowledge that the FPLC device used in this process was not exclusively isolated for mRNA procedures. The device had been employed for protein purifications, including the Cas13b ribonuclease for other ongoing projects in our laboratory. Despite the preparatory cleaning of the pipes, the potential for contamination remained, especially considering that the buffer alterations brought into the device were exposed to air due to the device's positioning.

3.1.3 Transfection of HEK293T Cells by IVT mRNAs

To assess the protein production efficiency of IVT mRNAs, we employed HEK293T cells known for their high translation efficiency, transfection success rate, and ease of maintenance. Lipofectamine3000, a commercial lipid-based transfection reagent, was used for the transfection experiments. Initial analysis was conducted using sfGFP IVT mRNAs, as the fluorescence emitted by sfGFP allowed for easy visualization under a fluorescence microscope. For each well of a 96-well plate, 10000 cells were seeded in 100 μ l of media. The aim of this experiment was to determine the optimal amounts of mRNA and lipid delivery agent (P3000 reagent and Lipofectamine 3000) for efficient transfection and translation.

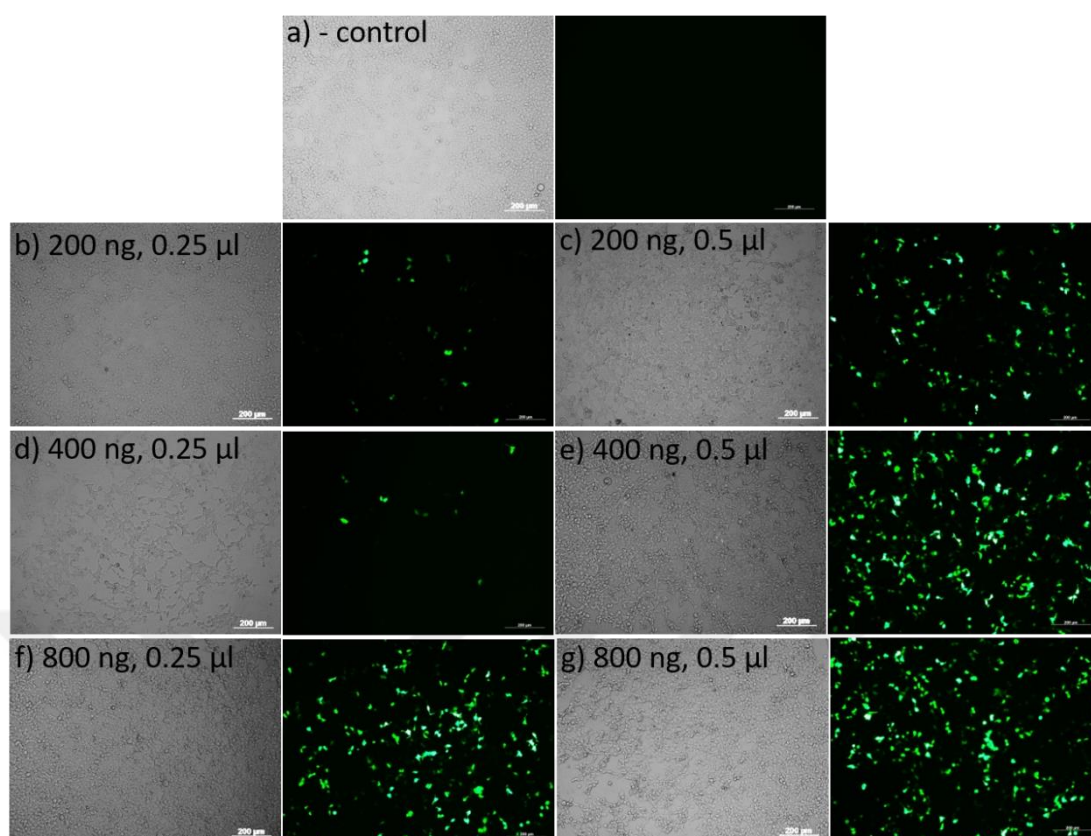


Figure 8: The transfection of HEK293T cells with 922 IVT mRNA. In every section, left image is the bright field visualization of the cells, right image is the fluorescence version of the same location at the left image. a) Negative control. b) 200 ng of mRNA with 0.25 µl of Lipofectamine3000. c) 200 ng of mRNA with 0.5 µl of Lipofectamine3000. d) 400 ng of mRNA with 0.25 µl of Lipofectamine3000. e) 400 ng of mRNA with 0.5 µl of Lipofectamine3000. f) 800 ng of mRNA with 0.25 µl of Lipofectamine3000. g) 800 ng of mRNA with 0.5 µl of Lipofectamine3000.

As shown in Figure 8, the most favorable transfection and translation results, achieved with the lowest mRNA quantity, were observed with 400 ng of mRNA, 0.8 µl of P3000 reagent, and 0.5 µl of Lipofectamine3000. Further increases in mRNA amounts did not significantly improve the results beyond the performance achieved with 400 ng. Following this experiment, these determined parameters were used consistently for all subsequent transfection trials.

We employed two distinct approaches to investigate the antibody interactions with antigens produced through the IVT mRNAs of 974, 976s, and 977. This was crucial as antibody interactions serve as a primary indicator of the potential of these mRNAs to initiate the desired immune response when used as vaccine candidates. The first approach employed was Western blotting, which demonstrated the expression of the full-length protein (Figure 9-10). The second approach was Dotblot, utilized to exhibit the interaction of potential epitopes with neutralizing antibodies (Figure 11). The dark regions of the Dotblot contained the negative controls or the samples that were not expected to be interacted. This interaction was examined independently from the complete translation of the protein. The rationale behind this approach is that a full-length protein is not necessary to trigger an immune response as a vaccine. The key requirement is the presentation of epitopes via antigen-presenting cells (APCs).

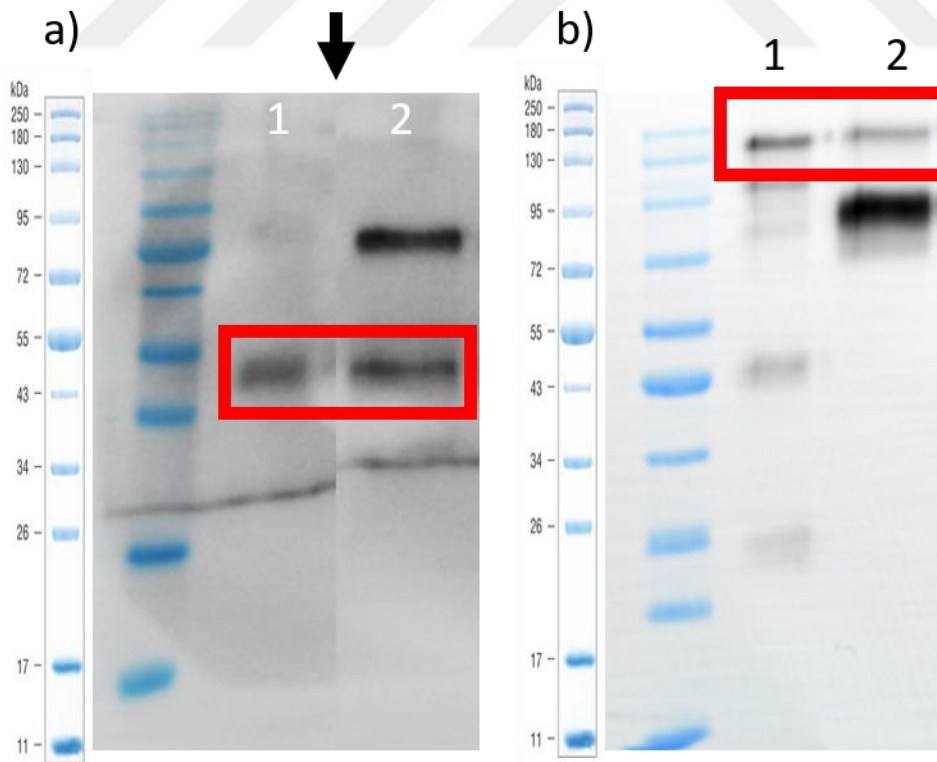


Figure 9: Western blot image of the proteins produced via mRNAs. Both sections were blotted via anti-RBD antibody. a) Lane 1 is RBD positive control and lane 2 is 976s

mRNA transfection lysate. Expected bands are 25 kDa. b) Lane 1 is Spike positive control and lane 2 is 974 mRNA transfection lysate. Expected bands are 134 kDa.

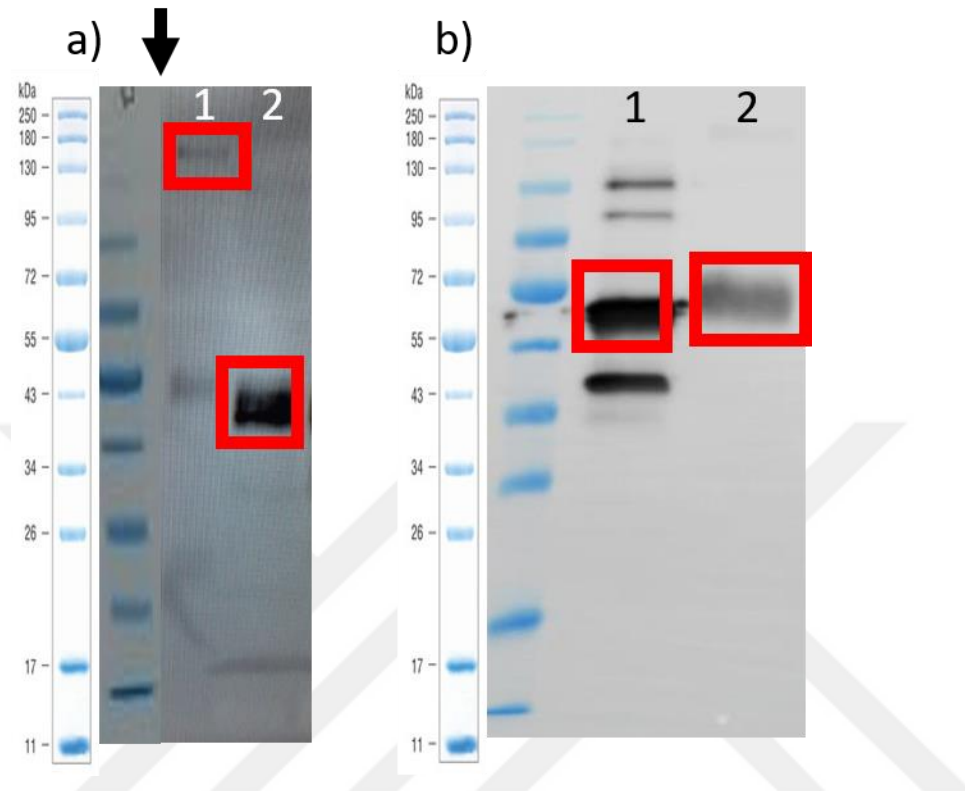


Figure 10: Western blot image of the proteins produced via 977 mRNA. a) This section was blotted via anti-RBD antibody. Lane 1 is Spike positive control for verification of antibody interaction and lane 2 is 977 mRNA transfection lysate. Expected bands are 134 kDa and 70 kDa, respectively. b) This section was blotted via anti-N antibody. Lane 1 is N positive control for verification of antibody interaction and lane 2 is 977 mRNA transfection lysate. Expected bands are 45 kDa and 70 kDa, respectively.

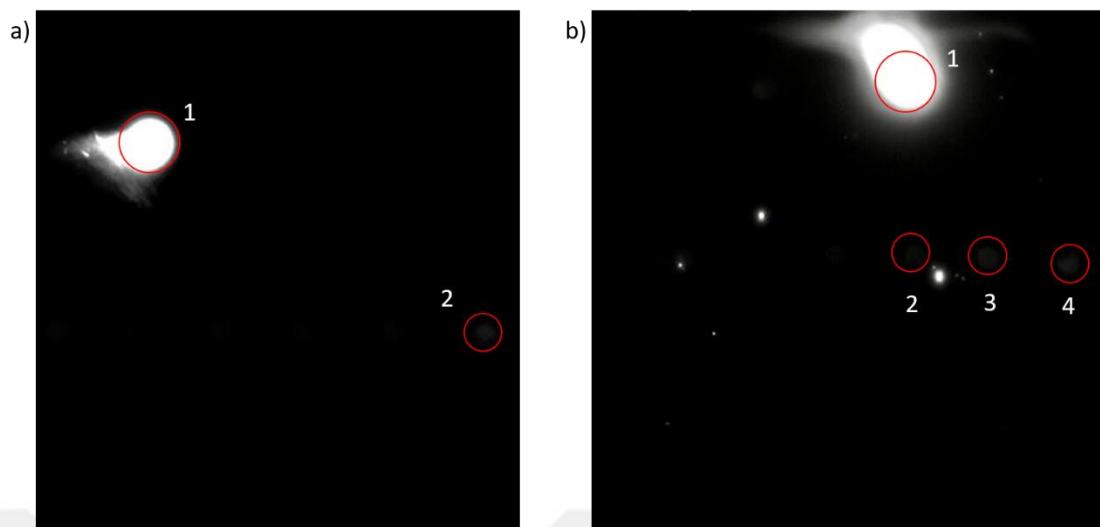


Figure 11: The dot blot image for IVT transfected cell lysates. a) Anti-N antibody was used. Circle 1 is N positive control (200 ng) and circle 2 is 977 mRNA transfection lysate. b) Anti-RBD antibody was employed. Circle 1, 2, 3 and 4 are RBD positive control (200 ng), 974 mRNA transfection lysate, 976s mRNA transfection lysate and 977 mRNA transfection lysate, respectively..

3.2 Synthetic 5' UTR Library Creation for Differential Expression in Healthy and Cancer Breast Cells

The 5' untranslated region (5' UTR) of an mRNA molecule plays a pivotal role in gene expression regulation. It serves as the docking site for the translational machinery, influencing translation initiation and efficiency. Additionally, the structural motifs and high translation rates within the 5' UTR can impact transcript stability. In the field of gene expression, transcript abundance or the mean ribosomal load of a transcript is often used to predict gene expression levels. In this thesis, a novel hypothesis is proposed: the protein and transcript quantities of a given gene are linked through their

interaction with the 5' UTR, indicating intrinsic information about the quality of this region. In simpler terms, if a gene exhibits high protein levels with relatively low transcript amounts, this may be attributed to the characteristics of its 5' UTR (Figure 12). While other factors, such as protein stability and the effects of the 3' UTR, play a role, a comprehensive genome-wide analysis with a large sample size can minimize these influences.

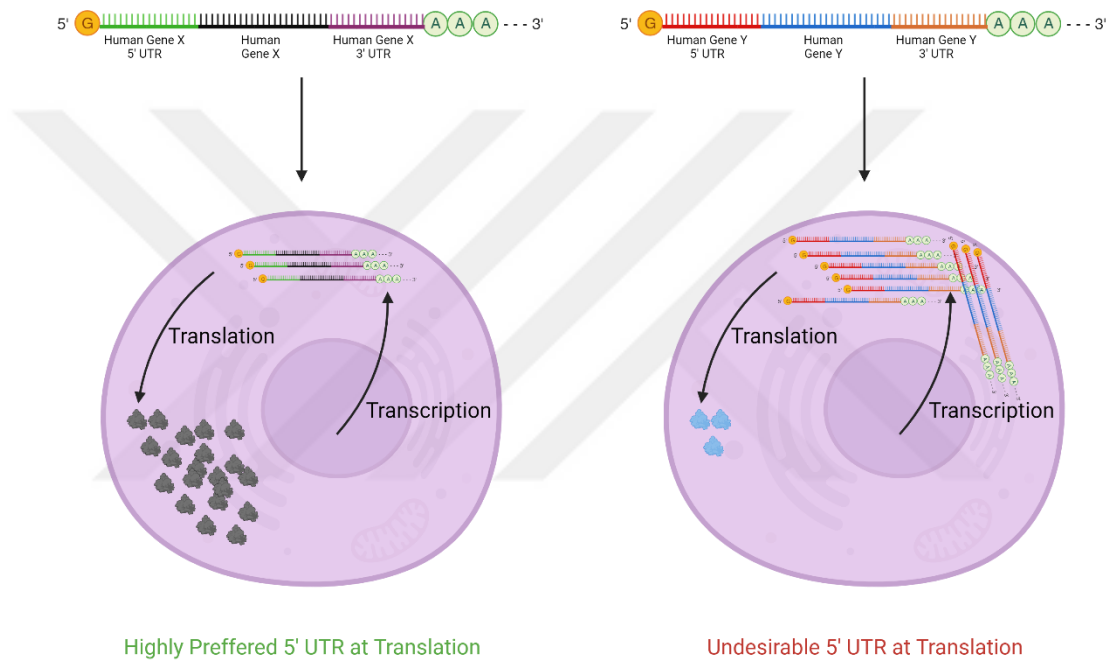


Figure 12: The illustration of scoring idea for human 5' UTRs.

To test this hypothesis, a scoring table was constructed using human tissue-based transcriptomic and proteomic data. This scoring approach aimed to highlight differences in expression profiles across different genes. Following this, the approach was applied to compare the scores of the same gene in two distinct cell lines. Specifically, the focus was on differentially expressed genes' 5' UTRs in cancerous and healthy breast cell lines (MCF7 and MCF10A). The selected 5' UTRs demonstrated a

high tendency for increased translation with low transcript levels in cancer cells, as opposed to healthy cells. The consensus sequences within these genes were identified and employed to create a synthetic library of 5' UTRs.

Through this methodology, the developed synthetic 5' UTRs can introduce a novel dimension to mRNA-based cancer vaccines. These synthetic 5' UTRs have the potential to selectively enhance the production of desired proteins solely in cancer cells or to significantly reduced production in healthy cells. This approach could revolutionize cancer vaccine design by providing a way to target cancer-specific protein expression, while minimizing the impact on healthy cells.

3.2.1 Initial Analysis of Human 5' UTRs

The transcriptomic data used for human genes were obtained from the Genotype-Tissue Expression (GTEx) database, which categorizes gene expression in various human tissues. The data were represented in transcript per million (TPM), a normalized unit that accounts for variations in transcript length and quantity. This normalization ensures that all transcripts are brought to the same scale for accurate comparison.

On the other hand, proteomic data were gathered in the form of IBAQ values, which reflect protein abundance in different tissues. By calculating the ratio of proteomic values to transcriptomic TPM values (IBAQ/TPM), specific scores for different genes were derived (Table 1). These scores provide a measure of how efficiently a gene's transcript leads to protein production within a specific tissue context.

Table 1: The scoring of chosen human 5' UTR with their corresponding plasmid names.

The total number of the genes scored was 7791.

Plasmid Name	Gene	Protein Amount in Brain Tissue (Normalized iBAQ Value)	Transcript Amount in Brain Tissue (All Cell Types Average) (Gene Median TPM)	Score (iBAQ / TPM)
AH1	APOA2	884245488.5	0.23498035	3763061477
AH3	LYZ	16845607.3	2.02699231	8310641.948
AH5	RNF219	37152.75649	8.80760769	4218.257419
AH7	HIST1H4H	26907944185	2.40116231	11206216297
AH9	PGM2	218902924.4	3.95590692	55335711.55
922	HBB	66105101250	241.470231	273760873.3

With the intention of creating a candidate of 5' UTR for potential vaccine production, 5' UTRs of various lengths ranging from 30 to 600 nucleotides were considered to ease mRNA production. These chosen 5' UTRs were assessed based on their iBAQ/TPM scores, allowing for the selection of genes with varying translation efficiencies. In total, six human 5' UTRs were selected, representing a range of translation efficiencies from very high to very low, which would be used in further studies for potential mRNA-based vaccine design.

3.2.2 Cloning of Chosen UTRs

To construct the new mRNA expression plasmids containing different 5' UTRs, the backbone plasmid 922, which harbored the human β globulin gene 5' UTR followed by the sfGFP coding region, was utilized. However, as the poly(A) tail region of this plasmid proved challenging to amplify through PCR without losing repetitive nucleotides, a different strategy was employed.

The process involved digesting the 922 plasmid with KpnI and XbaI restriction enzymes to remove both the 5' UTR and a specific segment of the sfGFP coding region (Figure 13). This step helped prevent the unnecessary amplification of the poly(A) tail region during subsequent PCR steps.

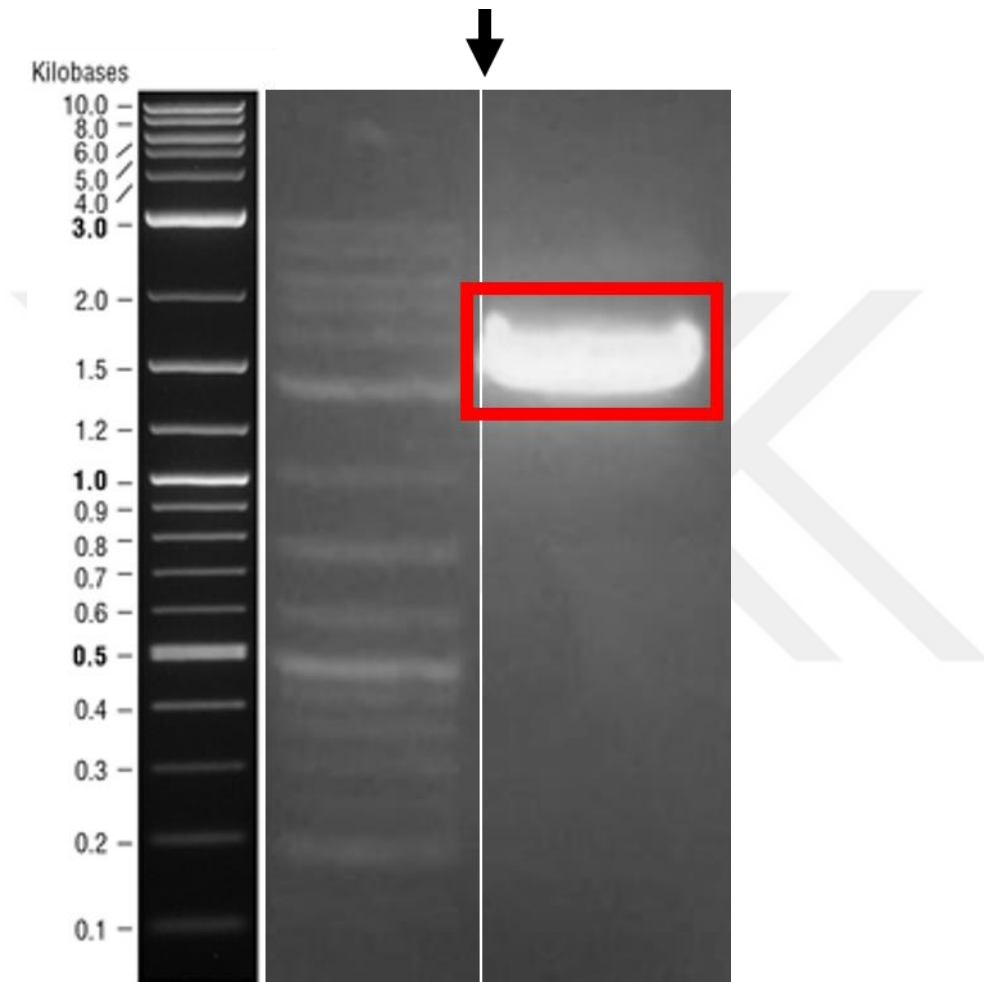


Figure 13: The agarose gel image of 922 digestion. Expected band was 3519 bps.

To insert the new 5' UTRs, PCR amplifications were carried out using Q5 DNA polymerase. Overhangs were incorporated into the primers to facilitate the cloning process. Due to the varying lengths of the chosen 5' UTRs, two consecutive PCR reactions were conducted to achieve full-length insertion (Figure 14). The resulting

amplicons were subjected to KpnI and XbaI digestions to generate compatible ends for the subsequent ligation step.



Figure 14: The second PCR products to clone the chosen 5' UTRs. Lane 1, 2, 3, 4 and 5 are the amplicons for AH1, AH3, AH5, AH7 and AH9, respectively. Expected bands were approximately 1100 bps for all reactions.

The plasmid fragments (cut backbone and amplicons) were then ligated using T4 ligase. The ligation products were introduced into *E. coli* PRO cells through chemical transformation. Following an overnight incubation, colonies on agar plates containing the antibiotic ampicillin were examined. The presence of the desired DNA inserts was verified through colony PCR using PFU DNA polymerase, which is produced in our laboratory (Figure 15).

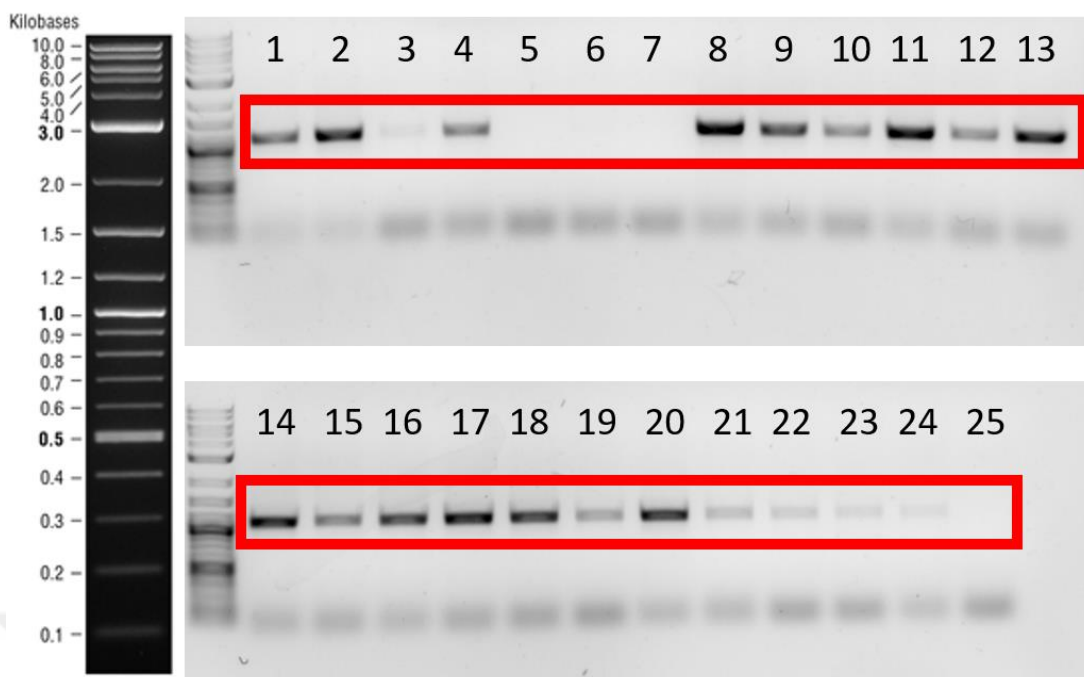


Figure 15: The colony PCR results for AH1, AH3, AH5, AH7 and AH9 constructs. 5 colonies were chosen for each construct (Lanes 1-5 AH1, lanes 6-10 AH3, lanes 11-15 AH5, lanes 16-20 AH7 and lanes 21-25 AH9). Expected bands were nearly 1100 bps.

To ensure the accuracy of the plasmid constructs, the plasmids from selected colonies were subjected to next-generation sequencing. This sequencing process provided a detailed analysis of the DNA sequence to confirm the successful insertion of the desired 5' UTRs and other components (Appendix D).

3.2.3 Expression Profiles at HEK293T Cells

The next steps in the experimental process involved the preparation and validation of IVT mRNAs containing the chosen 5' UTRs (AH1, AH3, AH5, AH7, AH9) and the backbone (922) for cell line transfection experiments.

The sequence-verified plasmids AH1, AH3, AH5, AH7, AH9, and 922 were digested using HindIII and SapI restriction enzymes to create linear templates suitable for the *in vitro* transcription (IVT) reaction. The ARCA IVT kit was then utilized to produce the

mRNAs from these templates. Notably, certain downstream modifications such as cap1 structure addition, dsRNA removal, and phosphatase treatment were omitted in this context since these mRNAs were intended for cell line experiments and not for vaccination studies.

To ensure the successful production of the desired mRNAs, an agarose gel electrophoresis was performed, confirming their lengths and verifying their presence (Figure 16). Subsequently, these mRNAs were used in transfection experiments involving the HEK293T cell line.

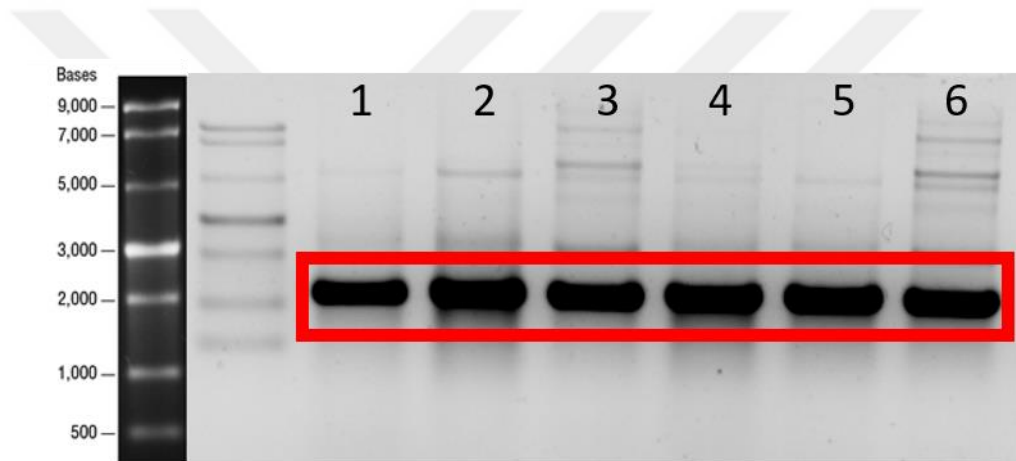


Figure 16: IVT reaction results for 922 (lane 1), AH1 (lane 2), AH3 (lane 3), AH5 (lane 4), AH7 (lane 5) and AH9 (lane 6). Expected bands are approximately 1050 bps.

Images of the cells post-transfection were captured using a fluorescence microscope to observe the expression of the fluorescent protein encoded by the transfected mRNAs (Figure 17).

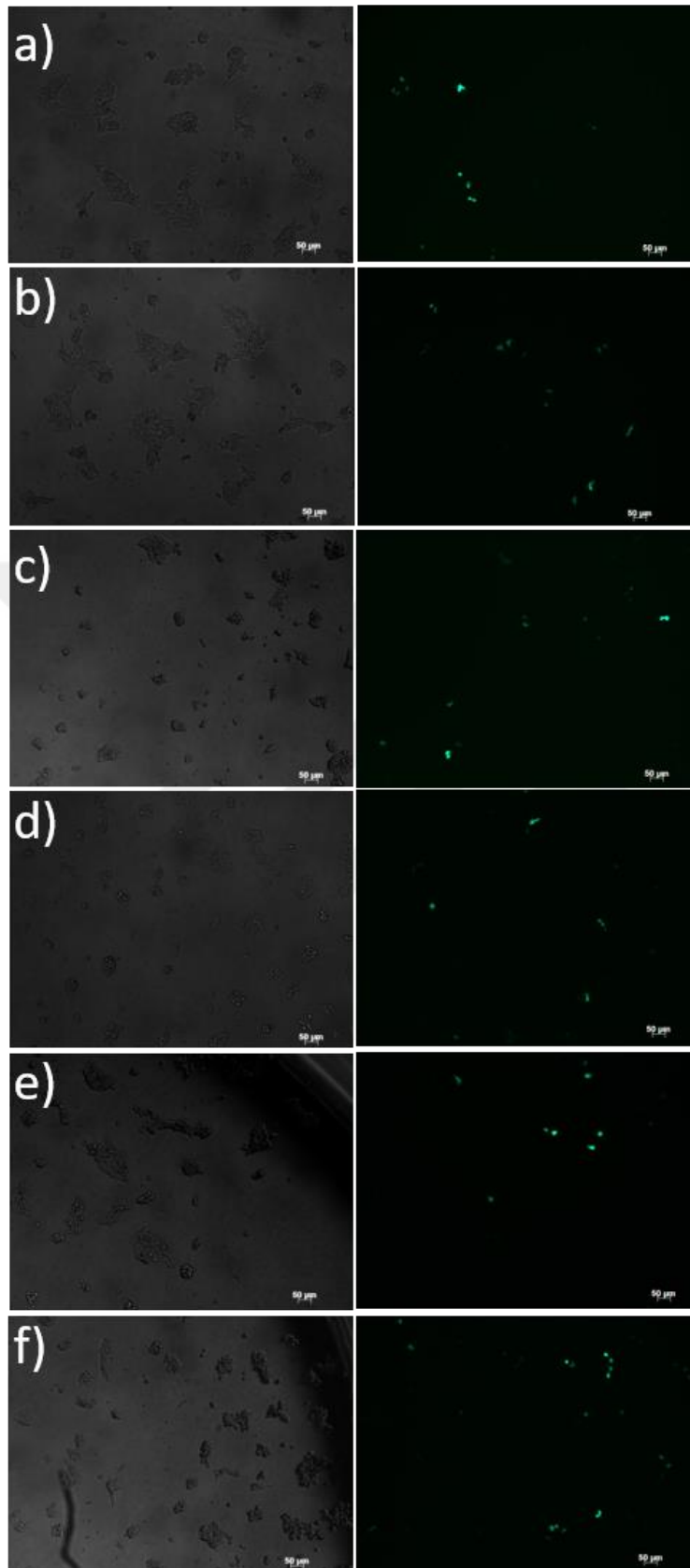


Figure 17: The transfection of HEK293T cells with IVT mRNAs with chosen UTRs. In every section, left image is the bright field visualization of the cells, right image is

the fluorescence version of the same location at the left image. a) 922. b) AH1. c) AH3. d) AH5. e) AH7. f) AH9.

In order to quantitatively assess the translation efficiency and success of the different mRNAs with distinct 5' UTRs, flow cytometry analysis was employed (Figure 18). This analysis provided crucial data in the form of mean fluorescence values. The scattered fluorescence values among the cell population were indicative of the heterogeneous distribution of the transfected mRNAs. However, the mean fluorescence value of the transfected cells gave insight into the overall translation levels and could serve as a metric to compare the performance of different UTRs.

The Percentage of Expression

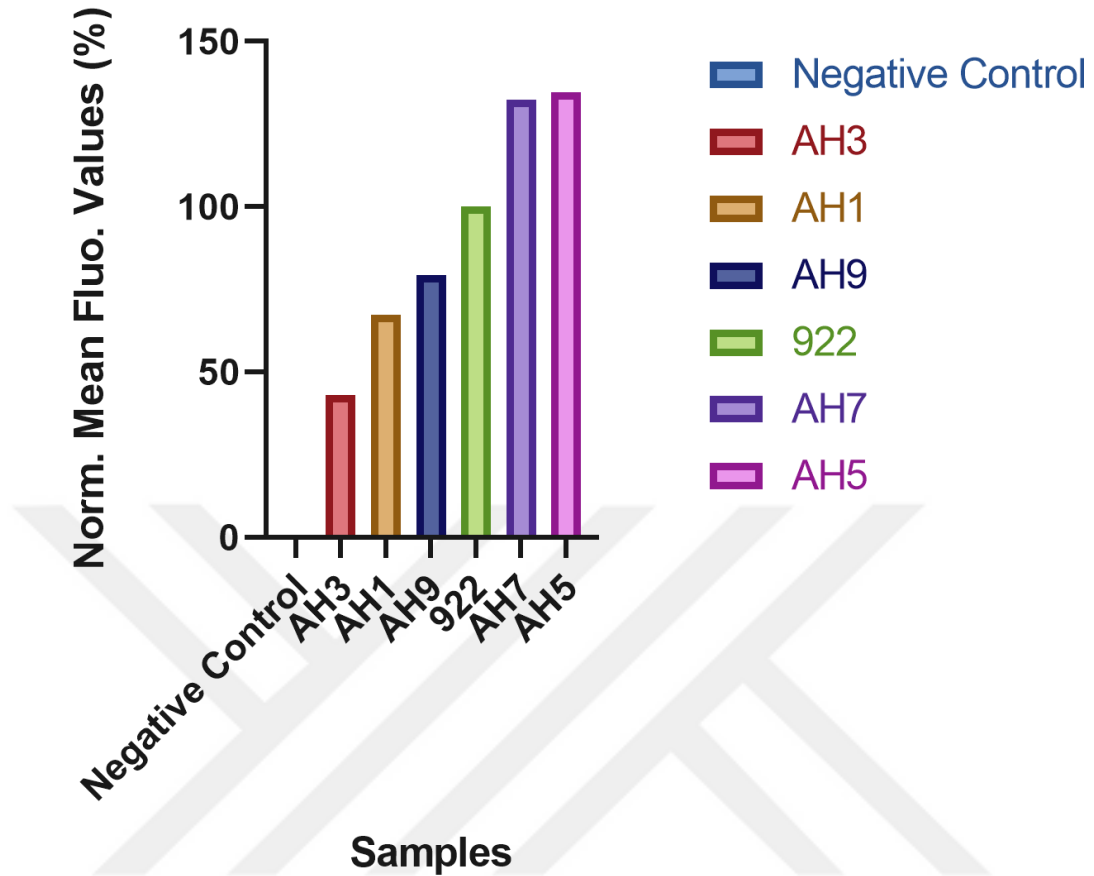


Figure 18: The expression percentage of sfGFP obtained from flow cytometry for chosen UTRs. The mean fluorescence values are normalized between 0% (negative control) to 100% (922) (n=3).

HEK293T cells are known for their origin as a heterogeneous mixture of cell types, including fibroblasts, endothelial cells, and epithelial cells. However, their exposure to adenovirus 5 DNA during transfection tends to induce expression profiles resembling neuronal cells and can result in cytogenetic instability that influences gene expression[71]. Therefore, a direct comparison of the HEK293T cell experimental data with the scoring derived from neuronal tissues might lead to misinterpretation (Table 2).

Table 2: The accuracy of the expression profiles of chosen 5' UTRs.

Plasmid Name	Gene	Theoretical Magnitude of Expression	Experimental Magnitude of Expression	Fitness of Theoretical to Experimental Data
AH7	HIST1H4H	Very High	Very High	Match
AH1	APOA2	High	Medium	No match
922	HBB	Medium	Medium	Match
AH9	PGM2	Medium	Medium	Match
AH3	LYZ	Low	Low	Match
AH5	RNF219	Very Low	Very High	No match

Recognizing these challenges, adjustments were made to the scoring methodology. Rather than categorizing based on tissues, the scoring was made specific to cell lines to account for the cellular mechanisms of translation. This refinement aimed to enhance the accuracy of the conclusions drawn from the experimental data.

3.2.4 Shadowing Effect Analysis to Shift mRNA Usage to DNA

Usage

The consideration of manufacturing costs between DNA and IVT mRNA, driven by the price of ARCA kits, as well as the aim to generate a high quantity of library elements for differential expression, prompted a shift in the choice of nucleic acid type for transfection experiments. The key criterion for this transition was ensuring that the transcripts derived from DNA remained equivalent to IVT mRNA counterparts. As a result, the transcription start site of the CMV promoter, known for its activity in mammalian cells, was pinpointed through a combination of the ElemeNT tool and various pertinent studies. Subsequently, a plasmid containing the CMV promoter followed by the identical regions found at the transcribed location of 922 was constructed, bearing the name CMV7.

In the context of protein production from DNA templates, the process necessitates passage through the cell nucleus for transcription. The cumulative amount of transcribed mRNA from DNA is intricately linked to the potency of the promoter. These intricacies, which stem from the transcriptional machinery, are referred to as the "shadowing effect." With the quantity of IVT mRNA for cell transfection experiments already determined, the challenge lay in ascertaining the total amount of DNA that would yield similar production efficiency within the cells.

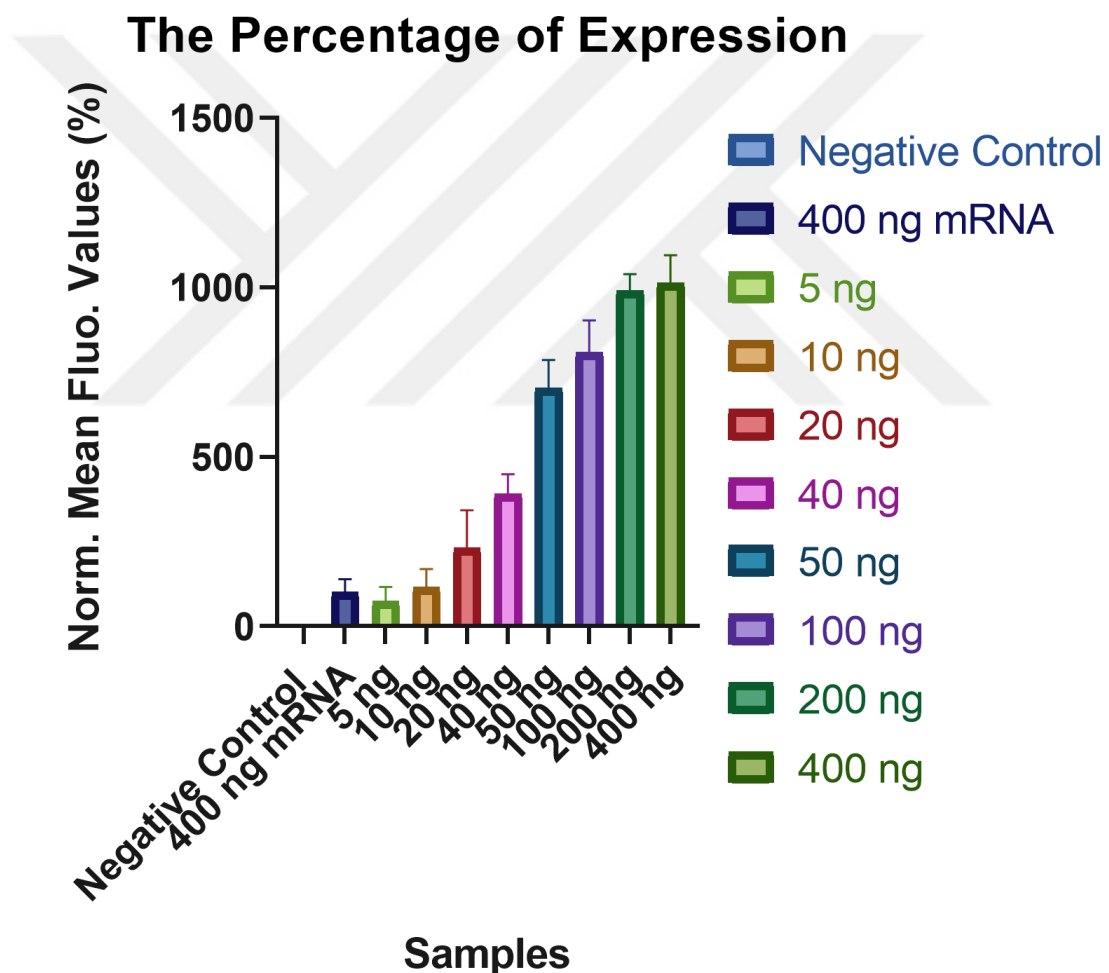


Figure 19: The expression percentage of sfGFP obtained from flow cytometry for mRNA and DNA transfections. The mean fluorescence values are normalized between 0% (negative control) to 100% (400 ng 922 mRNA) (n=3).

To address this, various quantities of DNA were employed for transfection into HEK293T cells, alongside 400 ng of the 922 IVT mRNA. The mean fluorescence values resulting from flow cytometry analysis were normalized to establish a baseline of 0% for the negative control and 100% for the 922 mRNA. As depicted in Figure 19, transfecting 10 ng of CMV7 DNA yielded an equivalent effect to that of 922 mRNA transfection. Consequently, for the upcoming experiments following this thesis, 10 ng of library elements will be the designated quantity for transfection.

3.2.5 Final Analysis of Human 5' UTRs and Consensus Production

We developed two R scripts and two corresponding R functions to identify specific patterns within the 5' UTRs of human genes, which could potentially contribute to differential expression between cancerous and healthy breast cells (Appendix F). To achieve this, we aligned combined proteomic and transcriptomic data for human genes with their respective 5' UTR sequences. Each gene was assigned a total ratio value, indicating its degree of differential expression. This value was computed by dividing the ratio of protein abundance in MCF7 cells by the transcript amount in MCF7 cells, by the ratio of protein abundance in MCF10A cells by the transcript amount in MCF10A cells. This approach enabled us to pinpoint genes with highly active translational capacity in MCF7 cells and decreased translational capacity in MCF10A cells, as hypothesized (Table 3). The threshold for this distinction was set at three times greater translational capacity in MCF7 cells compared to MCF10A cells.

Table 3: The differential expression scoring of the human 5' UTRs between MCF7 and MCF10A cells. The total gene amount obtained was 4496.

gene	mcf10a_mean_normalized_expression	mcf7_mean_normalized_expression	mcf10a_tpm	mcf7_tpm	mcf7ratio	mcf10aratio	totalratio
VIM	6.078574	3.797938	579.6007	0.841014502	4.515900175	0.010487519	430.5976
TGM1	2.616728	3.568119	413.2087	3.516938633	1.014552538	0.006332702	160.2085
KRT16	6.258502	4.249191	281.5611	2.44618301	1.737069802	0.022227863	78.1483
CD82	4.277298	4.751306	204.2952	3.03870011	1.563598281	0.020936848	74.68165
PPP1R14C	5.261004	3.840646	79.70097	0.877000456	4.379297397	0.066009281	66.34366
ANXA1	7.177533	3.909239	1245.05	11.99520857	0.325900037	0.005764856	56.5322
LAMB1	4.463888	4.772627	169.0121	3.210748881	1.486452882	0.02641165	56.2802
RCN1	6.379154	6.327671	261.7868	5.243546372	1.206754001	0.024367746	49.52259
TUBA4A	7.022789	4.161827	126.5608	1.916408712	2.17168018	0.055489445	39.13682
PGM1	6.036758	4.867554	143.3586	2.978888466	1.634017023	0.042109502	38.804
CPS1	3.182886	3.158809	19.42512	0.539679976	5.853115606	0.163854108	35.72151
CAPN2	5.946132	4.198622	655.6532	15.60967868	0.26897555	0.00906902	29.65872
TPM2	6.481121	6.047446	259.8414	9.743061147	0.620692577	0.024942605	24.88483
ITGA5	4.628938	3.902561	140.0644	4.780201318	0.816400995	0.033048632	24.70302

Next, we conducted a consensus analysis on the 5' UTR sequences of the genes surpassing this threshold (Figure 20). This allowed us to identify recurring patterns of various lengths within this subset of UTRs. Subsequently, each pattern was systematically counted within the 5' UTRs of the genes.

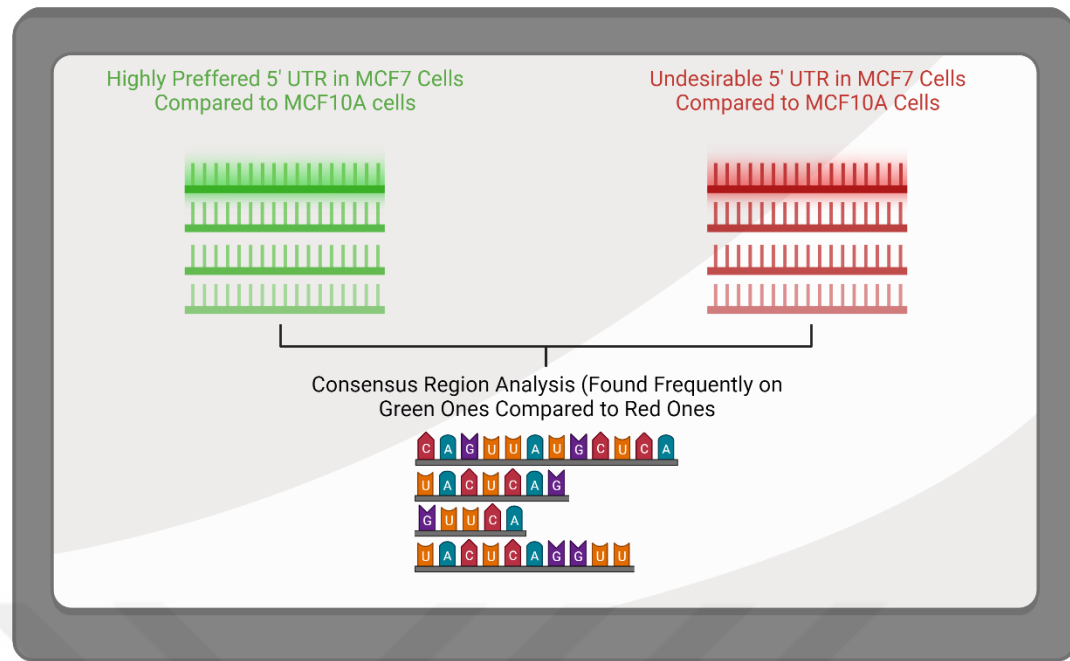


Figure 20: The illustration of the consensus region analysis according to the differential expression scores of human 5' UTRs.

For enhanced insight, we divided the counts into two categories: those present above and below the total ratio value of 1. The "above" region encompassed genes that exhibited a heightened translational capacity preference in MCF7 cells, whereas the "below" region indicated a bias towards MCF10A cells. To ascertain the significance of these findings, we applied Fisher's Exact Test, a statistical method, to demonstrate that the consensus sequences, formed based on translational activity tendencies in distinct cell lines, were not occurring randomly. Furthermore, we employed Odd's Ratio analysis to quantify the magnitude of likelihood of a sequence being favored by the translational machinery in MCF7 cells over MCF10A cells.

Table 4: Consensus sequence analysis of the scored human 5' UTRs. Total number of found patterns was 433.

pattern	len	countinchosen	inchosenrest	countNOTinchosen	notinchosenrest	probabilityinchosen	probabilityNOTinchosen	differentiality_by_hand	odds	p_val	GC_content
CTTTAAATG	9	10	2045	1	2440	0.488997555	0.040983607	11.93154034	11.92661122	0.00372897	22.22222
GGAGAGGGAG	10	10	2045	1	2440	0.488997555	0.040983607	11.93154034	11.92661122	0.00372897	70
TTGCGCGCG	9	10	2045	1	2440	0.488997555	0.040983607	11.93154034	11.92661122	0.00372897	77.77778
GTCCAGGTG	9	9	2046	1	2440	0.439882698	0.040983607	10.73313783	10.7287593	0.00729181	66.66667
GCGGAGTTC	9	9	2046	1	2440	0.439882698	0.040983607	10.73313783	10.7287593	0.00729181	77.77778
GCGTGAGCC	10	9	2046	1	2440	0.439882698	0.040983607	10.73313783	10.7287593	0.00729181	80
CGCTGAGCG	10	9	2046	1	2440	0.439882698	0.040983607	10.73313783	10.7287593	0.00729181	80
CCGGAGGTGG	10	9	2046	1	2440	0.439882698	0.040983607	10.73313783	10.7287593	0.00729181	80
TTGGCACT	8	9	2046	1	2440	0.439882698	0.040983607	10.73313783	10.7287593	0.00729181	50
TCACCTCA	8	16	2039	2	2439	0.784698382	0.08200082	9.569396763	9.56551998	0.00019238	50

As a culmination of these comprehensive analyses, our observations indicated that certain consensus sequences with a higher likelihood of expression in MCF7 cells compared to MCF10A cells were not random occurrences. Instead, they exhibited an elevated probability of being identified in the 5' UTRs of genes favored by the translational machinery in MCF7 cells (Table 4).

3.2.6 Random Assembly PCR and Library Creation

The primers designed using the chosen consensus sequences, as described in section 2.15, were utilized in the process of random assembly PCR. Figure 21 illustrates the smear bands and their shifts in length between the initial 7 cycles of PCR and the subsequent 30 cycles. Notably, the composition of assembled amplicons varied during the initial 7 cycles of PCR, resulting in the creation of distinct pools of synthetic 5' UTR sequences.

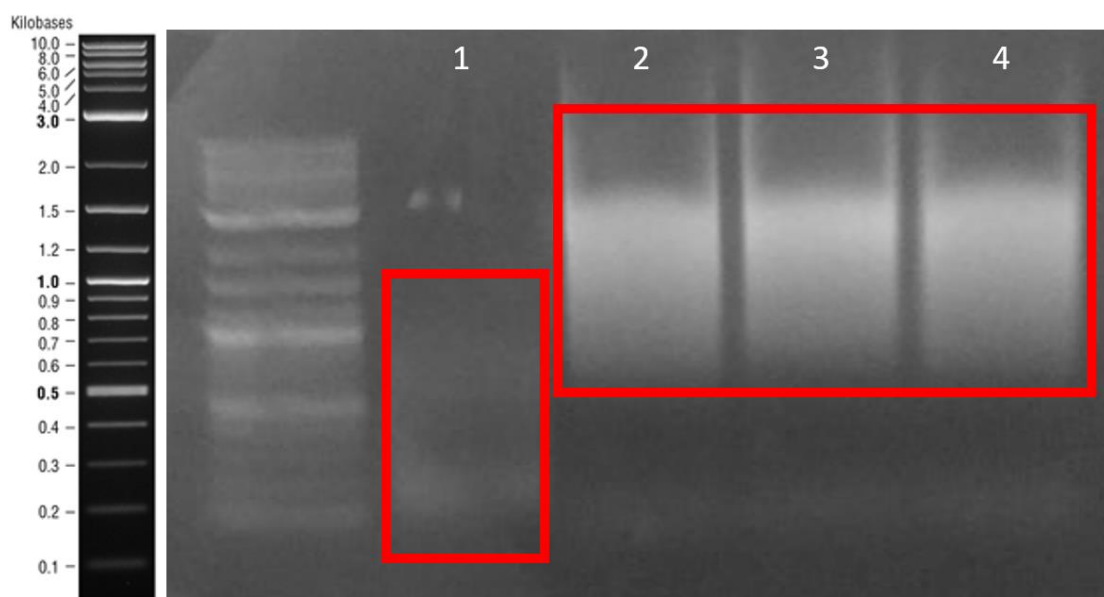


Figure 21: The agarose gel image of random assembly PCR. Lane 1 is the result of first 10 cycles whereas lane 2, 3 and 4 are the end reactions of different pooling attempts. The expected bands were smear due to various lengths of amplicons.

The most significant challenge encountered during the library element creation was the removal of the 60-nucleotide-long human β -globulin 5' UTR from the 6135-nucleotide-long CMV7 backbone due to the disparities in their lengths. The single-cut version of the plasmid did not differ significantly from the double-cut version (Figure 22). Attempts were made to ligate the double-cut version with the digested smear, yielding nearly equal numbers of colonies in both the expected and the negative control plates. The results of colony PCR from the expected plate colonies revealed self-ligation of the backbone alone (Figure 23). To address this, the CMV7 plasmids were subjected to digestion with BamHI-PstI and EcoRI-PstI (Figure 22). Bands lacking the 60-nucleotide sequence were isolated, and an attempted ligation was performed with these bands and the digested smear. Unfortunately, no colonies were observed as a result.

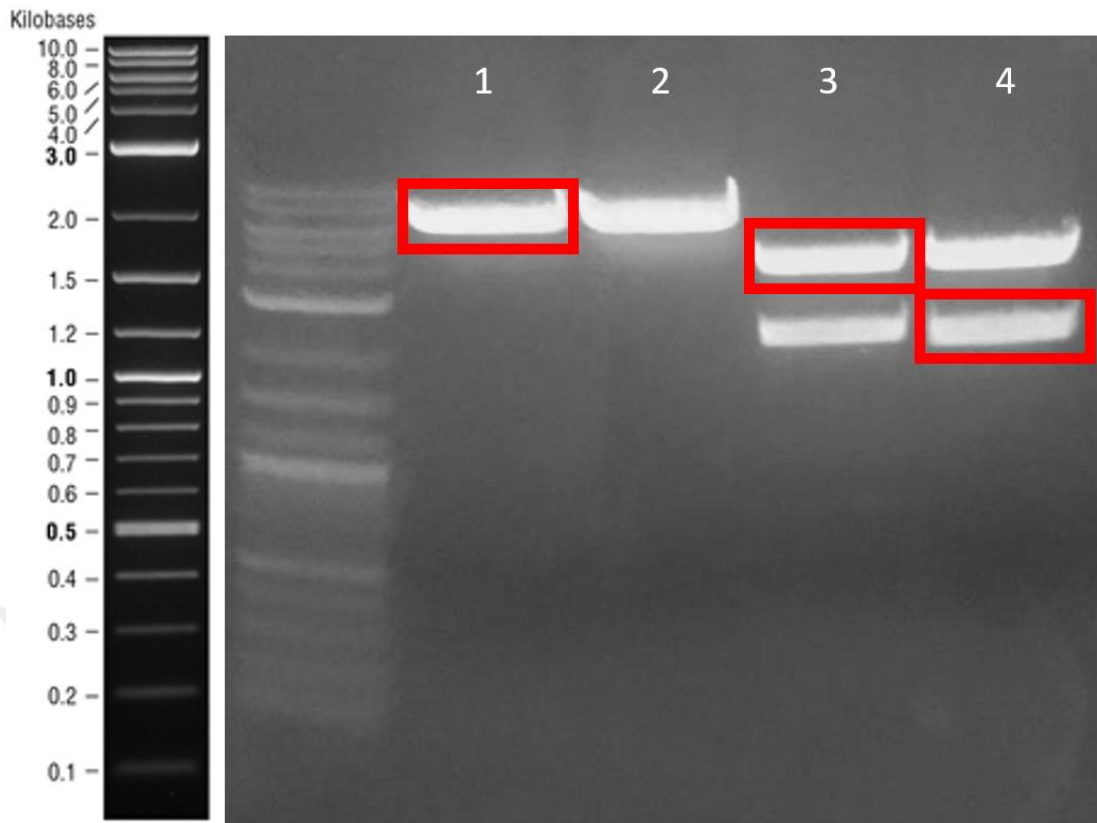


Figure 22: The agarose gel image of the CMV7 digestion. BamHI-EcoRI digestion (lane 1), only BamHI digestion (lane 2), EcoRI-PstI digestion (lane 3) and BamHI-PstI digestion (lane 4) were executed. The expected bands were 6139 (lane 1), 6195 (lane 2), 3959 and 2236 (lane 3) and 4015 and 2180 (lane 4) bps.

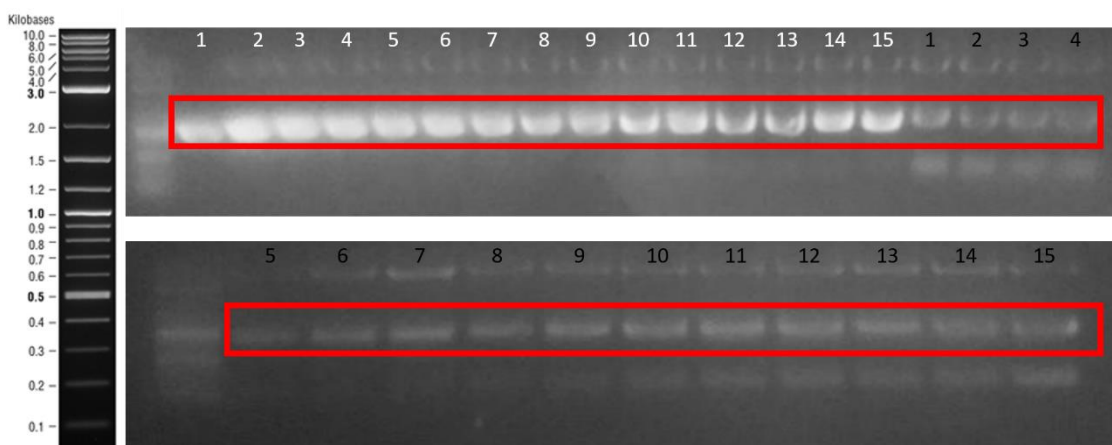


Figure 23: The colony PCR results for library creation with ligation. Lanes numbered with white color show the presence of sfGFP whereas black colored ones demonstrate

the presence of human β globulin 5' UTR. Lane 1 is CMV7 positive control. Lanes 2-7 are the colonies from negative control plate. Lanes 8-15 are the colonies from expected plate. Expected bands for both conditions were approximately 1020 bps (small increase for black colored ones).

To overcome this challenge, the incubation time for the CMV7 double digestion was extended, and the resulting product was used in a Gibson assembly reaction with the uncut smear. Ultimately, the negative control plate yielded no colonies, while the expected plate contained 52 distinct colonies. To confirm the presence of the sfGFP region without the human β -globulin 5' UTR in the possible library elements from the expected plate, two separate colony PCR reactions were conducted. Through the strategic choice of primer pairs, one reaction indicated the presence of sfGFP, while the other revealed the absence of the human β -globulin 5' UTR. This distinction was possible because the forward primers of these pairs specifically interacted either with the beginning of sfGFP or the UTR region.

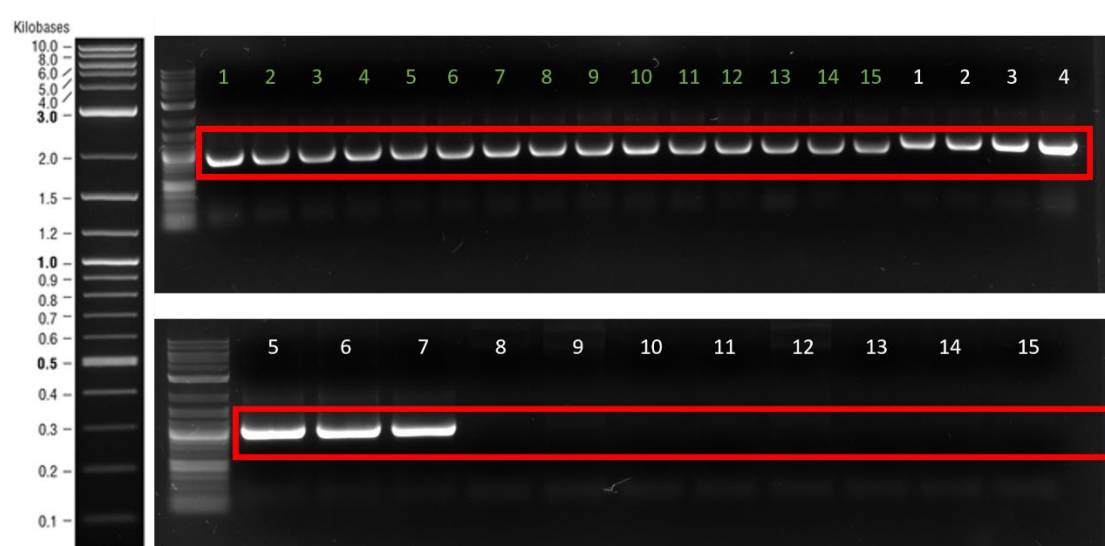


Figure 24: The colony PCR results for library creation with Gibson assembly. Lanes numbered with green color show the presence of sfGFP whereas white colored ones

demonstrate the presence of human β globulin 5' UTR. Lane 1 is CMV7 positive control. Lanes 2-7 are the colonies from negative control plate of ligation. Lanes 8-15 are the colonies from expected plate. Expected bands for both conditions were approximately 1020 bps (small increase for white colored ones).

Furthermore, to assess the success of this approach, six negative plate colonies from the initial ligation experiment were included. Additionally, a selection of eight randomly chosen library elements underwent analysis, confirming the successful creation and integration of the synthetic 5' UTRs. The disappearance of the bands demonstrated it (Figure 24). Subsequently, 52 colonies containing library elements were cultivated, and plasmid isolations were performed to enable further investigations through MCF7 and MCF10A transfection experiments, as well as flow cytometry analyses.

CHAPTER IV: CONCLUSION

mRNA-based therapeutics have gained extensive traction as versatile candidates for a wide spectrum of applications, encompassing studies involving protein replacement, genetic reprogramming, and the establishment of novel vaccine platforms targeting infectious diseases and cancer. The emergence of the SARS-CoV-2 pandemic underscored the potential prowess of mRNA in vaccine production. In contrast to DNA-based, attenuated virus, and adenoviral vaccines, the inherent attributes of mRNA molecules and their interaction dynamics with cells confer distinct advantages to mRNA-based vaccine strategies. Despite nearly three decades having passed since the initial exploration of externally introduced mRNA translation in cells, the complete potential of mRNA-based methodologies remained somewhat latent due to limited insights into the manipulation and modulation of this molecule. Precisely targeting mRNA to specific bodily locales or orchestrating its behavior upon cell entry constitute pivotal prerequisites for unlocking its manifold applications.

The scope of this thesis was anchored in elucidating post-cellular entry mRNA regulations, with a particular emphasis on the coding regions. The primary objective of the initial segment was to engineer an mRNA vaccine candidate for SARS-CoV-2 endowed with dual-antigenic properties. This involved the successful creation of a single mRNA molecule housing combined antigens from distinct SARS-CoV-2 compartments. Western and dot blot analyses corroborated the interaction of this composite antigen with two disparate antibodies concurrently. This interaction highlighted the potential for eliciting a dual immune response, suggesting its viability as a vaccination agent. Moreover, the techniques crucial for evoking a robust immune

response in vaccination trials were practiced, in anticipation of subsequent animal experimentation.

The subsequent facet of this thesis proposed an innovative methodology to discern between cell types. Capitalizing on the mRNA molecule's ribosome docking centers located in the 5' UTRs, we devised a scoring framework to identify genes displaying elevated translational capacity despite their relatively low transcript abundance. By comparing these scores between cancerous and healthy breast cells, we aimed to ascertain whether common sequences found in the 5' UTRs of genes being highly active in cancer cells rather than healthy cells demonstrated a proclivity that was caused by maybe the change of translational machinery of a cancer cell. We unearthed varying lengths of sequences preferred by cancer cells compared to their healthy counterparts, which served as the foundation for constructing a synthetic 5' UTR library. The forthcoming step involves examining the differential expression capabilities of these sequences in cancer and healthy cells.

In conclusion, this research delved into the regulatory domains of mRNA molecules, offering fresh insights for advancing mRNA-based therapeutics. Enhanced control over the tunability of mRNA therapies holds the promise of specialized treatments for a spectrum of diseases, ranging from cancer to viral infections.

BIBLIOGRAPHY

- [1] F. Crick, Central Dogma of Molecular Biology, *Nature*. 227 (1970) 561–563.
<https://doi.org/10.1038/227561a0>.
- [2] A. Liljas, Messenger RNA, in: S. Maloy, K. Hughes (Eds.), *Brenners Encycl. Genet.* Second Ed., Academic Press, San Diego, 2013: pp. 369–370.
<https://doi.org/10.1016/B978-0-12-374984-0.00929-3>.
- [3] J. Guo, Transcription: the epicenter of gene expression, *J. Zhejiang Univ. Sci. B*. 15 (2014) 409–411. <https://doi.org/10.1631/jzus.B1400113>.
- [4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, From DNA to RNA, in: *Mol. Biol. Cell* 4th Ed., Garland Science, 2002.
<https://www.ncbi.nlm.nih.gov/books/NBK26887/> (accessed August 13, 2023).
- [5] J. Feher, 2.2 - DNA and Protein Synthesis, in: J. Feher (Ed.), *Quant. Hum. Physiol.* Second Ed., Academic Press, Boston, 2017: pp. 120–129.
<https://doi.org/10.1016/B978-0-12-800883-6.00011-2>.
- [6] H.S. Bernhardt, The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a, *Biol. Direct*. 7 (2012) 23.
<https://doi.org/10.1186/1745-6150-7-23>.
- [7] Ribonucleic Acid (RNA), *Genome.Gov.* (n.d.).
<https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid> (accessed August 23, 2023).
- [8] Defining the Genetic Coding Problem, 1954-1957, Francis Crick - *Profiles Sci.* (2019). <https://profiles.nlm.nih.gov/spotlight/sc/feature/defining> (accessed August 23, 2023).

- [9] R. Carter, G. Drouin, Structural differentiation of the three eukaryotic RNA polymerases, *Genomics*. 94 (2009) 388–396. <https://doi.org/10.1016/j.ygeno.2009.08.011>.
- [10] Noncoding RNA and Gene Expression | Learn Science at Scitable, (n.d.). <https://www.nature.com/scitable/topicpage/regulation-of-transcription-and-gene-expression-in-1086/> (accessed August 23, 2023).
- [11] J. Soutourina, Transcription regulation by the Mediator complex, *Nat. Rev. Mol. Cell Biol.* 19 (2018) 262–274. <https://doi.org/10.1038/nrm.2017.115>.
- [12] T.W.R. Kelso, K. Baumgart, J. Eickhoff, T. Albert, C. Antrecht, S. Lemcke, B. Klebl, M. Meisterernst, Cyclin-Dependent Kinase 7 Controls mRNA Synthesis by Affecting Stability of Preinitiation Complexes, Leading to Altered Gene Expression, Cell Cycle Progression, and Survival of Tumor Cells, *Mol. Cell. Biol.* 34 (2014) 3675–3688. <https://doi.org/10.1128/MCB.00595-14>.
- [13] A.R. Jurado, D. Tan, X. Jiao, M. Kiledjian, L. Tong, Structure and Function of Pre-mRNA 5'-End Capping Quality Control and 3'-End Processing, *Biochemistry*. 53 (2014) 1882–1898. <https://doi.org/10.1021/bi401715v>.
- [14] T. Schlake, A. Thess, M. Fotin-Mleczek, K.-J. Kallen, Developing mRNA-vaccine technologies, *RNA Biol.* 9 (2012) 1319–1330. <https://doi.org/10.4161/rna.22269>.
- [15] D. Weissman, mRNA transcript therapy, *Expert Rev. Vaccines*. 14 (2015) 265–281. <https://doi.org/10.1586/14760584.2015.973859>.
- [16] A. Ramanathan, G.B. Robb, S.-H. Chan, mRNA capping: biological functions and applications, *Nucleic Acids Res.* 44 (2016) 7511–7526. <https://doi.org/10.1093/nar/gkw551>.

- [17] M. Ziemniak, M. Strenkowska, J. Kowalska, J. Jemielity, Potential therapeutic applications of RNA cap analogs, *Future Med. Chem.* 5 (2013) 1141–1172. <https://doi.org/10.4155/fmc.13.96>.
- [18] L. Bossi, J.R. Roth, The influence of codon context on genetic code translation, *Nature*. 286 (1980) 123–127. <https://doi.org/10.1038/286123a0>.
- [19] G. Cannarozzi, G. Cannarozzi, N.N. Schraudolph, M. Faty, P. von Rohr, M.T. Friberg, A.C. Roth, P. Gonnet, G. Gonnet, Y. Barral, A role for codon order in translation dynamics, *Cell*. 141 (2010) 355–367. <https://doi.org/10.1016/j.cell.2010.02.036>.
- [20] T. Tuller, Y.Y. Waldman, M. Kupiec, E. Ruppin, Translation efficiency is determined by both codon bias and folding energy, *Proc. Natl. Acad. Sci.* 107 (2010) 3645–3650. <https://doi.org/10.1073/pnas.0909910107>.
- [21] A. Narula, J. Ellis, J.M. Taliaferro, O.S. Rissland, Coding regions affect mRNA stability in human cells, *RNA*. 25 (2019) 1751–1764. <https://doi.org/10.1261/rna.073239.119>.
- [22] L.A. Passmore, J. Collier, Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression, *Nat. Rev. Mol. Cell Biol.* 23 (2022) 93–106. <https://doi.org/10.1038/s41580-021-00417-y>.
- [23] C. Mayr, What Are 3' UTRs Doing?, *Cold Spring Harb. Perspect. Biol.* 11 (2019) a034728. <https://doi.org/10.1101/cshperspect.a034728>.
- [24] K. Leppek, R. Das, M. Barna, Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them, *Nat. Rev. Mol. Cell Biol.* 19 (2018) 158–174. <https://doi.org/10.1038/nrm.2017.103>.
- [25] S.C. Kim, S.S. Sekhon, W.-R. Shin, G. Ahn, B.-K. Cho, J.-Y. Ahn, Y.-H. Kim, Modifications of mRNA vaccine structural elements for improving mRNA

- stability and translation efficiency, *Mol. Cell. Toxicol.* 18 (2022) 1–8.
<https://doi.org/10.1007/s13273-021-00171-4>.
- [26] R.W. Malone, mRNA transfection of cultured eukaryotic cells and embryos using cationic liposomes, *Focus*. 11 (n.d.) 61–66.
- [27] J.A. Wolff, R.W. Malone, P. Williams, W. Chong, G. Acsadi, A. Jani, P.L. Felgner, Direct gene transfer into mouse muscle in vivo, *Science*. 247 (1990) 1465–1468. <https://doi.org/10.1126/science.1690918>.
- [28] Y. Wang, Z. Zhang, J. Luo, X. Han, Y. Wei, X. Wei, mRNA vaccine: a potential therapeutic strategy, *Mol. Cancer*. 20 (2021) 33. <https://doi.org/10.1186/s12943-021-01311-z>.
- [29] K. Karikó, H. Muramatsu, F.A. Welsh, J. Ludwig, H. Kato, S. Akira, D. Weissman, Incorporation of Pseudouridine Into mRNA Yields Superior Nonimmunogenic Vector With Increased Translational Capacity and Biological Stability, *Mol. Ther. J. Am. Soc. Gene Ther.* 16 (2008) 1833–1840. <https://doi.org/10.1038/mt.2008.200>.
- [30] O. Andries, S. Mc Cafferty, S.C. De Smedt, R. Weiss, N.N. Sanders, T. Kitada, N1-methylpseudouridine-incorporated mRNA outperforms pseudouridine-incorporated mRNA by providing enhanced protein expression and reduced immunogenicity in mammalian cell lines and mice, *J. Controlled Release*. 217 (2015) 337–344. <https://doi.org/10.1016/j.jconrel.2015.08.051>.
- [31] U. Sahin, A. Muik, I. Vogler, E. Derhovanessian, L.M. Kranz, M. Vormehr, J. Quandt, N. Bidmon, A. Ulges, A. Baum, K.E. Pascal, D. Maurus, S. Brachtendorf, V. Lörks, J. Sikorski, P. Koch, R. Hilker, D. Becker, A.-K. Eller, J. Grützner, M. Tonigold, C. Boesler, C. Rosenbaum, L. Heesen, M.-C. Kühnle, A. Poran, J.Z. Dong, U. Luxemburger, A. Kemmer-Brück, D. Langer, M. Bexon, S. Bolte, T.

- Palanche, A. Schultz, S. Baumann, A.J. Mahiny, G. Boros, J. Reinholz, G.T. Szabó, K. Karikó, P.-Y. Shi, C. Fontes-Garfias, J.L. Perez, M. Cutler, D. Cooper, C.A. Kyriatsous, P.R. Dormitzer, K.U. Jansen, Ö. Türeci, BNT162b2 vaccine induces neutralizing antibodies and poly-specific T cells in humans, *Nature*. 595 (2021) 572–577. <https://doi.org/10.1038/s41586-021-03653-6>.
- [32] A.B. Vogel, I. Kanevsky, Y. Che, K.A. Swanson, A. Muik, M. Vormehr, L.M. Kranz, K.C. Walzer, S. Hein, A. Güler, J. Loschko, M.S. Maddur, A. Ota-Setlik, K. Tompkins, J. Cole, B.G. Lui, T. Ziegenhals, A. Plaschke, D. Eisel, S.C. Dany, S. Fesser, S. Erbar, F. Bates, D. Schneider, B. Jesionek, B. Sängler, A.-K. Wallisch, Y. Feuchter, H. Junginger, S.A. Krumm, A.P. Heinen, P. Adams-Quack, J. Schlereth, S. Schille, C. Kröner, R. de la Caridad Güimil Garcia, T. Hiller, L. Fischer, R.S. Sellers, S. Choudhary, O. Gonzalez, F. Vascotto, M.R. Gutman, J.A. Fontenot, S. Hall-Ursone, K. Brasky, M.C. Griffor, S. Han, A.A.H. Su, J.A. Lees, N.L. Nedoma, E.H. Mashalidis, P.V. Sahasrabudhe, C.Y. Tan, D. Pavliakova, G. Singh, C. Fontes-Garfias, M. Pride, I.L. Scully, T. Ciolino, J. Obregon, M. Gazi, R. Carrion, K.J. Alfson, W.V. Kalina, D. Kaushal, P.-Y. Shi, T. Klamp, C. Rosenbaum, A.N. Kuhn, Ö. Türeci, P.R. Dormitzer, K.U. Jansen, U. Sahin, BNT162b vaccines protect rhesus macaques from SARS-CoV-2, *Nature*. 592 (2021) 283–289. <https://doi.org/10.1038/s41586-021-03275-y>.
- [33] U. Sahin, K. Karikó, Ö. Türeci, mRNA-based therapeutics — developing a new class of drugs, *Nat. Rev. Drug Discov.* 13 (2014) 759–780. <https://doi.org/10.1038/nrd4278>.
- [34] N. Pardi, M.J. Hogan, F.W. Porter, D. Weissman, mRNA vaccines — a new era in vaccinology, *Nat. Rev. Drug Discov.* 17 (2018) 261–279. <https://doi.org/10.1038/nrd.2017.243>.

- [35] S. Xu, K. Yang, R. Li, L. Zhang, mRNA Vaccine Era-Mechanisms, Drug Platform and Clinical Prospection, *Int. J. Mol. Sci.* 21 (2020) E6582. <https://doi.org/10.3390/ijms21186582>.
- [36] J.D. Beck, D. Reidenbach, N. Salomon, U. Sahin, Ö. Türeci, M. Vormehr, L.M. Kranz, mRNA therapeutics in cancer immunotherapy, *Mol. Cancer*. 20 (2021) 69. <https://doi.org/10.1186/s12943-021-01348-0>.
- [37] O.V. Sergeeva, V.E. Koteliansky, T.S. Zatsepin, mRNA-Based Therapeutics - Advances and Perspectives, *Biochem. Biokhimiia*. 81 (2016) 709–722. <https://doi.org/10.1134/S0006297916070075>.
- [38] L. Warren, C. Lin, mRNA-Based Genetic Reprogramming, *Mol. Ther.* 27 (2019) 729–734. <https://doi.org/10.1016/j.ymthe.2018.12.009>.
- [39] H.-X. Zhang, Y. Zhang, H. Yin, Genome Editing with mRNA Encoding ZFN, TALEN, and Cas9, *Mol. Ther.* 27 (2019) 735–746. <https://doi.org/10.1016/j.ymthe.2019.01.014>.
- [40] D. Di Fusco, V. Dinallo, I. Marafini, M.M. Figliuzzi, B. Romano, G. Monteleone, Antisense Oligonucleotide: Basic Concepts and Therapeutic Application in Inflammatory Bowel Disease, *Front. Pharmacol.* 10 (2019). <https://www.frontiersin.org/articles/10.3389/fphar.2019.00305> (accessed July 20, 2023).
- [41] S. Qin, X. Tang, Y. Chen, K. Chen, N. Fan, W. Xiao, Q. Zheng, G. Li, Y. Teng, M. Wu, X. Song, mRNA-based therapeutics: powerful and versatile tools to combat diseases, *Signal Transduct. Target. Ther.* 7 (2022) 1–35. <https://doi.org/10.1038/s41392-022-01007-w>.

- [42] E. Blanco, H. Shen, M. Ferrari, Principles of nanoparticle design for overcoming biological barriers to drug delivery, *Nat. Biotechnol.* 33 (2015) 941–951. <https://doi.org/10.1038/nbt.3330>.
- [43] K.J. Kauffman, J.R. Dorkin, J.H. Yang, M.W. Heartlein, F. DeRosa, F.F. Mir, O.S. Fenton, D.G. Anderson, Optimization of Lipid Nanoparticle Formulations for mRNA Delivery in Vivo with Fractional Factorial and Definitive Screening Designs, *Nano Lett.* 15 (2015) 7300–7306. <https://doi.org/10.1021/acs.nanolett.5b02497>.
- [44] K. Karikó, M. Buckstein, H. Ni, D. Weissman, Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA, *Immunity.* 23 (2005) 165–175. <https://doi.org/10.1016/j.immuni.2005.06.008>.
- [45] J. Rehwinkel, M.U. Gack, RIG-I-like receptors: their regulation and roles in RNA sensing, *Nat. Rev. Immunol.* 20 (2020) 537–551. <https://doi.org/10.1038/s41577-020-0288-3>.
- [46] J. Jemielity, T. Fowler, J. Zuberek, J. Stepinski, M. Lewdorowicz, A. Niedzwiecka, R. Stolarski, E. Darzynkiewicz, R.E. Rhoads, Novel “anti-reverse” cap analogs with superior translational properties, *RNA.* 9 (2003) 1108–1122. <https://doi.org/10.1261/rna.5430403>.
- [47] J.M. Henderson, A. Ujita, E. Hill, S. Yousif-Rosales, C. Smith, N. Ko, T. McReynolds, C.R. Cabral, J.R. Escamilla-Powers, M.E. Houston, Cap 1 Messenger RNA Synthesis with Co-transcriptional CleanCap® Analog by In Vitro Transcription, *Curr. Protoc.* 1 (2021) e39. <https://doi.org/10.1002/cpz1.39>.

- [48] A.G. Hinnebusch, I.P. Ivanov, N. Sonenberg, Translational control by 5'-untranslated regions of eukaryotic mRNAs, *Science*. 352 (2016) 1413–1416. <https://doi.org/10.1126/science.aad9868>.
- [49] K.H. Asrani, J.D. Farelli, M.R. Stahley, R.L. Miller, C.J. Cheng, R.R. Subramanian, J.M. Brown, Optimization of mRNA untranslated regions for improved expression of therapeutic mRNA, *RNA Biol.* 15 (2018) 756–762. <https://doi.org/10.1080/15476286.2018.1450054>.
- [50] N. Ryczek, A. Łyś, I. Makałowska, The Functional Meaning of 5'UTR in Protein-Coding Genes, *Int. J. Mol. Sci.* 24 (2023) 2976. <https://doi.org/10.3390/ijms24032976>.
- [51] L. Jia, Y. Mao, Q. Ji, D. Dersh, J.W. Yewdell, S.-B. Qian, Decoding mRNA translatability and stability from the 5' UTR, *Nat. Struct. Mol. Biol.* 27 (2020) 814–821. <https://doi.org/10.1038/s41594-020-0465-x>.
- [52] P.J. Sample, B. Wang, D.W. Reid, V. Presnyak, I.J. McFadyen, D.R. Morris, G. Seelig, Human 5' UTR design and variant effect prediction from a massively parallel translation assay, *Nat. Biotechnol.* 37 (2019) 803–809. <https://doi.org/10.1038/s41587-019-0164-5>.
- [53] S. Brogna, J. Wen, Nonsense-mediated mRNA decay (NMD) mechanisms, *Nat. Struct. Mol. Biol.* 16 (2009) 107–113. <https://doi.org/10.1038/nsmb.1550>.
- [54] E.M. Zhao, A.S. Mao, H. de Puig, K. Zhang, N.D. Tippens, X. Tan, F.A. Ran, I. Han, P.Q. Nguyen, E.J. Chory, T.Y. Hua, P. Ramesh, D.B. Thompson, C.Y. Oh, E.S. Zigon, M.A. English, J.J. Collins, RNA-responsive elements for eukaryotic translational control, *Nat. Biotechnol.* 40 (2022) 539–545. <https://doi.org/10.1038/s41587-021-01068-2>.

- [55] M. Baierdörfer, G. Boros, H. Muramatsu, A. Mahiny, I. Vlatkovic, U. Sahin, K. Karikó, A Facile Method for the Removal of dsRNA Contaminant from In Vitro-Transcribed mRNA, *Mol. Ther. - Nucleic Acids*. 15 (2019) 26–35. <https://doi.org/10.1016/j.omtn.2019.02.018>.
- [56] M.R. Green, J. Sambrook, Precipitation of RNA with Ethanol, *Cold Spring Harb. Protoc.* 2020 (2020) pdb.prot101717. <https://doi.org/10.1101/pdb.prot101717>.
- [57] S.C. Devarkar, C. Wang, M.T. Miller, A. Ramanathan, F. Jiang, A.G. Khan, S.S. Patel, J. Marcotrigiano, Structural basis for m⁷G recognition and 2'-O-methyl discrimination in capped RNAs by the innate immune receptor RIG-I, *Proc. Natl. Acad. Sci. U. S. A.* 113 (2016) 596–601. <https://doi.org/10.1073/pnas.1515152113>.
- [58] GTEx Portal, (n.d.). <https://www.gtexportal.org/home/datasets> (accessed June 22, 2022).
- [59] D. Wang, B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D.P. Zolg, J. Zecha, A. Asplund, L.-H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, B. Kuster, A deep proteome and transcriptome abundance atlas of 29 healthy human tissues, *Mol. Syst. Biol.* 15 (2019) e8503. <https://doi.org/10.15252/msb.20188503>.
- [60] G. Pesole, S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi, C. Saccone, UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002, *Nucleic Acids Res.* 30 (2002) 335–340.
- [61] GEO Accession viewer, (n.d.). <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE208731> (accessed April 3, 2023).

- [62] GEO Accession viewer, (n.d.).
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114332> (accessed August 23, 2023).
- [63] Ensembl Genome Browser, (n.d.). <https://grch37.ensembl.org/index.html> (accessed August 23, 2023).
- [64] Homo_sapiens - Ensembl genome browser 110, (n.d.).
https://www.ensembl.org/Homo_sapiens/Info/Index (accessed August 23, 2023).
- [65] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinformatics*. 12 (2011) 323.
<https://doi.org/10.1186/1471-2105-12-323>.
- [66] Assembly PCR synthesis of optimally designed, compact, multi-responsive promoters suited to gene therapy application | *Scientific Reports*, (n.d.).
<https://www.nature.com/articles/srep29388> (accessed August 23, 2023).
- [67] L. Velten, S. Haas, A. Rademacher, H. Meyer, R. Eils, RA-PCR, a method for the generation of randomized promoter libraries, (2009).
<https://dspace.mit.edu/handle/1721.1/49502> (accessed August 23, 2023).
- [68] R.L. Hajnik, J.A. Plante, Y. Liang, M.-G. Alameh, J. Tang, S.R. Bonam, C. Zhong, A. Adam, D. Scharon, G.H. Rafael, Y. Liu, N.C. Hazell, J. Sun, L. Soong, P.-Y. Shi, T. Wang, D.H. Walker, J. Sun, D. Weissman, S.C. Weaver, K.S. Plante, H. Hu, Dual spike and nucleocapsid mRNA vaccination confer protection against SARS-CoV-2 Omicron and Delta variants in preclinical models, *Sci. Transl. Med.* 14 (2022) eabq1945. <https://doi.org/10.1126/scitranslmed.abq1945>.
- [69] X. Mu, E. Greenwald, S. Ahmad, S. Hur, An origin of the immunogenicity of in vitro transcribed RNA, *Nucleic Acids Res.* 46 (2018) 5239–5249.
<https://doi.org/10.1093/nar/gky177>.

- [70] Y. Gholamalipour, W.C. Johnson, C.T. Martin, Efficient inhibition of RNA self-primed extension by addition of competing 3'-capture DNA-improved RNA synthesis by T7 RNA polymerase, *Nucleic Acids Res.* 47 (2019) e118. <https://doi.org/10.1093/nar/gkz700>.
- [71] Human Embryonic Kidney 293, (n.d.). <https://www.beckman.com/resources/product-applications/lead-optimization/cell-line-development/human-embryonic-kidney-293> (accessed August 23, 2023).



APPENDIX A

DNA SEQUENCES OF CONSTRUCTS

Table A. 1: DNA sequences used in this thesis

Gene or Regulatory Region	Sequence
T7 Promoter IVT Source	TAATACGACTCACTATAGGGAGAGAGAGAG AATTACCCTCACTAAAGGGAGGAG
5' UTR of Human β Globulin	ACATTTGCTTCTGACACAACCTGTGTTCCTAG CAACCTCAAACAGACACC
3' UTR of Human β Globulin	GCTCGCTTTCTTGCTGTCCAATTTCTATTAAA GGTTCCTTTGTTCCCTAAGTCCAACCTACTAAA CTGGGGGATATTATGAAGGGCCTTGAGCATC TGGATTCTGCCTAATAAAAAACATTTATTTTC ATTGCAA
Poly(A) Tail	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Kozak Sequence	GCCGCCACC
sfGFP Gene	ATGAGCAAGGGCGAGGAAGTGTTCACCGGCGT GGTGCCCATCCTGGTGGAACTGGACGGAGACG TGAACGGCCACAAATTCTCCGTGCGGGGAGAG GGCGAAGGCGACGCTACAAACGGAAAGCTGAC CCTGAAATTCATCTGCACAACAGGCAAACTGCC

	CGTGCCTTGGCCTACCCTCGTTACAACCCTGAC ATACGGCGTGCAATGTTTTTCTAGATACCCAGA CCACATGAAAAGACATGATTTTTTCAAAAGCGC CATGCCCCGAGGGCTACGTGCAGGAGCGGACCA TCAGCTTCAAGGACGACGGAACCTACAAGACC AGAGCCGAGGTCAAGTTCGAGGGCGACACCCT GGTGAATAGAATCGAGCTGAAGGGCATCGACT TTAAGGAAGATGGCAACATCCTGGGCCACAAG CTGGAGTACAACCTTCAACAGCCACAACGTGTA CATCACCGCCGACAAGCAGAAGAACGGCATCA AGGCCAATTTCAAGATCAGACACAATGTGGAAG ATGGCAGCGTGCAGCTGGCCGATCACTATCAGC AGAACACACCTATCGGCGATGGACCTGTGCTGC TGCCTGACAACCACTACCTGAGCACCCAGTCTG TGCTGTCCAAGGACCCTAACGAGAAGCGGGAC CACATGGTCCTTCTGGAGTTCGTGACCGCCGCT GGCATTACCCACGGCATGGACGAGCTGTACAAG
Spike Gene	ATGTTTCGTGTTTCCTGGTGCTGCTGCCCCTGGTCT CCAGCCAATGTGTGAACCTGACAACCAGAACCC AACTGCCACCTGCTTACACCAATAGCTTTACCA GAGGCGTGTACTIONACCCTGACAAGGTGTTTCAGAA GCAGCGTGCTCCACAGCACCCAGGACCTGTTCC TGCCCTTCTTCAGCAATGTGACCTGGTTCCACG CCATCCACGTGAGCGGTACAAACGGCACGAAGC GGTTCGACAACCCCGTGCTGCCCTTCAACGACG

GCGTGTACTTTGCCTCGACAGAAAAGAGCAACA TCATTAGAGGATGGATCTTCGGCACCACACTGG ACAGCAAGACGCAGAGCCTGCTGATCGTGAACA ACGCCACAAACGTGGTGATCAAAGTCTGCGAGT TCCAGTTTTGCAACGACCCTTTTCTGGGCGTGTA CTACCACAAGAACAACAAGTCCTGGATGGAAA GCGAGTTCCGGGTGTATTCTTCGGCCAACAAC TG CACATTGAGTACGTGAGCCAGCCTTTTCTGATG GACCTGGAAGGCAAGCAGGGGAAC TTCAAGAAC CTGAGGGAGTTCGTGTTCAAGAATATCGACGGCT ACTTCAAGATCTACTCTAAGCACACCCCTATCAA CCTGGTCAGAGACCTCCCTCAGGGCTTCAGCGCT CTGGAGCCTCTGGTGGACCTGCCCATCGGCAT CAACATCACCCGGTTCCAAACTCTGCTGGCCC TGCACCGGTCCTACCTGACCCCAGGAGATAGC AGCTCTGGATGGACCGCCGGCGCTGCAGCTTA CTACGTGGGATATCTGCAGCCTAGAACATTCC TGCTTAAGTACAACGAAAACGGCACCATAAC CGACGCCGTGGATTGCGCTCTGGACCCTCTGT CTGAAACCAAGTGCACCCTGAAGTCTTTTACA GTGGAGAAGGGCATCTACCAGACATCCAATTT CCGAGTGCAGCCAACCGAAAGTATCGTGCGAT TTCCTAACATCACCAACCTGTGCCCTTTCGGAG AGGTCTTCAACGCCACCCGCTTCGCTTCCGTCT ACGCCTGGAATAGAAAGCGGATCTCCAAC TGC

GTGGCTGATTACTCAGTGCTGTACAATTCTGCTA GCTTTAGCACATTCAAATGCTACGGTGTGAGCCC TACCAAGCTAAACGACCTGTGCTTCACCAACGTG TACGCCGACAGCTTCGTGATCAGAGGCGACGAA GTGCGGCAGATCGCTCCCGGCCAGACAGGCAAG ATTGCCGATTACAACACTACAAGTTGCCTGATGATT TCACTGGTTGCGTGATCGCCTGGAACAGCAACAA TCTGGACAGCAAGGTCGGCGGCAACTACAACACTAC CTGTATCGGCTGTTTAGAAAGAGCAACCTGAAGC CCTTCGAGAGAGATATCTCTACAGAAATCTACCA GGCCGGCAGCACCCCTTGTAATGGAGTGGAGGGA TTCAACTGTTACTTCCCTCTGCAGAGCTACGGGTT CCAGCCTACCAACGGTGTGGATAACCAGCCTTACA GAGTGGTGGTTCTGAGCTTTGAGCTGCTGCATGCC CCTGCCACAGTGTGTGGCCCTAAGAAGTCTACCA ATCTGGTAAAGAACAAGTGTGTGAATTTCAACTT CAACGGCCTGACTGGCACGGGCGTGTTGACAGA GAGCAACAAAAAGTTTCTCCCTTTCCAGCAGTTC GGCAGAGATATCGCCGACACAACCGACGCTGTG AGAGACCCTCAAACCCTTGAAATCCTGGACATT ACACCTTGCAGCTTCGGCGGAGTGAGCGTAATC ACCCAGGCACCAACACCAGCAATCAGGTGGCC GTACTGTACCAGGACGTGAACTGCACCGAGGT GCCCCGTCGCCATCCACGCTGACCAGCTGACCC CTACATGGCGGGTGTACAGCACTGGCAGCAAT
--

GTGTTCCAGACAAGAGCCGGCTGTCTGATAGGA
GCCGAACACGTGAACAACAGCTACGAGTGCGAC
ATTCCCATCGGCGCGGGCATCTGCGCCTCTTACC
AGACCCAGACGAACAGCCCTAGACGGGGCCAGAA
GCGTGGCCAGCCAGAGCATCATCGCCTATACCAT
GAGCCTGGGCGCTGAAAACAGCGTGGCCTATT
CTAACAACAGCATCGCCATTCCCACAAACTT
CACCATCTCAGTGACCACCGAGATCCTGCCG
GTCTCCATGACCAAGACCTCCGTGATTGCA
CCATGTACATCTGCGGCGACAGCACCGAATG
TAGCAACCTGCTGCTGCAGTACGGCAGCTTC
TGCACCCAGCTGAACAGAGCCCTGACCGGCA
TCGCTGTTGAGCAGGACAAGAACACCCAGGA
GGTGTTGCGCACAGGTGAAGCAGATCTACAAGA
CACCCCCCATCAAGGACTTCGGAGGCTTTAAC
TTCTCTCAGATCCTGCCTGATCCAAGCAAGCCT
TCTAAACGGAGCTTTATCGAGGACCTGCTGTTC
AACAAGGTGACCCTGGCCGATGCCGGCTTCAT
CAAGCAGTACGGCGATTGCCTGGGCGACATCG
CCGCTAGAGACCTGATTTGTGCCCAGAAGTTC
AACGGTCTTACCGTGTTGCCTCCTCTGCTGACA
GACGAGATGATCGCCCAGTATACAAGCGCCC
TGCTGGCTGGAACCATCACATCCGGCTGGACC
TTGGGCGCCGGCGCCGCCCTGCAGATCCCCTT
CGCCATGCAGATGGCCTACAGATTTAACGGCA

TCGGCGTGACACAAAACGTGCTGTACGAAAAC CAGAAACTGATAGCCAACCAGTTCAACTCCGC CATCGGCAAAATCCAGGACAGCCTGTCTTCTA CCGCCAGCGCCCTGGGCAAACGTGCAAGACGTC GTGAATCAGAACGCGCAGGCTCTGAACACTTT AGTGAAGCAGCTGAGCTCTAATTTTCGGCGCCA TCAGCTCCGTGCTCAATGATATCCTGTCAAGAC TGGATAAAGTCGAGGCCGAGGTGCAGATCGAT AGACTGATCACCGGAAGACTGCAATCCCTCCA AACATACGTGACCCAGCAACTGATCAGGGCTG CCGAAATCAGAGCCTCTGCTAATCTCGCCGCC ACCAAGATGAGCGAATGCGTGCTGGGCCAGA GCAAAAGAGTGGACTTTTGCGGCAAGGGCTA TCACCTGATGTCCTTCCCCCAGAGCGCCCTC ACGGCGTGGTGTTCTGTCATGTGACATACGTG CCCGCCCAGGAGAAAAACTTCACCACAGCCC CTGCCATCTGCCACGACGGCAAAGCCCCTTC CCCAGAGAGGGCGTGTTTCGTGTCCAACGGCAC CCACTGGTTCGTACCCAGAGAACTTCTACG AGCCCCAGATTATCACCACCGACAACACCTTTG TGTCCGGCAACTGTGATGTGGTCATCGGAATCG TGAACAACACCGTGTACGACCCTCTGCAGCCCG AGCTGGATAGCTTCAAAGAGGAACTGGACAAG TATTTCAAGAACCACACATCTCCAGACGTGGAC CTGGGCGACATCAGCGGCATCAACGCCAGCGTA
--

	GTGAACATCCAGAAGGAAATCGACAGACTGAAC GAGGTGGCCAAGAATCTCAATGAATCCCTGATCG ATCTGCAGGAGCTGGGAAAATACGAGCAGTACA TCAAGTGGCCTTAATAA
RBD Gene	ATGTTTCGTGTTCTGGTGCTGCTGCCCCCTGGTCT CCAGGGTGCAGCCCACCGAGAGCATCGTGAGG TTCCCCAACATCACCAACCTGTGCCCCCTTCGGC GAGGTGTTCAACGCCACCAGGTTCCGCCAGCGTG TACGCCTGGAACAGGAAGAGGATCAGCAACTG CGTGGCCGACTACAGCGTGCTGTACAACAGCGC CAGCTTCAGCACCTTCAAGTGCTACGGCGTGAG CCCCACCAAGCTGAACGACCTGTGCTTCACCAA CGTGTACGCCGACAGCTTCGTGATCAGGGGCGA CGAGGTGAGGCAGATCGCCCCCGGCCAGACCG GCAAGATCGCCGACTACAACTACAAGCTGCCCCG ACGACTTCACCGGCTGCGTGATCGCCTGGAACA GCAACAACCTGGACAGCAAGGTGGGCGGCAAC TACAACTACCTGTACAGGCTGTTTCAAGAGAGC AACCTGAAGCCCTTCGAGAGGGACATCAGCAC CGAGATCTACCAGGCCGGCAGCACCCCCTGCAA CGGCGTGGAGGGCTTCAACTGCTACTTCCCCCT GCAGAGCTACGGCTTCCAGCCCACCAACGGCG TGGGCTACCAGCCCTACAGGGTGGTGGTGCTGA GCTTCGAGCTGCTGCACGCCCCCGCCACCGTGT GCGGCCCCAAGAAGTCCACCAACCTGGTGAAG

	AACAAGTGCGTGAAC TTCTAATAA
5' UTR of APOA2 Gene	AGGCACAGACACCAAGGACAGAGACGCTGGC TAGGCCGCCCCTCCCCACTGTTACCAAC
5' UTR of LYZ Gene	AAATACTGGGGCCAGCTCACCTGGTCAGCCT AGCACTCTGACCTAGCAGTCAAC
5' UTR of RNF219 Gene	GGCGAAGAGTAGCGGTAGGTCGGCGGGACT TCCGTGTTGGCGGGATTCTGAACGCTGCC
5' UTR of HIST1H4H Gene	CGGCATATAAGGGCGTTGCTTTGCCTTGGC GTTTAGGTTTCTTAAGTTGGTTTTAGAAGTTG CTTAGTC
5' UTR of PGM2 Gene	CACGTCTTGGGGCCGGGCCGGAAGGCAGA TCTCACCGCCTGCTTCCCTCTGCAGCGGTA GCACAAGCTCAGCG
CMV Enhancer	GACATTGATTATTGACTAGTTATTAATAGTAA TCAATTACGGGGTCATTAGTTCATAGCCCATA TATGGAGTTCCGCGTTACATAACTTACGGTAA ATGGCCCGCCTGGCTGACCGCCCAACGACCCC CGCCCATTGACGTCAATAATGACGTATGTTCC CATAGTAACGCCAATAGGGACTTTCCATTGAC GTCAATGGGTGGACTATTTACGGTAAACTGCC CACTTGGCAGTACATCAAGTGTATCATATGCC AAGTACGCCCCCTATTGACGTCAATGACGGTA AATGGCCCGCCTGGCATTATGCCCAGTACATG ACCTTATGGGACTTTCCTACTTGGCAGTACATC TACGTATTAGTCATCGCTATTACCATG

CMV Promoter	GTGATGCGGTTTTGGCAGTACATCAATGGGC GTGGATAGCGGTTTGACTCACGGGGATTTC AAGTCTCCACCCCATGACGTCAATGGGAGT TTGTTTTGGCACCAAAATCAACGGGACTT TCCAAAATGTCGTAACAACCTCCGCCCCATTG ACGCAAATGGGCGGTAGGCGTGTACGGTGG GAGGTCTATATAAGCAGAGCT
bGH Poly(A) Signal	CTGTGCCTTCTAGTTGCCAGCCATCTGTTGT TTGCCCCCTCCCCCGTGCCTTCCTTGACCCTG GAAGGTGCCACTCCCCTGTCCTTTCCTAAT AAAATGAGGAAATTGCATCGCATTGTCTGA GTAGGTGTCATTCTATTCTGGGGGGTGGGG TGGGGCAGGACAGCAAGGGGGAGGATTGG GAAGACAATAGCAGGCATGCTGGGGATGC GGTGGGCTCTATGG

APPENDIX B

LIST OF PRIMERS

Table B. 1: Primer sequences used in both cloning of the constructs and library creation

Primer Name	Sequence (5'-3')
AH1_1	CCAAGGACAGAGACGCTGGCTAGGC CGCCCTCCCCACTGTTACCAACGCCG CCACCATGAGCAAG
AH1_2	GAATTACCCTCACTAAAGGGAGGAG GGTACCAGGCACAGACACCAAGGAC AGAGACGCTG
AH3_1	GGGCCAGCTCACCTGGTCAGCCT AGCACTCTGACCTAGCAGTCAACG CCGCCACCATGAGCAAG
AH3_2	GAATTACCCTCACTAAAGGGAGGA GGGTACCAAATACTGGGGCCAGCT CACCTGGTCAGCC
AH5_1	CGGTAGGTCGGCGGGACTTCCGTGT TGGCGGGATTCTGAACGCTGCCGC CGCCACCATGAGCAAG
AH5_2	GAATTACCCTCACTAAAGGGAGGAG GGTACCGGCGAAGAGTAGCGGTAGG TCGGCGGGAC
AH7_1	GCCTTGGCGTTTtaggtttCTTAAGTT

	GGTTTTAGAAGTTGCTTAGTCGCCG CCACCATGAGCAAG
AH7_2	GAATTACCCTCACTAAAGGGAGGAG GGTACCCGGCATATAAGGGCGTTGC TTTGCCTTGGCGTTTAGGTTTCTTAA GTTG
AH9_1	AGATCTCACCGCCTGCTTCCCTCTGC AGCGGTAGCACAAGCTCAGCGGCCG CCACCATGAGCAAG
AH9_2	GAATTACCCTCACTAAAGGGAGGAG GGTACCCACGTCTTGGGGCCGGGCC GGAAGGCAGATCTCACCGCCTGCTTC
AH10	ATCGGGATCGCGGCGCGGGCGGG AAACAGCTATGACCATGATTACGC
AH13	ACATTTGCTTCTGACACAACCTGTGTT CACTAGCAACCTCAAACAGACACCG GATCCGCCGCCACCATGAGCAAG
AH14	TGTACGGTGGGAGGTCTATATAAGC AGAGCTCTCTGGCTAACTAGAGAAC CCACTGCGAATTCACATTTGCTTCTG ACAC
AH15	GCAGCCGAGGGGTATTAGGCAGAAT CCAGATGCTCAAGGC
AH16	CAACTAGAAGGCACAGTCGAGGCTG ATCAGCGAGCTCTTGCAATGAAAAT

	AAATGTTTTTTATTAGGCAGAATCCA GATG
pAH_L0	GTGAGGAGAAGTGCCAACAACCTTCT GGAGGACACTTGAGGTGA
pAH_L1	TTGGCACTTCTCCTCACGGAGAGGG AGTCACCTCAAGTGTCTC
pAH_L2	GTGAGGAGAAGTGCCAACCGGCTT CTGAGGACACTTGAGGTGA
pAH_L3	TTGGCACTTCTCCTCACTTGCGCGC GTCACCTCAAGTGTCTC
pAH_L4	GTGAGGAGAAGTGCCAAAAACAA GAAGAGGACACTTGAGGTGA
pAH_L5	TTGGCACTTCTCCTCACGCGCTGA GCCTCACCTCAAGTGTCTC
pAH_L6	GTGAGGAGAAGTGCCAAATATTCT CGAGGACACTTGAGGTGA
pAH_L7	TTGGCACTTCTCCTCACCGCTGAGC CGTCACCTCAAGTGTCTC
pAH_L8	GTGAGGAGAAGTGCCAAACAGGG GCTGAGGACACTTGAGGTGA
pAH_L9	TTGGCACTTCTCCTCACCCGGAGG TGGTCACCTCAAGTGTCTC
pAH_L10	GTGAGGAGAAGTGCCAAGTCTTTT GGAGGACACTTGAGGTGA
pAH_L11	TTGGCACTTCTCCTCACGTCCAGGT

	GTCACCTCAAGTGTCTC
pAH_L12	GTGAGGAGAAGTGCCAAATGCTTT TGAGGACACTTGAGGTGA
pAH_L13	TTGGCACTTCTCCTCACGGCGGAG GGAGTCACCTCAAGTGTCTC
pAH_L14	GTGAGGAGAAGTGCCAACGCCGGC CTGGAGGACACTTGAGGTGA
pAH_L15	TTGGCACTTCTCCTCACGCCGCCGC CGCATCACCTCAAGTGTCTC
pAH_L16	GTGAGGAGAAGTGCCAAGCTGGGC GCAGAGGACACTTGAGGTGA
pAH_L17	TTGGCACTTCTCCTCACCGGCGGC GGCGGCGGCGACTCACCTCAAGT GTCCTC
pAH_L18	GTGAGGAGAAGTGCCAACGAGCT CCTGAGGACACTTGAGGTGA
pAH_L19	TTGGCACTTCTCCTCACGGAGGAG GAGGAGAGTCACCTCAAGTGTCC TC
pAH_L20	GTGAGGAGAAGTGCCAACTTCTG AAAGAGGACACTTGAGGTGA
pAH_L21	TTGGCACTTCTCCTCACGAGAGG CTTTCACCTCAAGTGTCTC
pAH_5end_second	CAGAGCTCTCTGGCTAACTAGAG AACCCACTGCGAATTCTCACCTCA

	AGTGTCTC
pAH_5end_final	CTGGCTAACTAGAGAACCCACTGC
pAH_3end_second	CTCGCCCTTGCTCATGGTGGCGG CGGATCCGAGGACACTTGAGGTGA
pAH_3end_final	CTCGCCCTTGCTCATGGTG



APPENDIX C

PLASMID MAPS OF CONSTRUCTS USED IN THIS THESIS

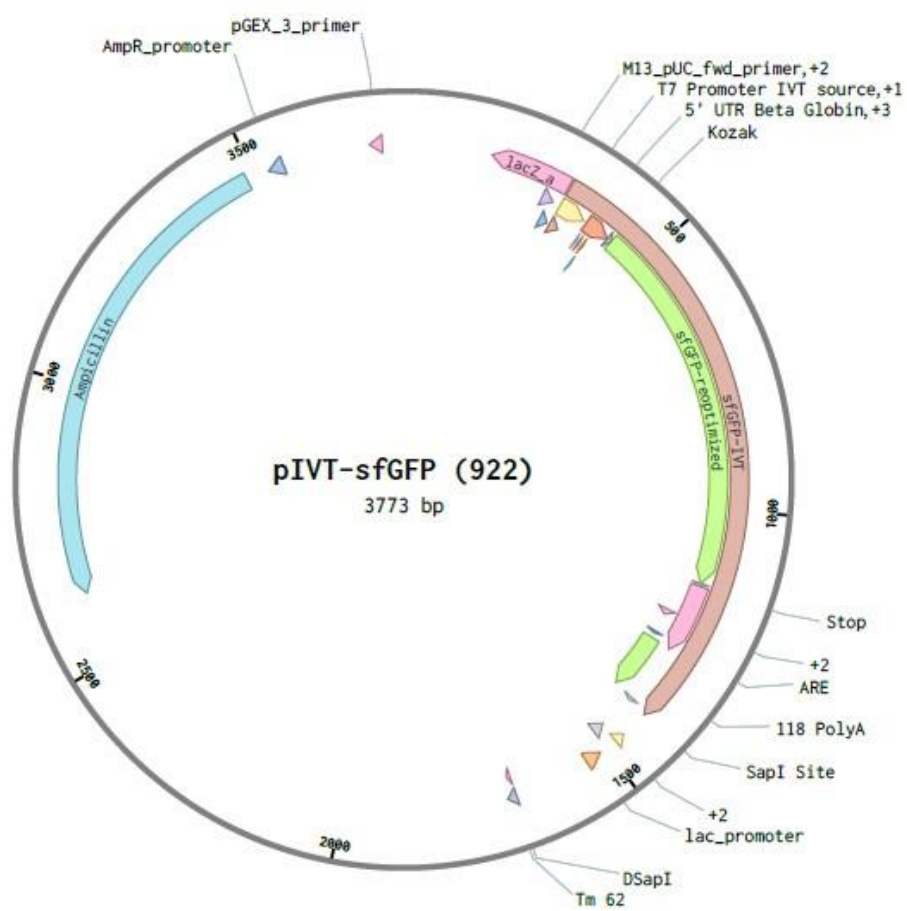


Figure C. 1: The schematic demonstration of 922 plasmid

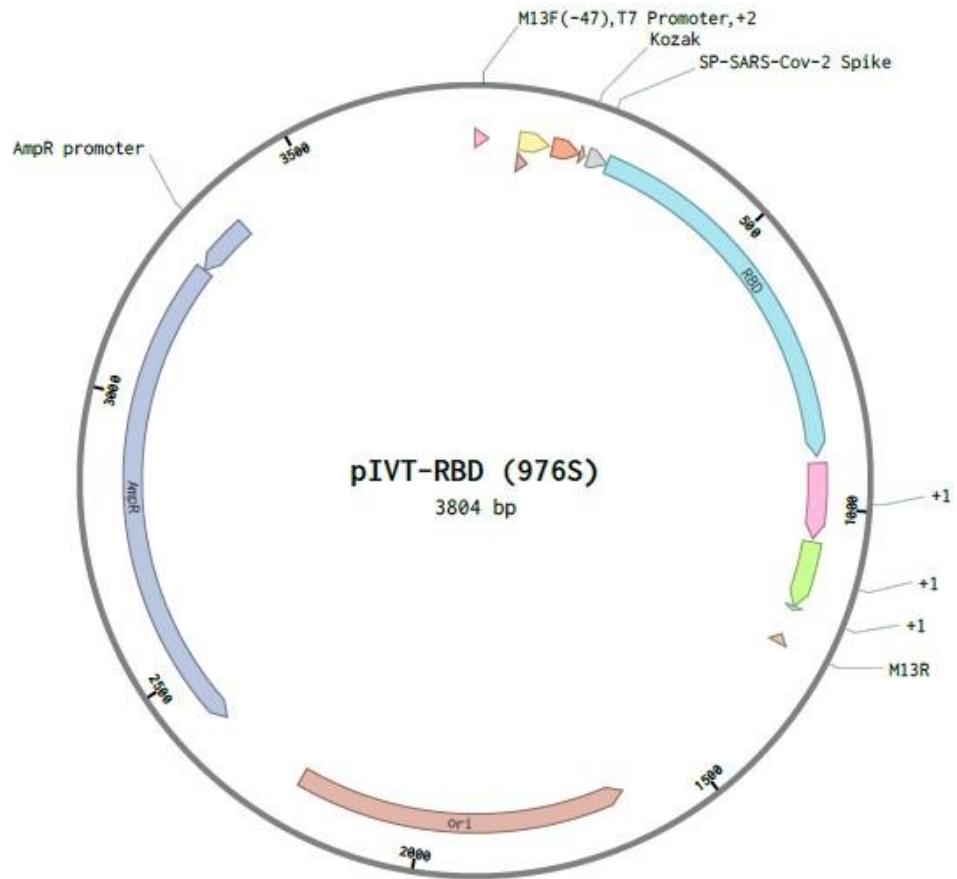


Figure C. 3: The schematic demonstration of 976S plasmid

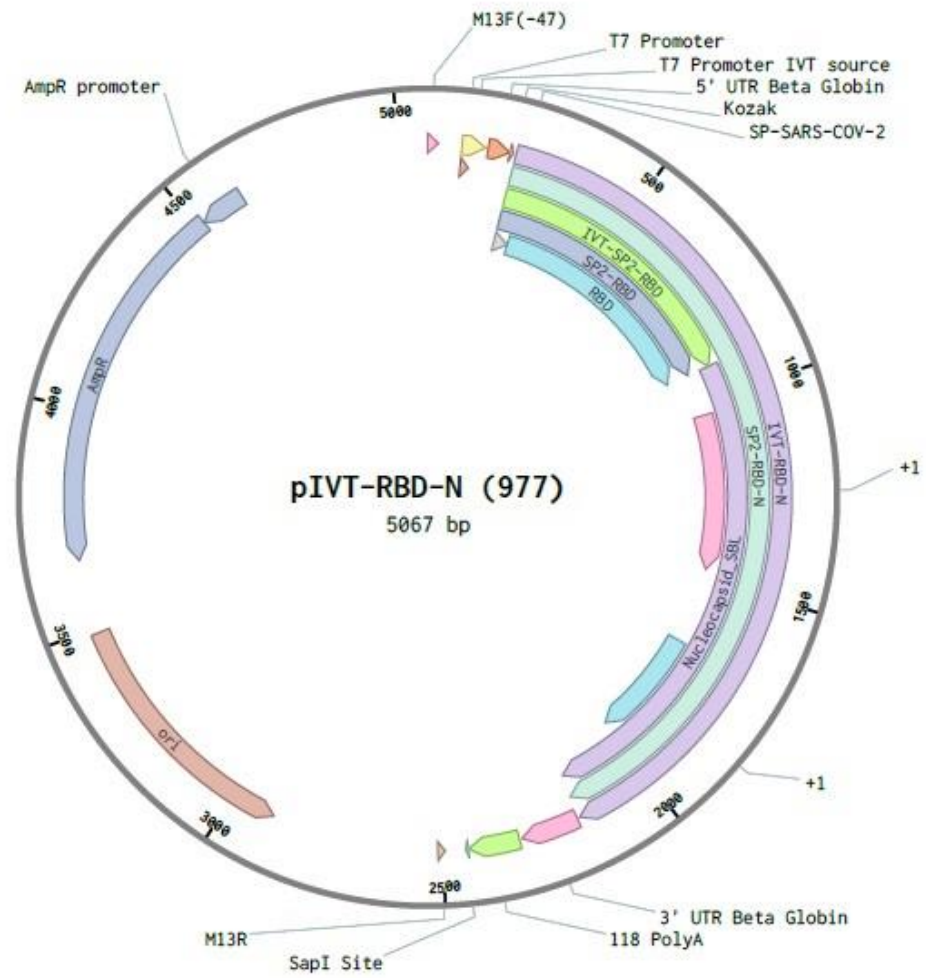


Figure C. 4: The schematic demonstration of 977 plasmid

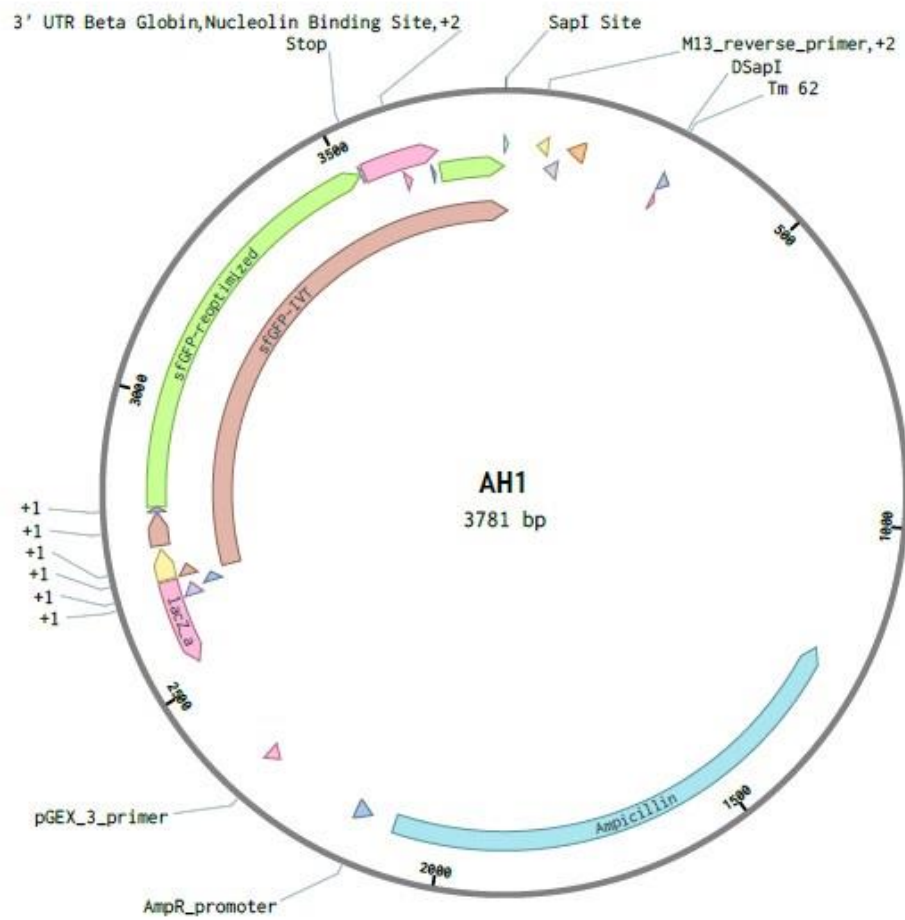


Figure C. 5: The schematic demonstration of AH1 plasmid

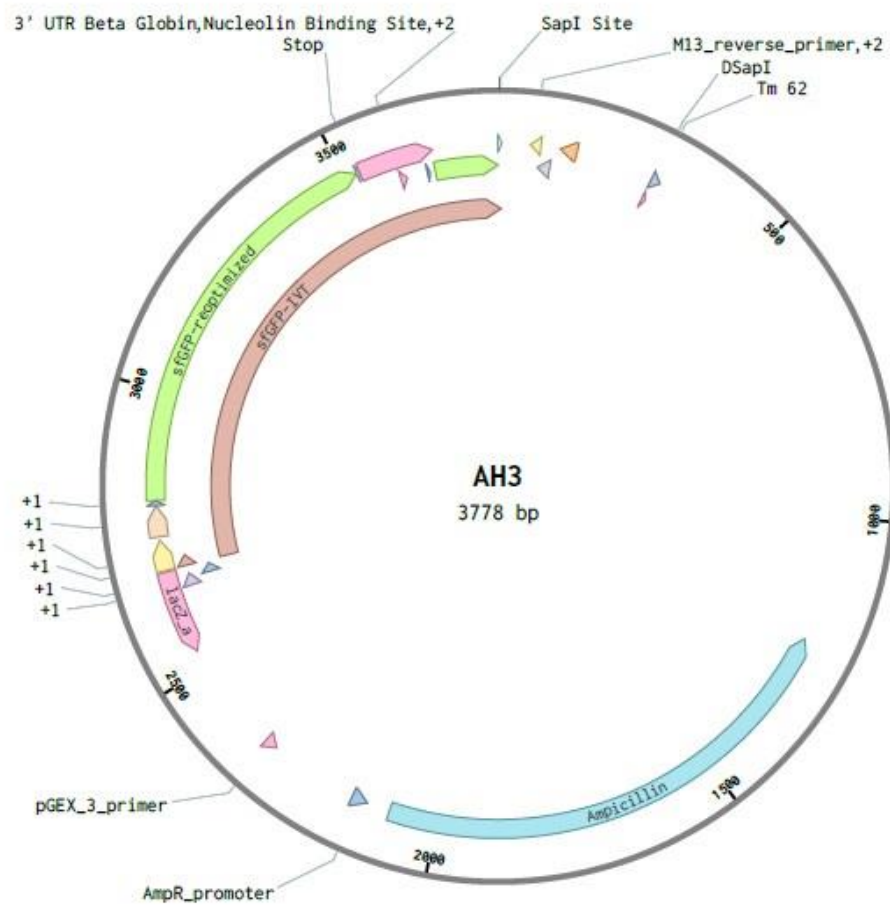


Figure C. 6: The schematic demonstration of AH3 plasmid

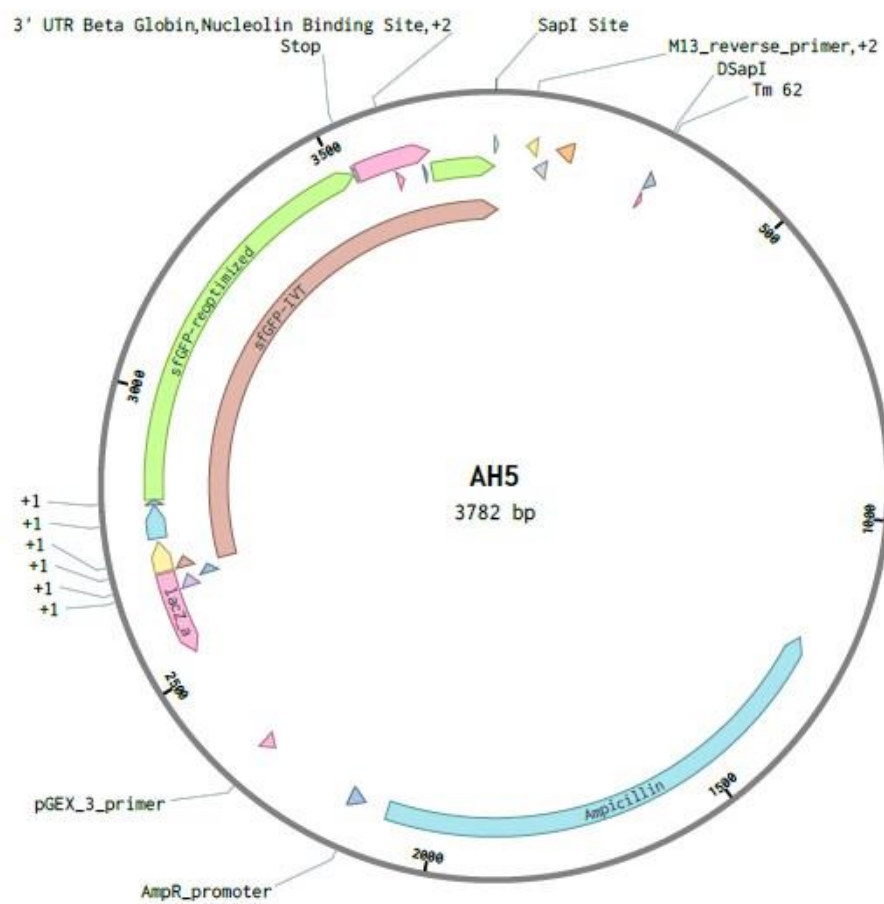


Figure C. 7: The schematic demonstration of AH5 plasmid

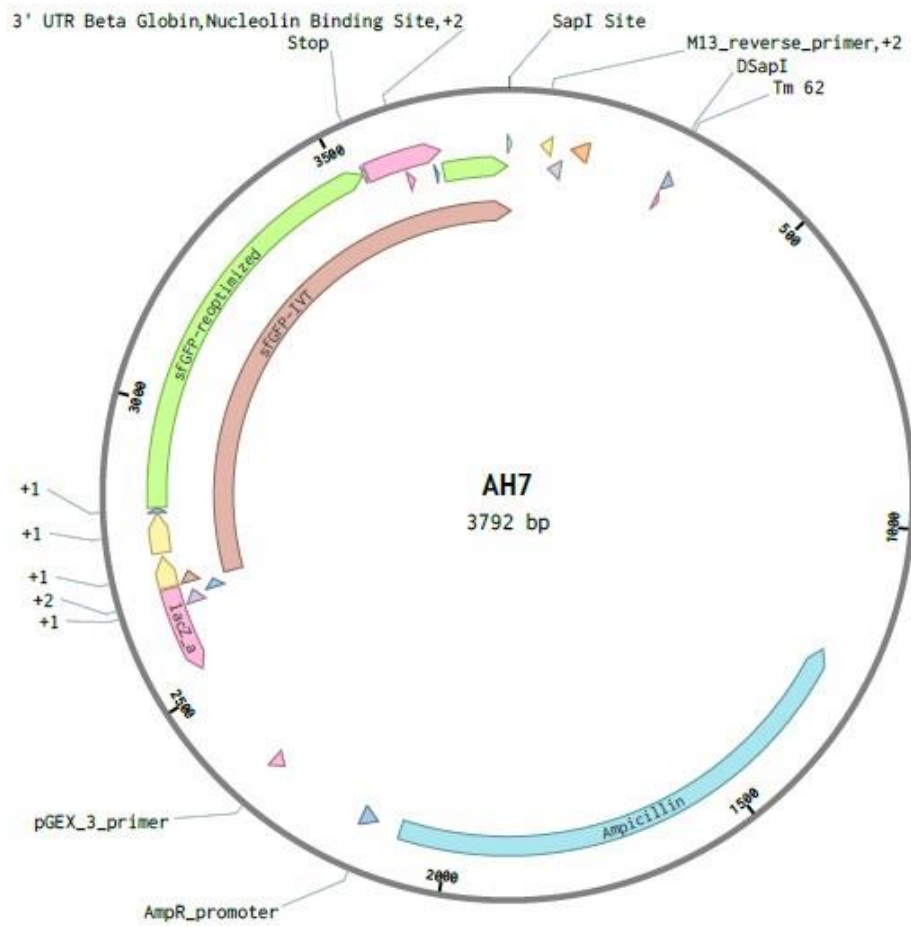


Figure C. 8: The schematic demonstration of AH7 plasmid

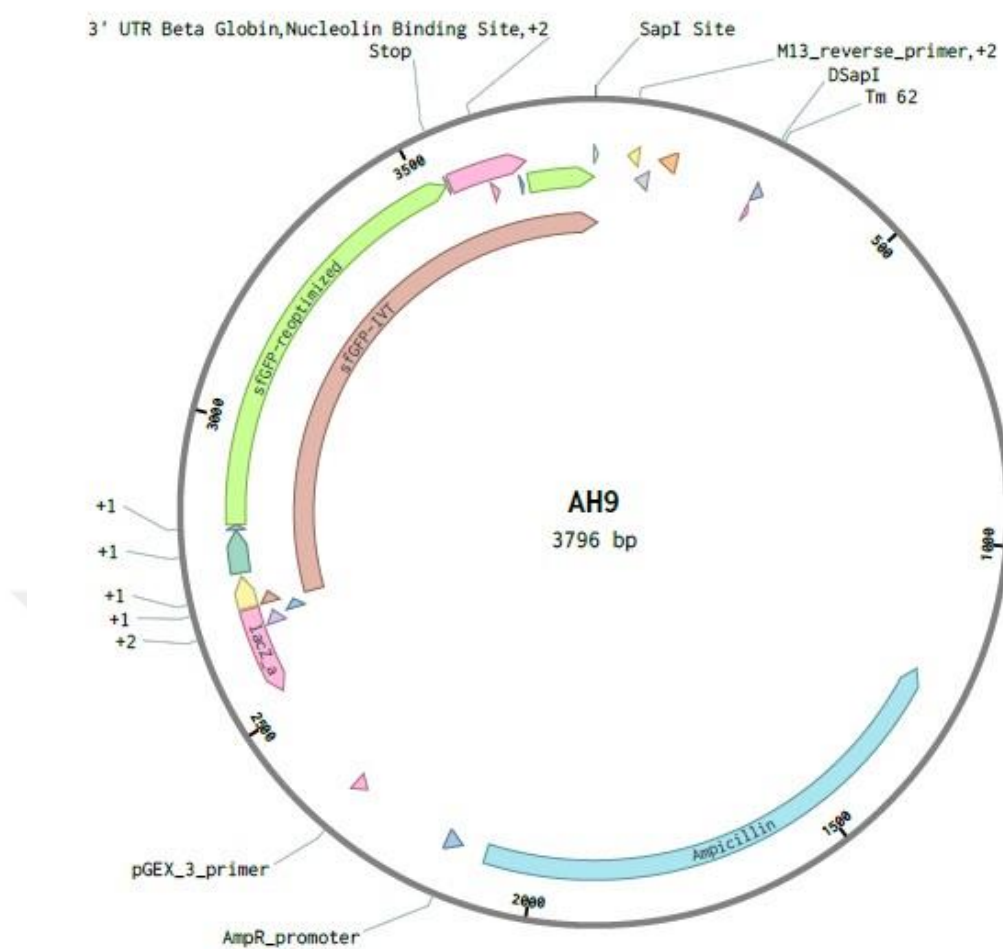


Figure C. 9: The schematic demonstration of AH9 plasmid

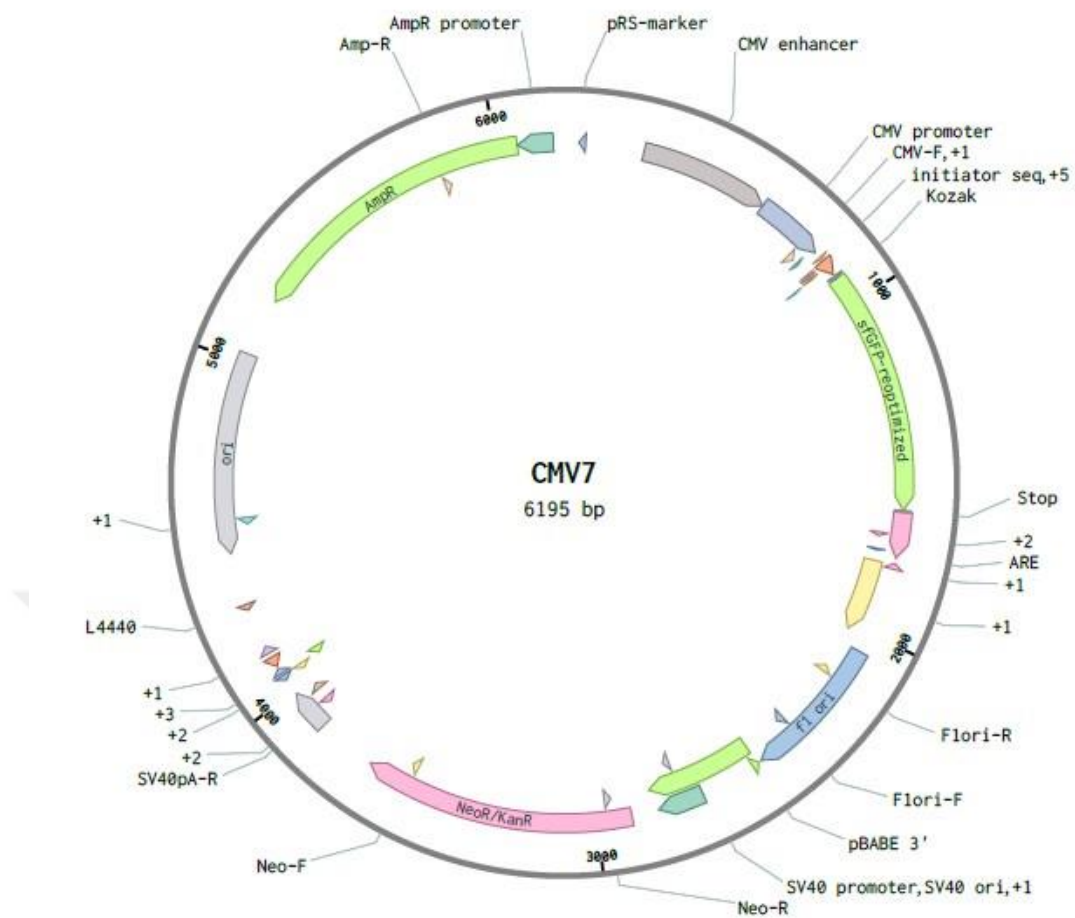


Figure C. 10: The schematic demonstration of CMV7 plasmid

APPENDIX D

SEQUENCING RESULTS

Next generation sequencing (NGS) was performed for sequence analysis. For the ease of observing mismatched bases, the consensus sequence obtained from NGS results (Geneious Application for map to reference alignment) was uploaded to Benchling to see matching status of plasmid sequence. Only the regions which were newly cloned were demonstrated.

```

2748                                     2829
APOA2 to ... ACTAAAGGGAGGAGGGTACCaggcacagacaccaaggacagagacgctggctagggccgcccctccccactgttaccacacGCCG
ah1_2_con... ACTAAAGGGAGGAGGGTACCAGGCACAGACACCAAGGACAGAGACGCTGGCTAGGCCGCCCTCCCCACTGTTACCAACGCCG
.....

2830                                     2911
APOA2 to ... CCACCATGAGCAAGGGCGAGGAACTGTTACCCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGGCCACAA
ah1_2_con... CCACCATGAGCAAGGGCGAGGAACTGTTACCCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGGCCACAA
.....

2912                                     2993
APOA2 to ... ATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAAGCTGACCCTGAAATTCATCTGCACAACAGGCAAACTG
ah1_2_con... ATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAAGCTGACCCTGAAATTCATCTGCACAACAGGCAAACTG
.....

2994                                     3075
APOA2 to ... CCCGTGCCTTGGCCTACCCTCGTTACAACCCTGACATACGGCGTGCAATGTTTTTCTAGATACCCAGACCACATGAAAAGAC
ah1_2_con... CCCGTGCCTTGGCCTACCCTCGTTACAACCCTGACATACGGCGTGCAATGTTTTTCTAGATACCCAGACCACATGAAAAGAC
.....
```

Figure D. 1: Sequencing result of AH1

```

2752                                     2833
LYZ to 92... GGAGGAGGGTACCAaatactggggccagctcacccctggctcagcctagcactctgacctagcagtcacGCCGCCACCATGAG
ah3_1_con... GGAGGAGGGTACCAAACTACTGGGGCCAGCTCACCCCTGGTCAGCCTAGCACTCTGACCTAGCAGTCAACGCCGCCACCATGAG
.....

2834                                     2915
LYZ to 92... CAAGGGCGAGGAACTGTTACCCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGGCCACAAATTCCTCGTG
ah3_1_con... CAAGGGCGAGGAACTGTTACCCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGGCCACAAATTCCTCGTG
.....

2916                                     2997
LYZ to 92... CGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAAGCTGACCCTGAAATTCATCTGCACAACAGGCAAACTGCCCGTGCCTT
ah3_1_con... CGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAAGCTGACCCTGAAATTCATCTGCACAACAGGCAAACTGCCCGTGCCTT
.....
```

Figure D. 2: Sequencing result of AH3

```

2747                                     2828
RNF219 to... CACTAAAGGAGGAGGTACCGgcaagagtagcggtaggtcggcgggacttccgtgttgccgggattctgaacgctgccGC
ah5_5_con... CACTAAAGGAGGAGGTACCGGCGAAGAGTAGCGGTAGGTGCGCGGACTTCCGTGTTGGCGGGATTCTGAACGCTGCCGC
.....

2829                                     2910
RNF219 to... CGCCACCATGAGCAAGGGCGAGGAAGTGTTCACCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGCCAC
ah5_5_con... CGCCACCATGAGCAAGGGCGAGGAAGTGTTCACCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGCCAC
.....

2911                                     2992
RNF219 to... AAATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAGCTGACCTGAAATTCATCTGCACAACAGGCAAAC
ah5_5_con... AAATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAGCTGACCTGAAATTCATCTGCACAACAGGCAAAC
.....

2993                                     3074
RNF219 to... TGCCCGTGCCCTTGGCCTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCACATGAAAAG
ah5_5_con... TGCCCGTGCCCTTGGCCTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCACATGAAAAG
.....

```

Figure D. 3: Sequencing result of AH5

```

2753                                     2834
HIST1H4H ... AGGGAGGAGGTACCGgcatataagggcggttgctttgccttgccgttttaggtttcttaagttgggttttagaagttgcttag
ah7_3_con... AGGGAGGAGGTACCGGCATATAAGGGCGTTGCTTTGCCTTGCGCTTTAGGTTTCTTAAGTTGGTTTTAGAAGTTGCTTAG
.....

2835                                     2916
HIST1H4H ... tcGCCGCCACCATGAGCAAGGGCGAGGAAGTGTTCACCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGG
ah7_3_con... TCGCCGCCACCATGAGCAAGGGCGAGGAAGTGTTCACCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTGAACGG
.....

2917                                     2998
HIST1H4H ... CCACAAATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAGCTGACCTGAAATTCATCTGCACAACAGGC
ah7_3_con... CCACAAATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAGCTGACCTGAAATTCATCTGCACAACAGGC
.....

2999                                     3080
HIST1H4H ... AAAGTGGCCGTGCTTGGCCTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCACATGA
ah7_3_con... AAAGTGGCCGTGCTTGGCCTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCACATGA
.....

```

Figure D. 4: Sequencing result of AH7

```

2749                                     2830
PGM2 to 9... AAGGGAGGAGGTACCGcagctcttggggcgccggcggaaggcagatctcaccgcctgcttccctctgcagcggtagcacaag
ah9_3_con... AAGGGAGGAGGTACCCACGTCTTGGGGCCGGGCCGGAAGGCAGATCTCACCGCCTGCTTCCTCTGCAGCGGTAGCACAA
.....

2831                                     2912
PGM2 to 9... ctcagcgGCCGCCACCATGAGCAAGGGCGAGGAAGTGTTCACCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTG
ah9_3_con... CTCAGCGGCCGCCACCATGAGCAAGGGCGAGGAAGTGTTCACCGGCGTGGTGCCCATCCTGGTGGAAGTGGACGGAGACGTG
.....

2913                                     2994
PGM2 to 9... AACGGCCACAAATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAGCTGACCTGAAATTCATCTGCACAA
ah9_3_con... AACGGCCACAAATTCTCCGTGCGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAGCTGACCTGAAATTCATCTGCACAA
.....

2995                                     3076
PGM2 to 9... CAGGCAAACTGCCCGTGCCCTTGGCCTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCA
ah9_3_con... CAGGCAAACTGCCCGTGCCCTTGGCCTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCA
.....

```

Figure D. 5: Sequencing result of AH9

```

820                                     901
pCDNA3.1=... tctggcttaactagagagacccActgggaattcACATTGCTTCTGACACAACGTGTGTTCACTAGCAACCTCAAAACAGACACCG
cmv2_cmv1... TCTGGCTAACTAGAGAACCCACTGGGAATTCACATTGCTTCTGACACAACGTGTGTTCACTAGCAACCTCAAAACAGACACCG

-----

902                                     983
pCDNA3.1=... GATCCGCCGCCACCATGAGCAAGGGCGAGGAACGTTCACCGGGCGTGGTGGCCATCCTGGTGGAACTGGACGGAGACGTGAA
cmv2_cmv1... GATCCGCCGCCACCATGAGCAAGGGCGAGGAACGTTCACCGGGCGTGGTGGCCATCCTGGTGGAACTGGACGGAGACGTGAA

-----

984                                     1065
pCDNA3.1=... CGGCCACAAATTCCTCGTGGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAAGCTGACCTGAAATTCATCTGCACAACA
cmv2_cmv1... CGGCCACAAATTCCTCGTGGGGGAGAGGGCGAAGGCGACGCTACAAACGGAAAGCTGACCTGAAATTCATCTGCACAACA

-----

1066                                     1147
pCDNA3.1=... GGCAAACTGCGCGTGGCTTGGCTTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCCACA
cmv2_cmv1... GGCAAACTGCGCGTGGCTTGGCTTACCTCGTTACAACCTGACATACGGCGTGCAATGTTTTCTAGATACCCAGACCCACA

-----

1148                                     1229
pCDNA3.1=... TGAAAAGACATGATTTTTTCAAAAGCGGCATGCCCGAGGGCTACGTGCAGGAGCGGACCATCAGCTTCAGGACGACGGAAC
cmv2_cmv1... TGAAAAGACATGATTTTTTCAAAAGCGGCATGCCCGAGGGCTACGTGCAGGAGCGGACCATCAGCTTCAGGACGACGGAAC

-----

1230                                     1311
pCDNA3.1=... CTACAAGACCCAGAGCGGAGGTCAAGTTGAGGGGCGACACCTGGTGAATAGAAATCGAGCTGAAGGGCATCGACTTTAAGGAA
cmv2_cmv1... CTACAAGACCCAGAGCGGAGGTCAAGTTGAGGGGCGACACCTGGTGAATAGAAATCGAGCTGAAGGGCATCGACTTTAAGGAA

-----

1312                                     1393
pCDNA3.1=... GATGGCAACATCCTGGGCCACAAGCTGGAGTACAACCTCAACAGCCACAACGTGTACATCAACCGCGACAAGCAGAGAAACG
cmv2_cmv1... GATGGCAACATCCTGGGCCACAAGCTGGAGTACAACCTCAACAGCCACAACGTGTACATCAACCGCGACAAGCAGAGAAACG

-----

1394                                     1475
pCDNA3.1=... GCATCAAGGCCAATTTCAAGATCAGACACAATGTGGAAGATGGCAGCGTGCAGCTGGCCGATCACTATCAGCAGAACACACCC
cmv2_cmv1... GCATCAAGGCCAATTTCAAGATCAGACACAATGTGGAAGATGGCAGCGTGCAGCTGGCCGATCACTATCAGCAGAACACACCC

-----

1476                                     1557
pCDNA3.1=... TATCGGCGATGGAACCTGTGCTGTGCTGACAAACCACTACCTGAGCACCCAGTCTGTGCTGTCCAGGACCTTAACGAGAG
cmv2_cmv1... TATCGGCGATGGAACCTGTGCTGTGCTGACAAACCACTACCTGAGCACCCAGTCTGTGCTGTCCAGGACCTTAACGAGAG

-----

1558                                     1639
pCDNA3.1=... CGGGACCATGAGTCCCTTCTGGAGTTCGTGACCGCGCTGGCATTACCCACGGCATGGACGAGCTGTACAAGTAATAGCTC
cmv2_cmv1... CGGGACCATGAGTCCCTTCTGGAGTTCGTGACCGCGCTGGCATTACCCACGGCATGGACGAGCTGTACAAGTAATAGCTC

-----

1640                                     1721
pCDNA3.1=... GCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCC
cmv2_cmv1... GCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCC

```

Figure D. 6: Sequencing result of CMV7

APPENDIX E

ADDITIONAL RECIPES

Table E. 1: LB Medium and Agar Plate Recipe

Ingredients	Amounts
Tryptone	10 g
Sodium Chloride	10 g
Yeast Extract	5 g
Agar (Only for agar plates)	15 g
Water	1000 ml

Table E. 2: PCR reaction with Q5

Ingredients	Amounts (25 µl)
5x Q5 Reaction Buffer	5 µl
10 mM dNTP	0.5 µl
10 µM Forward Primer	1.25 µl
10 µM Reverse Primer	1.25 µl
Template DNA	Less than 1 µg
Q5 DNA Polymerase	0.5 µl
Nuclease Free Water	Up to 25 µl

Table E. 3: Double Digestion Reaction

Ingredients	Amounts (50 µl)
10x Cut Smart Buffer	5 µl
DNA	2 µl
Restriction Endonuclease 1	1 µl
Restriction Endonuclease 2	1 µl
Nuclease Free Water	Up to 50 µl

Table E. 4: Gibson Assembly Mixture (1.33X)

Ingredients	Amounts
Taq Ligase	50 µl
5X Isothermal Buffer	100 µl
T5 Exonuclease	2 µl
Phusion Polymerase	6.25 µl
Nuclease Free Water	216.75 µl

Table E. 5: TSS Buffer Composition

Ingredients	Amounts
PEG 8000	5 g
1M MgCl ₂	1.5 ml
2.5 ml DMSO	2.5 ml
LB Medium	Up to 50 ml

Table E. 6: T4 Ligation Reaction

Ingredients	Amounts
10x T4 DNA Ligase Buffer	2 µl
Vector DNA	50 ng
Insert DNA	1:1, 3:1 or 5:1 molar ratio to vector
T4 DNA Ligase	1 µl
Nuclease Free Water	Up to 20 µl

Table E. 7: 10X TBE Buffer Recipe

Ingredients	Amounts
Tris Base	121.1 g
Boric Acid	61.8 g
EDTA	7.1 g

APPENDIX F

R SCRIPTS & API LINKS FOR BIOINFORMATIC ANALYSIS

Combination Script

```
library(readxl)

library(stringr)

library(xlsx)

library(DGEobj.utils)

library(edgeR)

#Transcriptomic

all_trans=read_excel('transcriptomic1.xlsx','transcriptomic') #GSE208731

all_trans$length=NA

all_trans_before_omitting_unnecessary=all_trans[,c(1,14,15,16,17,18,19,26,27,28,29,
30,31)]

all_trans_before_omitting_unnecessary$mean=rowMeans(all_trans_before_omitting_
unnecessary[,-1])

all_trans_required=all_trans_before_omitting_unnecessary[all_trans_before_omitting
_unnecessary$mean!=0,]

all_trans_required$length=NA

all_trans_required$UTR='Not found'

all_trans_required$sens=NA
```

```

supply_name1=read_excel('grch3738.xlsx','trans2')#GSE114332

supply_name2=read_excel('grch3738.xlsx','gtex') #directly from gtex v8 mean tpm
folder

supply_name=rbind(supply_name1,supply_name2)

tr_length=read_excel('grch3738.xlsx','grch38') #ensemble gene name, Trans length,
gene id, gene id ver

tr_length2=read_excel('grch3738.xlsx','grch37') #ensemble37 gene name, Trans
length, gene id, gene id ver

tr_length_known=tr_length[!is.na(tr_length$Id),]
tr_length_unknown=tr_length[is.na(tr_length$Id),]

nam=NA
for (i in 1:dim(tr_length_unknown)[1]){
  nam=supply_name$Id[supply_name$Id_ver==tr_length_unknown$Id_ver[i]]
  if (length(nam)){
    tr_length_unknown$Id[i]=unique(nam)
  }
  nam=NA
}

tr_length_known=rbind(tr_length_known,tr_length_unknown)

tr_length_all=tr_length_known[!is.na(tr_length_known$Id),]

len_vec=c()

```

```

seq_vec=c()

unknown=c()

unknown2=c()


for (i in 1:dim(all_trans_required)[1]){

  len_vec=na.omit(tr_length_all$len[tr_length_all$id==all_trans_required$Gene[i]])

  if (length(len_vec)){

    ind=which.max(len_vec)

    all_trans_required$length[i]=len_vec[ind]

  }

  else{

    unknown=c(unknown,all_trans_required$Gene[i])

  }

  len_vec=c()

}


all_trans_unknown=all_trans_required[is.na(all_trans_required$length),]

all_trans_known=all_trans_required[!is.na(all_trans_required$length),]

for (i in 1:dim(all_trans_unknown)[1]){

  len_vec=na.omit(tr_length2$len[tr_length2$id==all_trans_unknown$Gene[i]])

  if (length(len_vec)){

    ind=which.max(len_vec)

    all_trans_unknown$length[i]=len_vec[ind]

  }

}

```

```

else{

  unknown2=c(unknown2,all_trans_unknown$Gene[i])

}

len_vec=c()

}

all_trans_known=rbind(all_trans_known,all_trans_unknown)

all_trans_unknown=all_trans_known[is.na(all_trans_known$length),]

all_trans_known=all_trans_known[!is.na(all_trans_known$length),]

ens=c()

len_kontrol_eksik=c()

for (i in 1:dim(all_trans_unknown)[1]){

  ens=unique(supply_name$Id_ver[supply_name$Id==all_trans_unknown$Gene[i]])

  if (length(ens)>1){

    len_kontrol_eksik=c(len_kontrol_eksik,all_trans_unknown$Gene[i])

    next

  }

  len_vec=na.omit(tr_length$len[tr_length$Id_ver==ens])

  if (length(len_vec)){

    ind=which.max(len_vec)

    all_trans_unknown$sens[i]=ens

    all_trans_unknown$length[i]=len_vec[ind]

  }

  else {

```

```

len_vec=na.omit(tr_length2$len[tr_length2$id_ver==ens])

if (length(len_vec)) {

  ind=which.max(len_vec)

  all_trans_unknown$ens[i]=ens

  all_trans_unknown$length[i]=len_vec[ind]

}

else{

  len_kontrol_eksik=c(len_kontrol_eksik,all_trans_unknown$Gene[i])

}

}

len_vec=c()

ens=c()

}

all_trans_known=rbind(all_trans_known,all_trans_unknown)

all_trans_unknown=all_trans_known[is.na(all_trans_known$length),]

all_trans_known=all_trans_known[!is.na(all_trans_known$length),]


mcf10a_counts=all_trans_known[,c(2,3,4,5,6,7)]

mcf7_counts=all_trans_known[,c(8,9,10,11,12,13)]

mcf10a_counts_dataframe=as.data.frame(mcf10a_counts)

mcf7_counts_dataframe=as.data.frame(mcf7_counts)

rownames(mcf10a_counts_dataframe)=all_trans_known$Gene

rownames(mcf7_counts_dataframe)=all_trans_known$Gene

mcf10a_counts_matrix=as.matrix.data.frame(mcf10a_counts_dataframe)

```

```

mcf7_counts_matrix=as.matrix.data.frame(mcf7_counts_dataframe)

mcf10a_tpm=convertCounts(mcf10a_counts_matrix,unit="TPM",geneLength =
all_trans_known$length)

mcf7_tpm=convertCounts(mcf7_counts_matrix,unit="TPM",geneLength =
all_trans_known$length)

mcf10a_counts_dataframe$tpm_ave=rowMeans(mcf10a_tpm)

mcf7_counts_dataframe$tpm_ave=rowMeans(mcf7_tpm)


#Proteomic

mcf10a_pro=read_excel('15.03.2023_proteomicsdb.xlsx','mcf10a_proteomics')

mcf10a_pro$gene=NA

mcf10a_exp=read_excel('15.03.2023_proteomicsdb.xlsx','mcf10a_exp_2_<0.05_unique')

mcf7_pro=read_excel('15.03.2023_proteomicsdb.xlsx','mcf7_proteomics')

mcf7_pro$gene=NA

mcf7_exp=read_excel('15.03.2023_proteomicsdb.xlsx','mcf7_exp_2_<0.05_unique')

gene_name=NA

eliminated_genes_mcf10a=c()

for (i in 1:dim(mcf10a_pro)[1]){

gene_name=mcf10a_exp$Value.GENE_NAME[mcf10a_exp$Value.UNIQUE_IDENTIFI
TIFIER==mcf10a_pro$Value.UNIQUE_IDENTIFIER[i]]

if (length(gene_name)==0){

```

```

eliminated_genes_mcf10a=c(eliminated_genes_mcf10a,mcf10a_pro$Value.UNIQUE
_IDENTIFIER[i])

  gene_name=NA

  next

}

mcf10a_pro$gene[i]=gene_name

gene_name=NA
}

gene_name=NA
eliminated_genes_mcf7=c()
for (i in 1:dim(mcf7_pro)[1]){

gene_name=mcf7_exp$Value.GENE_NAME[mcf7_exp$Value.UNIQUE_IDENTIFI
ER==mcf7_pro$Value.UNIQUE_IDENTIFIER[i]]

  if (length(gene_name)==0){

eliminated_genes_mcf7=c(eliminated_genes_mcf7,mcf7_pro$Value.UNIQUE_IDEN
TIFIER[i])

  gene_name=NA

  next

}

mcf7_pro$gene[i]=gene_name

gene_name=NA

}

```

```

mcf10a_pro_known=mcf10a_pro[!is.na(mcf10a_pro$gene),]
mcf7_pro_known=mcf7_pro[!is.na(mcf7_pro$gene),]
uni_mcf10a=data.frame(gene=unique(mcf10a_pro_known$gene))
uni_mcf10a$mcf10a_mean_normalized_expression=NA
uni_mcf7=data.frame(gene=unique(mcf7_pro_known$gene))
uni_mcf7$mcf7_mean_normalized_expression=NA
for (i in 1:dim(uni_mcf7)[1]){

uni_mcf7$mcf7_mean_normalized_expression[i]=mean(as.double(mcf7_pro_known$
Value.NORMALIZED_EXPRESSION[mcf7_pro_known$gene==uni_mcf7$gene[i]])
)
}

for (i in 1:dim(uni_mcf10a)[1]){

uni_mcf10a$mcf10a_mean_normalized_expression[i]=mean(as.double(mcf10a_pro_
known$Value.NORMALIZED_EXPRESSION[mcf10a_pro_known$gene==uni_mcf
10a$gene[i]]))
}

common_mcf7=uni_mcf7[which(uni_mcf7$gene %in% uni_mcf10a$gene),]
common_mcf10a=uni_mcf10a[which(uni_mcf10a$gene %in% uni_mcf7$gene),]
common_both=cbind(common_mcf10a,common_mcf7)
common_both=common_both[,c(1,2,4)]

#Combination of both

```

```

common_both$mcf10a_tpm=NA

common_both$mcf7_tpm=NA

tpm_mcf10a=NA

tpm_mcf7=NA

not_found_mcf10a=c()

not_found_mcf7=c()

for (i in 1:dim(common_both)[1]){

  tpm_mcf10a=mcf10a_counts_dataframe$tpm_ave[rownames(mcf10a_counts_dataframe)==common_both$gene[i]]

  tpm_mcf7=mcf7_counts_dataframe$tpm_ave[rownames(mcf7_counts_dataframe)==common_both$gene[i]]

  if (length(tpm_mcf10a)==0 | length(tpm_mcf7)==0 ){

    if(length(tpm_mcf10a)==0 ){

      not_found_mcf10a=c(not_found_mcf10a,common_both$gene[i])

      tpm_mcf10a=NA

    }

    if(length(tpm_mcf7)==0 ){

      not_found_mcf7=c(not_found_mcf7,common_both$gene[i])

      tpm_mcf7=NA

    }

    next

  }

  common_both$mcf10a_tpm[i]=tpm_mcf10a

```

```

common_both$mcf7_tpm[i]=tpm_mcf7

tpm_mcf10a=NA

tpm_mcf7=NA

}

#ratio calculations

common_both$mcf7ratio=common_both$mcf7_mean_normalized_expression/comm
on_both$mcf7_tpm

common_both$mcf10aratio=common_both$mcf10a_mean_normalized_expression/c
ommon_both$mcf10a_tpm

common_both$totalratio=common_both$mcf7ratio/common_both$mcf10aratio

common_both=common_both[!is.na(common_both$mcf10a_tpm),]

common_both=common_both[!is.na(common_both$mcf7_tpm),]

common_both=common_both[common_both$mcf7_tpm>0,]

common_both=common_both[common_both$mcf10a_tpm>0,]


#UTR matching


ch38_utr=read_excel('grch38_utr.xlsx','utr')

ch38_utr_known=ch38_utr[ch38_utr$seq!='Sequence unavailable',]

ch38_utr_known=ch38_utr_known[!is.na(ch38_utr_known$id),]


common_both$UTR=NA

utr_vec=NA

len_vec=NA

ind=NA

```

```

ens=NA

new_name=NA

unknown_UTR=c()

for (i in 1:dim(common_both)[1]){

  len_vec=ch38 utr_known$len[ch38 utr_known$id==common_both$gene[i]]

  utr_vec=ch38 utr_known$seq[ch38 utr_known$id==common_both$gene[i]]

  if (length(len_vec)){

    ind=which.max(len_vec)

    common_both$UTR[i]=utr_vec[ind]

  }

  else{

    ens=unique(supply_name$id_ver[supply_name$id==common_both$gene[i]])

    new_name=unique(na.omit(tr_length$id[tr_length$id_ver==ens]))

    len_vec=ch38 utr_known$len[ch38 utr_known$id==new_name]

    utr_vec=ch38 utr_known$seq[ch38 utr_known$id==new_name]

    if (length(len_vec)){

      ind=which.max(len_vec)

      common_both$UTR[i]=utr_vec[ind]

    }

    else{

      unknown_UTR=c(unknown_UTR,common_both$gene[i])

    }

  }

}

utr_vec=NA

len_vec=NA

```

```

ind=NA

ens=NA

new_name=NA

}

#discarding low ratio 20 genes due to lack of UTR sequence information
common_both=common_both[!is.na(common_both$UTR),]

#according to ensemble cutoff:0.5
common_both_over_0.5=common_both[common_both$mcf10a_tpm>0.5,]
common_both_over_0.5=common_both[common_both$mcf7_tpm>0.5,]
common_both_over_0.5=common_both_over_0.5[order(common_both_over_0.5$totalratio,decreasing = TRUE),]

write.xlsx(common_both_over_0.5,'Final_combine_data_filtered.xlsx',sheetName =
'combine')

```

Counting_statistics Script

```

library(tictoc)

source("consensus_found.R")

source("consensus_counter.R")

tic('total')

cc=consensus_found(len=2,location =

common_both_over_0.5$UTR[common_both_over_0.5$totalratio>3])

c_all=data.frame(pattern=cc,len=2)

for (i in 3:19){

```

```

tic('for loop duration')

cc=consensus_found(pattern_list=cc,len=i,location =
common_both_over_0.5$UTR[common_both_over_0.5$totalratio>3])

c_all=rbind(data.frame(pattern=cc,len=i),c_all)

toc()

}

toc()

final_table=consensus_counter(c_all$pattern,common_both_over_0.5$UTR,common
_both_over_0.5$totalratio,1)

final_table_sign=final_table[final_table$p_val<0.05,]

final_table_sign=final_table_sign[final_table_sign$odds>3,]

final_table_contains=data.frame(pattern=character(0),found=integer(0))

for (i in final_table_sign$pattern){

  tt=length(grep(i,final_table_sign$pattern,ignore.case = TRUE))

  final_table_contains[nrow(final_table_contains)+1,]=list(i,tt)

}

final_table_contains_ordered=final_table_contains[order(final_table_contains$found,
decreasing = TRUE),]

write.xlsx(final_table_sign,'21.03.2023_found patterns.xlsx',sheetName =
"p_val<0.05,odds>3")

write.xlsx(final_table_contains_ordered,'21.03.2023_found patterns.xlsx',sheetName
= "repetitives",append = TRUE)

```

Consensus Finder Function

```
consensus_found <- function(pattern_list=c("A","T","C","G"),len,location){

  bases=c("A","T","C","G")

  pattern_list2=c()

  #pattern_all=c()

  sword=character(1)

  z=location

  while (length(pattern_list)!=0 & nchar(sword)<len){
    for(i in pattern_list){
      for (j in 1:4){

        sword=paste(i,bases[j],sep="")

        c=sum(grepl(sword,z,fixed=TRUE,useBytes = TRUE))

        if (c>1){

          pattern_list2=c(sword,pattern_list2)

        }

      }

    }

    pattern_list=pattern_list2

    #pattern_all=c(pattern_list,pattern_all)

    pattern_list2=c()

    print(nchar(sword))

  }

  return(pattern_list)
```

```
}
```

Consensus Counter Function

```
consensus_counter<-function(pattern_file,location,location_for_cutoff,cutoff){  
  final_file=  
  data.frame(pattern=character(0),len=integer(0),countinchosen=integer(0),inchosenrest  
  =integer(0),countNOTinchosen=integer(0),notinchosenrest=integer(0),probabilityinch  
  osen=numeric(0),probabilityNOTinchosen=numeric(0),differentiality_by_hand=nume  
  ric(0),odds=numeric(0),p_val=numeric(0))  
  for (i in pattern_file){  
    c=length(grep(i,location[location_for_cutoff>cutoff],ignore.case = TRUE))  
    c_not=length(location[location_for_cutoff>cutoff])-c  
    c2=length(grep(i,location[location_for_cutoff<=cutoff],ignore.case = TRUE))  
    c2_not=length(location[location_for_cutoff<=cutoff])-c2  
    c_ratio=c/c_not  
    c2_ratio=c2/c2_not  
    c_c2ratio=c_ratio/c2_ratio  
    dat_fr=data.frame('higher'=c(c,c_not),'lower'=c(c2,c2_not),row.names =  
    c('found','not found'),stringsAsFactors = FALSE)  
    ss=fisher.test(dat_fr)  
  
    final_file[nrow(final_file)+1,]=list(i,as.integer(nchar(i)),c,c_not,c2,c2_not,c_ratio*10  
    0,c2_ratio*100,c_c2ratio,ss$estimate[[1]],ss$p.value)  
  }  
  return(final_file)  
}
```

Table F. 1: API links for the proteomic data obtainment

Name of folder and its sheets	Content	Links
15.03.2023_proteomicsdb.xlsx / mcf10a_proteomics	MCF10a_proteomic	https://www.proteomicsdb.org/proteomicsdb/logic/api/proteinspertissue.xsodata/InputParams(TISSUE_ID='BTO:0001939',CALCULATION_METHOD=0,SWISSPROT_ONLY=1,NO_ISOFORM=1,TAXCODE=9606)/Results?\$select=ENTRY_NAME,UNIQUE_IDENTIFIER,DATABASE,PROTEIN_DESCRIPTION,PEPTIDES,TISSUE_ID,SAMPLE_NAME,SAMPLE_DESCRIPTION,UNNORMALIZED_EXPRESSION,NORMALIZED_EXPRESSION&\$format=json
15.03.2023_proteomicsdb.xlsx / mcf10a_experiment_2_<0.05_unique	MCF10a_data_experiment (FD R<0.05, evidence=	https://www.proteomicsdb.org/proteomicsdb/logic/api/proteinsperexperiment.xsodata/InputParams(EXPERIMENTFILTER=116)/Results?\$select=ENTRY_NAME,UNIQUE_IDENTIFIER,PROTEIN_NAME,GENE_NAME,PROTEIN_ID,PROTEIN_Q_VALUE,OBSERVED_SEQUENCE_COVERAGE_ALL,EVIDENCE,SEARCH_ENGINE&\$filter=ISUNIQUE%20eq%201&\$format=json

	2, unique(unique_identifier))	
15.03.2023_proteomicsdb.xlsx / mcf7_proteomics	MCF7_proteomics	https://www.proteomicsdb.org/proteomicsdb/logic/api/proteinspertissue.xsodata/InputParams(TISSUE_ID='BTO:0000093',CALCULATION_METHOD=0,SWISSPROT_ONLY=1,NO_ISOFORM=1,TAXCODE=9606)/Results?\$select=ENTRY_NAME,UNIQUE_IDENTIFIER,DATABASE,PROTEIN_DESCRIPTION,PEPTIDES,TISSUE_ID,SAMPLE_NAME,SAMPLE_DESCRIPTION,UNNORMALIZED_EXPRESSION,NORMALIZED_EXPRESSION&\$format=json
15.03.2023_proteomicsdb.xlsx / mcf7_exp_2_5_unique	MCF7_data_experiment (FD R<0.05, evidence=2, unique)	https://www.proteomicsdb.org/proteomicsdb/logic/api/proteinsperexperiment.xsodata/InputParams(EXPERIMENTFILTER=122)/Results?\$select=ENTRY_NAME,UNIQUE_IDENTIFIER,PROTEIN_NAME,GENE_NAME,PROTEIN_ID,PROTEIN_Q_VALUE,OBSERVED_SEQUENCE_COVERAGE_ALL,EVIDENCE,SEARCH_ENGINE&\$filter=ISUNIQUE%20eq%201&\$format=json

	<code>_identifier</code>)	
--	----------------------------	--

