Engineering Multi-Protein Assemblies with Machine Learning and Experimental Optimization

Thesis by

Ishaan Joshan Dev

In Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

in

Chemical Engineering

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2027

(Defended July 10th, 2026)

Ishaan Joshan Dev
ORCID: 0000-0001-6305-9841

Acknowledgements

I am very grateful to my friends, family, and mentors who have supported me throughout my career. First and foremost, I would like to thank my advisor, Dr. Mikhail Shapiro. Two years before I even came to Caltech, Mikhail provided me with a lot of guidance on fellowship proposal writing, graduate school selection, and career decisions. Even though he does science at the highest level imaginable, I admire how much time he carves out to mentor younger folks in the field. Throughout my time in graduate school, Mikhail always made himself available to chat about projects, science, and career trajectory while sufficiently dropping me in the deep end so I could learn to swim on my own. It has been a privilege to be a member of the Shapiro Lab, where Mikhail has fostered an atmosphere that blends old-school collegiality with more progressive and inclusive principles. In addition, I would like to thank the members of my Thesis Committee, Dr. Frances Arnold, Dr. Shu-ou Shan, and Dr. Matt Thomson.

Despite having earned a Ph.D., I cannot say I am good at “school.” Quite the contrary, I have always been decently challenged by academics. But I have had amazing mentors along the way that have taught me strategies to tackle problems that seem out of my reach. My first academic mentor is my mother, Niroo Dev. I struggled in math and English in grade school. I am grateful to my mother who recognized where I needed help and brainstormed with my teachers on ways to help me improve. And she came up with some fun ideas! We would sing along (not well) to a CD about multiplication tables. On weekends, we would play a math version of Go Fish. And at bedtime, we would read together. Most importantly, when the math and the English got harder and a little farther out of her understanding, she helped connect me with tutors and other resources to make sure I got the support I needed. I have been able to incorporate feedback and instruction from all the mentors in my life because my mother taught me at a young age that it is okay to get input from others and to admit where you are struggling.

With weekends spent at the science museum and trips to his chemistry lab, I am thankful to my dad who, whether knowingly or unknowingly, incepted me with an interest in the sciences. For better or worse, my dad has often taken an approach of “yes and” with me. When I wanted the full Batman cave for Christmas at the age of six, he said yes…and then also got me the matching Batmobile. When I would tell him I was too tired from schoolwork to empty the dishwasher or do my chores, he would do them in my place…and make me a snack. He spoiled me and I will never know true hardship in life because of him. However, in all that spoiling, he also showed me the most nourishing moments in life can come when lending someone a hand and supporting those around you not in the manner you think they should be supported but in the way they ask to be supported. He may have sheltered me from day-to-day hardships, but he always suggested being rigorous with my education. In high school, my dad would come home from work in the evening just to be bombarded with chemistry and physics questions I was struggling with. He never just told me the answers. He encouraged me to teach myself concepts through careful reading and re-reading of my textbooks. And only then, when I had grasped the fundamentals for myself, did he help me take my understanding further.

Although they may be the best folks around, my parents can still be…parents. One warrior has always protected me from their more annoying and nosy tendencies. To my sister, Dr. Sheena Dev, I credit you with not only my childhood sanity but as a constant source of inspiration. My sister is the most independent and coolest person I know. She is my favorite person to be around. She is an exceptional strategist and a fighter. From watching her do hard things, I have found the courage to do hard things. The greatest luck I’ve had in my life is being born to the same parents as her.

I also would like to thank some folks from my childhood who helped me get to this point. My time in Scouts taught me about discipline, how to work with others, and how to organize. My Scoutmaster, Mr. Henry Pena, poured so much of his time and effort to raise us boys to be independent and reliable individuals. He never lost his patience with us. He only ever thought about how he could help us grow. There may never be another Scoutmaster as dedicated as him. My aunt, Shalini, and uncle, Vinny, took me in for a summer during high school so that I could work at a lab in Palo Alto. My time with them was my first immersive experience into the world of research. Dr. Vinod Bhutani has always challenged my perspective of academia and what is possible as an academic, and I am truly grateful for that. Thank you to Mr. Mike Danielson and Mr. Kevin Sherrill for teaching me that through consistent hard work and thoughtful preparation one can earn the right to enjoy the show regardless of the outcome. To my high school teachers that shaped my love of the sciences, Mr. Jorge Munoz, Mrs. Cherryl Mynster, and Ms. Dorothy Daniel, thank you for every lesson plan and every creative metaphor. To Ashok Mittal, thank you for all your guidance at the major crossroads of my career.

I got my start in undergraduate research in the laboratory of Dr. Jay Keasling thanks to my mentors Dr. Leo d'Espaux and Dr. Jeff Wong. My time in the Keasling lab taught me all the foundational molecular biology and wet lab techniques that I use to this very day. Leo's "cook book" is still a resource I refer to often. Thank you to all my fabulous mentors in the Keasling Lab. In addition, thanks to Dr. Joshua Leonard for his mentorship in science policy and for helping me brainstorm about exciting directions at the forefront of synthetic biology. My time at Ginkgo Bioworks accelerated my scientific abilities in many avenues. I want to thank Dr. Sudeep Agarwala, Ruchita Chawla and Dr. Roy Marella for teaching me to "think in controls" and to approach every experiment in a meticulous and detailed manner. In addition, I owe my start in computational biology to the chance Dr. Joshua Dunn gave me when he allowed me to join the Design Team. The mentorship and guidance of Dr. Justin Gardin, Dr. Sandrine Boissel, and Dr. Lucas Hartsough provided me with the computational skillset to tackle many of the problems I explore in this thesis. Everyone I worked with at Ginkgo showed me you can both do great science and be a great person and collaborator.

I came to Caltech for the world-class professors and laboratories. But I stayed at Caltech because of the thought and effort Allison Kinnard and Dr. Mike Vicic pour into the Chemical Engineering Department. Allison and Mike kept me grounded throughout this experience, even when I could not see the light at the end of the tunnel. Thank you to Dr. Sabine Brinkmann-Chen for advising me over the years, whether it be editing research proposals or helping me get trained on laboratory equipment.

In the Shapiro Lab, I am very grateful to Dr. Robert Hurt, Dr. Erik Schrunk, and Dr. Mengtong Duan, who took me under their wings when I joined the lab. Erik's creative project ideas, kindness, and genuine personality made me want to join the lab. Rob saved me from a lonely second floor office and helped me find a home in the lab. Tom showed me all the wet-lab skills I needed to be successful with mammalian biology and pushed me to get out of my own way as a scientist. To Dr. Pradeep Ramesh, I felt graduate school lacked an "intellectual momentum" until you started mentoring me. You have truly helped me grow as a scientist in the directions that I am passionate about. I am thankful for Dr. Ernesto Criado-Hidalgo, who showed me that being an ambitious scientist never needs to come at the cost of being yourself and a decent human being. And thank you to the many postdoctoral mentors I have had over the years. Dr. Ernesto Criado-Hidalgo, Dr. Nivin Nystrom, Dr. Brian Zhong, Dr. William Benman, Dr. Di Wu, Dr. Josh Chen, Dr. Zhiyang Jin, Dr. Alen Pavlic, and Dr. Ahyoung Kim, you all approach science in such unique, creative, and wonderfully weird and amazing ways. If there were a trading card game for academics, I would put all of yours in plastic sleeves and store them in a temperature-controlled environment.

I come into the lab every day for the chance to work alongside Ann Liu, Liz Hughes, Pierina Barturen, Hao Shen, Hengyu Li, Daniel Tang, Anthony Phung, Akanksha Yadev, George Daghlian, Jimmy Kim, Sunho Lee, Kaiwen Luo, Dr. Rohit Nayak, and Dr. Jee Won Yang. You all never cease to make my day. Special thanks to Anthony Phung for his golden hands. Rosie Zedan, your consistency, your care for us, your cooking, and your work ethic really have set the tone for our lab over the years and for that I am so grateful. Thank you to Grace Shin and Margaret Swift for managing our crazy lab along with our RAs Julio Revilla, Trevor Fochtman, Haein An, Goar Ayrapetyan, Ruby Zhang, and Rosmelle Zelgar. No work would get done without our lab management team.

To our undergraduate researchers who have trusted me to fill their brains with mostly nonsense and some scientific principles, I can say there is nothing I have loved more than working with you all. Arul Goel, Tammy Sujaritchai, Noor Ibrahim, and Jamie Kwon thank you for all your contributions. I am so excited to see where you all end up in the future. I'm sure I'll end up working for at least one of you!

I would also like to thank our core facility leaders, Dr. Michael Gregory and Dr. Igor Antoshechkin for their support and guidance in collecting flow cytometry and sequencing data. In addition, thank you to Dr. James Linton and Miki Yun for all their help accessing tools and equipment in their respective laboratories. Thank you to Dr. Natasha George for your guidance and writing tips in preparing this work and thank you to all the members of the HPC that are running the smoothest operation on campus. Thank you to other technical contributors, Dr. Yueming Long, Dr. Andrew Lu and Mei Yi You.

To my dearest friends from Berkeley, Dr. Ravi Lal, Dr. Alec Glisman, Kayla Panora, and Eric Tran, I love you all so much and I am so lucky you chose me to be your friend a decade ago. To Abraham Polayes and Dr. Katie Sicinski, thank you for keeping me mobile, loose, engaged, and filled with gratitude. Maia Helterbrand, Vignesh Bethanabotla, Grace Holbrook and Ojas Pradhan, thank you for all your support and making Caltech a worthwhile place to be. To my roommates Mark Legendre and Julio Revilla, I enjoy coming home every day to you both, our countless movie marathons, and all our banter. I have never had a better set of roommates. Thanks to Dr. Miya Erickson and Dr. Marcus Soto for providing me with the physical and mental tools that have transformed how I approach my everyday life. Thanks to Morgan Mapstone and Robert Hurt for small conversations that gave me the courage to get through the hardest of moments. Everyone I have ever played softball with, thanks for putting up with my crazy energy.

Abstract

Mammalian Acoustic Reporter Gene (mARG) expression systems are used to express gas vesicle (GV) acoustic contrast agents, enabling visualization of mammalian cell dynamics in living organisms using ultrasound. Further advancement of mARGs is encumbered by all-too-common problems in synthetic biology for mammalian systems; there are limited frameworks to compactly express multiple protein components in distinct stoichiometries in a manner amenable for gene delivery. Here, we present new methods of polycistronic expression in mammalian cells through exploitation of the natural scanning behavior of ribosomes. We term this system Stoichiometric Expression of mRNA Polycistrons from Eukaryotic Ribosomes (SEMPER). The tunability of SEMPER systems is greatly reduced when proteins contain methionines within their structure. Therefore, we go on to demonstrate how zero-shot predictions yielded by a set of protein language models and structure-based methods can be used to generate methionine-free proteins with minimal loss in activity. As demonstrated on mARGs, SEMPER combined with our zero-shot prediction workflow provides a new and efficient toolkit by which synthetic biologists can encode multi-protein expression systems for complex synthetic biology applications. We complement these advances by characterizing cellular responses to gas vesicle expression using RNA-seq and by developing new methods to screen cellular acoustic properties.

Published Content and Contributions

Duan, M.*, **Dev, I.*,** Lu, A., Ayrapetyan, G., You, M.Y., and Shapiro, M.G. (2024). SEMPER: Stoichiometric expression of mRNA polycistrons by eukaryotic ribosomes for compact, ratio-tunable multi-gene expression. Cell Syst *15*, 597-609.e4. https://doi.org/10.1016/j.cels.2024.06.001.

I.D. helped conceive and execute the study, analyze data, and write the manuscript. I.D. constructed the fluorescent protein SEMPER breadboarding system. I.D. developed the SEMPER IVT mRNA breadboarding system. I.D. performed and analyzed IVT mRNA experiments and 2-ORF and 3-ORF FP expression experiments and developed the computational model.

Dev, I., Baskota, A., Shapiro, M., and Lal, A. (2024). Combining gas vesicle acoustic contrast agents with field-deployable GHz frequency ultrasonic chips to detect gene expression in sentinel microorganisms. In 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS), pp. 1–4. https://doi.org/10.1109/UFFC-JS60046.2024.10793589.

I.D. helped conceive and execute the study, analyze data, and write the manuscript. I.D. conducted gas vesicle induction experiments in bacteria. I.D. helped with ultrasound data processing.

Dev, I., … Ramesh, P., Shapiro, M.G., et al. (2026, In preparation). Zero-shot mutagenesis enables stoichiometric expression of multi-protein complexes.

I.D. helped conceive the study. I.D. led the execution of the study, analyzed the data, and wrote the manuscript. I.D. created the computational pipeline for conducting zero-shot predictions, the assay methodology for luminescence, fluorescence, and ultrasound measurements, and the construction and sequencing strategy for plasmids. I.D. curated all model inputs.

* Equal Contribution

Table of Contents

List of Illustrations and Tables

Nomenclature

Halo: *Halobacterium salinarum*

Ana: *Anabaena flos-aquae*

Mega: *Bacillus megaterium*

FP: Fluorescent protein

GV: Gas vesicle

ARG: Acoustic Reporter Genes

mARG: Mammalian Acoustic Reporter Genes

pgvpA: Vector containing. GvpA

pgvpN-W: Vector containing Gvp chaperones

CDS: Coding sequence

ORF: Open reading frame

GOI: Gene of interest

POI: Protein of interest

ROI: Region of Interest

rtTA: reverse tetracycline-controlled transactivator.

TRE: tetracycline response element

CMV: cytomegalovirus

Chapter 1

Motivation and General Background

Imaging and Controlling Cells Inside Organisms

The broader motivation of this work is to improve methods of noninvasive imaging and control of cells inside genetic models and eventually clinical patients. Specifically, we are interested in using ultrasound to dynamically visualize and communicate with engineered cells *in vivo* to enable richer study of natural cellular processes and to yield new breakthroughs in human health in domains such as neurodegeneration, immunology, oncology, and the gut microbiome. To exemplify these concepts, we will first discuss the growing interest in the use of molecular imaging to track and control engineered cells, such as chimeric antigen receptor (CAR) T cells, that carry out therapeutic functions in the body.[1] With CAR T-cell therapy, circulating T cells are extracted from a patient's blood and engineered to express a chimeric antigen receptor which endows these cells with cancer targeting abilities (**Figure 1-1A**). These CAR T-cells are then returned to the patient's circulation, allowing them to identify and destroy cancers. However, CAR T cells and similar cell-based therapies have limitations. For example, once reintroduced into the patient's body, these cells can generate an excess of therapeutic molecules at undesired times or locations, resulting in overdose or damage to healthy tissues.[2] In addition, CAR T cell therapy is still not efficient for targeting solid tumors. To mitigate these harmful side effects and study the limits of CAR T cell targeting, molecular imaging technologies may be used to track and precisely control the genetic circuitry of these living therapeutics *in vivo* to improve their study and activate their therapeutic function at desired times and locations in the body. The example provided for tracking and controlling CAR T cells is easily extrapolated to other biological contexts, including tracking and controlling the colonization of various microbes in the gut microbiome and observing how cancers metastasize or grow over time.

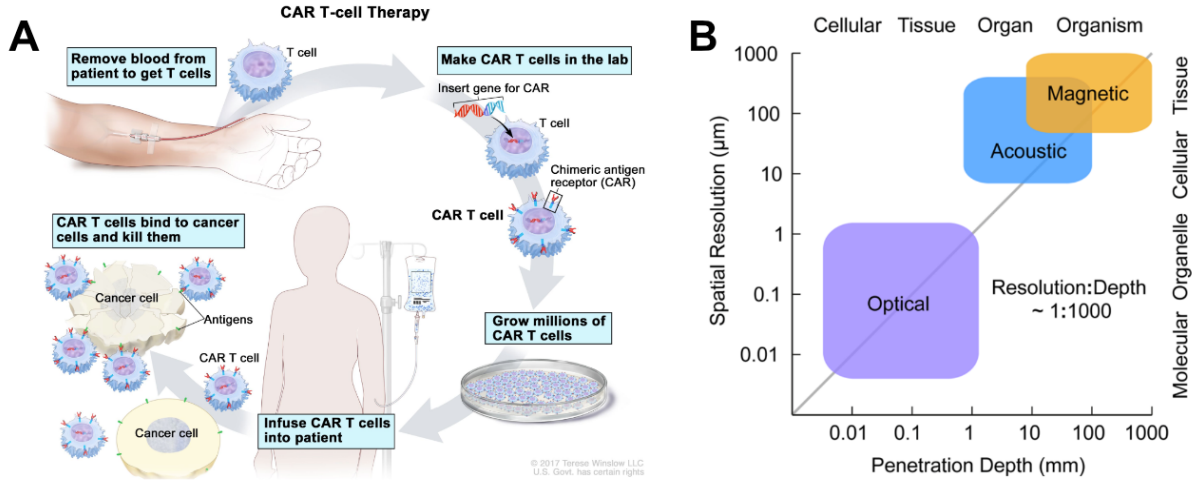

Figure 1-1: Molecular imaging for tracking and controlling cells in vivo.
A) Chimeric antigen receptor (CAR) Therapy development and administration (Reprinted from Winslow).[3] **B**) Comparison of penetration depth and spatial resolution for three common methods of molecular imaging (Reprinted from Piraner et al.).[4]

To enable precise tracking and control of an engineered cell in genetic models or even patients, researchers and clinicians must choose a molecular imaging modality that i) has sufficient

penetration depth, ii) has sufficient spatial resolution, iii) has the ability to supply heat to and do work on a specific region of interest (ROI), and iv) is affordable and accessible (**Figure 1-1B**).[4] For biological research, a useful molecular imaging modality must be able to penetrate through ~1-10 cm of tissue to image ROIs deep within organisms. As visible light used for optical imaging is scattered within 1 mm of most tissues, its use is limited to thin *in vitro* samples or genetic models with significant optical clarity, like zebrafish. A useful molecular imaging modality for our goals must also be able to resolve structures at the level of whole tissues, neuronal pathways, and organs (~5 cm) all the way down to the scale of single cells (~10 μm). Molecular imaging tools that use acoustic or magnetic waves (i.e., medical ultrasound and magnetic resonance imaging (MRI)) have significant penetration depth and spatial resolution on the order of ~100 μm. These modalities can also deposit heat and deliver mechanical forces to cells and tissues. Satisfying our final criteria, ultrasound is compact, affordable, and widely accessible compared to MRI, allowing it to be readily deployed for both research and point-of-care clinical applications.[5] With adequate spatial resolution and significant penetration depth at frequencies of 3-10 MHz, ultrasound is an ideal candidate for imaging and controlling cells *in vivo.* However, new acoustic contrast agents and biomolecular tools must be developed to further improve the sensitivity of medical ultrasound to detect single cells and enable gene expression in target cells using mechanical energy or heat delivered by acoustic waves.

Our laboratory has been engineering genetically encoded gas vesicle acoustic reporter proteins to increase the sensitivity of ultrasound for the detection of microorganisms and mammalian cells *in vivo*. Furthermore, members of our lab are seeking to use gas vesicles (GV) as sensors for biologically relevant ions and molecules, reporters of cell state and environmental changes (e.g., hypoxia), and actuators of gene expression in response to specific ultrasound parameters. The main contributions of this work are to improve genetic expression systems for producing gas vesicles in individual mammalian cells to advance sensitive, noninvasive, and dynamic imaging of these cells *in vivo*. The following section provides an overview of the gas vesicle and genetically encoded acoustic reporter genes (ARGs).

Gas Vesicles: An Acoustic Contrast Agent

Discovery and Evolutionary Purpose

First discovered in cyanobacteria by German microbiologists, gas vesicles are hollow, gas filled proteinaceous structures common to over 150 species of prokaryotes, including multiple phyla of mostly aquatic bacteria and archaea (**Figure 1-2A, 1-2B**).[6] Gas vesicle structural proteins along with GV-specific assembly factors and chaperones are genetically encoded in these microorganisms by a cluster of 8-20+ genes referred to as “gvps” (**g**as **v**esicle **p**roteins). Gas vesicles are predominantly composed of many repeating units of a 7-8 kDa protein called GvpA. The GvpA protein makes up the ~2-4 nm thick shell of the gas vesicle, where the hydrophobic regions of GvpA are facing into the center of the shell to create a hydrophobic interior that excludes water and forms a gaseous pocket (**Figure 1-2C**).[7] In many microorganisms, GVs take on a cylindrical structure with conical ends, however, other gas vesicle geometries have been observed. Formed GVs are light-refracting structures that often increase the opacity of a microorganism upon production. Gas vesicles endow microbes with flotation-based motility, as their production decreases the density of cells and increases their buoyancy in the surrounding fluid. The phototropic bacteria, *Anabaena flos-aquae*, is known to produce gas vesicles in low-light conditions to increase its buoyancy, allowing it to float closer to the surface of a water body to access sunlight.[7] Some microorganisms that use GVs for motility have been observed to generate dense carbohydrates to act in opposition to the gas-filled structures, allowing the cells

to sink in the surrounding fluid. Other microorganisms, such as *Serratia sp. ATCC 39006*, produce gas vesicles in response to low oxygen conditions.[8]

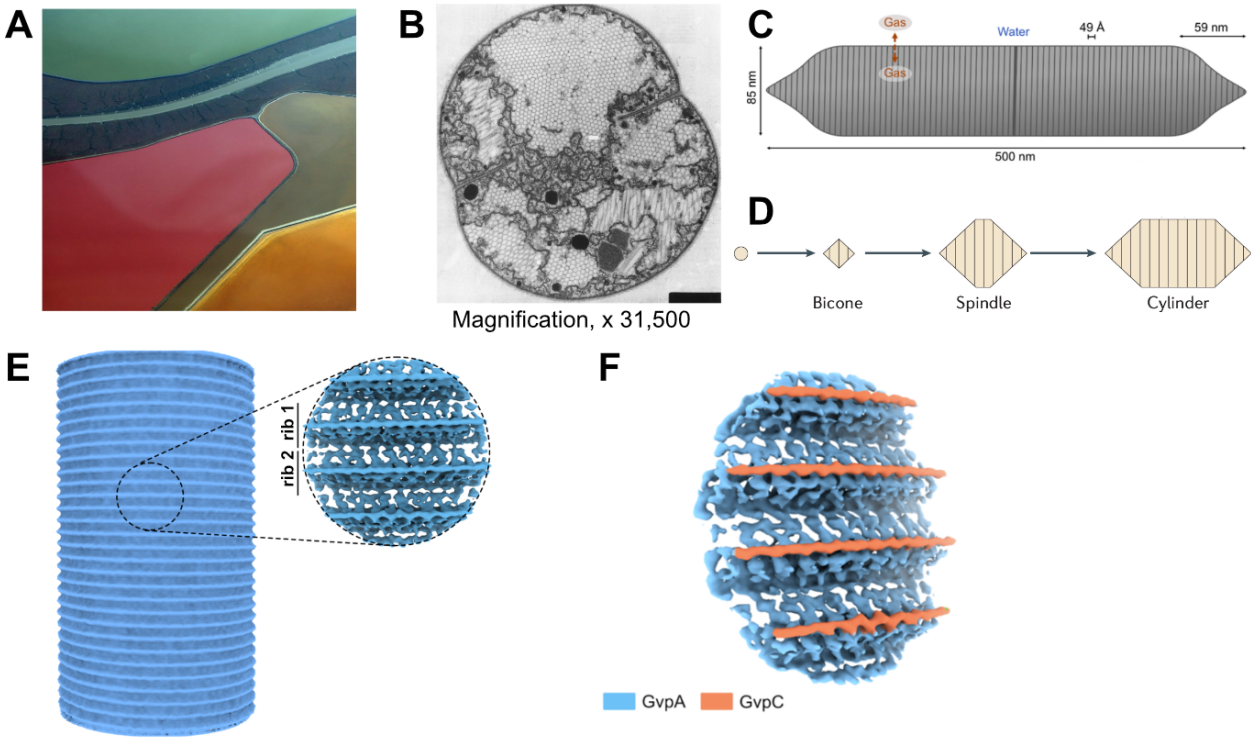

Figure 1-2: Overview of the gas vesicle protein structure.
A) Salt pond with gas vesicle producing microbes (Adapted from Atlas Obscura).[9] **B**) A dividing cell of cyanobacterium *Microcystis* sp. containing gas vesicles (Adapted from Walsby).[6] **C**) Average dimensions of a cylindrical gas vesicle produced by *Anabaena flos-aquae*. The protein shell creates a phase boundary between gas and water (Reprinted from Dutka et al.).[10] **D**) The process of gas vesicle growth (Adapted from Pfeifer et al.).[7] **E**) Subtomogram average of the gas vesicle shell structure of *Anabaena flos-aquae* obtained by cryo-electron tomography. The GvpA and GvpC proteins make up the majority of the shell and create a corrugated pattern or ribbing (Adapted from Dutka et al.).[10] **F**) The GvpC strengthening protein wraps radially around the outer shell of the gas vesicle creating structural rigidity (Reprinted from Dutka et al.).[10]

Orchestrated by multiple genetically encoded chaperones and assembly factor proteins, the formation of a gas vesicle in microorganisms is considered to begin with an initial nucleation event which leads to the formation of a small biconical structure (bicone), which then grows in width and diameter into an intermediary structure known as a spindle (**Figure 1-2D**). GvpA subunits are added to the spindle to increase the axial length of the vesicle through a process termed elongation. The GvpA subunits form the "ribs" of the gas vesicle shell, as confirmed by cryoelectron tomography (**Figure 1-2E**).[10] Formed GVs are quite stable but are known to collapse once the net pressure across the gas vesicle wall exceeds a critical collapse pressure, usually between 0.1 and 1.4 MPa.[11,12] These collapse pressures vary for gas vesicles derived from different organisms and are inversely correlated with the diameters of the vesicles.[7] Some gas vesicles are fortified by repeating units of a removable structural protein known as GvpC, which wraps around the gas vesicle shell (**Figure 1-2F**). Gas vesicles strengthened with GvpC have been shown to have higher collapse pressures than those with GvpC removed.[13]

Gas Vesicle Gene Clusters

To understand how we will encode gas vesicle genes into genetic circuits for mammalian cells, we first review how gvps are organized and regulated in their native hosts. As mentioned, the GvpA and GvpC proteins as well as other GV-specific chaperones and assembly factors are encoded by a cluster of gvp genes. Certain microbes also contain gvr genes, which produce

regulatory proteins that can control the production of gas vesicles in response to specific environmental conditions.[8] We will focus our discussion to particular GV gene clusters from three organisms, *Halobacterium salinarum* (domain: archaea), *Bacillus megaterium* (domain: bacteria), *and Anabaena flos-aquae* (domain: bacteria).

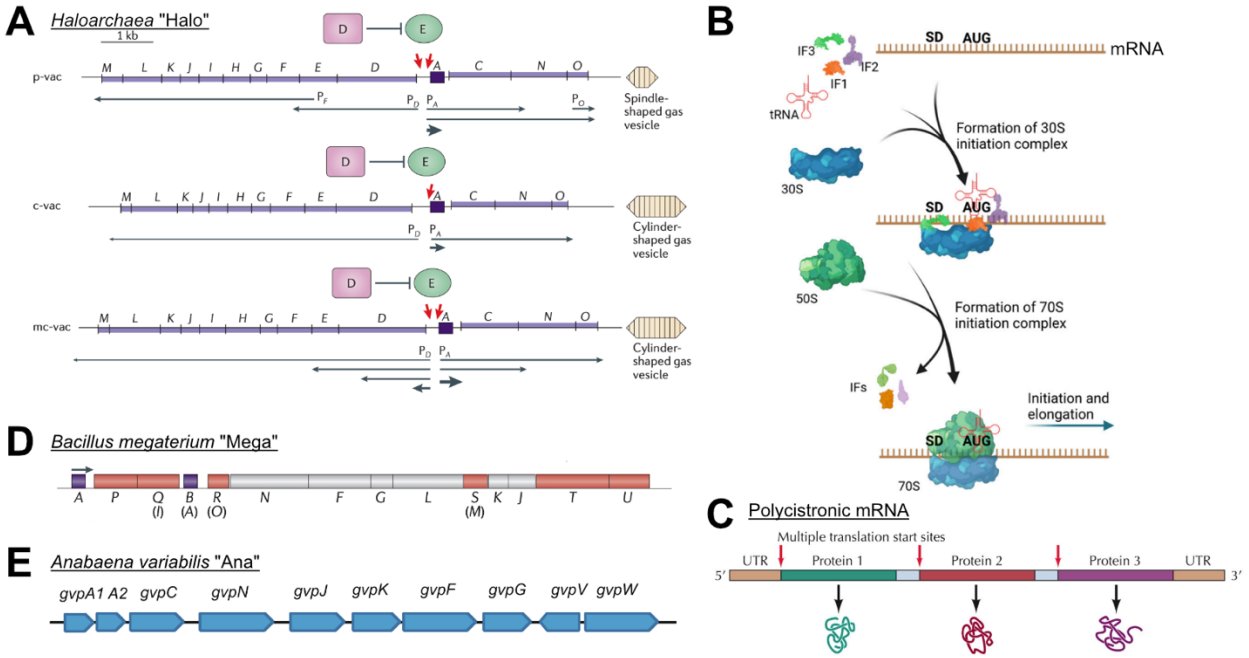

Figure 1-3: Gene clusters and prokaryotic translation of gas vesicles.
A) Gene clusters for three species of *Haloarchaea*. GvpE is a transcriptional activator that can upregulate transcription of gvpA. GvpD sequesters GvpE. (Adapted from Pfeifer et al.).[7] **B**) The canonical mechanism of prokaryotic translation includes the formation of the 30S complex that identifies the Shine-Dalgarno sequence upstream of a gene. The 70S complex then forms and translation begins. (Adapted from Leiva et al.).[14] **C**) Prokaryotic cells produce polycistronic mRNAs in which ribosomes can bind at different ribosome binding sites (RBS, red arrows) to begin translation. Depending on their sequence, RBSs can yield different levels of translation. (Adapted from Cooper and Hausman et al.).[15] **D**) Gene cluster for *Bacillus megaterium*. Parenthetical letters below the gene depict closest homolog to Halo gvps (Adapted from Pfeifer et al.).[7] **E**) Gene cluster for *Anabaena variabilis*, a close relative of *Anabaena flos-aquae*. Ana contains multiple copies of gvpA (Adapted from Wang et al.).[16]

Three similar gene clusters for two strains of *Halobacterium salinarum* and the closely related *Haloferax mediterranei* produce polycistronic and monocistronic transcripts in the expression of gas vesicles (**Figure 1-3A**).[7] The cluster is organized into two halves, producing transcripts containing gvpDEFGHIJKLM and transcripts containing gvpACNO. Transcriptional start sites are denoted by the letter "P" and the full transcripts are depicted by the narrow arrows. GvpE encoded by gvpE, acts as a transcriptional activator and can increase the transcription of GV genes. In the p-vac and mc-vac clusters, multiple transcripts are found to contain gvpA and these transcripts are found in greater abundance, yielding higher translation of GvpA structural proteins needed to form the GV shell. In addition, these polycistronic transcripts contain putative Shine-Dalgarno (SD) sequences which allow ribosomes to identify and translate multiple open reading frames (ORF) from single transcripts.[17] *Escherichia coli* expression hosts are able to express gas vesicles using nearly native GV gene cluster sequences, further suggesting native hosts and expression hosts use the canonical prokaryotic translation mechanism reliant on putative SD sequences and ribosome binding sites (RBS) (**Figures 1-3B, 1-3C**).[11,14] *Halobacterium* and similar haloarchaea are referred to as "Halo" in this text.

The soil bacterium *Bacillus megaterium,* which we will refer to as "Mega", contains GV genes with significant homology to those found in archaea (**Figure 1-3D**). Mega interestingly does not readily

produce GVs as it does not reside in water and does not readily produce gas vesicles. However, expressing the Mega GV gene cluster in *E. coli* leads to the production of gas vesicles. Due to parallel research efforts, some of the gene names for Mega are different than those used for Halo. Mega encodes gvpB, which has high sequence similarity to gvpA. In addition, plasmid-based expression studies of the Mega cluster in *E. coli* suggest there are multiple transcriptional start sites as well as RBS sites that regulate the transcription and translation level of each gvp.[18] The Mega cluster does not contain the GV strengthening protein, GvpC.

The cyanobacteria *Anabaena flos-aquae,* referred to here as "Ana"*,* readily produces gas vesicles to enable floatation-based motility. The Ana cluster is one of the most compact, with the fewest number of distinct genes required to express GVs (**Figure 1-3E**). In addition to having multiple gene copies of gvpA (some strains are reported to have up to seven copies),[16] the Ana gene cluster is also expected to contain multiple transcriptional start sites and variable strength RBS sequences to produce different stoichiometries of the Gvp proteins needed for efficient assembly. Pictured above is a gene cluster from a similar species, *Anabaena variabilis* ATCC 29413.

As mentioned, gvpA and gvpC encode the major structural proteins while the remaining genes encode accessory proteins (chaperones and assembly factors) that work together to build the gas vesicle. Using split-FP assays, Volkner and Jost *et al.* confirmed that Gvps from *Haloarchaea* do interact with one another to different degrees during the assembly process.[19] However, the exact mechanisms of assembly are still under investigation. In addition, the structure and function of all the accessory proteins have not yet been determined. Notably, GvpN has sequence homology to AAA+ ATPase proteins which use energy derived from hydrolyzing ATP to undergo conformational changes that can exert mechanical forces on a target substrate.[20] In the case of GvpN, such conformational changes are believed to contribute to the elongation of the GV shell.[21] In the absence of GvpN, prokaryotic expression hosts solely produce bicones.[22] Multiple sources have demonstrated for archaeal and bacterial gene clusters that the gvp accessory proteins must be expressed in a fine balance of stoichiometries to yield GV production.[23,24]

Through the above discussion, we have demonstrated how three prokaryotic species have evolved to express GvpA in excess of GV accessory proteins, either by increasing copy numbers of gvpA or its homologs (gvpB), upregulating transcripts producing gvpA, or potentially through the use of stronger RBSs compared to those used to translate accessory genes. We have also touched on the importance of relative abundances of the remaining accessory proteins that contribute to GV assembly in these cells. In future chapters, we will continue to explore the complex set of Gvps and how they can be expressed in beneficial stoichiometries in mammalian cells.

Ultrasound Imaging Methods for Gas Vesicles

To evaluate new genetic circuits for the expression of GVs, we first must identify the methods by which we will confirm their assembly and acoustic properties. In this section, we will discuss various ultrasound imaging methods that enable the detection of gas vesicles both inside and outside of cellular organisms. In ultrasound transducers, acoustic waves are generated when a voltage is applied across a set of piezoelectric crystals or elements. The emitted sound waves travel through a fluid until they encounter an interface between mediums with different acoustic impedances (equal to the product of the density of the medium and the speed of sound in the medium). At such interfaces, ultrasound waves are reflected back towards the transducer. The echoes are received by the piezoelectric elements and converted into an electrical signal which can be processed into an image. Interfaces with a large acoustic impedance mismatch will reflect more of the incident ultrasound waves. The gas-liquid interface that exists at the surface of the

gas vesicle has an inherently large acoustic impedance mismatch, allowing the GV to reflect ultrasound waves, producing strong echoes. The Shapiro Lab uses two main categories of ultrasound imaging to detect gas vesicles, linear methods and nonlinear methods.

With linear imaging methods, the magnitude of the measurable return echoes from GVs increases linearly with the magnitude of the transmitted ultrasound wave. However, materials in the body produce linear responses to pressure waves just as gas vesicles do, making it difficult to tease apart which return echoes are due to tissue and which are due to gas vesicles. To increase the imaging sensitivity for solely gas vesicles using linear imaging methods, members of the Shapiro Lab image a sample before and after sending a high acoustic pressure wave to collapse the gas vesicles. These pre/post collapse frames can be subtracted to yield a final image of gas vesicles with minimal background noise.[25] Unfortunately, this form of "destructive" imaging destroys the gas vesicles and is therefore not useful for dynamic imaging of these acoustic contrast agents *in vivo.*

With nonlinear imaging methods, the measured acoustic pressure from return echoes does not increase linearly with increasing transmit pressure. Depending on the rigidity of their structure, some gas vesicles undergo a process of reversible buckling when exposed to specific acoustic waves, producing a distinct nonlinear signal (**Figure 1-4**). Gas vesicles with the GvpC strengthening protein generally have minimal nonlinear contrast as their structures are more rigid. Imaging techniques that utilize amplitude modulation (AM) can detect the nonlinear signal produced by buckling GVs. AM imaging methods are "nondestructive," allowing gas vesicles to be tracked dynamically *in vivo.* AM modalities also effectively cancel out the linear contrast produced by tissues, leaving only the nonlinear contrast produced by GVs. However, published AM imaging modalities created by Maresca and Sawyer *et al*.—parabolic AM (pAM) and cross-propagating wave AM (xAM)—are unable to detect a single GV-expressing mammalian cell within a single volumetric pixel (voxel).[11,26] Multiple GV-expressing cells are required in a single voxel to create detectable nonlinear contrast. Members of our lab are currently working on increasing the sensitivity of nondestructive AM-based imaging methods for detecting nonlinear signals from GVs.[27,28]

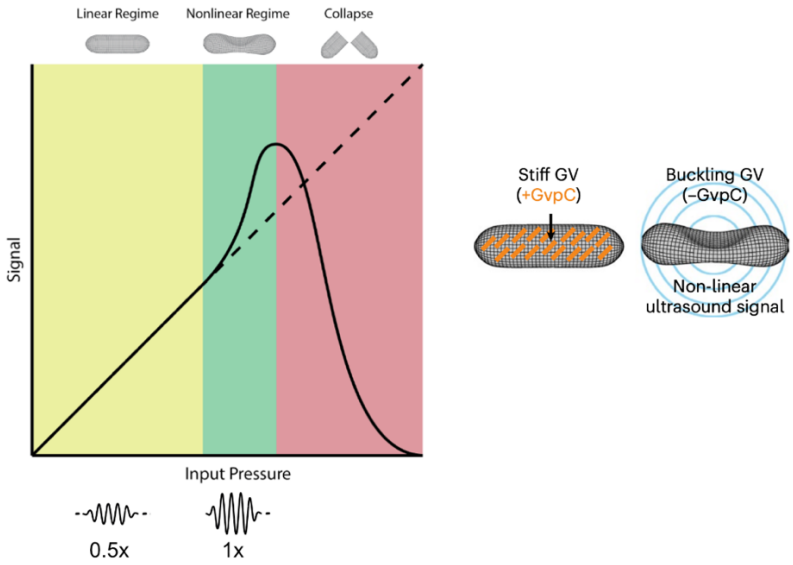

Figure 1-4: Nonlinear imaging for detecting gas vesicles.
Amplitude modulation imaging visualizes scatters that produce nonlinear responses to linear increases in input pressure. Gas vesicles can behave as nonlinear scatters for certain input pressures up to a collapse pressure. The removal of GvpC from Ana GVs allows gas vesicles to buckle, yielding nonlinear contrast. AM imaging methods are nondestructive and allow dynamic imaging of gas vesicles. (Adapted from Hurt, Buss, Duan et al.).[11]

The last imaging method of importance to this discussion is burst ultrasound reconstructed with signal templates (BURST). Developed by Sawyer *et al.,* BURST is an ultrasensitive GV-specific imaging method, achieving detection of single bacterial and mammalian cells expressing gas vesicles.[29] However, it relies on the strong signal generated by gas vesicles when they irreversibly collapse. BURST is therefore a destructive imaging modality that cannot be used for dynamic imaging of gas vesicles.

Heterologous Expression of ARGs in Mammalian Cells

This section discusses methods for expressing mammalian acoustic reporter genes (mARGs) to produce gas vesicle acoustic contrast agents within mammalian cells. Farhadi *et al.* created the first generation of mARGs, mARG1.0, using gvps derived from *Bacillus megaterium* (**Figure 1-5**).[30] The mARG1.0 system is composed of three vectors, one encoding the main structural protein GvpB (a homolog of GvpA), another encoding all the accessory proteins linked together with P2A self-cleaving peptides, and a third vector encoding additional copies of certain accessory gvps. This third plasmid is vital to tuning the stoichiometries of the Gvps to yield GV production in HEK293T cells. Although no GvpC strengthening proteins are encoded in mARG1.0, the resultant Mega gas vesicles do not produce significant nonlinear signal and can only be imaged using destructive imaging methods.

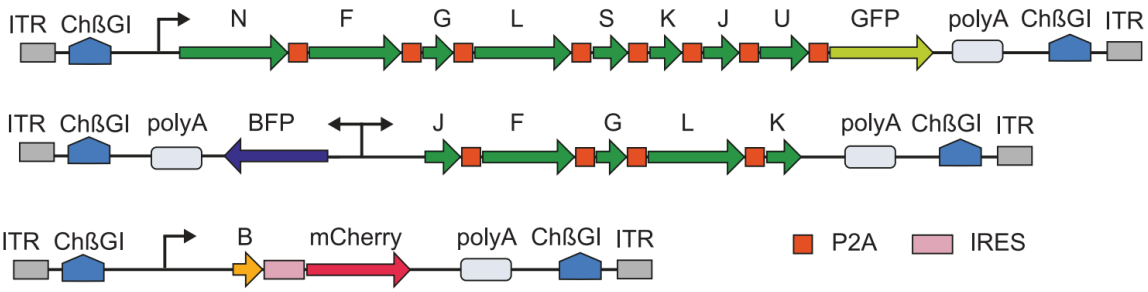

Figure 1-5: mARG 1.0 vectors for expression of gas vesicles in mammalian cells.
Three-vector system produces Mega gas vesicles in mammalian cells. Distinct fluorescent protein reporters are used in these vectors to enable cell sorting and flow cytometry analyses. ChBGI (insulator), ITR (inverted terminal repeat), IRES (internal ribosome entry site). (Adapted from Farhad et al.)[30]

To enable nonlinear imaging of mARGs, Hurt, Buss, and Duan *et al*. created the mARG2.0 system. The mARG2.0 system uses the gvps from *Anabaena flos-aquae* without the GvpC structural protein to produce Ana gas vesicles in mammalian cells (**Figure 1-6**). These Ana GVs produce nonlinear contrast and can be imaged with nondestructive AM imaging modalities. The mARG2.0 system comprises two vectors, one encoding the main structural protein GvpA and the other encoding the accessory Gvps linked together by P2A elements. The plasmids in this two-vector system are referred to as pgvpA and pgvpN-W. pgvpA is commonly transfected or transduced in excess of pgvpN-W to tune the relative protein stoichiometries and boost GV production. mARG2.0 can also be used with tetracycline-inducible promoter systems (Tet-ON). In these systems (referred to as mARG2.0$^{Tet\text{-}ON}$), mARG2.0 vectors with Tet-responsive (TRE) promoters are introduced into a cell line with a constitutively expressed transactivator. In the presence of doxycycline (dox), this transactivator induces transcription at the Tet-ON promoter. Notably, the mARG2.0$^{Tet\text{-}ON}$ system has been introduced into epithelial human breast cancer cells (MDA-MB-231) in culture using the Piggybac transposase system, a pseudo-random method for integrating genes into mammalian cells. These mARG2.0$^{Tet\text{-}ON}$ MDA cells have also been implanted in mice, resulting in tumors that could be studied *in vivo* using xAM ultrasound imaging.

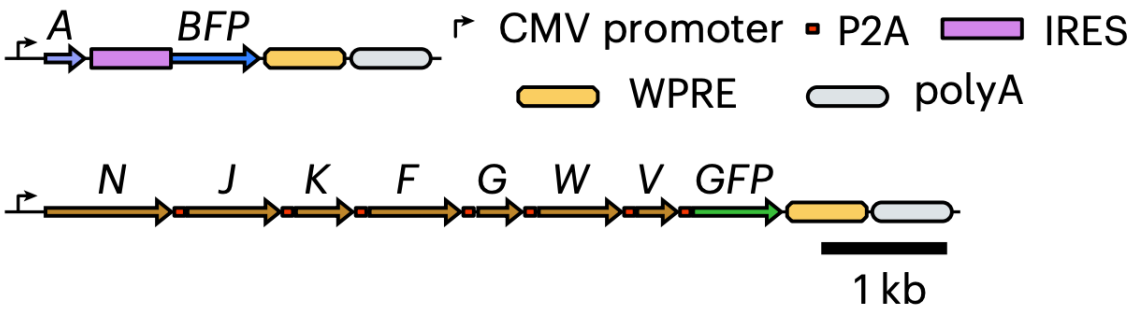

Figure 1-6: mARG2.0 vectors enable dynamic imaging in mammalian cells.
This two-vector system produces Ana gas vesicles in mammalian cells. The pgvpA is often delivered in excess of pgvpN-W. WPRE signals extend the half-life of the transcribed mRNA. We refer to the dox-inducible Tet-ON version of mARG2.0 as mARG2.0$^{Tet\text{-}ON}$ (Adapted from Hurt, Buss, Duan et al.).[11]

The engineered MDA cells containing the dox inducible mARG2.0$^{Tet\text{-}ON}$ system described above are a polyclonal population (i.e., not all the cells have the same genotype or the even the same number of copies of integrated pgvpA and pgvpN-W). As expected, the polyclonal population is comprised of MDA cells containing a distribution of GV abundances. This is further confirmed by conducting buoyancy enrichment of the polyclonal cells (**Figure S1-1**) Ideally, future mARGs will tighten this GV abundance distribution, reducing cell-to-cell variance across the population. Sources of stochasticity that widen this distribution likely include gene delivery, payload integration, and protein expression.

Gaps in Gene Expression Methods for Mammalian Cells

We have thus far reviewed gas vesicle gene cluster organization in microbes and introduced mARG expression systems. Next, we discuss the shortcomings of current synthetic biology gene expression technologies and their impact on mARG circuits. When designing complex genetic circuits for mammalian biology, a trifecta of limitations must be considered. A complex circuit often requires 1) expression of multiple genes in 2) defined ratios within a 3) genetic footprint amenable for diverse delivery methods (**Figure 1-7A**).

As discussed in the previous sections, gas vesicle protein expression requires fine tuning of the relative stoichiometries of the component Gvp proteins. Prokaryotes have evolved to have compact operonic structures, where multiple proteins can be expressed in desired ratios from single mRNA transcripts. In the case of the Ana gene cluster in its native context, expression tuning is achieved predominantly at the translational level, using variable strength RBS sequences within a single transcript (**Figure 1-1B**).

Conversely, eukaryotic genomes generally follow a single coding sequence per transcript rule (**Figure 1-1C**). Thus, tuning multi-gene expression systems in mammalian cells is often done at the transcriptional level by titrating gene copy numbers or using variable strength promoters. However, encoding multiple transcriptional units can lead to lengthy constructs while gene copy number titration suffers from a lack of precise tuning across a population due to the inherent stochasticity of multi-vector co-delivery.[32] In the case of GV expression, these mismatches in co-delivery stoichiometries can result in an overabundance of unchaperoned GvpA monomers—suspected to form cytotoxic aggregates due to their high hydrophobicity.[33]

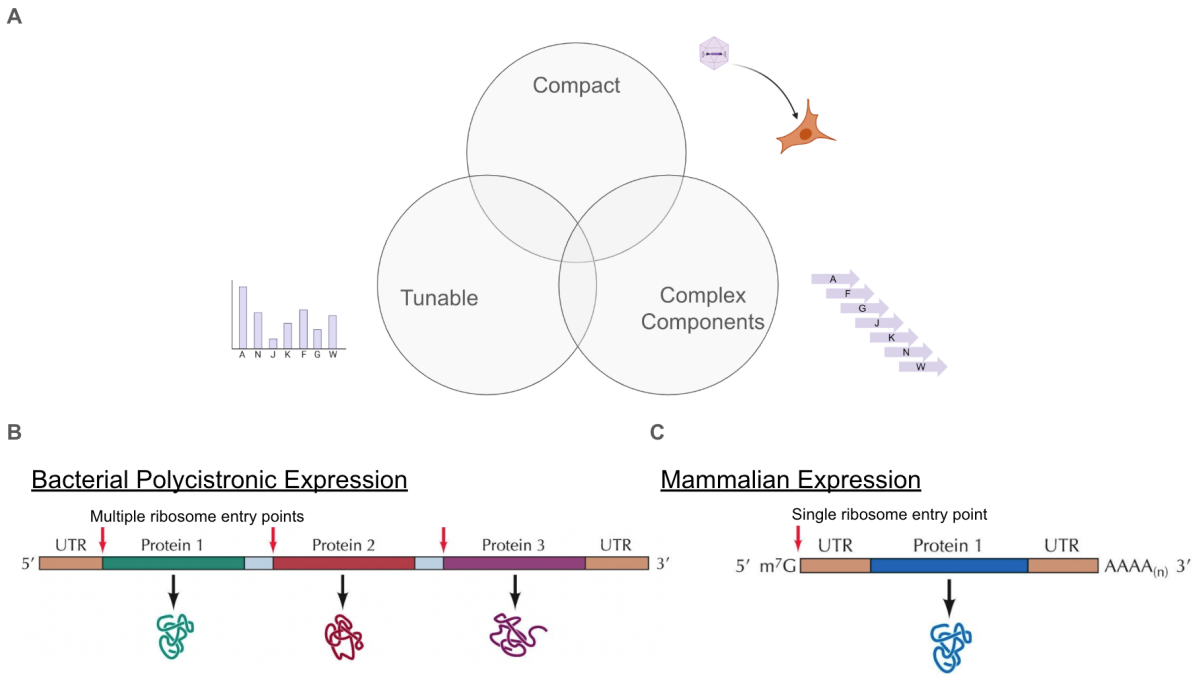

Figure 1-7: Limitations in synthetic biology and eukaryotic translation.
A) Limitations to be considered in mammalian synthetic biology. Circuits often need to be compact to enable delivery through viral methods. For complex applications like gas vesicles, these systems also need to be able to encode all of the necessary protein parts and tune their stoichiometries to achieve robust assembly of GVs. B) Bacterial translation allows for multiple proteins to be expressed from a single mRNA transcript. The expression level of each protein can be tuned by the RBS at each ribosome entry site (Adapted from Cooper et al.)[31] C) Mammalian translation canonically produces a single protein product from a single transcript. Ribosomes canonically have one entry site, 5' cap of the mRNA (Adapted from Cooper et al.)[31].

Translational methods of protein expression tuning in mammalian systems have gained significant traction in the field and heavily rely on non-native genetic elements derived from viruses known as Internal Ribosome Entry Sites (IRES) or p2A/t2A self-cleaving peptide elements.[34,35] Akin to SD sequences in bacteria, IRES sequences enable direct loading of ribosomes directly upstream of a desired open reading frame to be translated. These genetic parts are commonly ~650 base pairs (bp) in size, although smaller versions have now been engineered to be less than 200 bp.[36] Some efforts have successfully created variable strength IRES sequences.[37]

P2A genetic parts were first identified in picornaviruses.[38] These relatively small genetic parts (~60 bp) can be used as linker sequences between two or more CDS sequences. During translation, these P2A or T2A sequences cause the ribosome to miss a peptide bond on the growing protein chain, leading to production of distinct protein products. However, these protein products contain unnatural amino acid "scars" from the P2A or T2A.[39] Such scarring may impact protein activity or lead to identification by immunoregulatory systems, although severe immune responses have not yet been reported.[40] Although these provide multi-gene expression, P2A elements are not easily tunable. It has been shown that ORFs later in a P2A chain may be expressed in lower quantities than those earlier in the chain.[39] Cleaving orchestrated by these elements is mostly reliable, but with some frequency can fail, leading to the production of inactive fusion proteins.

Through translational methods of tuning, synthetic biologists are recapitulating bacterial operons in mammalian cells as these methods balance the challenges described in **Figure 1-7A**. Yet further improvements are needed in this area for gas vesicle gene expression in mammalian cells, where 1) peptide scars must be avoided, 2) transcript size must be limited, and 3) precise

stoichiometric control is required. In the following sections, we outline our approach to improve upon translational tuning mechanisms in mammalian biology.

Constructing and Evaluating Next-Generation mARG Expression Systems

In summary, mARG systems express a set of gas vesicle proteins that work collaboratively to assemble larger gas vesicle assemblies within mammalian cells. The inherent acoustic impedance mismatch between the cytosol and the vesicles gaseous inner core yields strong reflection of ultrasound waves. The ultimate goal of mARGs is to increase the acoustic contrast of a cell past the limit required for single-cell detection using non-destructive ultrasound imaging methods. All the while, the implementation of mARGs should not impact the underlying cellular phenotypes and responses that researchers seek to observe and study *in situ*. Towards the development of better mARGs, we 1) create a new synthetic biology framework to express gas vesicles from compact gene circuits, 2) showcase a methodology to sensitively evaluate cellular responses to gas vesicle expression, and 3) conduct proof-of-concept experiments to explore higher-throughput screening methods for mARG circuits. These efforts are expanded upon in the following chapters:

Chapter 2: To encode the many genes of Gvps into a compact circuit, we develop a new synthetic biology framework for tunable, multi-protein expression from single mRNA transcripts. We demonstrate this methodology for the expression of gas vesicles, showcasing that tuning the expression ratios of Gvps can improve cellular fitness. We call this framework Stoichiometric Expression of mRNA Polycistrons by Eukaryotic Ribosomes (SEMPER).

Chapter 3: The SEMPER methodology is difficult to implement with methionine containing proteins. To maximize the use of SEMPER for expressing GVs and other multi-protein systems, we use zero-shot predictions from protein language models to identify active protein variants without methionines. We apply this computational pipeline to generate mutant Gvps that can be used for multi-protein expression with SEMPER. ***Chapter Temporarily Unavailable.***

Chapter 4: By implementing an iterative experimental workflow guided by Bayesian Optimization, we quantify the relative importance of individual Gvps to acoustic contrast. In addition, we identify expression ratios amongst the Gvps that improve acoustic contrast, yielding design rules that can be used to advance future mARG systems. ***Chapter Temporarily Unavailable.***

Chapter 5: Complementing the engineering of the preceding chapters, here we demonstrate a methodology to assess the impacts of gas vesicle expression on mammalian cells using RNA-seq and differential expression analysis. We suggest metrics for evaluating future mARG systems on the task of reducing cellular burden due to gas vesicle expression. ***Chapter Temporarily Unavailable.***

Chapter 6: We explore methods to increase screening throughput for mARG circuits. Through a proof-of-concept experiment, we demonstrate the use of an ultrasonic chip for detecting gas vesicles inside living cells and argue that complementing this system with on-chip microfluidics could yield a single-cell acoustic cytometer.

References

1. Volpe, A., Lang, C., Lim, L., Man, F., Kurtys, E., Ashmore-Harris, C., Johnson, P., Skourti, E., de Rosales, R.T.M., and Fruhwirth, G.O. (2020). Spatiotemporal PET Imaging Reveals Differences in CAR-T Tumor Retention in Triple-Negative Breast Cancer Models. Mol. Ther. *28*, 2271–2285. https://doi.org/10.1016/j.ymthe.2020.06.028.
2. Flugel, C.L., Majzner, R.G., Krenciute, G., Dotti, G., Riddell, S.R., Wagner, D.L., and Abou-el-Enein, M. (2023). Overcoming on-target, off-tumour toxicity of CAR T cell therapy for solid tumours. Nat. Rev. Clin. Oncol. *20*, 49–62. https://doi.org/10.1038/s41571-022-00704-3.
3. Terese Winslow (2011). Definition of chimeric antigen receptor T-cell therapy. Natl. Cancer Insitute. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/chimeric-antigen-receptor-t-cell-therapy.
4. Piraner, D.I., Farhadi, A., Davis, H.C., Wu, D., Maresca, D., Szablowski, J.O., and Shapiro, M.G. (2017). Going Deeper: Biomolecular Tools for Acoustic and Magnetic Imaging and Control of Cellular Function. Biochemistry *56*, 5202–5209. https://doi.org/10.1021/acs.biochem.7b00443.
5. Wang, S., Hossack, J.A., and Klibanov, A.L. (2020). From Anatomy to Functional and Molecular Biomarker Imaging and Therapy: Ultrasound Is Safe, Ultrafast, Portable, and Inexpensive. Invest. Radiol. *55*, 559–572. https://doi.org/10.1097/RLI.0000000000000675.
6. Walsby, A.E. (1972). Structure and function of gas vacuoles. Bacteriol. Rev. *36*, 1–32.
7. Pfeifer, F. (2012). Distribution, formation and regulation of gas vesicles. Nat. Rev. Microbiol. *10*, 705–715. https://doi.org/10.1038/nrmicro2834.
8. Ramsay, J.P., Williamson, N.R., Spring, D.R., and Salmond, G.P.C. (2011). A quorum-sensing molecule acts as a morphogen controlling gas vesicle organelle biogenesis and adaptive flotation in an enterobacterium. Proc. Natl. Acad. Sci. *108*, 14932–14937. https://doi.org/10.1073/pnas.1109169108.
9. Salt Ponds of San Francisco Bay Atlas Obscura. http://www.atlasobscura.com/places/salt-ponds-san-francisco.
10. Dutka, P., Metskas, L.A., Hurt, R.C., Salahshoor, H., Wang, T.-Y., Malounda, D., Lu, G., Chou, T.-F., Shapiro, M.G., and Jensen, G.J. (2022). Structure of Anabaena flos-aquae gas vesicles revealed by cryo-ET. Preprint at bioRxiv, https://doi.org/10.1101/2022.06.21.496981 https://doi.org/10.1101/2022.06.21.496981.
11. Hurt, R.C., Buss, M.T., Duan, M., Wong, K., You, M.Y., Sawyer, D.P., Swift, M.B., Dutka, P., Barturen-Larrea, P., Mittelstein, D.R., et al. (2023). Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. Nat. Biotechnol., 1–13. https://doi.org/10.1038/s41587-022-01581-y.
12. Dutka, P., Malounda, D., Metskas, L.A., Chen, S., Hurt, R.C., Lu, G.J., Jensen, G.J., and Shapiro, M.G. (2021). Measuring gas vesicle dimensions by electron microscopy. Protein Sci. *30*, 1081–1086. https://doi.org/10.1002/pro.4056.
13. Walsby, A.E., and Hayes, P.K. (1988). The Minor Cyanobacterial Gas Vesicle Protein, GVPc, Is Attached to the Outer Surface of the Gas Vesicle. Microbiology *134*, 2647–2657. https://doi.org/10.1099/00221287-134-10-2647.
14. Leiva, L.E., and Katz, A. (2022). Regulation of Leaderless mRNA Translation in Bacteria. Microorganisms *10*, 723. https://doi.org/10.3390/microorganisms10040723.
15. Geoffrey M Cooper, N.W. and Robert E Hausman. (2007). The Cell: A Molecular Approach. Fourth Edition. By Geoffrey M Cooper and Robert E Hausman. Q. Rev. Biol. *82*, 44–44. https://doi.org/10.1086/513338.
16. Wang, H., Sivonen, K., Rouhiainen, L., Fewer, D.P., Lyra, C., Rantala-Ylinen, A., Vestola, J., Jokela, J., Rantasärkkä, K., Li, Z., et al. (2012). Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium Anabaena sp. strain 90. BMC Genomics *13*, 613. https://doi.org/10.1186/1471-2164-13-613.

17. Sartorius-Neef, S., and Pfeifer, F. (2004). In vivo studies on putative Shine–Dalgarno sequences of the halophilic archaeon Halobacterium salinarum. Mol. Microbiol. *51*, 579–588. https://doi.org/10.1046/j.1365-2958.2003.03858.x.
18. Li, N., and Cannon, M.C. (1998). Gas Vesicle Genes Identified in Bacillus megaterium and Functional Expression in Escherichia coli. J. Bacteriol. *180*, 2450–2458.
19. Völkner, K., Jost, A., and Pfeifer, F. (2020). Accessory Gvp Proteins Form a Complex During Gas Vesicle Formation of Haloarchaea. Front. Microbiol. *11*.
20. Erzberger, J.P., and Berger, J.M. (2006). Evolutionary Relationships and Structural Mechanisms of Aaa+ Proteins. Annu. Rev. Biophys. Biomol. Struct. *35*, 93–114. https://doi.org/10.1146/annurev.biophys.35.040405.101933.
21. Pfeifer, F. (2022). Recent Advances in the Study of Gas Vesicle Proteins and Application of Gas Vesicles in Biomedical Research. Life Basel Switz. *12*, 1455. https://doi.org/10.3390/life12091455.
22. Ling, B., Gungoren, B., Yao, Y., Dutka, P., Vassallo, R., Nayak, R., Smith, C.A.B., Lee, J., Swift, M.B., and Shapiro, M.G. (2024). Truly Tiny Acoustic Biomolecules for Ultrasound Imaging and Therapy. Adv. Mater. *36*, 2307106. https://doi.org/10.1002/adma.202307106.
23. Tavlaridou, S., Faist, K., Weitzel, K., and Pfeifer, F. (2013). Effect of an overproduction of accessory Gvp proteins on gas vesicle formation in Haloferax volcanii. Extrem. Life Extreme Cond. *17*, 277–287. https://doi.org/10.1007/s00792-013-0515-y.
24. Monson, R.E., Tashiro, Y., and Salmond, G.P.C. (2016). Overproduction of individual gas vesicle proteins perturbs flotation, antibiotic production and cell division in the enterobacterium Serratia sp. ATCC 39006. Microbiol. Read. Engl. *162*, 1595–1607. https://doi.org/10.1099/mic.0.000347.
25. Bourdeau, R.W., Lee-Gosselin, A., Lakshmanan, A., Farhadi, A., Kumar, S.R., Nety, S.P., and Shapiro, M.G. (2018). Acoustic reporter genes for noninvasive imaging of microbes in mammalian hosts. Nature *553*, 86–90. https://doi.org/10.1038/nature25021.
26. Maresca, D., Lakshmanan, A., Lee-Gosselin, A., Melis, J.M., Ni, Y.-L., Bourdeau, R.W., Kochmann, D.M., and Shapiro, M.G. (2017). Nonlinear ultrasound imaging of nanoscale acoustic biomolecules. Appl. Phys. Lett. *110*, 073704. https://doi.org/10.1063/1.4976105.
27. Nayak, R., Duan, M., Ling, B., Jin, Z., Malounda, D., and Shapiro, M.G. (2024). Harmonic imaging for nonlinear detection of acoustic biomolecules. APL Bioeng. *8*, 046110. https://doi.org/10.1063/5.0214306.
28. Smith, C.A.B., Duan, M., Yan, J., Taylor, L., Shapiro, M., and Tang, M.-X. (2025). Achieving single cell acoustic localisation with deactivation super resolution. Npj Acoust. *1*, 5. https://doi.org/10.1038/s44384-025-00008-7.
29. Sawyer, D.P., Bar-Zion, A., Farhadi, A., Shivaei, S., Ling, B., Lee-Gosselin, A., and Shapiro, M.G. (2021). Ultrasensitive ultrasound imaging of gene expression with signal unmixing. Nat. Methods *18*, 945–952. https://doi.org/10.1038/s41592-021-01229-w.
30. Farhadi, A., Ho, G.H., Sawyer, D.P., Bourdeau, R.W., and Shapiro, M.G. (2019). Ultrasound Imaging of Gene Expression in Mammalian Cells. Science *365*, 1469–1475. https://doi.org/10.1126/science.aax4804.
31. Cooper, G.M. (2000). Translation of mRNA. In The Cell: A Molecular Approach. 2nd edition (Sinauer Associates).
32. Kim, Y., Ding, X., Yang, C., Lee, G., Gradinaru, V., and Park, J. (2026). Controlled linking of AAV capsids enables coordinated multi-vector delivery. Preprint at bioRxiv, https://doi.org/10.64898/2026.04.13.716639 https://doi.org/10.64898/2026.04.13.716639.
33. Li, Z., Ho, C.-Y., Barr, D.E., Jeong, S., Ma, L., and Lu, G.J. (2025). Temporal gene regulation enables controlled expression of gas vesicles and preserves bacterial viability. Nat. Commun. *17*, 953. https://doi.org/10.1038/s41467-025-67667-8.
34. Ho, S.C.L., Bardor, M., Li, B., Lee, J.J., Song, Z., Tong, Y.W., Goh, L.-T., and Yang, Y. (2013). Comparison of Internal Ribosome Entry Site (IRES) and Furin-2A (F2A) for Monoclonal

Antibody Expression Level and Quality in CHO Cells. PLOS ONE *8*, e63247. https://doi.org/10.1371/journal.pone.0063247.
35. Douin, V., Bornes, S., Creancier, L., Rochaix, P., Favre, G., Prats, A.-C., and Couderc, B. (2004). Use and comparison of different internal ribosomal entry sites (IRES) in tricistronic retroviral vectors. BMC Biotechnol. *4*, 16. https://doi.org/10.1186/1472-6750-4-16.
36. Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A.A., Stern-Ginossar, N., Yakhini, Z., Weinberger, A., and Segal, E. (2016). Systematic discovery of cap-independent translation sequences in human and viral genomes. Science *351*, aad4939. https://doi.org/10.1126/science.aad4939.
37. Koh, E.Y.C., Ho, S.C.L., Mariati, Song, Z., Bi, X., Bardor, M., and Yang, Y. (2013). An Internal Ribosome Entry Site (IRES) Mutant Library for Tuning Expression Level of Multiple Genes in Mammalian Cells. PLoS ONE *8*, e82100. https://doi.org/10.1371/journal.pone.0082100.
38. Luke, G.A., de Felipe, P., Lukashev, A., Kallioinen, S.E., Bruno, E.A., and Ryan, M.D. (2008). Occurrence, function and evolutionary origins of “2A-like” sequences in virus genomes. J. Gen. Virol. *89*, 1036–1042. https://doi.org/10.1099/vir.0.83428-0.
39. Liu, Z., Chen, O., Wall, J.B.J., Zheng, M., Zhou, Y., Wang, L., Ruth Vaseghi, H., Qian, L., and Liu, J. (2017). Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. Sci. Rep. *7*, 2193. https://doi.org/10.1038/s41598-017-02460-2.
40. Arber, C., Abhyankar, H., Heslop, H.E., Brenner, M.K., Liu, H., Dotti, G., and Savoldo, B. (2013). The immunogenicity of virus-derived 2A sequences in immunocompetent individuals. Gene Ther. *20*, 958–962. https://doi.org/10.1038/gt.2013.25.

Chapter 2

Tunable Polycistronic Expression Using Leaky Ribosomal Scanning

Specific Background

mARG2.0 and other mARG expression systems used in the lab are encoded across two or more plasmid vectors, generally pgvpA and pgvpN-W. To achieve the necessary stoichiometries for GV production, researchers aim to deliver copies of these genes to cells in desired ratios, a form of transcriptional level tuning. However, co-delivering two plasmids into cells with transient transfection is an inherently stochastic process, yielding cells with undesired plasmid ratios. Furthermore, the actual copy number of each plasmid that enters the cell cannot be controlled, resulting in cells that have too many burdensome plasmid copies. Similar issues occur for stable transfection methods. Viral transduction of multiple plasmids can better reduce cell-to-cell variability and control copy numbers but still has inherent stochasticity. Due to the inherent stochasticity of gene delivery through transfection and transduction, it is advantageous to consolidate the plasmids that carry mARGs into a single vector to increase the likelihood a cell receives all gvp genes and expresses them at the desired stoichiometry.

In addition, many commonly used transfection and viral transduction methods for stable integration will integrate payloads randomly or pseudo-randomly. Due to epigenetic mechanisms, certain integration loci may be more or less suited for transcription. Therefore, achieving protein stoichiometries by tuning transcript abundances using gene copy numbers may also yield unexpected performance or significant cell-to-cell variability. Bacterial synthetic biologists often encode multiple proteins into a single transcript and tune their relative translation levels by using varying strength RBS sites. Encoding multiple proteins and their relative stoichiometries into a single transcript can make expression systems more robust to variations in transcription. In the next sections, we present a method for expressing multiple open reading frames (ORFs) from single transcripts using the leaky scanning model of translation initiation. In this approach termed Stoichiometric Expression of mRNA Polycistrons by Eukaryotic Ribosomes (SEMPER), adjacent ORFs are translated from a single mRNA in tunable ratios determined by their order in the sequence and the strength of their translation initiation sites. We validate this approach by expressing up to three fluorescent proteins from one plasmid in two different cell lines. We then use it to encode a stoichiometrically tuned polycistronic construct encoding gas vesicle acoustic reporter genes that enables efficient formation of the multi-protein complex while minimizing cellular toxicity. We also demonstrate that SEMPER enables polycistronic expression of recombinant monoclonal antibodies from plasmid DNA and of two fluorescent proteins from single mRNAs made through *in vitro* transcription. Finally, we provide a probabilistic model to elucidate the mechanisms underlying SEMPER. Overall, this expression strategy for mammalian cells 1) reduces the mARG expression system to one compact vector to mitigate inherent challenges with gene delivery and 2) allows us to coarsely tune the stoichiometries of Gvp structural subunits to cognate chaperones at a translational level.

This section is adapted from a publication produced by Ishaan Joshan Dev, co-author Dr. Mengtong Duan, and corresponding author Dr. Mikhail Shapiro—with contributions from other co-authors Andrew Lu, Goar Ayrapetyan, and Mei Yi You.

Introduction

Mammalian cell engineering promises to enable treatment of age-related degeneration, reversal of genetic diseases, and even transformation of cells into living therapeutics and sensors.[1] Realizing this promise requires the development of genetic circuits capable of finely tuning the relative expression stoichiometries of multiple proteins to produce functional multimeric protein assemblies, multi-component signaling systems, or multi-enzyme biosynthetic pathways.[2–6] Current approaches to doing so at the DNA level (e.g., varying promoter strength[7,8] or titrating copy numbers of each gene[2,9]) yield lengthy DNA constructs that often must be packaged into multiple delivery vectors and lead to undesirable cell-to-cell variability due to both stochastic gene delivery and integration across the population. Additionally, attainable protein expression stoichiometries are limited by the transcriptional strengths of a relatively small set of curated promoters that often demonstrate cell-to-cell variability.[7,8,10]

Post-transcriptionally, sequence motifs such as internal ribosome entry sites (IRES) or 2A self-cleaving peptides may be used to encode multiple open reading frames (ORF) into a single transcript and tune protein stoichiometries using relative translation levels.[11,12] Such post-transcriptional mechanisms are also useful for mRNA-based, multi-gene expression systems, which are of relevance to mRNA vaccine and therapeutic development.[13] Although powerful tools, these genetic parts have their disadvantages. For instance, IRES sequences have a considerable genetic footprint (~200-600 bps),[14] leading to lengthier genetic constructs which may reduce viral packaging efficiency. Likewise, 2A self-cleaving peptides leave peptide scars and can yield undesired fusion proteins, both of which can be detrimental to protein function.[15] In addition, when employed in bicistronic vectors, these genetic parts can strongly attenuate the expression of the second ORF relative to the expression of the first ORF.[12,16] Here, we present an alternative approach that enables compact and robustly tunable polycistronic expression in mammalian cells using plasmid- and mRNA-based expression systems. We call this approach “Stoichiometric Expression of mRNA Polycistrons by Eukaryotic Ribosomes” (SEMPER).

SEMPER uses the canonical cap-dependent ribosome recruitment and translation mechanism in mammalian systems, which begins when the 43S preinitiation complex (PIC) of the ribosome is loaded onto the 5’ end of mRNA.[17,18] This complex then scans the 5’ untranslated region (5’ UTR) until it encounters a translation initiation site (TIS), consisting of the start codon (AUG) and ~3-10 neighboring nucleotides.[19] The TIS sets the translational reading frame and initiates translation by engaging with the 60S ribosomal subunit. The full ribosome then translates the mRNA into protein until it encounters a stop codon, where it terminates translation and disengages the transcript. With some frequency, the 43S PIC may scan through the first TIS and initiate translation from a downstream ORF starting at another TIS, a phenomenon called leaky ribosomal scanning (LRS).[20] As determined by its sequence, a strong TIS (e.g. the Kozak consensus sequence) will reliably initiate translation while weaker ones will more frequently allow the 43S PIC to scan past.[21]

By employing a short ORF (uORF) upstream of a gene of interest (GOI), mammalian cells naturally use LRS and alternate TISs to divert a portion of the ribosome flux away from a GOI, effectively downregulating its translation.[22] Recently, Ferreira and colleagues demonstrated that it is possible to use synthetic uORFs to regulate the expression of a downstream recombinant GOI.[23] They also empirically determined the strength of various translation initiation sequences and showed that it is possible to divert varying amounts of ribosomal flux away from the GOI by varying the uORF TIS strength. uORFs have also been used in synthetic systems to tune an endoribonuclease-based feedforward controller to manage resource limitation.[24] Others have also utilized translation initiation strength variability to tune the expression of a GOI in the context of

modular genetic design.[25] The SEMPER approach replaces non-protein-coding uORFs with longer, functional coding sequences. In this study, we show that by chaining together multiple ORFs while varying the translation initiation strength of each ORF, this approach achieves polycistronic expression of multiple proteins with tunable translation rates. We further demonstrate that this framework is functional in *in vitro* transcribed mRNA, paving the way for advances in mRNA-based protein therapeutics of higher complexity.

Results

Tunable, plasmid-based SEMPER framework for expressing two ORFs. To test tunable translation levels of two recombinant proteins from single transcripts, we encoded fluorescent proteins (FPs) with minimal spectral overlap into the first two ORFs of our SEMPER plasmid vector (**Figure 2-1A**). We included 11 bps of distance between these ORFs to ensure that read-through of the stop codon of the first ORF would not result in a fusion containing both proteins. We used the TIS sequence NNNAUGG to initiate translation of our ORFs, where NNN represents one of the following sequences: ACC, CCC, TTT. These sequences were described as having strong, medium, and weak translation initiation efficiency respectively by Ferreira et al.[23] We engineered all our FPs to contain a valine residue (GTG or GTT) following the N-terminal methionine to ensure that changes in translation initiation were due to the trinucleotide preceding the start codon and not the dinucleotide succeeding it. TIS sequences will be referred to by their variable NNN sequence. We used one other sequence (TTTCCAT), referred to as ***, to scrub the TIS entirely and prevent the ORF from being translated. For the first ORF, we generated a methionine-less monomeric Superfolder GFP (msfGFP[r5M]) in which all methionines—except the N-terminal one—were mutated to other amino acids.[26,27] In addition, out-of-frame AUGs were removed from the msfGFP[r5M] coding sequence using synonymous mutations. These mutations effectively removed all internal TISs within ORF 1 that could reduce ribosomal flux to downstream ORFs (**Figure 2-1B**). In the second ORF, we encoded mEBFP2 with all its natural methionines. Finally, downstream of the SEMPER ORFs, we included an IRES followed by mCherry (IRES-mCherry) for normalization. As ribosomal binding to the IRES and subsequent translation of mCherry are conducted independently of ORF 1 and ORF 2 translation,[28,29] the mCherry allowed us to normalize single-cell fluorescence measurements for msfGFP[r5M] and mEBFP2, a strategy common to LRS-focused studies.[23,30] Because the analyte ORFs and IRES-mCherry were encoded on the same transcript, this normalization scheme accounted for variations in transfection efficiency, transcription, and mRNA decay. mCherry fluorescence also served as a proxy for transcript abundance in each cell (**Figure S2-1 and Table S2-1A-D**).

After cloning various combinations of TISs in front of our two ORFs, we transfected these “2-ORF SEMPER” constructs into human HEK293T cells—a widely used research model for mammalian cell biology. Three single-color control plasmids with msfGFP[r5M], mEBFP2, and mCherry were also transfected. We screened the transfected cells using flow cytometry, utilizing the single-color control plasmids for compensation and correction of fluorescence spillover emissions (**Figure S2-2, S2-3**). As hypothesized, our cell lines produced both msfGFP[r5M] and mEBFP2 from single transcripts (**Figure 2-1C**). Furthermore, the tested TIS combinations (TIS for ORF 1 / TIS for ORF 2) yielded unique relationships between the fluorescence of the first ORF and that of the second ORF, with the relative expression of the former vs the latter following the strength of the first TIS. To analyze the 2-ORF SEMPER performance as a function of transcript abundance or “copy number”, we binned cells into three categories (low copy, medium copy, and high copy) based on their mCherry fluorescence. This yielded distinct clusters. We then calculated the means of msfGFP[r5M]/mCherry, mEBFP2/mCherry, and msfGFP[r5M]/mEBFP2 for the cells in each bin for each tested TIS combination. We conducted min-max normalization for each ORF separately

using two constructs from our screen (***/ACC and ACC/***). We reasoned that ***/ACC would produce the maximum amount of mEBFP2 and the minimum amount of msfGFP[r5M], as all 43S PIC flux should bypass the first ORF and initiate translation at the second. Conversely, we expected ACC/*** to produce the maximum amount of msfGFP[r5M] and the minimum amount of mEBFP2. Conducted for each mCherry bin independently, this analysis enabled us to compare translation levels of a particular FP across various TIS combinations relative to its hypothesized maximum and minimum (**Figure 2-1D**). Pairwise statistical comparisons for these TIS combinations are provided in **Table S2-2**.

As we increased the TIS strength in front of msfGFP[r5M], we observed increases in msfGFP[r5M] translation levels and decreases in mEBFP2 translation levels for all mCherry bins. Across mCherry bins, we found that the rank order of relative translation levels for mEBFP2 for our different TIS combinations was conserved. Additional TIS combinations were also tested in HEK293T cells, showing that they can provide additional degrees of tuning in expression ratios (**Figure S2-4**).

To confirm generalizability across species and cell types, we also demonstrated that the 2-ORF SEMPER constructs yielded TIS combination-dependent relative translation levels of our ORFs in CHO-K1 cells, a widely used Chinese hamster ovary cell line for the production of biologics (**Figure S2-5, Table S2-3**).[31] Upon comparing the distributions of log10(msfGFP[r5M]/mEBFP2) values between cell types (**Figure 2-1E**), we determined that the three TIS combinations tested maintained the same rank order in both HEK293T and CHO-K1 cell lines.

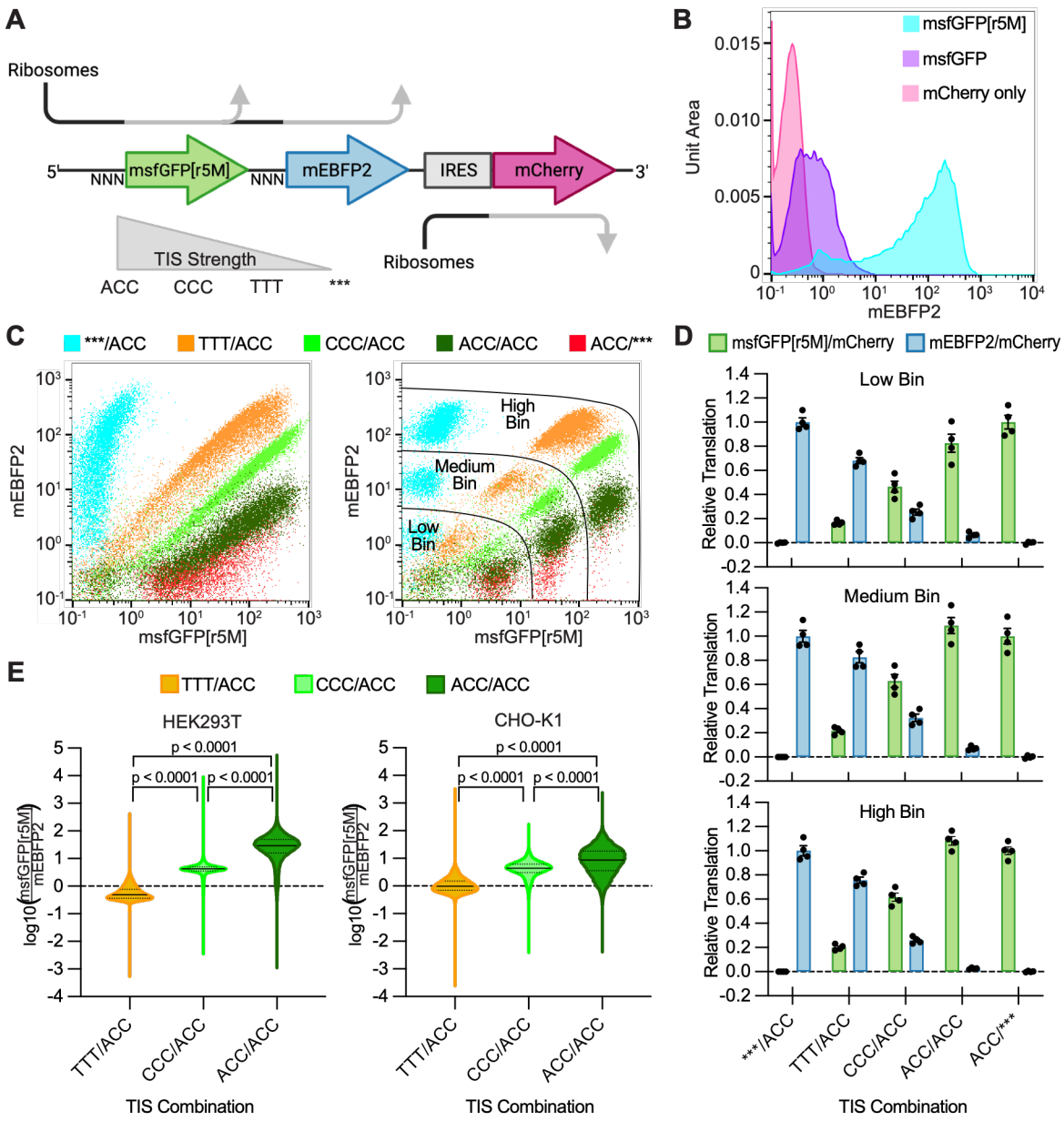

Figure 2-1: 2-ORF SEMPER constructs demonstrate tunable, bicistronic expression.
A) Architecture of an mRNA transcript produced by transfected 2-ORF SEMPER plasmid DNA. Cap-dependent ribosomes translate ORF 1, msfGFP[r5M], or ORF 2, mEBFP2, with frequencies dependent on the trinucleotide (NNN) upstream of each ORF. The relative strengths of the trinucleotides and TISs they represent are depicted. *** (a.k.a. TTTCCAT) does not contain a start codon, preventing translation of the ORF. IRES-mCherry is included to normalize fluorescence measurements. **B**) mEBFP2 distribution plots for constructs containing different versions of monomeric Superfolder GFP encoded in ORF 1. Removal of internal methionines from the msfGFP leads to much stronger expression of the downstream mEBFP2. Both ORFs used the strong ACC TIS (N=1, representative of four biological replicates). **C**) Flow cytometry plots of mCherry positive HEK293T cells transfected with 2-ORF SEMPER constructs. The legend contains the ORF 1 TIS and the ORF 2 TIS separated by a slash. (Left) Increasing the strength of the first TIS increases msfGFP[r5M] fluorescence relative to that of mEBFP2 (N=1, representative of four biological replicates). (Right) Binning cells into three mCherry fluorescence ranges produces unique clusters of cells (N=1, representative of four biological replicates). **D**) Min-max normalized fluorescence values for each GOI relative to ACC/*** and ***/ACC. Fluorescence measurements were first normalized by mCherry. Cluster means were then calculated followed by min-max normalization. Error bars indicate mean ± SEM (N=4 biological replicates). Welch's ANOVA tests were conducted for blue and green bars independently ($p < 0.0001$ for all colors across all bins) followed by Dunnett's T3 multiple comparisons tests. All pairwise comparisons are provided in **Table S2-2**. **E**) Violin plots of log10(msfGFP[r5M]/mEBFP2) values for all mCherry positive cells for four combined biological replicates of TTT/ACC, CCC/ACC, and ACC/ACC plasmids transfected into HEK293T and CHO-K1 cell lines. Each distribution is produced from the measurements of at least 15,000 cells. The median and quartiles of the distribution are represented by the solid and dotted lines respectively. Statistical analysis was conducted using one-way Welch's ANOVA tests ($p < 0.0001$ (Left) and $p < 0.0001$ (Right)) followed by Games-Howell's multiple comparisons test to yield pairwise comparisons.

Scaling plasmid-based SEMPER constructs to three ORFs. Enzymatic pathways and macromolecular assemblies are often composed of more than two proteins or peptide subunits.[2,6,32] To determine if the SEMPER system could be scaled beyond two ORFs to meet the needs of increasingly complex pathways and assemblies, we screened a variety of "3-ORF SEMPER" constructs. Following a similar strategy to the 2-ORF system, we first cloned and tested a methionine-less monomeric Superfolder BFP (msfBFP[r5M]).

We then encoded msfBFP[r5M], msfGFP[r5M], and emiRFP670—a far-red fluorescent protein—into a variety of 3-ORF SEMPER plasmids, maintaining IRES-mCherry for normalization (**Figure 2-2A**). To conduct min-max normalization, we cloned three plasmids where the starting TISs were removed from two of the ORFs using the *** sequence, while maintaining a strong TIS on the third: i) ***/***/ACC, ii) ***/ACC/***, iii) ACC/***/*** (**Figure S2-6**). The emiRFP670 coding sequence still encoded alternative translation initiation sites within it, yielding some observable far-red fluorescence in constructs ii) and iii). Additionally, we constructed and transiently transfected six other 3-ORF SEMPER plasmids with unique TIS combinations into HEK293T cells and measured their output using flow cytometry. Four single-color controls (msfBFP[r5M], msfGFP[r5M], emiRFP670, and mCherry) were used for fluorescence compensation (**Figure S2-7, S2-8**). Following an mCherry binning strategy similar to that employed for 2-ORF SEMPER constructs, we analyzed the mCherry-normalized translation levels of our three ORFs relative to their own theoretical maxima and minima from i), ii), iii). With the tested TIS combinations, we achieved a wide range of relative translation levels for each ORF, demonstrating tunable co-expression of three genes from a single transcriptional unit (**Figure 2-2B,** Left).

Notably, two combinations, ACC/TTT/ACC and ACC/ACC/ACC, yielded relative translation levels for msfBFP[r5M] greater than 1.0 in HEK293T cells across all mCherry bins (**Figure 2-2B**, Right). We also observed relative translation values for msfBFP[r5M] greater than 1.0 in CHO-K1 cells for TIS combinations in both medium and high bins (**Figure S2-9, Table S2-5**). Consistent with our observations, Wu et al. previously reported that the presence of downstream ORFs can increase the translation of upstream ones.[33] We hypothesize that constructs containing multiple ACC TISs have a higher prevalence of translating ribosomes, allowing these ribosomes and subunits—with their documented helicase activity—to maintain the mRNA secondary structure in an open conformation that is more amenable for ribosomal scanning, translation initiation, and subsequent protein expression.[34] However, more work is required to uncover the exact mechanisms that elicit this observed phenotype. For multiple TIS combinations, we found that the second and third ORFs achieved relative translation levels greater than 50% while still producing substantial levels of upstream ORF products. These results suggest there may be sufficient ribosomal flux to effectively scale the SEMPER paradigm to 4+ ORFs.

Based on the SEMPER mechanism, the relative expression of ORFs is expected to be tunable regardless of their order in the construct, although the exact expression levels could be affected by any resulting differences in RNA structure and stability. To examine this possibility, we interchanged the first and second ORFs (both of which lack methionines) in our 3-ORF constructs and found that the expected rank of relative expression was preserved (**Figure S2-10**). However, the exact expression levels varied with different ORF ordering, suggesting that some iterative tuning is required to achieve specific target values.

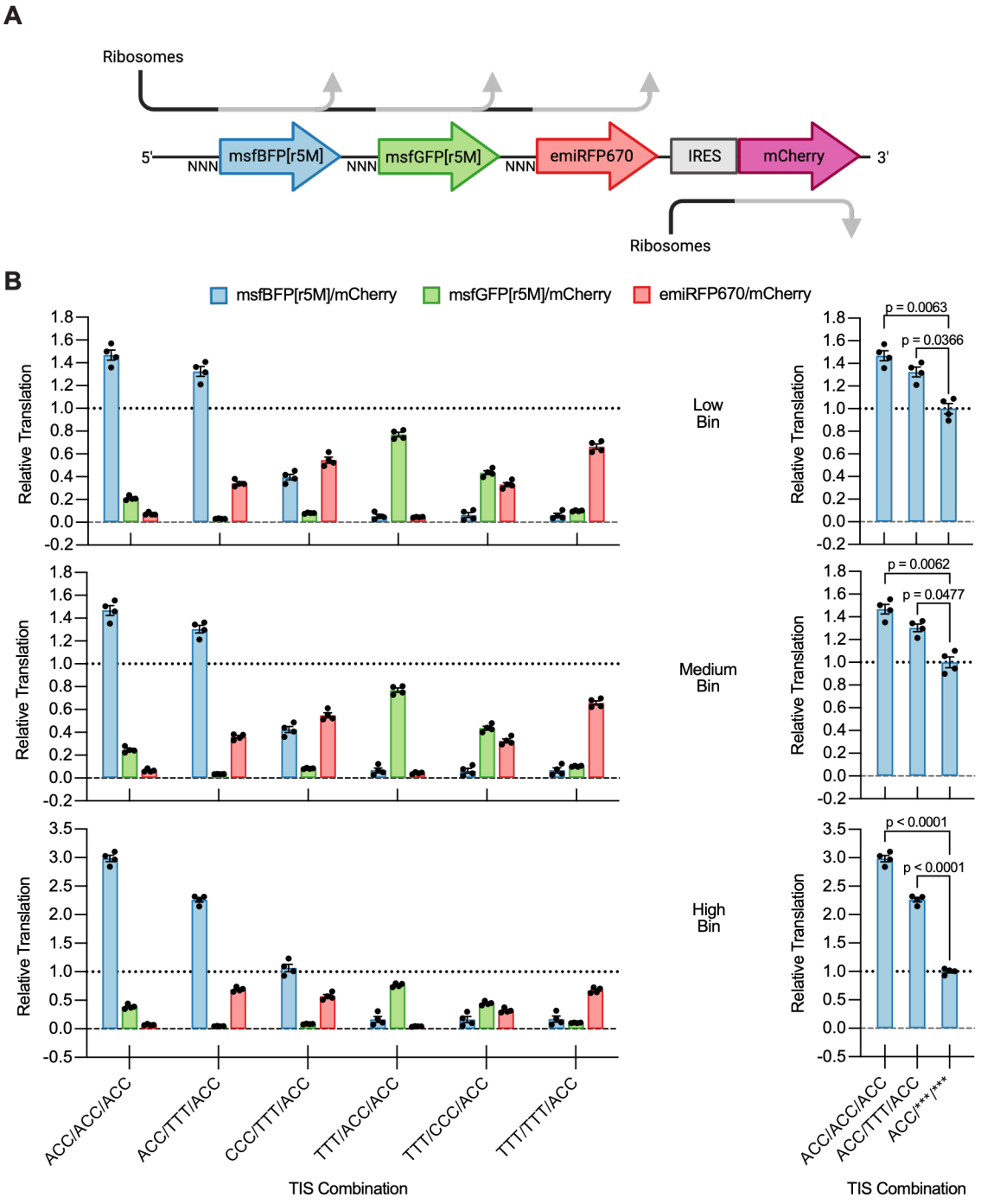

Figure 2-2: Scaling tunable polycistronic expression beyond two ORFs.
A) Architecture of an mRNA transcript produced by transfected 3-ORF SEMPER plasmid DNA. **B**) (Left) Min-max normalized relative translation levels for each fluorescent protein. Cells were first clustered into three mCherry fluorescence bins (low, medium, high). Error bars indicate mean ± SEM (N=4 biological replicates). Welch's ANOVA tests were conducted for blue, green, and red bars independently ($p < 0.0001$ for all colors across all bins). Post-hoc Dunnett's T3 multiple comparisons tests are provided in **Table S2-4**. (Right) Comparisons of msfBFP[r5M] relative translation levels between the max control and those TIS combinations with multiple ACC driven ORFs. Error bars indicate mean ± SEM (N=4 biological replicates). Post-hoc Dunnet's T3 comparisons tests demonstrate a significant upregulation in relative translation of msfBFP[r5M] for depicted constructs compared to that observed for ACC/***/***.

Producing gas vesicle ultrasound reporters using 2-ORF SEMPER. Next, we set out to demonstrate that 2-ORF SEMPER constructs could be applied to encoding a multimeric protein complex by expressing gas vesicles (GV) using mammalian acoustic reporter genes (mARGs).[35] Originally evolved in prokaryotes, GVs were recently introduced as genetically-encodable reporters for ultrasound imaging, enabling the noninvasive imaging of dynamic cellular processes in living organisms.[2,36] A single GV is made up of many GvpA structural units that are assembled together in a helical pattern through the cooperative activity of six heterologous assembly factors and minor constituents, referred to collectively as GvpNJKFGW.[37,38] Using a two-vector system, our group has previously expressed GVs in mammalian cells by co-transfecting one plasmid encoding the structural unit upstream of IRES-mCherry (pgvpA-IRES-mCherry) along with another plasmid encoding the assembly factors and a terminal Emerald GFP (EmGFP), all linked together by P2A elements (pgvpNJKFGW-EmGFP). These gene vectors are analogous to pgvpA and pgvpN-W referenced in the broader text. Additionally, we have found that co-transfecting the pgvpA-IRES-mCherry in excess of pgvpNJKFGW-EmGFP improves acoustic contrast.[35] Likewise, *Anabaena flos-aquae*—the organism from which mARGs used in this study are derived—contains more copies of gvpA relative to the other gvps in its GV gene cluster.[39]

Using the 2-ORF SEMPER strategy, we cloned single-vector systems for producing GVs in mammalian cells. We refer to these constructs as SEMPER mARGs. As gvpA does not contain any internal methionines, we encoded it directly into the first ORF. We tested our panel of TIS sequences in front of gvpA while maintaining the strong ACC TIS in front of the gvpNJKFGW-EmGFP ORF (**Figure 2-3A**). We compared these SEMPER constructs to our published two-vector expression system (referred to as mARG2.0 in the broader text), mixing the gvpA-IRES-mCherry plasmid in 4-fold molar excess of the gvpNJKFGW-EmGFP plasmid. We transfected these plasmids into HEK293T cells, maintaining the same total mass of plasmid for each transient transfection. As the strength of the TIS in front of gvpA increased, we found that the translation level of gvpNJKFGW-EmGFP, as measured by Emerald/mCherry, decreased (**Figure 2-3B**), as expected from our 2-ORF SEMPER FP experiments. As measured by BURST ultrasound imaging,[40] the ACC/ACC combination—predicted to yield the highest ratio of GvpA to GvpNJKFGW—produced the strongest acoustic contrast compared to the other SEMPER mARG plasmids (**Figure 2-3C, 2-3D**).

SEMPER gas vesicle expression system reduces cell toxicity. A critical issue with multimeric protein assemblies is the potential for cellular burden or toxicity due to imperfect stoichiometry or the absence of an essential assembly component or chaperone. In our experiments with mARGs, we observed that samples transfected with the two-vector system contained a larger fraction of cells that were positive for pgvpA-IRES-mCherry but negative for pgvpNJKFGW-EmGFP (39.03% ±1.42%) compared to cells transfected with SEMPER mARG plasmids (13.40%±0.85%) (**Figure S2-11**). This is expected due to the inherent stochasticity of transient co-transfection. As GvpA subunits have been speculated to nonspecifically aggregate when expressed without GV assembly factors,[41,42] we hypothesized that cells receiving a sub-optimal ratio of pgvpA-IRES-mCherry : pgvpNJKFGW-EmGFP or only pgvpA-IRES-mCherry may have higher incidences of apoptosis due to the formation of cytotoxic GvpA aggregates in the cytoplasm.

To test this hypothesis, we performed Annexin V-based apoptosis assays on cells transfected with either pgvpA-IRES-mCherry alone, the two-vector mARG expression system at optimal transfection ratio, or the ACC/ACC SEMPER mARG (**Figure 2-3E**). In each condition, molar amounts of gvpA gene were equalized in each transfection mixture. In addition, pgvpA-IRES-mCherry plasmid without a start codon in front of gvpA was transfected into HEK293T cells to establish a negative control. A subset of these samples was then treated with Raptinal to induce apoptosis, establishing a positive control. Notably, expressing GvpA without its assembly factors

led to high levels of apoptosis. While co-transfecting pgvpNJKFGW-EmGFP reduced some of the observed toxicity, the ACC/ACC SEMPER mARG transfected cells were healthier, with apoptosis levels indistinguishable from untreated negative controls. Taken together, these results suggest that the SEMPER mARG plasmid reduces cell toxicity by ensuring that assembly factors are consistently co-expressed with structural proteins in an appropriate ratio.

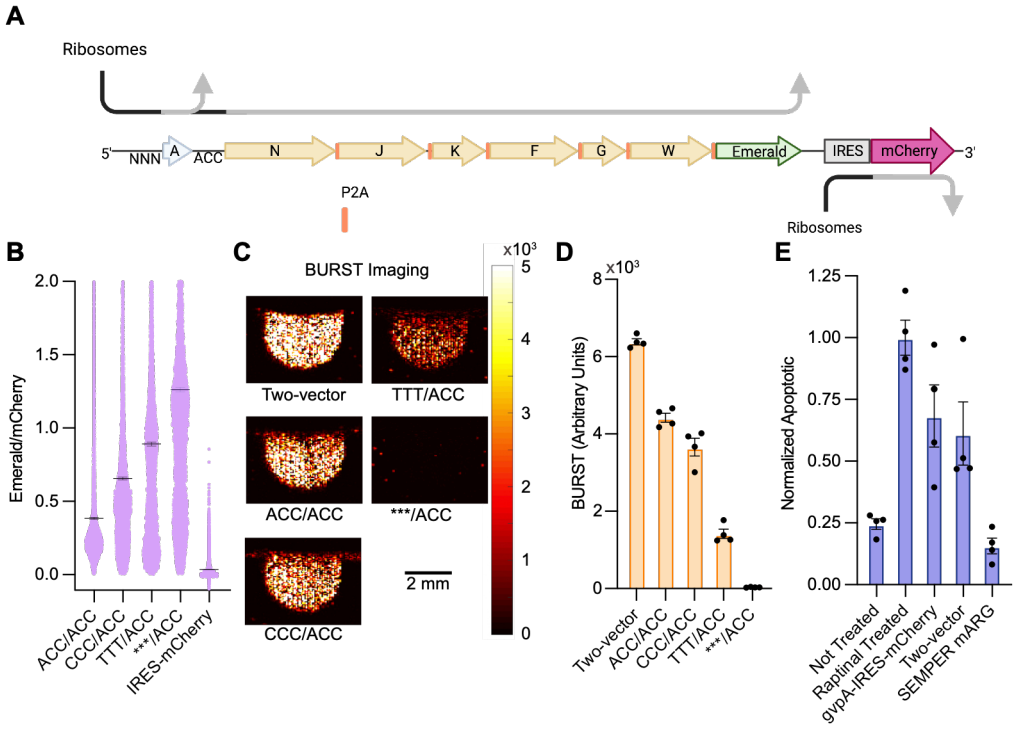

Figure 2-3: Utilizing 2-ORF SEMPER constructs to express gas vesicles in mammalian cells.
A) Architecture of an mRNA transcript transcribed from transfected SEMPER mARG plasmid DNA. The first ORF, gvpA, encodes the main structural protein, while the second ORF, gvpNJKFGW-EmGFP, encodes all necessary accessory proteins strung together with P2A self-cleaving peptides. **B**) Flow cytometry distributions of Emerald GFP normalized by mCherry values for each TIS combination tested in addition to an IRES-mCherry control. The thick horizontal lines depict the means of the distributions. Individual cells are represented as individual points. A minimum of 4,000 cells are depicted per condition. Error bars indicate mean ± SEM. p-value <0.0001 (Kruskal-Wallis test). p-values of multiple comparisons are found in **Table S2-6**. **C**) BURST images of acoustic contrast due to gas vesicle expression within HEK293T cells. Depicted are HEK293T cells loaded into agarose phantoms three days after transfection of SEMPER mARG plasmids or the leading two-plasmid system. The color bar represents the magnitude of BURST signal measured in linear arbitrary units. The floor and ceiling of the images are set to 0 and 5000 respectively (N=1, representative of four biological replicates). **D**) BURST signal quantification of gas vesicle acoustic contrast for HEK293T samples transfected with SEMPER mARG plasmids or the leading two plasmid system. Error bars indicate mean ± SEM (N=4 biological replicates). Statistical analysis was conducted using a one-way ANOVA ($p < 0.0001$) followed by Fisher's LSD post-hoc test with a single pooled variance to compare each treatment group with every other. Pairwise p-values are reported in **Table S2-7**. **E**) Annexin V staining assays to quantify the number of apoptotic cells following expression of mARG vectors. HEK293T cells transfected with pgvpA-IRES-mCherry with the start codon removed in front of gvpA were used to establish a baseline (Not Treated). A subset of these cells was treated with Raptinal to induce apoptosis (Raptinal Treated). Other cell populations were transfected solely with a fully functional pgvpA-IRES-mCherry (gvpA-IRES-mCherry), the two-vector system (Two-vector), or the ACC/ACC SEMPER mARG plasmid (SEMPER mARG). Error bars indicate mean ± SEM (N=4 biological replicates). Statistical analysis was conducted using a one-way ANOVA ($p < 0.0001$) followed by Fisher's LSD post-hoc test with a single pooled variance to compare each treatment group to the not treated control. Pairwise p-values are indicated above the bars.

SEMPER enables recombinant monoclonal antibody expression from a single construct. As a further demonstration of the generalizability and biotechnological utility of SEMPER, we used the system to express secreted recombinant monoclonal antibodies in HEK293T cells. As a model construct, we used the b12 antibody, which was discovered as part of early efforts to identify broadly neutralizing antibodies against HIV-1.[43,44] Contrary to the traditional approaches, the SEMPER-b12 plasmid was designed to express both the heavy and the light chains of the b12 IgG from a single mRNA transcript with tunable heavy chain (HC) to light chain (LC) stoichiometry

(**Figure 2-4A**). IgG production was quantified from a sample of media supernatant via an enzyme-linked immunosorbent assay (ELISA).

The SEMPER system was able to produce high levels of functional IgG due to the ability to tune the ratio of HC to LC expression, reaching a maximum using the CCC/ACC SEMPER-b12 construct (**Figure 2-4B**). This production level was similar to that achieved by co-transfecting two separate plasmids (**Figure S2-12**). This finding demonstrates the ability of SEMPER to streamline the production of a valuable biologic product by simplifying the genetic engineering required for complex protein expression, while highlighting the importance of being able to tune component protein stoichiometry.

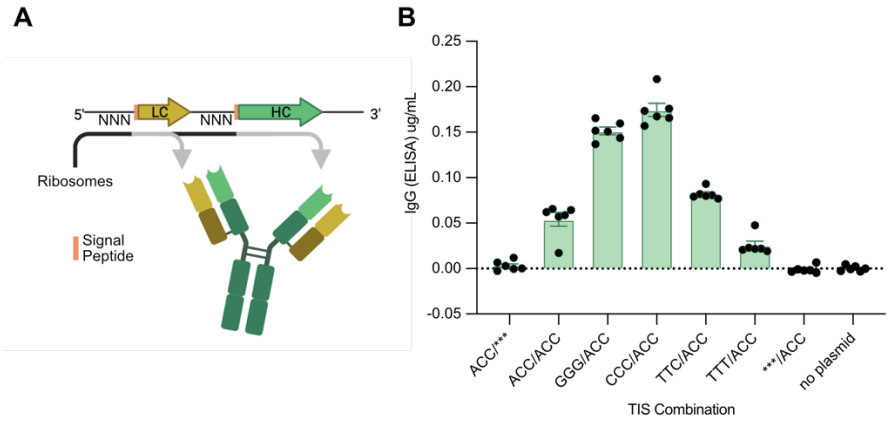

Figure 2-4: Recombinant expression of monoclonal antibodies using SEMPER.
A) Architecture of an mRNA transcript transcribed from transfected SEMPER-b12 plasmid DNA. The first ORF, LC, encodes the b12 IgG light chain, while the second ORF, HC, encodes the b12 IgG heavy chain. Both polypeptides contain a signal peptide for secretion. The TIS of the first ORF, NNN, was varied while the TIS of the second ORF, ACC, remained constant in different constructs. **B**) Total human IgG ELISA results following expression and secretion of SEMPER-b12 vectors. The total amount of transfected DNA was held constant across all conditions. HEK293T cells transfected with either the light chain (ACC/***) or heavy chain (***/ACC) produced no detectable IgG—a similar result to the no plasmid transfection condition. SEMPER-b12 constructs with varying expression stoichiometry produced intermediate amounts of IgG with CCC/ACC producing the highest yield. Error bars indicate mean ± SEM (N=6 biological replicates). Statistical analysis was conducted using a one-way ANOVA ($p < 0.0001$) followed by Fisher's LSD post-hoc test with a single pooled variance to compare each treatment group with every other. Pairwise p-values are reported in **Table S2-8.**

Demonstrating SEMPER 2-ORF efficacy from transfected synthetic mRNA. With accelerating advancements in mRNA vaccine technology and growing development efforts in mRNA-based therapeutics,[13,45] we sought to demonstrate the efficacy of the SEMPER 2-ORF system to tune relative protein expression stoichiometries using mRNA-based expression systems (**Figure 2-5A**). Our *in vitro* transcribed (IVT) mRNA constructs do not include IRES-mCherry, as we observed reduced expression of our GOIs in IVT mRNA constructs containing this element (**Figure S2-13**).

Through IVT and subsequent purification steps, we produced concentrated mRNA samples with the same TIS combinations and ORFs as tested for the plasmid-based 2-ORF SEMPER constructs. As nucleotide chemistry has been shown to impact translation efficiency,[46] we hypothesized that changing the nucleotide chemistry of our IVT mRNA may lead to changes in expression of our two ORFs. Therefore, we created a set of mRNA samples using standard uridine triphosphate (UTP) nucleotides in the IVT reaction, while another set was made using N^1-

methylpseudouridine-5'-triphosphate (m1Ψ), commonly used to manufacture mRNA for clinical use.[47] We then transiently transfected these mRNA samples into HEK293T cells and screened them using flow cytometry. Additionally, we transfected and collected flow cytometry data for two plasmid-based single-color controls for compensation.

Flow cytometry showed that some TIS combinations yielded unique relationships between msfGFP[r5M] and mEBFP2 fluorescence (**Figure 2-5B, 2-5C**). When using standard UTPs, the relationship for TTT/ACC was substantially different than those of CCC/ACC and ACC/ACC, demonstrating that multiple product stoichiometries are accessible from IVT mRNA SEMPER constructs. However, the specific expression ratios for CCC/ACC and ACC/ACC were not as distinct as those achieved using plasmid constructs, suggesting that further IVT-specific tuning is required, potentially in combination with additional mRNA engineering to improve expression efficiency (**Figure S2-13**). In mRNA constructs produced with m1Ψ, the mEBFP2 vs msfGFP[r5M] relationships for TTT/ACC, CCC/ACC, and ACC/ACC were less distinct, confirming that nucleotide chemistry impacts translation initiation.

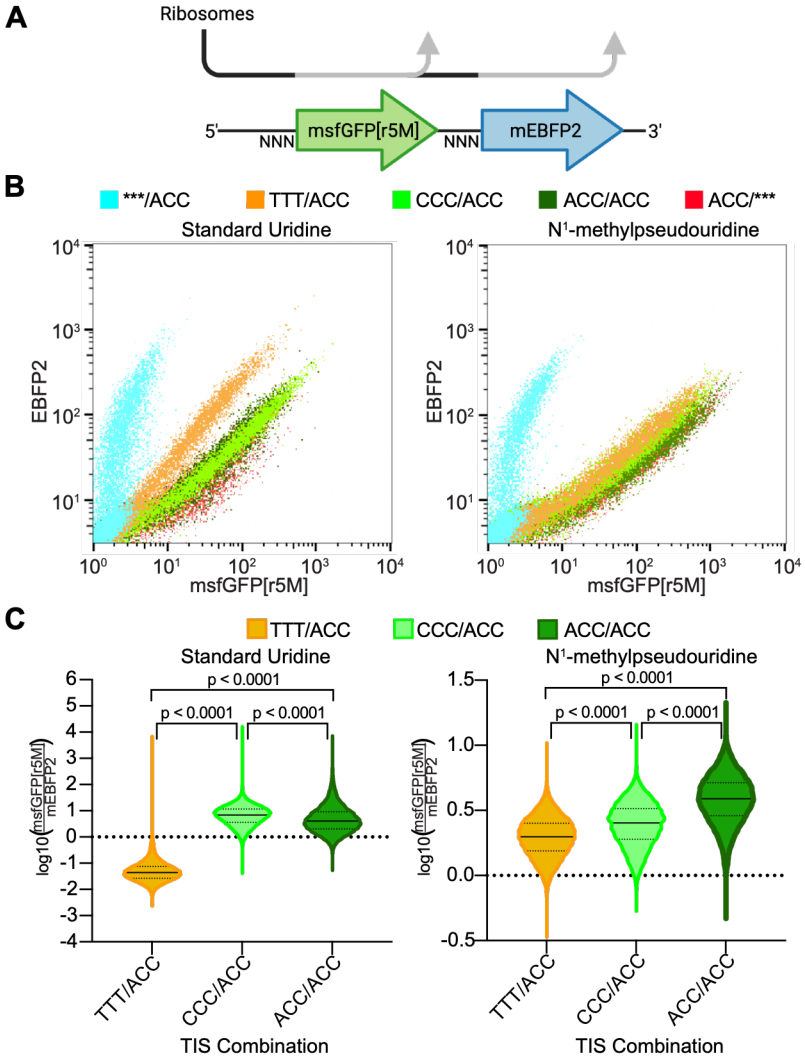

Figure 2-5: 2-ORF SEMPER using *in vitro* transcribed mRNA.
A) Schematic of 2-ORF SEMPER mRNA produced from *in vitro* transcription. The first and second ORFs encode msfGFP[r5M] and mEBFP2 respectively. **B**) Flow cytometry data from transient transfection of IVT mRNA encoding various TIS combinations into HEK293T cells (N=1, representative of four biological replicates). IVT mRNA constructs were made with (Left) standard uridine triphosphate or (Right) N^1-methylpseudouridine-5'-triphosphate. **C**) Violin plots of log10(msfGFP[r5M]/mEBFP2) values for msfGFP[r5M] and mEBFP2 double-positive cells transfected with TTT/ACC, CCC/ACC, or ACC/ACC IVT mRNA produced with different uridine nucleotides. Each distribution represents measurements of at least 25,000 cells from four combined biological

replicates. The median and quartiles of the distribution are represented by the solid and dotted lines respectively. Statistical analysis was conducted using a one-way Welch's ANOVA ($p < 0.0001$ (Left) and $p < 0.0001$ (Right)) followed by Games-Howell's multiple comparisons test to compare TIS combinations.

Basic performance of the SEMPER system is captured by a simple Monte Carlo model. To provide insights on the essential SEMPER mechanism and facilitate future construct design, we developed a simple Monte Carlo simulation that captures the probabilistic nature of tunable, multi-ORF expression based on the LRS model. Briefly, for each Monte Carlo mRNA simulation, the k^{th} scanning ribosome can be consumed by one of three cases (i, ii, iii) as it traverses from the 5' end to the 3' end of the simulated transcript (**Figure 2-6A**). Implemented using sequential Boolean statements, the k^{th} scanning ribosome will first encounter i) a TIS in front of the first ORF. A Bernoulli random variable is then sampled with success probability p_TIS1 dictated by the strength of the TIS. This is equivalent to flipping a biased coin. If the instance of the random variable denotes successful initiation, the ribosome translates the first ORF, producing a single unit of Protein 1 and the simulation moves on to load the next k+1 scanning ribosome. However, if the instance of the random variable denotes non-initiation, the k^{th} scanning ribosome moves on to ii) the TIS sequence in front of the second ORF. Similarly, by sampling a Bernoulli random variable with success probability p_TIS2, the scanning ribosome will either translate a single unit of Protein 2, advancing the simulation to the k+1 iteration, or it will fail to initiate translation. In the absence of initiation, the k^{th} scanning ribosome is considered to iii) produce no translation product and the simulation moves on to load the next k+1 scanning ribosome. The aggregate totals of Protein 1 and Protein 2 from each simulated mRNA were then stored. The model does not implement IRES-mCherry expression as mCherry was only used for experimental normalization.

p_TIS success probabilities for ***, TTT, CCC, and ACC were chosen as 0.0001, 0.1, 0.5, and 0.9 respectively. p_TIS for *** was chosen to be nonzero to represent minute translation of functional protein initiated at in-frame, non-canonical start sites.[48] For each TIS combination depicted in **Figure 2-1**, we conducted nearly 500,000 Monte Carlo mRNA simulations. An average of 100 scanning ribosomes were simulated for each Monte Carlo mRNA simulation. Using standard resampling techniques, the simulated expression results of 100 mRNAs, on average, were then aggregated together into approximately 10,000 "single cells" to model transient transfection and then transcription of plasmid DNA. The results of our Monte Carlo simulations depicted in **Figure 2-6B** capture the major observed expression differences seen in **Figure 2-1C** caused by changing the TIS in front of the first ORF.

To extend our model's ability to capture practical design rules, we simulated results for a TTT/ACC construct containing one internal methionine (iMet) within ORF 1. As seen in **Figure 2-1B**, internal methionines within ORF 1 reduced the flux of scanning ribosomes to ORF 2, thus shifting the expression of Protein 2 (mEBFP2) to lower values. To implement this in our simulation, we added an additional TIS (with p_TIS=0.5) between ORF 1 and ORF 2 in our model to represent an iMet within ORF 1. As expected, the presence of an iMet shifted the expression distribution of Protein 2 to lower values (**Figure 2-6C,** Left). For experimental validation, we added one canonical methionine back into our msfGFP[r5M] CDS, producing msfGFP[r4M]. The TIS created by the inclusion of this iMet was predicted to have a TIS strength similar to that of CCC based on quantitative results and modeling conducted by Noderer et al.[30] In HEK293T cells, we then compared two TTT/ACC plasmids, both of which contained mEBFP2 in ORF 2, while ORF 1 was varied between msfGFP[r5M] and msfGFP[r4M]. Consistent with modeling, the construct encoding msfGFP[r4M] yielded less mEBFP2 expression compared to the one without iMet (**Figure 2-6C,** Right), although differences in the magnitude of this effect suggest that future optimization of p_TIS values and other parameters in the model are needed to improve its quantitative prediction accuracy. For more detailed descriptions of model architecture, design choices, parameters, and assumptions, please see **Methods**.

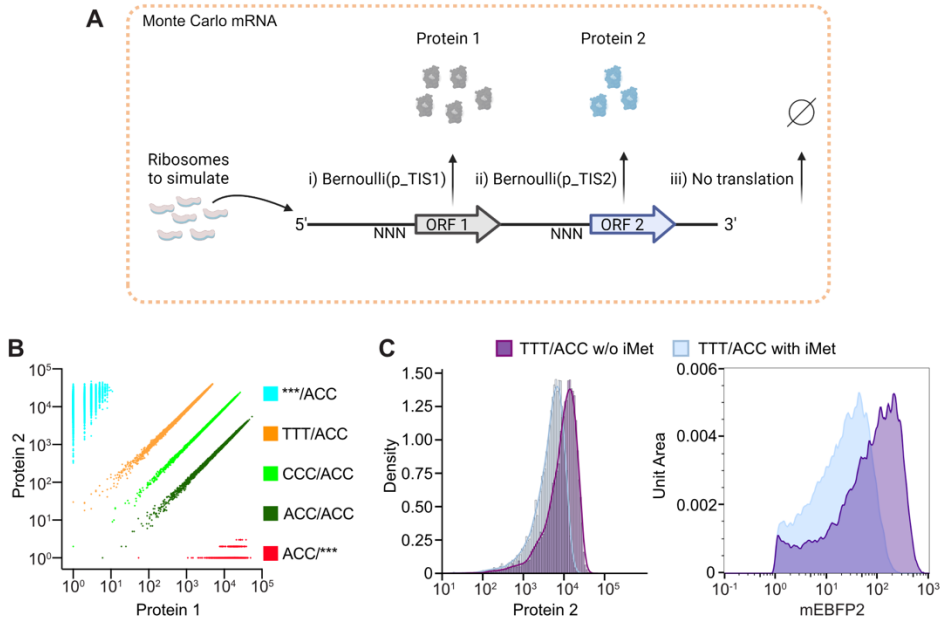

Figure 2-6: Simulating SEMPER-based gene expression.
A) Depiction of an individual Monte Carlo simulation unit referred to as a "Monte Carlo mRNA". In each mRNA simulation, scanning ribosomes are sequentially loaded onto the mRNA. Each ribosome traverses the mRNA until it is consumed by one of three Boolean stopping conditions. i) The ribosome may successfully initiate translation of ORF 1 at the first TIS to produce a single unit of Protein 1. If successful, the simulation loads the next ribosome. If unsuccessful, the ribosome continues to the TIS in front of ORF 2 and ii) the ribosome may successfully initiate translation of ORF 2 at the second TIS to produce a single unit of Protein 2. If successful, the simulation loads the next ribosome. If unsuccessful, the ribosome is discarded iii) without producing any translation product. The probability the ribosome will translate a given ORF is modeled by a Bernoulli random variable with success probability p_TIS. p_TIS values are based on the identity of the TIS in front of the ORF. This process is repeated for multiple scanning ribosomes. **B)** Simulated flow cytometry plots for TIS combinations depicted in **Figure 2-1C**. For each TIS combination, hundreds of thousands of Monte Carlo mRNAs were simulated. The results of these simulations were aggregated into 10,000 smaller groups or "cells" and then plotted. **C**) Modeling expression changes in ORF 2 due to the presence of an internal methionine (iMet) in ORF 1. (Left) Simulated results of the expression distributions for cells transfected with plasmids utilizing TTT and ACC for ORF 1 and ORF 2 respectively. The left-shifted, light-blue distribution is caused by the consumption of ribosomes that initiate translation at the iMet, reducing ribosomal flux to ORF 2. 10,000 cells were simulated for each distribution. (Right) TTT/ACC constructs were tested experimentally, one with msfGFP[r5M] encoded in ORF 1 and one where an iMet was added back into a canonical position of msfGFP[r5M]. The presence of the iMet led to a reduction in ORF 2 translation, represented by mEBFP2 expression (N=1, representative of four biological replicates).

Discussion

Our results demonstrate that SEMPER enables tunable expression of multiple recombinant ORFs from single transcripts in mammalian cells by using leaky ribosomal scanning and TISs of varying strength. The SEMPER paradigm is applicable across cells from multiple species and can yield a range of user-tunable expression stoichiometries for at least three proteins on a single compact mRNA. Our results with SEMPER mARGs in GV production exemplify its application in encoding complex multi-gene constructs, allowing for precise modulation of expression ratios to optimize output while reducing cellular toxicity. Such tunability could have similar value in other scenarios demanding specific subunit/chaperone or toxin/antitoxin ratios, where co-transfection would result in a wider distribution of ratios. Additionally, the application of SEMPER to secreted mAb expression highlights the method's versatility and potential utility in therapeutic production. Our results also show that the SEMPER framework can be applied directly to mRNA-based genetic circuits by demonstrating polycistronic expression with 2-ORF SEMPER mRNA constructs using IVT. Furthermore, we show that nucleotide chemistry can alter the translation levels of encoded ORFs in IVT mRNA, opening avenues to further tune protein stoichiometries with synthetic

mRNA. The basic performance of SEMPER is captured by a relatively simple Monte Carlo model, which will help elucidate core design rules for future users.

While SEMPER has demonstrated tunability and rank-order relationships between TIS sets across multiple cell types, gene delivery methods, and protein outputs, the precise expression levels are context-dependent. Differences in these levels may arise from complex gene expression dynamics influenced by factors such as ORF length, RNA secondary structure, and RNA processing.[49] Nevertheless, the consistent rank order of expression ratio maintained across the set of TIS combinations tested in this work provides a convenient starting point for construct design, necessitating only limited empirical screening (which is required in most other synthetic biology approaches). The ability of SEMPER to provide tunability of polycistronic expression within a defined range is supported by our experiments with diverse ORFs, including fluorescent proteins, prokaryotic GVs, and the heavy and light chains of recombinant IgG.

The translation-based stoichiometry control provided by SEMPER complements transcription-based approaches that control relative expression using synthetic promoters and programmable transcriptional systems[7,8]. SEMPER's translation-focused approach allows tuning to be more compact and encodable at both DNA and mRNA levels. Combining the two classes of methods would enable the construction of gene circuits with greater complexity if needed for synthetic biology applications. SEMPER can also be combined with advanced circuits designed to overcome resource limitation and DNA copy number variability.[24,50]

Although the genetic constructs described in this work should be immediately useful in a variety of contexts, future work is needed to demonstrate the SEMPER framework's utility upon integration into the genome through stable transfection and viral transduction methods. Another future step is to make SEMPER compatible with methionine-containing proteins. Internal methionine codons create TISs that may undesirably consume ribosomal flux from downstream open reading frames. If mutating these methionines is not possible, an alternative approach may be to alter the nucleotide context surrounding the in-frame AUG using rationally chosen degenerate codons to effectively mask the TIS. Future work should also elucidate the maximum number of ORFs that are encodable using SEMPER at both the plasmid and mRNA levels. Additional work focused on IVT mRNA applications should examine the influence of different base chemistries, capping reagents, and untranslated regions. Lastly, future studies can further determine the mechanisms by which strong TIS sequences can enhance translation initiation at upstream and even downstream ORFs, as observed in our 3-ORF SEMPER screen with multiple ACC-containing ORFs.

We believe SEMPER represents a substantial advancement in the field of recombinant multi-protein expression and synthetic biology, enabling compact, user-friendly tuning of multiple proteins within mammalian systems. Researchers can readily adopt this framework to rapidly screen libraries relevant to expression and tuning of enzymatic pathways and circuits, assembly of multimeric protein structures, and other endeavors.[51] As we have demonstrated in this study, this technology can enable simultaneous expression of a protein of interest and its folding chaperones from a single transcriptional unit at an optimal ratio, offering a strategy to tackle protein misfolding precisely. Moreover, as shown in some of the constructs used in this study, SEMPER can be used in combination with IRES and 2A elements for versatile encoding of more complex genetic constructs.

SEMPER also holds potential in the field of RNA vaccines and therapeutics. It could enhance the development of polyvalent RNA vaccines as well as the expression of mosaic virus-like particles from delivered mRNA, thereby broadening the immune response against highly variable

viruses.[52,53] Further, it could improve RNA-delivered monoclonal antibody therapeutics by optimizing the ratio of heavy and light chain production in each cell, and by allowing for simultaneous production of multiple or bi-specific antibodies from a single mRNA.[54–56] This technology also paves the way for the production of cytokine cocktails through the co-expression of multiple cytokines from a single mRNA, which could offer synergistic effects to modulate immune responses. Taken together, SEMPER provides an option that can always be considered in developing polycistronic constructs for synthetic biology and medicine.

Methods

Cell lines

HEK293T (female human embryonic kidney) cells (American Type Culture Collection (ATCC), CRL-3216) and CHO-K1 (female, Chinese hamster ovary) cells (American Type Culture Collection (ATCC), CCL-61) were cultured in 24-well plates at 37 °C and 5% CO_2 in a humidified incubator in 0.5 mL of DMEM (Corning, 10-013-CV) with 10% FBS (Gibco) and 1X penicillin–streptomycin.

Vector construction

Plasmids were constructed using standard cloning techniques, including Gibson assembly and conventional restriction and ligation. All final sequences were verified using whole-plasmid sequencing through Primordium Labs. Plasmids containing two or three fluorescent protein ORFs were constructed in the following way: individual FP sequences were ordered from IDT or TWIST Biosciences as synthetic gBlocks. Modified TIS sequences and the *** sequence were introduced using overhang PCR primers. Finally, all components were subcloned into pCMV-Sport-gvpA-IRES-mCherry using NEB HiFi DNA Assembly replacing the gvpA ORF. Some modifications were made to the 5' UTR sequences. Plasmid sequences for the 2-ORF TTT/ACC construct and the 3-ORF ACC/***/*** construct as well as sequences for individual genetic parts in these plasmids are provided in **Table S2-9** to aid in reproduction of this work. A subset of plasmids used in this work are available through Addgene.

SEMPER mARG plasmids containing gvpA and gvpNJKFGW-EmGFP were constructed as follows: the ORF containing gvpNJKFGW-EmGFP was PCR amplified and subcloned into pCMV-Sport-gvpA-IRES-mCherry using NEB HiFi DNA Assembly between gvpA and IRES. Different gvpA TIS sequences were introduced using single-stranded oligo bridge primers with NEB HiFi Assembly.

To produce plasmids amenable for *in vitro* transcription of mRNA, sequences containing the 5' UTR through the stop codon of the mEBFP2 ORF from the 2-ORF SEMPER plasmids were cloned into an IVT plasmid backbone containing a T7 promoter, a synthetic 3' UTR sequence, and a 100 bp polyA track (**Table S2-9**) using PCR amplification and Gibson assembly methods.

Production of mRNA by *in vitro* transcription

Linear templates were PCR amplified from the IVT plasmids (described in Vector construction) using primer p101 and "Ultramer" p54 synthesized by Integrated DNA Technologies. Primer p101 introduced an AG dinucleotide following the T7 promoter sequence on linear templates to ensure efficient 5' capping of IVT mRNA. Linear templates were purified using standard gel electrophoresis and DNA cleanup methods. IVT mRNA was produced using the HiScribe T7 High Yield RNA Synthesis Kit (NEB # E2040) along with 500 ng of linear template and 4 mM CleanCap AG Reagent (Trilink, N-7113). 5 mM N^1-Methylpseudouridine-5'-triphosphate (Trilink, N-1081) was substituted for the standard uridine triphosphate included in the HiScribe Kit for certain

reactions. Following the IVT incubation, reactions were additionally incubated with 2 units of DNAse I to remove remaining DNA template. Purification of IVT mRNA was conducted using the Monarch RNA Cleanup Kit (NEB #T2040L). To confirm IVT mRNAs were of the correct lengths, aliquots of purified IVT mRNA were denatured at 70°C for 15 minutes and then subjected to gel electrophoresis on a 1% agarose gel in Tris acetate EDTA (TAE) stained with SYBR Safe. ssRNA ladder (NEB #N0362S) was used to confirm band sizes. mRNA concentrations for each IVT mRNA were quantified using a Qubit Fluorometer (Invitrogen). Completed IVT mRNA samples were stored at -80°C.

Cell culture and transfection

HEK293T cells (American Type Culture Collection (ATCC), CRL-3216) and CHO-K1 cells (American Type Culture Collection (ATCC), CCL-61) were cultured in 24-well plates at 37 °C and 5% CO_2 in a humidified incubator in 0.5 mL of DMEM (Corning, 10-013-CV) with 10% FBS (Gibco) and 1X penicillin–streptomycin until about 80% confluency before transfection with plasmid DNA or IVT mRNA.

For plasmid DNA transfection, transient transfection mixtures were created by mixing 500 ng of plasmid with polyethyleneimine (PEI-MAX, linear 40 kD #24765-2, Polysciences) at 4.12 µg of polyethyleneimine per microgram of DNA in 150 mM NaCl. The mixture was incubated for 12 minutes at room temperature and added dropwise to HEK293T or CHO-K1 adherent cells. Media was changed after 12–16 hours and daily thereafter. Cells were analyzed with flow-cytometry 48 hours post-transfection. Deviations from this protocol are described in the supplementary information where applicable.

For IVT mRNA transfection, transient transfection mixtures were created by mixing 500 ng of mRNA, 1.5 µL of Lipofectamine MessengerMAX Reagent (Invitrogen #LMRNA008), and 50 µL of Opti-Mem media (Gibco #31985070) according to the Lipofectamine MessengerMAX Reagent standard operating procedure. The mixture was incubated for 5 minutes at room temperature and added dropwise to HEK293T cells. Media was changed after 12–16 hours and daily thereafter. Cells were analyzed with flow-cytometry 48 hours post-transfection.

For transfection of plasmids encoding mARGs, transient transfection mixtures were created as above except that 600 ng of total DNA was mixed together as follows: 56 fmol of gvpA-IRES-mCherry plasmid with or without 14 fmol gvpNJKFGW-EmGFP plasmid or 56 fmol of a 2-ORF SEMPER mARG plasmid. Plasmid mixtures were normalized with addition of pUC19 up to 600 ng before complexing with PEI-MAX. The mixture was incubated for 12 minutes at room temperature and added dropwise to HEK293T adherent cells. Media was changed after 12–16 hours and daily thereafter. Cells were harvested 3-days post-transfection.

Expression of b12 IgG in HEK293T cells

HEK293T cells were plated at a density of 200,000 cells per well in 48-well plates using 250 µL of Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1X penicillin/streptomycin, 20 mM HEPES (pH 7.0), 1X MEM Non-Essential Amino Acids (NEAA), and 1X GlutaMAX™. The cells were cultured at 37˚C in a humidified atmosphere containing 5% CO_2 for 24 hours prior to transfection.

A DNA-polyethylenimine (PEI) Max transfection mixture was prepared by dissolving 206 mg of PEI-MAX (40 kDa, Polysciences, #24765-2) in 500 mL of tissue culture (TC)-grade phosphate-buffered saline (PBS) to achieve a final concentration of 0.412 mg/mL (10.3 µM). The solution was adjusted to pH 7.0 with NaOH and filter-sterilized. Aliquots were stored at -80˚C until needed. For each transfection, 350 ng of the SEMPER-b12 plasmid (7485 bp) was combined with 17.5 µL

of a sterile 150 mM NaCl solution. Separately, the PEI-Max solution was diluted with 150 mM NaCl in a 1:4 ratio, and 17.5 µL of this dilution was added to the plasmid-NaCl mixture. The combined mixture was gently mixed and incubated at room temperature for 12 minutes to allow complex formation, after which it was added dropwise to each well. For the two-plasmid control, 175 ng of heavy chain (HC) and light chain (LC) plasmids were combined to total 350 ng and then complexed with the PEI-Max solution as described above.

The media in the wells were replaced after 16 hours with fresh DMEM. An additional 250 µL of the same medium was supplemented after another 24 hours, bringing the total volume in each well to 500 µL. Three days post-transfection, 5 µL of the cell culture supernatant was transferred to 95 µL of assay buffer.

IgG yields were quantified using the Human Total IgG Coated ELISA Kit (Invitrogen, #BMS2091), adhering to the manufacturer's protocol. Absorbance was measured at 450 nm with a reference wavelength of 620 nm using the Tecan SPARK plate reader.

Flow cytometry and apoptosis assays

To harvest the cells, cells were dissociated using trypsin/EDTA and centrifuged at 300 g for 6 minutes at room temperature. For experiments involving mARGs, 2-ORF SEMPER plasmid transfections, and 2-ORF SEMPER mRNA transfections, cells were analyzed with MACSQuant VYB (Miltenyi Biotec) except for data depicted in **Figure 2-6C** (Right) which was gathered using a MACSQuant10 (Miltenyi Biotec). For experiments involving 3-ORF SEMPER, cells were analyzed using Cytoflex S (Beckman Coulter). Single-color controls were used to allow for compensation. In plasmid transfection experiments, cells were gated for size and doublet-discriminated before being gated and further binned by mCherry fluorescence. For IVT mRNA transfection experiments, cells were gated as above except they were gated for msfGFP[r5M] and mEBFP2 double-positivity instead of by mCherry fluorescence.

For the Annexin V apoptosis assay, we transfected cells as described above but did not perform media changes. For the positive control, cells transfected with pgvpA-IRES-mCherry (w/ start codon removed in front of gvpA) were treated with 10 µM Raptinal 18 hours before harvest. Cells were harvested two days after transfection. To do this, we collected supernatant from cells to recover the dead cells. We then trypsinized the adhered cells and added this trypsinized cell fraction to the original supernatant for maximum cell recovery. Then, we spun the cells down at 300 g for 6 minutes, carefully removed the supernatant, and resuspended the cells in 90 µL of cold annexin binding buffer with no EDTA and supplemented Ca^{2+}, followed by 10 µL of Annexin V Pacific Blue stain (A35122). Following a 15-minute incubation, we added 150 µL of the same annexin binding buffer and then performed flow cytometry to obtain fluorescence values for the pacific blue stain. We gated in FlowJo by selecting the largest population in the FCS/SSC plot, keeping the lower left corner in the cell gate to include smaller apoptotic cells. We selected a 10^4 threshold for the Pacific Blue Annexin V stain, corresponding to the right tail of the unstained control so that the Annexin V+ population of the unstained control was around 0%. Values were normalized by setting the mean value of Raptinal-treated samples to 100%.

In vitro ultrasound imaging of HEK293T cells

Cells were harvested as described above. Cells were resuspended with 1% low-melt agarose (GoldBio) in PBS at 40°C at concentrations of ~15 million cells per milliliter and then loaded into wells of pre-formed phantoms consisting of 1% molecular biology-grade agarose (Bio-Rad) in PBS. Phantoms were imaged using L22-14vX transducer (Verasonics) at 15.625 MHz while submerged in PBS on top of an acoustic absorber pad. For BURST imaging, wells were centered around the 8 mm natural focus of the transducer and a BURST pulse sequence was applied in

pAM acquisition mode with the focus set to 8 mm, and the voltage was set to 2V for the first 10 frames and 15V for the remaining frames.

BURST images were produced by pixel-wise subtraction of the 11th (collapse) frame from the 54th (post-collapse) frame. Resulting differences were divided by 100. Images were quantified as follows: The sample ROIs were drawn inside the well of the agarose phantom. Average pixel value inside the ROI was calculated for each replicate.

Simulations

The RAM and CPU power of a standard laptop computer were sufficient to run the simulations depicted in **Figure 2-6**. Simulation code along with example usage and a list of dependencies are accessible through GitHub: https://github.com/shapiro-lab/semper-simulation and through the following DOI: 10.5281/zenodo.11136551.

The model assumes scanning ribosomes (also referred to below as just "ribosomes") load at the 5' end of an mRNA transcript and will only move across the mRNA from 5' to 3'. The p_TIS values for ***, TTT, CCC, and ACC were estimated based on our observed results and the results of Ferreira et al. and Noderer et al.[23,30] Estimates of these relative p_TIS values can be made for other TIS sequences through analysis of the predicted and measured strengths of TISs made by Noderer et al. We assume there is inherent stochasticity with the number of scanning ribosomes that will traverse an mRNA and the number of mRNA that may exist in a cell. This stochasticity was captured by sampling these values from probabilistic distributions. Specifically, normality was used to describe these processes. The model does not consider the sequence identity of the ORFs other than the TIS sequence in front of it. The model does not account for ribosomes falling off the transcript during scanning, the length of the transcript, the length of the 5' UTR, local structure of the mRNA, or global structure of the mRNA.

The architecture of the modeling pipeline is described below:

1. *Choosing the number of mRNA to model.* The number of unique mRNA simulation results needed to fill the desired number of cells to be modeled is estimated.
2. *Estimating the number of ribosomes that will traverse each mRNA*. Once the number of mRNA are determined, the number of ribosomes that will traverse each mRNA are sampled from a normal distribution. We sampled from a normal distribution with mean and standard deviation both equal to 100. The mean value of 100 was motivated by estimates of the number of observed translation events from single CD147 mRNAs in mammalian cells.[57]
3. *Simulating ribosomes traversing an mRNA.* Each mRNA then undergoes a Monte Carlo simulation where R ribosomes will traverse the mRNA. To do this we iterate through a set of Boolean conditions k=1:R times, as described in Results. We say the k^{th} ribosome has been loaded onto the 5' end of the mRNA when the k^{th} iteration of the loop begins.
4. *Loading/aggregating mRNA simulation results into cells.* Once all mRNAs have been simulated, the results of individual mRNA simulations are aggregated together as a proxy for transient transfection and implicit transcription of plasmid-DNA. Here, the number of individual mRNA simulation results that are aggregated together is sampled from a normal distribution with mean 100 and standard deviation of 100 to capture the stochasticity of transient transfection. This mean value is motivated by observed endogenous CD147 mRNA counts in single cells. Resampling of the mRNA simulations was used in this step as mRNA simulations were computationally expensive to calculate. Specifically, from ~500,000 individual mRNA simulations, 2,500 cells were "loaded" with aggregated results. The simulations were then randomized, and another 2,500 cells were loaded. This process

was conducted four times to yield 10,000 simulated cells per condition, as plotted in **Figure 2-6B**.

The model can readily be used to consider systems with more than two ORFs by adding additional p_TIS values in the input parameters. The addition of a p_TIS can be considered as the addition of a desired ORF or an internal methionine within an ORF. The latter case was considered to produce **Figure 2-6C**, as showcased in the publicly available code. The model does not differentiate between a TIS that may exist at the canonical 5' end of an ORF or a TIS that may be encoded in the middle of an ORF.

Quantification and statistical analysis

All statistical analysis was performed in GraphPad Prism 10. Pairwise statistical analysis, number of datapoints, and statistical methods used can be found in figure legends and in the referenced tables provided in the Supplementary Information. As used throughout the text, we define a single biological replicate as the transient transfection of cells with DNA or mRNA in a single culture well.

References

1. Lienert, F., Lohmueller, J.J., Garg, A., and Silver, P.A. (2014). Synthetic biology in mammalian cells: next generation research tools and therapeutics. Nat. Rev. Mol. Cell Biol. *15*, 95–107. https://doi.org/10.1038/nrm3738.
2. Farhadi, A., Ho, G.H., Sawyer, D.P., Bourdeau, R.W., and Shapiro, M.G. (2019). Ultrasound Imaging of Gene Expression in Mammalian Cells. Science *365*, 1469–1475. https://doi.org/10.1126/science.aax4804.
3. Kitada, T., DiAndreth, B., Teague, B., and Weiss, R. (2018). Programming gene and engineered-cell therapies with synthetic biology. Science *359*, eaad1067. https://doi.org/10.1126/science.aad1067.
4. Bosma, B., du Plessis, F., Ehlert, E., Nijmeijer, B., de Haan, M., Petry, H., and Lubelski, J. (2018). Optimization of viral protein ratios for production of rAAV serotype 5 in the baculovirus system. Gene Ther. *25*, 415–424. https://doi.org/10.1038/s41434-018-0034-7.
5. Schlatter, S., Stansfield, S.H., Dinnis, D.M., Racher, A.J., Birch, J.R., and James, D.C. (2005). On the Optimal Ratio of Heavy to Light Chain Genes for Efficient Recombinant Antibody Production by CHO Cells. Biotechnol. Prog. *21*, 122–133. https://doi.org/10.1021/bp049780w.
6. Trolle, J., McBee, R.M., Kaufman, A., Pinglay, S., Berger, H., German, S., Liu, L., Shen, M.J., Guo, X., Martin, J.A., et al. (2022). Resurrecting essential amino acid biosynthesis in mammalian cells. eLife *11*, e72847. https://doi.org/10.7554/eLife.72847.
7. Qin, C., Xiang, Y., Liu, J., Zhang, R., Liu, Z., Li, T., Sun, Z., Ouyang, X., Zong, Y., Zhang, H.M., et al. (2023). Precise programming of multigene expression stoichiometry in mammalian cells by a modular and programmable transcriptional system. Nat. Commun. *14*, 1500. https://doi.org/10.1038/s41467-023-37244-y.
8. Patel, Y.D., Brown, A.J., Zhu, J., Rosignoli, G., Gibson, S.J., Hatton, D., and James, D.C. (2021). Control of Multigene Expression Stoichiometry in Mammalian Cells Using Synthetic Promoters. ACS Synth. Biol. *10*, 1155–1165. https://doi.org/10.1021/acssynbio.0c00643.
9. Hurt, R.C., Buss, M.T., Duan, M., Wong, K., You, M.Y., Sawyer, D.P., Swift, M.B., Dutka, P., Barturen-Larrea, P., Mittelstein, D.R., et al. (2023). Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. Nat. Biotechnol. *41*, 919–931. https://doi.org/10.1038/s41587-022-01581-y.
10. Ferreira, J.P., Peacock, R.W.S., Lawhorn, I.E.B., and Wang, C.L. (2011). Modulating ectopic gene expression levels by using retroviral vectors equipped with synthetic promoters. Syst. Synth. Biol. *5*, 131–138. https://doi.org/10.1007/s11693-011-9089-0.
11. Koh, E.Y.C., Ho, S.C.L., Mariati, Song, Z., Bi, X., Bardor, M., and Yang, Y. (2013). An Internal Ribosome Entry Site (IRES) Mutant Library for Tuning Expression Level of Multiple Genes in Mammalian Cells. PLoS ONE *8*, e82100. https://doi.org/10.1371/journal.pone.0082100.
12. Liu, Z., Chen, O., Wall, J.B.J., Zheng, M., Zhou, Y., Wang, L., Ruth Vaseghi, H., Qian, L., and Liu, J. (2017). Systematic comparison of 2A peptides for cloning multi-genes in a polycistronic vector. Sci. Rep. *7*, 2193. https://doi.org/10.1038/s41598-017-02460-2.
13. Schlake, T., Thess, A., Fotin-Mleczek, M., and Kallen, K.-J. (2012). Developing mRNA-vaccine technologies. RNA Biol. *9*, 1319–1330. https://doi.org/10.4161/rna.22269.
14. Douin, V., Bornes, S., Creancier, L., Rochaix, P., Favre, G., Prats, A.-C., and Couderc, B. (2004). Use and comparison of different internal ribosomal entry sites (IRES) in tricistronic retroviral vectors. BMC Biotechnol. *4*, 16. https://doi.org/10.1186/1472-6750-4-16.
15. Ho, S.C.L., Bardor, M., Li, B., Lee, J.J., Song, Z., Tong, Y.W., Goh, L.-T., and Yang, Y. (2013). Comparison of Internal Ribosome Entry Site (IRES) and Furin-2A (F2A) for Monoclonal Antibody Expression Level and Quality in CHO Cells. PLOS ONE *8*, e63247. https://doi.org/10.1371/journal.pone.0063247.

16. Mizuguchi, H., Xu, Z., Ishii-Watabe, A., Uchida, E., and Hayakawa, T. (2000). IRES-Dependent Second Gene Expression Is Significantly Lower Than Cap-Dependent First Gene Expression in a Bicistronic Vector. Mol. Ther. *1*, 376–382. https://doi.org/10.1006/mthe.2000.0050.
17. Jackson, R.J., Hellen, C.U.T., and Pestova, T.V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat. Rev. Mol. Cell Biol. *11*, 113–127. https://doi.org/10.1038/nrm2838.
18. Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell *136*, 731–745. https://doi.org/10.1016/j.cell.2009.01.042.
19. Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. *109*, E2424–E2432. https://doi.org/10.1073/pnas.1207846109.
20. Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. Gene *299*, 1–34. https://doi.org/10.1016/S0378-1119(02)01056-9.
21. Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell *44*, 283–292. https://doi.org/10.1016/0092-8674(86)90762-2.
22. Barbosa, C., Peixeiro, I., and Romão, L. (2013). Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. PLOS Genet. *9*, e1003529. https://doi.org/10.1371/journal.pgen.1003529.
23. Ferreira, J.P., Overton, K.W., and Wang, C.L. (2013). Tuning gene expression with synthetic upstream open reading frames. Proc. Natl. Acad. Sci. *110*, 11284–11289. https://doi.org/10.1073/pnas.1305590110.
24. Jones, R.D., Qian, Y., Siciliano, V., DiAndreth, B., Huh, J., Weiss, R., and Del Vecchio, D. (2020). An endoribonuclease-based feedforward controller for decoupling resource-limited genetic modules in mammalian cells. Nat. Commun. *11*, 5690. https://doi.org/10.1038/s41467-020-19126-9.
25. O'Connell, R.W., Rai, K., Piepergerdes, T.C., Wang, Y., Samra, K.D., Wilson, J.A., Lin, S., Zhang, T.H., Ramos, E.M., Sun, A., et al. (2023). Ultra-high throughput mapping of genetic design space (Synthetic Biology) https://doi.org/10.1101/2023.03.16.532704.
26. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T.C., and Waldo, G.S. (2006). Engineering and characterization of a superfolder green fluorescent protein. Nat. Biotechnol. *24*, 79–88. https://doi.org/10.1038/nbt1172.
27. Soundrarajan, N., Sokalingam, S., Raghunathan, G., Budisa, N., Paik, H.-J., Yoo, T.H., and Lee, S.-G. (2012). Conjugation of Proteins by Installing BIO-Orthogonally Reactive Groups at Their N-Termini. PLoS ONE *7*, e46741. https://doi.org/10.1371/journal.pone.0046741.
28. Kaminski, A., Howell, M.T., and Jackson, R.J. (1990). Initiation of encephalomyocarditis virus RNA translation: the authentic initiation site is not selected by a scanning mechanism. EMBO J. *9*, 3753–3759. https://doi.org/10.1002/j.1460-2075.1990.tb07588.x.
29. Jackson, R.J., Howell, M.T., and Kaminski, A. (1990). The novel mechanism of initiation of picornavirus RNA translation. Trends Biochem. Sci. *15*, 477–483. https://doi.org/10.1016/0968-0004(90)90302-R.
30. Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., and Wang, C.L. (2014). Quantitative analysis of mammalian translation initiation sites by FACS -seq. Mol. Syst. Biol. *10*, 748. https://doi.org/10.15252/msb.20145136.
31. Walsh, G. (2018). Biopharmaceutical benchmarks 2018. Nat. Biotechnol. *36*, 1136–1145. https://doi.org/10.1038/nbt.4305.
32. Wu, C.-Y., Yeh, Y.-C., Yang, Y.-C., Chou, C., Liu, M.-T., Wu, H.-S., Chan, J.-T., and Hsiao, P.-W. (2010). Mammalian Expression of Virus-Like Particles for Advanced Mimicry of

Authentic Influenza Virus. PLoS ONE *5*, e9784. https://doi.org/10.1371/journal.pone.0009784.
33. Wu, Q., Wright, M., Gogol, M.M., Bradford, W.D., Zhang, N., and Bazzini, A.A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. EMBO J. *39*, e104763. https://doi.org/10.15252/embj.2020104763.
34. Takyar, S., Hickerson, R.P., and Noller, H.F. (2005). mRNA Helicase Activity of the Ribosome. Cell *120*, 49–58. https://doi.org/10.1016/j.cell.2004.11.042.
35. Hurt, R.C., Buss, M.T., Duan, M., Wong, K., You, M.Y., Sawyer, D.P., Swift, M.B., Dutka, P., Barturen-Larrea, P., Mittelstein, D.R., et al. (2023). Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. Nat. Biotechnol., 1–13. https://doi.org/10.1038/s41587-022-01581-y.
36. Bourdeau, R.W., Lee-Gosselin, A., Lakshmanan, A., Farhadi, A., Kumar, S.R., Nety, S.P., and Shapiro, M.G. (2018). Acoustic reporter genes for noninvasive imaging of microbes in mammalian hosts. Nature *553*, 86–90. https://doi.org/10.1038/nature25021.
37. Huber, S.T., Terwiel, D., Evers, W.H., Maresca, D., and Jakobi, A.J. (2023). Cryo-EM structure of gas vesicles for buoyancy-controlled motility. Cell *186*, 975-986.e13. https://doi.org/10.1016/j.cell.2023.01.041.
38. Dutka, P., Metskas, L.A., Hurt, R.C., Salahshoor, H., Wang, T.-Y., Malounda, D., Lu, G.J., Chou, T.-F., Shapiro, M.G., and Jensen, G.J. (2023). Structure of Anabaena flos-aquae gas vesicles revealed by cryo-ET. Structure *31*, 518-528.e6. https://doi.org/10.1016/j.str.2023.03.011.
39. Wang, H., Sivonen, K., Rouhiainen, L., Fewer, D.P., Lyra, C., Rantala-Ylinen, A., Vestola, J., Jokela, J., Rantasärkkä, K., Li, Z., et al. (2012). Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium Anabaena sp. strain 90. BMC Genomics *13*, 613. https://doi.org/10.1186/1471-2164-13-613.
40. Sawyer, D.P., Bar-Zion, A., Farhadi, A., Shivaei, S., Ling, B., Lee-Gosselin, A., and Shapiro, M.G. (2021). Ultrasensitive ultrasound imaging of gene expression with signal unmixing. Nat. Methods *18*, 945–952. https://doi.org/10.1038/s41592-021-01229-w.
41. Jost, A., Knitsch, R., Völkner, K., and Pfeifer, F. (2021). Effect of Mutations in GvpJ and GvpM on Gas Vesicle Formation of Halobacterium salinarum. Front. Microbiol. *12*.
42. Pfeifer, F. (2022). Recent Advances in the Study of Gas Vesicle Proteins and Application of Gas Vesicles in Biomedical Research. Life Basel Switz. *12*, 1455. https://doi.org/10.3390/life12091455.
43. Burton, D.R., Pyati, J., Koduri, R., Sharp, S.J., Thornton, G.B., Parren, P.W.H.I., Sawyer, L.S.W., Hendry, R.M., Dunlop, N., Nara, P.L., et al. (1994). Efficient Neutralization of Primary Isolates of HIV-1 by a Recombinant Human Monoclonal Antibody. Science *266*, 1024–1027. https://doi.org/10.1126/science.7973652.
44. Burton, D.R., Barbas, C.F., Persson, M.A., Koenig, S., Chanock, R.M., and Lerner, R.A. (1991). A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. Proc. Natl. Acad. Sci. *88*, 10134–10137. https://doi.org/10.1073/pnas.88.22.10134.
45. Qin, S., Tang, X., Chen, Y., Chen, K., Fan, N., Xiao, W., Zheng, Q., Li, G., Teng, Y., Wu, M., et al. (2022). mRNA-based therapeutics: powerful and versatile tools to combat diseases. Signal Transduct. Target. Ther. *7*, 1–35. https://doi.org/10.1038/s41392-022-01007-w.
46. Karikó, K., Muramatsu, H., Welsh, F.A., Ludwig, J., Kato, H., Akira, S., and Weissman, D. (2008). Incorporation of Pseudouridine Into mRNA Yields Superior Nonimmunogenic Vector With Increased Translational Capacity and Biological Stability. Mol. Ther. J. Am. Soc. Gene Ther. *16*, 1833–1840. https://doi.org/10.1038/mt.2008.200.
47. Nance, K.D., and Meier, J.L. (2021). Modifications in an Emergency: The Role of N1-Methylpseudouridine in COVID-19 Vaccines. ACS Cent. Sci. *7*, 748–756. https://doi.org/10.1021/acscentsci.1c00197.

48. Andreev, D.E., Loughran, G., Fedorova, A.D., Mikhaylova, M.S., Shatsky, I.N., and Baranov, P.V. (2022). Non-AUG translation initiation in mammals. Genome Biol. *23*, 111. https://doi.org/10.1186/s13059-022-02674-2.
49. Bohlen, J., Fenzl, K., Kramer, G., Bukau, B., and Teleman, A.A. (2020). Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells. Mol. Cell *79*, 561-574.e5. https://doi.org/10.1016/j.molcel.2020.06.005.
50. Du, R., Flynn, M.J., Honsa, M., Jungmann, R., and Elowitz, M.B. (2024). miRNA circuit modules for precise, tunable control of gene expression (Synthetic Biology) https://doi.org/10.1101/2024.03.12.583048.
51. Gao, X.J., Chong, L.S., Kim, M.S., and Elowitz, M.B. (2018). Programmable protein circuits in living cells. Science *361*, 1252–1258. https://doi.org/10.1126/science.aat5062.
52. Hoffmann, M.A.G., Yang, Z., Huey-Tubman, K.E., Cohen, A.A., Gnanapragasam, P.N.P., Nakatomi, L.M., Storm, K.N., Moon, W.J., Lin, P.J.C., West, A.P., et al. (2023). ESCRT recruitment to SARS-CoV-2 spike induces virus-like particles that improve mRNA vaccines. Cell. https://doi.org/10.1016/j.cell.2023.04.024.
53. Cohen, A.A., Gnanapragasam, P.N.P., Lee, Y.E., Hoffman, P.R., Ou, S., Kakutani, L.M., Keeffe, J.R., Wu, H.-J., Howarth, M., West, A.P., et al. (2021). Mosaic nanoparticles elicit cross-reactive immune responses to zoonotic coronaviruses in mice. Science *371*, 735–741. https://doi.org/10.1126/science.abf6840.
54. Van Hoecke, L., and Roose, K. (2019). How mRNA therapeutics are entering the monoclonal antibody field. J. Transl. Med. *17*, 54. https://doi.org/10.1186/s12967-019-1804-8.
55. Pardi, N., Secreto, A.J., Shan, X., Debonera, F., Glover, J., Yi, Y., Muramatsu, H., Ni, H., Mui, B.L., Tam, Y.K., et al. (2017). Administration of nucleoside-modified mRNA encoding broadly neutralizing antibody protects humanized mice from HIV-1 challenge. Nat. Commun. *8*, 14630. https://doi.org/10.1038/ncomms14630.
56. Stadler, C.R., Bähr-Mahmud, H., Celik, L., Hebich, B., Roth, A.S., Roth, R.P., Karikó, K., Türeci, Ö., and Sahin, U. (2017). Elimination of large tumors in mice by mRNA-encoded bispecific antibodies. Nat. Med. *23*, 815–817. https://doi.org/10.1038/nm.4356.
57. Albayrak, C., Jordi, C.A., Zechner, C., Lin, J., Bichsel, C.A., Khammash, M., and Tay, S. (2016). Digital Quantification of Proteins and mRNA in Single Mammalian Cells. Mol. Cell *61*, 914–924. https://doi.org/10.1016/j.molcel.2016.02.030.

Chapter 3

Using Protein Language Models to Domesticate Protein Parts for Polycistronic Circuits

This chapter is temporarily embargoed.

Chapter 4

Identifying Productive Gas Vesicle Stoichiometries using Bayesian Optimization

This chapter is temporarily embargoed.

Chapter 5

Evaluating Cellular Responses to Gas Vesicle Expression in Mammalian Cells

This chapter is temporarily embargoed.

Chapter 6

GHz Frequency Chips for Gas Vesicle Detection and Quantification

Specific Background

In Chapter 4, we evaluated GV acoustic contrast across samples or populations of cells transiently transfected with diverse plasmid mixtures. By measuring the average acoustic response of the population under ultrasound, we were able to draw conclusions on the relative contributions of Gvps to acoustic contrast and GV assembly. A significant portion of the engineering conducted on mARGs in the Shapiro Lab is evaluated by averaging across a population. However, being able to evaluate the acoustic contrast of single cells would increase the throughput with which we engineer and advance genetic circuits that use mARGs. For instance, single cell flow cytometry and cell sorting are routinely implemented for the detection and capture of GFP-expressing cells.[1] Synthetic biologists can tie these reporters to upstream logic and circuitry to report on phenomena of interest. These techniques can be deployed alongside one-pot or pooled experimental strategies, enabling researchers to both test hundreds of designs while capturing individual cells that have a desired response. Co-expression of mARGs with fluorescent protein reporters can be used for sorting cells with mARG circuitry. However, this implementation can lead to enrichment of cells that are solely good at the expression of fluorescent protein reporters and not mARGs. In summary, implementing single-cell acoustic cytometry with cell sorting can unlock screening methodologies of higher throughput, expediting the engineering of gas vesicle expression in mammalian cells as well as the construction of other genetic circuits that use mARGs.

Methods of acoustic cytometry and sorting have traditionally relied on acoustic radiation force (ARF).[2] Simply put, pressure gradients, such as those arising in an acoustic standing wave, can act upon a particle whose density and/or compressibility is distinct from the surrounding medium. Wu et al. demonstrated that cells expressing GVs can experience higher magnitudes of ARF, allowing for manipulation of these cells in physical space using ultrasonic standing waves.[3] The study demonstrated an acoustic cell sorter, where ARF was used to yield a population enriched with GV-expressing cells. Although an exciting proof-of-concept, the enriched population did contain a sizeable fraction of nonexpressers. In Wu's concept, the authors simultaneously use ARF to detect gas vesicles and sort them into bins. An alternative approach could decouple the acoustic measurement from the sorting mechanism. Akin to fluorescence-based cell sorters, future approaches could consider flowing cells through a microfluidic channel that sits on top of an acoustic detector. Cells with an acoustic contrast above a threshold value could then be binned into a separate chamber for downstream collection. The detector we explored for such a purpose was the Geegah GHz ultrasonic imager, which has been used in combination with microfluidic channels to study mixing between liquids with different acoustic impedances.[4]

The Geegah GHz imaging platform is currently employed for continuous multi-modal measurements of the physical, chemical, and biological properties of soil for agricultural use cases.[5–7] Here, we demonstrate that the Geegah GHz imaging platform can detect changes in the acoustic impedance of bacterial colonies due to the presence of gas vesicle contrast agents. We couple the production of these contrast agents to an external nutrient stimulus, further highlighting that the imager may be used to sense gene expression changes in living organisms. This proof-of-concept experiment unlocks future avenues to measure biological properties of soil by coupling the natural sensing capabilities of soil microorganisms with the Geegah GHz imaging platform by way of gas vesicle genetic circuits. Furthermore, this work demonstrates that the

Geegah GHz imaging platform can be used to detect acoustic impedance changes in gas vesicle expressing cells. When integrated with on-chip microfluidics, this imaging platform could be adapted for acoustic cytometry and cell sorting, enabling screening and enrichment of cells expressing mARG circuits.

The following is adapted from a manuscript that was accepted for publication in the proceedings of the IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium. The authors of the manuscript are Ishaan Dev, Anuj Baskota, Dr. Mikhail Shapiro, and Dr. Amit Lal.

Introduction

Soil is a complex environment with large spatial and temporal variability of physical, chemical, and biological properties. The physical properties of temperature, porosity, density, and texture are important to the structural integrity of soil.[8] A well-balanced chemical composition is necessary to maintain useful crop nutrients such as potassium, nitrates, phosphates, and other minerals.[9] Soil also harbors microbes and nematodes that are responsible for decomposing organic matter needed to regulate nutrient cycles.[10] A balance between these physical, chemical, and biological properties yields a healthy soil ecosystem which in turn improves crop growth and production.[11,12]

Over the last several decades, protocols for understanding these soil properties have been developed to improve and optimize farming practices. Currently, several chemical and biological measurements are performed in laboratories using spectroscopy techniques, chromatography, DNA sequencing, and other biological and chemical assays.[13] However, these techniques can be time-consuming and labor intensive, as they involve collecting soil samples from the field and shipping them to the respective analysis labs. With advancements in precision farming, internet-of-things (IoT) based technologies employed in farms have enabled remote sensing of measurements related to soil, air, and crops. While this has certainly improved efficiency of data acquisition across several soil sensors, most of these are still a single-modality measurement—each sensor measures a single soil parameter.[14] There is lack of integrated multi-modal sensing platforms that can sense across physical, chemical, and biological components of soil. Combining multiple sensing units with their individual packaging and interface into a single module can be bulky and expensive. Ideally, a cost-effective and lightweight multi-modal sensing platform should also be capable of continuous monitoring of soil parameters over long time periods.[15] This would allow farmers to implement real-time interventions, such as responding to pests, unfavorable weather conditions, and crop diseases.

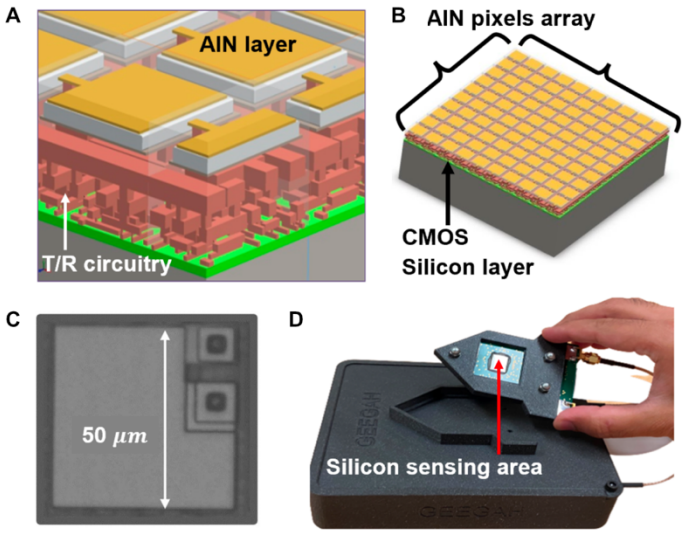

Figure 6-1: GHz chip schematic.
A) Schematic of AlN thin film deposited in Transmit/Receive (T/R) circuitry. **B**). An array of AlN transducers with integrated CMOS and silicon layer stacked on the backside. **C**). Individual pixel with T/R circuitry. **D**). Geegah imager unit where the silicon surface is the active sensing region.

In our previous works, we have shown the use of a GHz ultrasonic imager to measure soil moisture, morphology and temperature as well as to image soil nematodes.[6,7] The Geegah imager used here contains a 128 x 128 aluminum nitride (AlN) transducer array, with 50 µm pitch pixels (**Figures 6-1A, 6-1B, 6-1C, 6-1D**). The thin AlN film is integrated with 130 nm CMOS electronics and is stacked with a silicon layer on the backside. The transducers transmit ultrasonic waves at GHz frequency to the backside of the silicon where the sample is placed. The reflection of acoustic waves from this silicon/sample interface is received by the same transducers to obtain 128 x 128 matrix array of echoes. The compartmentalization of 128 x 128 transducer pixels allows for multi-modal sensing of different soil parameters simultaneously. The moisture was estimated by the ratio of pixels exposed to water with those exposed to dry regions. Temperature was measured by calculating the unwrapped phase of the return echoes from pixels that were not in direct contact with soil. The morphology of the soil and the organisms within it were visualized directly from the 2D maps of the magnitude data obtained at 13 fps. In addition, the same GHz ultrasonic imaging platform has been used to characterize the acoustic impedance of nematodes,[5] animal tissues,[16] and onion cells.[17] The ultrasonic pulses at GHz frequency has ~ 4.5 µm wavelength in silicon resulting in higher resolution images suitable for imaging large cells and tissues. In addition, the integration of the imager with GPS and RF systems can enable wireless communication between the device and a central server or node. This feature eliminates the need for soil sampling across various points and directly conveys important alerts to farmers for quick intervention.

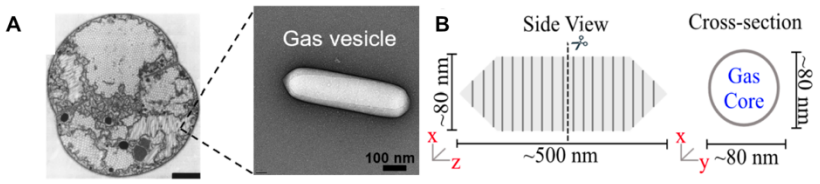

Figure 6-2. Gas vesicle schematics.
A) Image of a single-cell microorganism expressing GVs throughout the cell (Adapted from Walsby et al.).[18] An electron micrograph of a single GV is depicted as well. **B**) General schematic of an assembled gas vesicle. A GV has a proteinaceous shell and a gas core.

In this work, we demonstrate that the Geegah imager can also take measurements of biological properties by detecting gas vesicle (GV) acoustic contrast agents in bacteria in response to an environmental stimulus. Gas vesicles occur naturally within aquatic microbes and are composed of a gas-filled proteinaceous shell (**Figure 6-2**).[18,19] The gaseous contents of the balloon-like particles create an acoustic impedance mismatch with the surrounding aquatic environment of the cell cytosol.[20,21] Akin to microbubbles, the gas-liquid interface created by its protein shell enables imaging of these GVs and the organisms that contain them using ultrasound (US).[22] Importantly, the genes that encode GVs can be ported over to other microorganisms and mammalian cells using genetic engineering techniques, allowing these cellular systems to express or produce these ultrasonic contrast agents on command or in response to environmental stimuli. Previously, GVs expressed within bacteria have enabled 5-15 MHz non-invasive ultrasound imaging of bacterial colonization and inflammation sensing in the gut microbiome of living mice.[23,24] Furthermore, GVs expressed in tumor cells can be used to image and track tumor growth within living, complex organisms.[25,26] The integration of the Geegah imager with GV-producing microorganisms holds the potential to transform agricultural practices by paving the way for the creation of dynamic soil sensors. Such sensors would utilize genetically engineered bacteria or soil microorganisms that produce GVs in response to soil pests, nutrient limitations, or other environmental signals, leading to an acoustic impedance change detectable by a soil-embedded GHz ultrasonic chip.

As a proof-of-concept of these capabilities, we used the Geegah imager to observe changes in the acoustic impedance of bacteria before and after the expression of GV proteins in response to arabinose sugars (**Figure 6-3**).

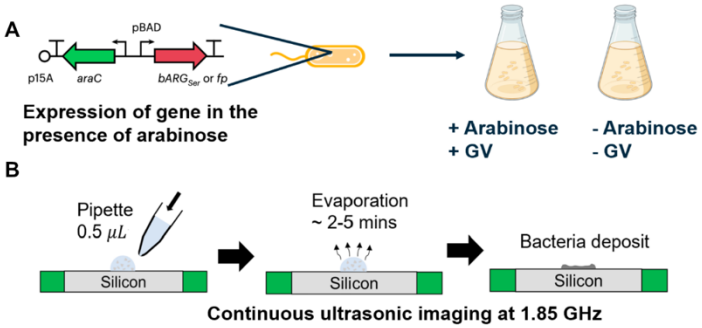

Figure 6-3. Experimental workflow for GHz imaging of GVs.
A) Culturing of *E. coli* expressing GV genes in response to arabinose in its environment. **B**) Experimental protocol showing deposition of a small volume of the cell solution on the GHz ultrasonic imager. Imaging was conducted throughout the evaporation cycle of the solvent.

Results

As described in Methods, bacterial cells grown with arabinose produce gas vesicles (GV+ cells) while those grown in the absence of arabinose do not (GV- cells). Droplets from these two cell populations were pipetted onto the surface of the imager and allowed to evaporate. Images at 1.85 GHz were continuously taken. We evaluated the acoustic impedance (Z) of the two cell samples both in the liquid phase and after drying (**Figure 6-4**).

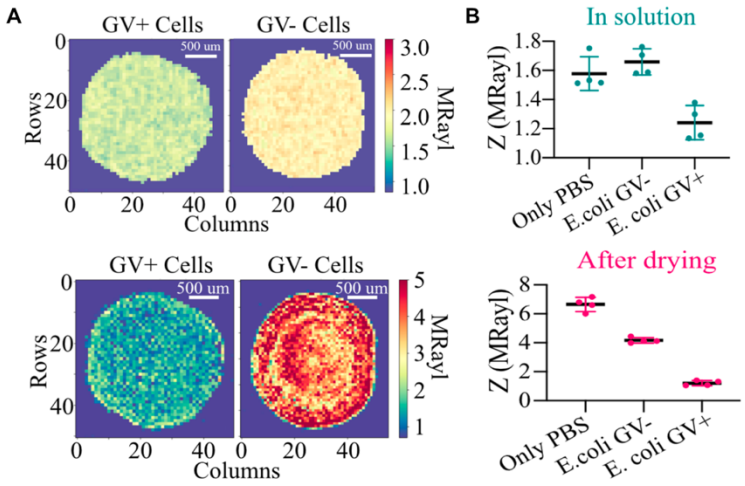

Figure 6-4: Acoustic impedance changes due to gas vesicles as measured at 1.85 GHz.
A) Acoustic impedance 2D maps showing larger reflection from bacteria culture droplets with GV both in solution (top) and dried states (bottom). **B**) Plotted are the mean acoustic impedances in the ROI measured for the phosphate-buffered saline (PBS) control droplets, *E. coli* with GV droplets, and *E. coli* without GV droplets for conditions before (top) and after the droplet dried (bottom). N=4. Error bars depict SEM.

The 2D maps show lower Z values for GV+ cells compared to the GV- cells in the solution phase (**Figure 6-4A**). The average Z measured for cells without GVs was 1.66 ± 0.08 *MRayls* which is consistent with Z measured for various cells by other sources.[27,28] The cells with GVs had an average Z of 1.24 ± 0.10 *MRayls* which demonstrates that gas vesicles increase the reflected signal of the acoustic waves

After evaporation, the GV+ cells still exhibit lower Z compared to GV- cells, suggesting gas vesicles remain intact in the cells even after complete evaporation of solution (**Figure 6-4B**). The Z for GV- cells increases during the drying process, as the dehydration likely causes an increase in cell density and stiffness. In addition, the high Z observed from the PBS control after drying was due to the formation of salt crystals on the imager surface. Therefore, salt crystals are expected to form amongst the cells after the drying of the solution, leading to increased Z values.

Discussion

This proof-of-concept experiment demonstrates the Geegah imager can measure gene expression events and GV contrast agents within bacteria caused by an environmental nutrient. To build upon the capabilities of such dynamic, multi-modal soil sensors, future works should implement GV genetic circuits in other important soil microorganisms, including nematodes and prominent soil bacteria. Furthermore, researchers should create genetic logic gates in these soil microorganisms that couple GV production to environmental stimuli sensed by these sentinel microbes. Combined with microfluidics,[4] the Geegah GHz imager may also be a viable option for detecting and capturing cells that have gas vesicles. Such a cell sorter could be used for high throughput screening of gene expression circuits that produce gas vesicles with improved properties.[29] However, the main drawback of the concept is the 1-2 GHz frequencies the imager operates at. This is much higher than the 5-15 MHz ultrasound that is generally used for biomedicine, one of the major application spaces for gas vesicle expression systems.[30–32] Therefore, this imager may be beneficial for general gas vesicle detection, but in its current form cannot measure important GV properties, like non-destructive buckling, at clinically relevant frequencies.[33]

Methods

Cell Culture: As described by the experiments of Hurt, Buss, and Duan [25], BL21(DE3) *Escherichia coli* (NEB) cells harboring an arabinose-inducible GV expression plasmid (pBAD-bARGSer-AxeTxe Addgene: #192473) were cultured in Luria Broth (LB) with 0.6% (v/v) glycerol, 0.3% (w/v) d-glucose, 0.5% l-arabinose (w/v), and 25 µg/mL chloramphenicol (**Figure 6-3A**). 25 mL cultures were grown for 24 hours in 250 mL baffled flasks at 30 ℃ in a shaking incubator. An additional flask of these cells was cultured similarly except l-arabinose was excluded, leading to no production of gas vesicles. The GV-producing (GV+) and non-producing (GV-) cultures were then centrifuged at 300 RCF to generate floating and sedimented cell layers respectively. Culture medium was separated from the cells. Cells were then washed with ice-cold phosphate-buffered saline (PBS) and finally concentrated into 1 mL of PBS using centrifugation at 300 RCF.

GHz imaging of cells: A volume of 0.5 µL of cells in PBS solution was pipetted onto the silicon surface of the imager. Imaging was performed at 1.85 GHz. The cell concentration in the droplet was estimated to have an OD600 of 75. The entire drying process until the solvent fully evaporated was imaged (**Figure 6-3B**). The *E. coli* with GVs and without GVs (N = 4) were imaged separately at an acquisition rate of 7 fps. The pulse signal received from the 128 x 128 transducers was demodulated into In-phase (I) and Out-of-phase (Q) components. A time point at the first reflection echo from the cells (I_echo and Q_echo) was captured along with a point after the echoes died off (I_noecho and Q_noecho). These no-echo signals were used to remove any DC offsets caused due to individual pixels T/R circuitry in the received signals. We derived the acoustic impedance (Z_cells) from these raw I and Q signals as shown in our previous work [16]. A median 2x2 kernel filter was applied to the 128 x 128 2D maps of Z_cells to decrease noise in the image. A mask of the cell droplet was then created by performing binary thresholding which was used to track the pixels exposed to the cell solution only. This ROI was then used to calculate the average Z_cells before and after evaporation of the solvent.

References

1. Papa, S., Ortolani, C., Fernández, P., and O'Connor, J.-E. (2023). Flow Cytometry and Its Applications to Molecular Biology and Diagnosis 2.0. Int. J. Mol. Sci. *24*, 16215. https://doi.org/10.3390/ijms242216215.
2. Ozcelik, A., Rufo, J., Guo, F., Gu, Y., Li, P., Lata, J., and Huang, T.J. (2018). Acoustic tweezers for the life sciences. Nat. Methods *15*, 1021–1028. https://doi.org/10.1038/s41592-018-0222-9.
3. Wu, D., Baresch, D., Cook, C., Ma, Z., Duan, M., Malounda, D., Maresca, D., Abundo, M.P., Lee, J., Shivaei, S., et al. (2023). Biomolecular actuators for genetically selective acoustic manipulation of cells. Sci. Adv. *9*, eadd9186. https://doi.org/10.1126/sciadv.add9186.
4. Baskota, A., Kuo, J., Ardanuç, S., and Lal, A. (2023). Visualization of Fluid Mixing Using GHz Ultrasonic Imaging. In 2023 22nd International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers), pp. 1109–1112.
5. Baskota, A., Kuo, J., and Lal, A. (2022). Real-time GHz Ultrasonic Imaging of Nematodes at Microscopic Resolution. Microsc. Microanal. *28*, 1594–1596. https://doi.org/10.1017/S1431927622006389.
6. Baskota, A., Kuo, J., and Lal, A. (2022). Gigahertz Ultrasonic Multi-Imaging of Soil Temperature, Morphology, Moisture, and Nematodes. In 2022 IEEE 35th International Conference on Micro Electro Mechanical Systems Conference (MEMS), pp. 519–522. https://doi.org/10.1109/MEMS51670.2022.9699813.
7. Kuo, J., Baskota, A., Zimmerman, S., Hay, F., Pethybridge, S., and Lal, A. (2021). Gigahertz Ultrasonic Imaging of Nematodes in Liquids, Soil, and Air. In 2021 IEEE International Ultrasonics Symposium (IUS), pp. 1–4. https://doi.org/10.1109/IUS52206.2021.9593762.
8. Lipiec, J., and Hatano, R. (2003). Quantification of compaction effects on soil physical properties and crop growth. Geoderma *116*, 107–136. https://doi.org/10.1016/S0016-7061(03)00097-1.
9. Regelink, I.C., Stoof, C.R., Rousseva, S., Weng, L., Lair, G.J., Kram, P., Nikolaidis, N.P., Kercheva, M., Banwart, S., and Comans, R.N.J. (2015). Linkages between aggregate formation, porosity and soil chemical properties. Geoderma *247–248*, 24–37. https://doi.org/10.1016/j.geoderma.2015.01.022.
10. Martin, J.P., and Focht, D.D. (1977). Biological Properties of Soils. In Soils for Management of Organic Wastes and Waste Waters (John Wiley & Sons, Ltd), pp. 113–169. https://doi.org/10.2134/1977.soilsformanagementoforganic.c6.
11. Delgado, A., and Gómez, J.A. (2016). The Soil. Physical, Chemical and Biological Properties. In Principles of Agronomy for Sustainable Agriculture, F. J. Villalobos and E. Fereres, eds. (Springer International Publishing), pp. 15–26. https://doi.org/10.1007/978-3-319-46116-8_2.
12. Banerjee, S., and van der Heijden, M.G.A. (2023). Soil microbiomes and one health. Nat. Rev. Microbiol. *21*, 6–20. https://doi.org/10.1038/s41579-022-00779-w.
13. Wilhelm, R.C., van Es, H.M., and Buckley, D.H. (2022). Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biol. Biochem. *164*, 108472. https://doi.org/10.1016/j.soilbio.2021.108472.
14. Karmakar, P., Teng, S.W., Murshed, M., Pang, S., Li, Y., and Lin, H. (2024). Crop monitoring by multimodal remote sensing: A review. Remote Sens. Appl. Soc. Environ. *33*, 101093. https://doi.org/10.1016/j.rsase.2023.101093.
15. Fan, Y., Wang, X., Funk, T., Rashid, I., Herman, B., Bompoti, N., Mahmud, M.S., Chrysochoou, M., Yang, M., Vadas, T.M., et al. (2022). A Critical Review for Real-Time Continuous Soil Monitoring: Advantages, Challenges, and Perspectives. Environ. Sci. Technol. *56*, 13546–13564. https://doi.org/10.1021/acs.est.2c03562.

16. Baskota, A., Kuo, J., Ardanuç, S., and Lal, A. (2023). Characterization of cells and tissues using a compact GHz ultrasonic imager. In 2023 IEEE International Ultrasonics Symposium (IUS), pp. 1–4. https://doi.org/10.1109/IUS51837.2023.10307657.
17. Baskota, A., Kuo, J., Ardanuç, S., and Lal, A. (2023). Compact GHz Ultrasonic Micro-Imager for Cells and Tissues. Microsc. Microanal. *29*, 1116–1117. https://doi.org/10.1093/micmic/ozad067.572.
18. Walsby, A.E. (1972). Structure and function of gas vacuoles. Bacteriol. Rev. *36*, 1–32.
19. Pfeifer, F. (2012). Distribution, formation and regulation of gas vesicles. Nat. Rev. Microbiol. *10*, 705–715. https://doi.org/10.1038/nrmicro2834.
20. Shapiro, M.G., Goodwill, P.W., Neogy, A., Yin, M., Foster, F.S., Schaffer, D.V., and Conolly, S.M. (2014). Biogenic gas nanostructures as ultrasonic molecular reporters. Nat. Nanotechnol. *9*, 311–316. https://doi.org/10.1038/nnano.2014.32.
21. Lakshmanan, A., Lu, G.J., Farhadi, A., Nety, S.P., Kunth, M., Lee-Gosselin, A., Maresca, D., Bourdeau, R.W., Yin, M., Yan, J., et al. (2017). Preparation of biogenic gas vesicle nanostructures for use as contrast agents for ultrasound and MRI. Nat. Protoc. *12*, 2050–2080. https://doi.org/10.1038/nprot.2017.081.
22. Maresca, D., Lakshmanan, A., Lee-Gosselin, A., Melis, J.M., Ni, Y.-L., Bourdeau, R.W., Kochmann, D.M., and Shapiro, M.G. (2017). Nonlinear ultrasound imaging of nanoscale acoustic biomolecules. Appl. Phys. Lett. *110*, 073704. https://doi.org/10.1063/1.4976105.
23. Bourdeau, R.W., Lee-Gosselin, A., Lakshmanan, A., Farhadi, A., Kumar, S.R., Nety, S.P., and Shapiro, M.G. (2018). Acoustic reporter genes for noninvasive imaging of microbes in mammalian hosts. Nature *553*, 86–90. https://doi.org/10.1038/nature25021.
24. Buss, M.T., Zhu, L., Kwon, J.H., Tabor, J.J., and Shapiro, M.G. (2024). Probiotic acoustic biosensors for noninvasive imaging of gut inflammation. Preprint at bioRxiv, https://doi.org/10.1101/2024.09.23.614598 https://doi.org/10.1101/2024.09.23.614598.
25. Hurt, R.C., Buss, M.T., Duan, M., Wong, K., You, M.Y., Sawyer, D.P., Swift, M.B., Dutka, P., Barturen-Larrea, P., Mittelstein, D.R., et al. (2023). Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. Nat. Biotechnol., 1–13. https://doi.org/10.1038/s41587-022-01581-y.
26. Bar-Zion, A., Nourmahnad, A., Mittelstein, D.R., Shivaei, S., Yoo, S., Buss, M.T., Hurt, R.C., Malounda, D., Abedi, M.H., Lee-Gosselin, A., et al. (2021). Acoustically triggered mechanotherapy using genetically encoded gas vesicles. Nat. Nanotechnol. *16*, 1403–1412. https://doi.org/10.1038/s41565-021-00971-8.
27. Fadhel, M.N., Berndl, E.S.L., Strohm, E.M., and Kolios, M.C. (2015). High-Frequency Acoustic Impedance Imaging of Cancer Cells. Ultrasound Med. Biol. *41*, 2700–2713. https://doi.org/10.1016/j.ultrasmedbio.2015.06.003.
28. Hozumi, N., Yoshida, S., and Kobayashi, K. (2019). Three-dimensional acoustic impedance mapping of cultured biological cells. Ultrasonics *99*, 105966. https://doi.org/10.1016/j.ultras.2019.105966.
29. Hurt, R.C., Jin, Z., Soufi, M., Wong, K.K., Sawyer, D.P., Shen, H.K., Dutka, P., Deshpande, R., Zhang, R., Mittelstein, D.R., et al. (2024). Directed Evolution of Acoustic Reporter Genes Using High-Throughput Acoustic Screening. ACS Synth. Biol. *13*, 2215–2226. https://doi.org/10.1021/acssynbio.4c00283.
30. Shivaei, S., Liu, A., Abedi, M.H., Revilla, J., Hurvitz, I.U., Swift, M.B., and Shapiro, M.G. (2024). Non-invasive imaging of cell-based therapies using acoustic reporter genes. Preprint at bioRxiv, https://doi.org/10.1101/2024.11.01.621111 https://doi.org/10.1101/2024.11.01.621111.
31. Shivaei, S., Cheung, K.Y.M., Yadav, A., Hurvitz, I.U., Lee, S., Revilla, J., Rabut, C., Criado-Hidalgo, E., Zhang, R.J., and Shapiro, M.G. (2025). Ultrasound imaging of in situ transcriptional activity in opaque tissue. Preprint at bioRxiv, https://doi.org/10.1101/2025.07.06.663365 https://doi.org/10.1101/2025.07.06.663365.

32. Feng, R., Lan, J., Goh, M.C., Du, M., and Chen, Z. (2024). Advances in the application of gas vesicles in medical imaging and disease treatment. J. Biol. Eng. *18*, 41. https://doi.org/10.1186/s13036-024-00426-3.
33. Salahshoor, H., Yao, Y., Dutka, P., Nyström, N.N., Jin, Z., Min, E., Malounda, D., Jensen, G.J., Ortiz, M., and Shapiro, M.G. (2022). Geometric effects in gas vesicle buckling under ultrasound. Biophys. J. *121*, 4221–4228. https://doi.org/10.1016/j.bpj.2022.09.004.

Supplementary Information 1

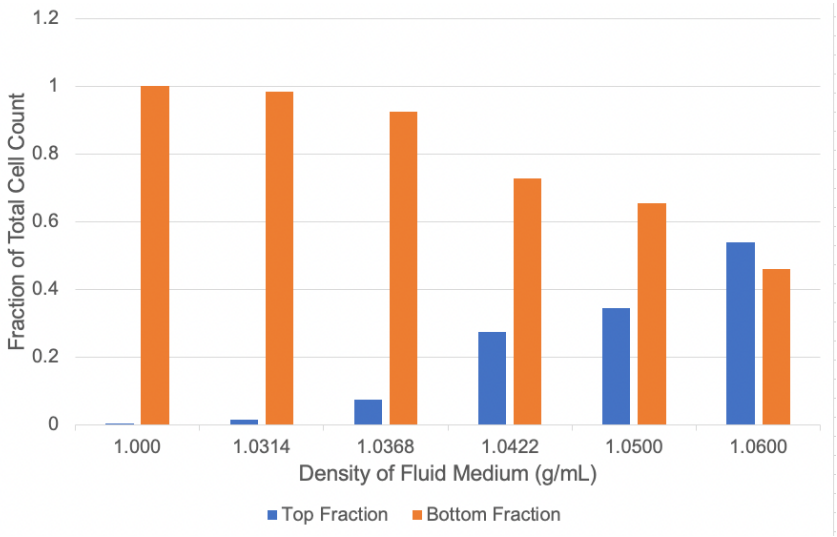

Figure S1-1: Buoyancy enrichment tests demonstrate variability of gas volumes in cellulo.
By changing the density of a medium containing GV-expressing MDA cells, we can enrich the surface (top) fraction of the medium with cells producing high amounts of gas vesicles. These cells also produce greater nonlinear acoustic contrast under ultrasound. By varying the density of the medium, varying amounts of cells can be collected from the top fraction, indicating that there is a distribution of GV abundances (and gas volumes) within these cells. We attribute these differences to the polyclonal nature of the population. This distribution could also be due to cell division, where daughter cells split the formed GVs across their cell bodies. The unpublished data below was provided by Pierina Barturen of the Shapiro Lab. (N=1)

Supplementary Information 2

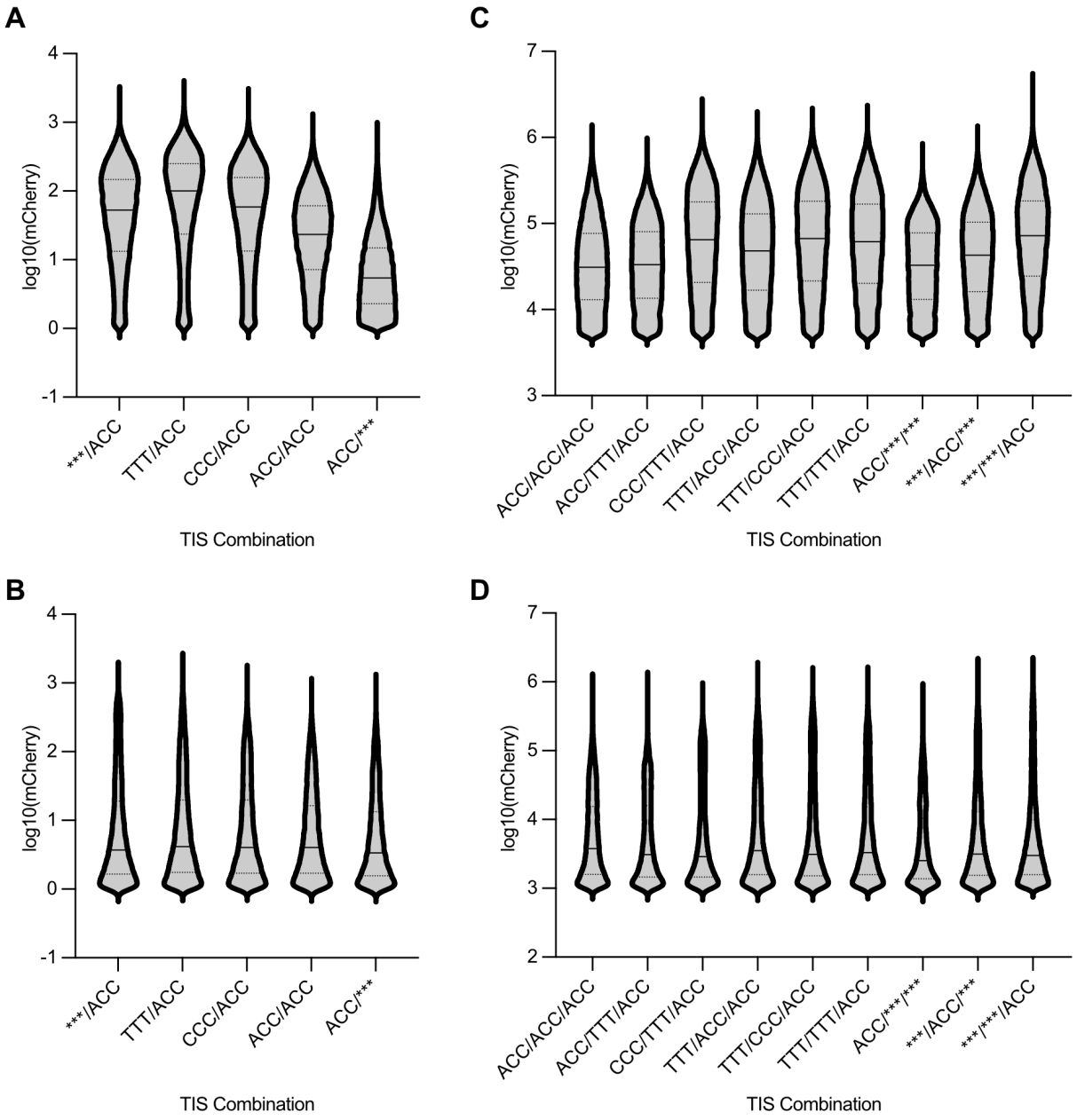

Figure S2-1: mCherry distribution plots for 2-ORF and 3-ORF fluorescent protein SEMPER libraries.
Acting at the level of each mRNA, this IRES mCherry normalization strategy aimed to account for variability in transfection and transcription that could occur between tested constructs. **A**) mCherry distributions for 2-ORF SEMPER constructs (for expression of msfGFP[r5M] and mEBFP2) after transfection in HEK293T cells. **B**) mCherry distributions for 2-ORF SEMPER constructs (for expression of msfGFP[r5M] and mEBFP2) after transfection in CHO-K1 cells. **C**) mCherry distributions for 3-ORF SEMPER constructs (for expression of msfBFP[r5M], msfGFP[r5M], and emiRFP670) after transient transfection in HEK293T cells. **D**) mCherry distributions for 3-ORF SEMPER constructs (for expression of msfBFP[r5M], msfGFP[r5M], and emiRFP670) after transient transfection in CHO-K1 cells. mCherry distributions for CHO-K1 cells vary less than those for HEK293T, suggesting differences in transfection efficiency and/or burden. However, despite this variation, the rank orders of ORF expression ratios and other trends of mCherry-normalized expression titration were consistent between the two cell types in **Figures 2-1, 2-2, S2-5, and S2-9** for various TIS combinations tested. Each violin plot contains data from four combined replicates of the depicted TIS combination. The median and quartiles of the distribution are represented by the solid and dotted lines respectively.

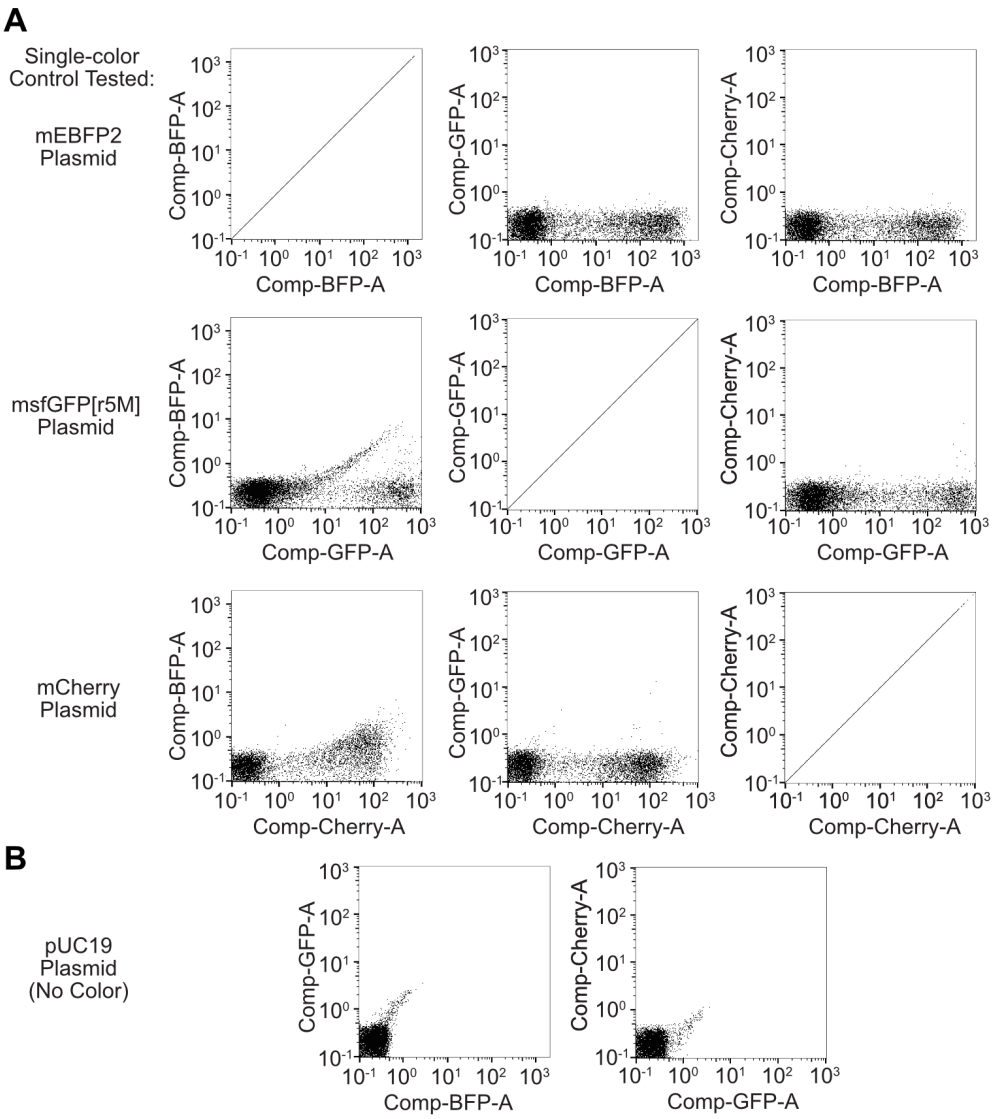

Figure S2-2: HEK293T 2-ORF compensation controls.
A) Flow cytometry plots for single-color control plasmids after transient transfection and compensation. Compensation was conducted using FlowJo. Depicted are cells that have been gated for by size and then doublet-discriminated. Comp-BFP-A, Comp-GFP-A, and Comp-Cherry-A, represent the channels used for mEBFP2, msfGFP[r5M], and mCherry respectively. mCherry and msfGFP[r5M] single-color controls showed some, yet not substantial, fluorescence crossover into the Comp-BFP-A channel. Measurements were taken using a MACSQuant VYB cytometer. **B**) Flow cytometry plots for pUC19 plasmid transfections containing no fluorescent marker. Minimal fluorescence was observed in all channels.

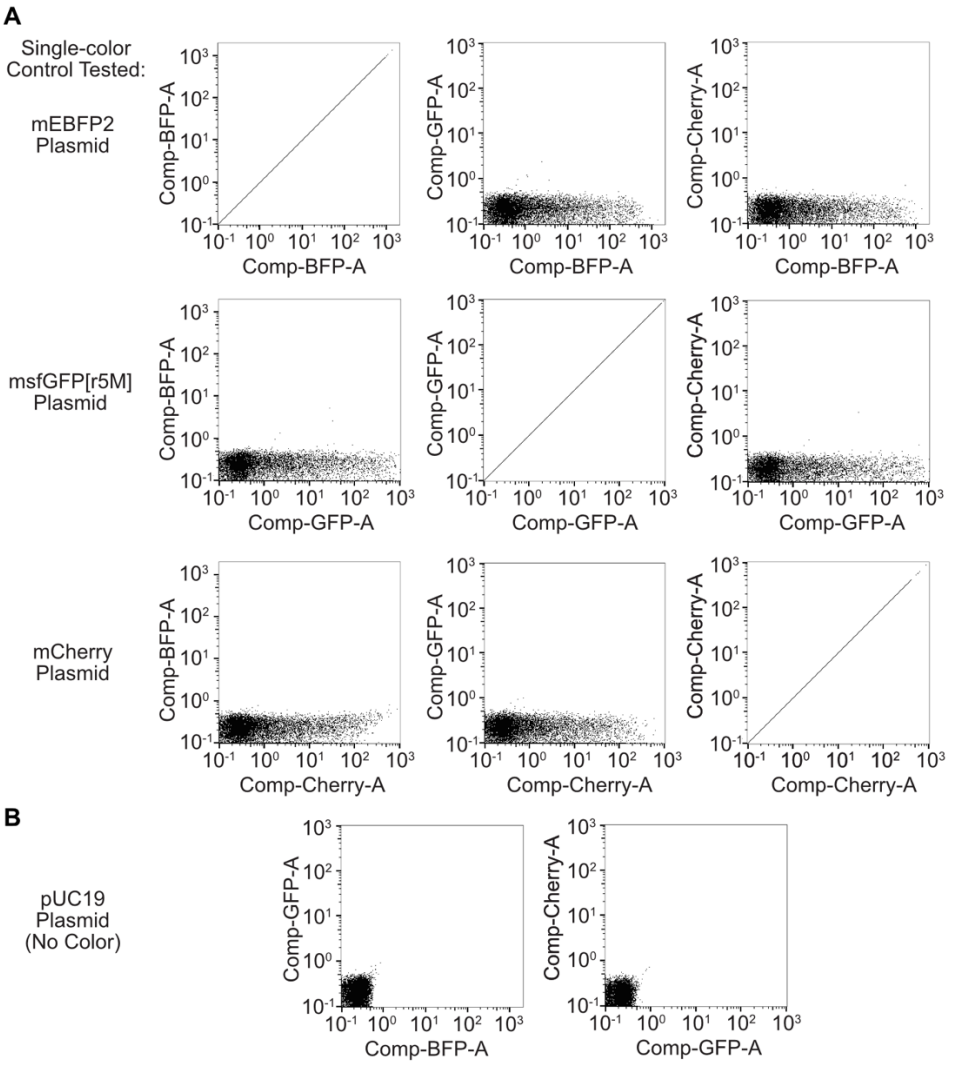

Figure S2-3: CHO-K1 2-ORF compensation controls.
A) Flow cytometry plots for single-color control plasmids after transient transfection and compensation. Compensation was conducted using FlowJo. Depicted are cells that have been gated for by size and then doublet-discriminated. Comp-BFP-A, Comp-GFP-A, and Comp-Cherry-A, represent the channels used for EBFP2, msfGFP[r5M], and mCherry respectively. Minimal to no fluorescence spillover was observed in the Comp-BFP-A channel for cells transfected with msfGFP[r5M] or mCherry single-color controls. Measurements were taken using a MACSQuant VYB cytometer. **B**) Flow cytometry plots for pUC19 plasmid transfections containing no fluorescent marker.

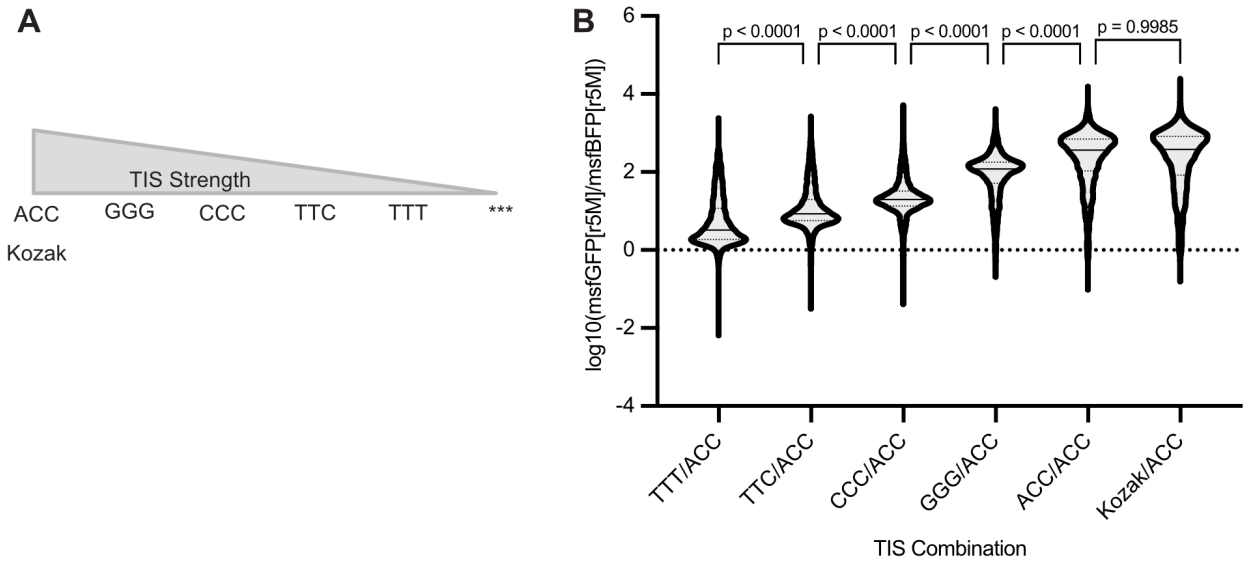

Figure S2-4: Additional TIS combinations for 2-ORF SEMPER.
A) In HEK293T cells**,** we tested three additional TIS sequences (TTC, GGG, and Kozak) in front of ORF 1 while maintaining a strong ACC TIS on ORF 2. ORF 1 encoded msfGFP[r5M] and ORF 2 encoded msfBFP[r5M]. IRES-mCherry was included on the plasmid as done in our other designs. **B)** Violin plots of log10(msfGFP[r5M]/msfBFP[r5M]) values for all mCherry positive cells for four combined replicates of the depicted TIS combinations. The median and quartiles of the distribution are represented by the solid and dotted lines respectively. TTC and GGG provide additional tunability of the system, accessing new expression stoichiometries. ACC TIS used for ORF 1 performs similarly to the Kozak sequence (GCCGCCACCATGG) when implemented for ORF 1. The data shown here were collected using the transfection techniques described in Methods for plasmid transfection of 2-ORF SEMPER fluorescent protein constructs, except these experiments were conducted in 96-well format and therefore transfection reagents and DNA quantities were scaled down by a factor of 4 to accommodate. The Games-Howell's multiple comparisons tests depicted above were conducted as a post-hoc analysis of Welch's ANOVA Test ($p < 0.0001$).

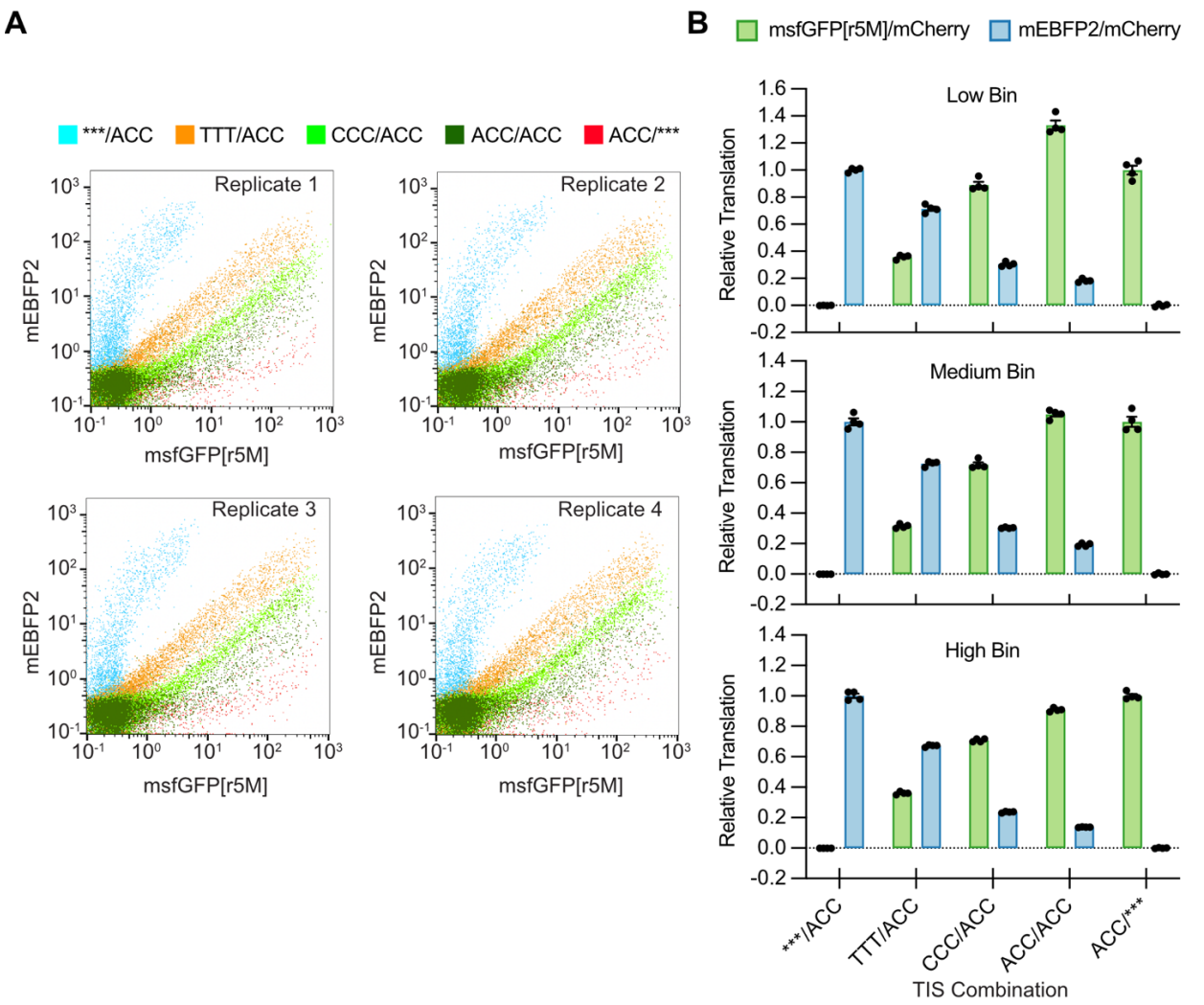

Figure S2-5: CHO-K1 plasmid-based 2-ORF SEMPER results.
A) Flow cytometry data for CHO-K1 cells transfected with 2-ORF SEMPER constructs. mCherry positive cells are depicted. Transfection conditions were optimized for HEK293T cells and therefore lower transfection efficiency was observed for CHO-K1 cells. Replicates for both cell lines demonstrated reproducibility, as shown here for CHO-K1. **B**) Min-max normalized fluorescence values for each GOI relative to ACC/*** and ***/ACC. Fluorescence measurements were first normalized by mCherry before min-max normalization. Error bars depict SEM (N=4). Welch's ANOVA tests were conducted for blue and green bars independently ($p < 0.0001$ for all colors across all bins) followed by Dunnett's T3 multiple comparisons tests. All pairwise comparisons are provided in **Table S2-3**.

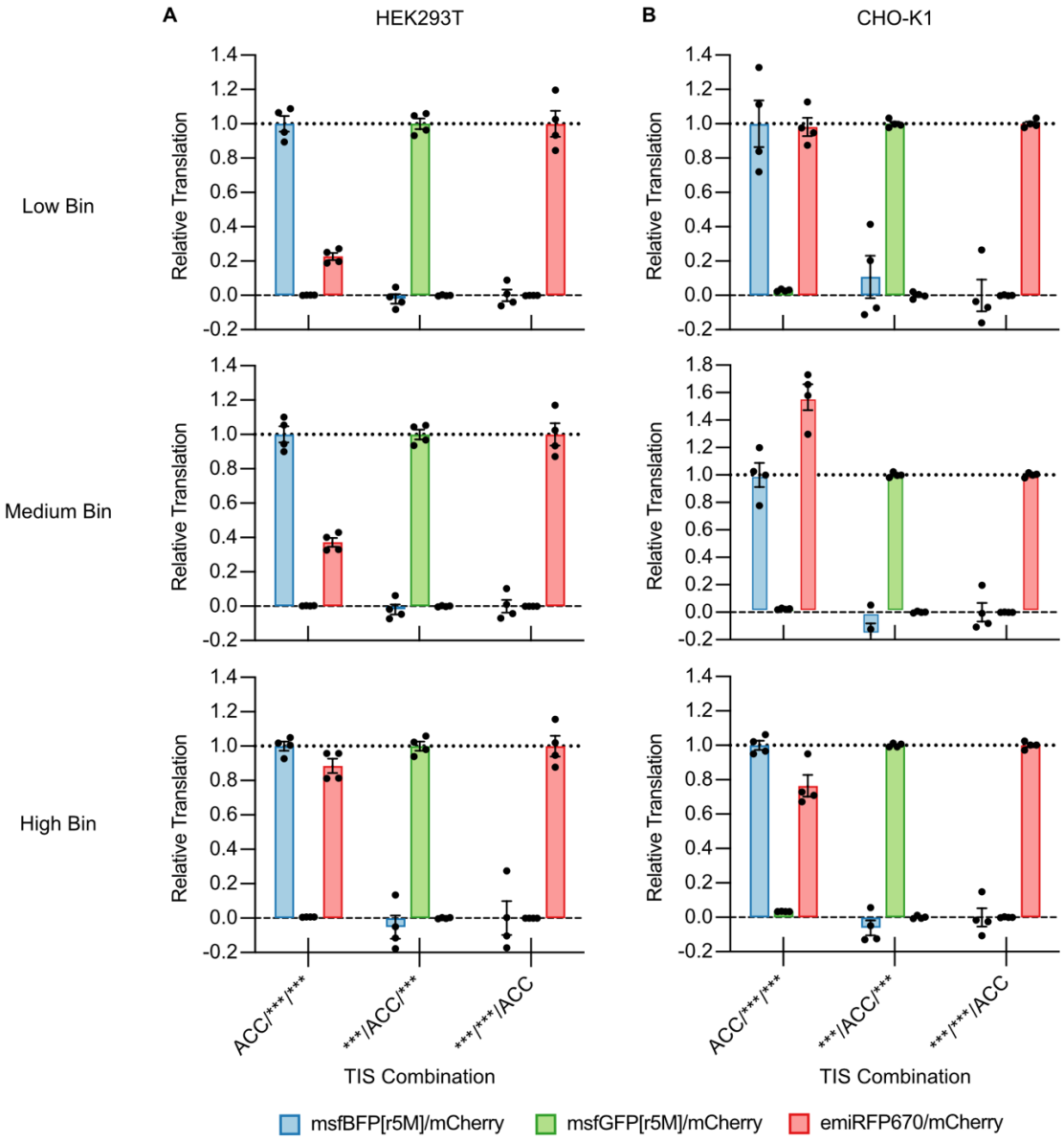

Figure S2-6: Min-max controls for 3-ORF SEMPER library.
A) HEK293T min-max controls after mCherry normalization and min-max normalization. Error bars depict standard error of the mean (SEM) (N=4). **B**) CHO-K1 min-max controls after mCherry normalization and min-max normalization. Error bars depict standard error of the mean (SEM) (N=4). For both cell types, min-max normalization for msfBFP[r5M] was conducted using the fluorescence values in the BFP fluorescence channel for ACC/***/*** (max) and ***/***/ACC (min). Min-max normalization for msfGFP[r5M] was conducted using the fluorescence values in the GFP fluorescence channel for ***/ACC/*** (max) and ***/***/ACC (min). Min-max normalization for emiRFP670 was conducted using the fluorescence values in the iRFP fluorescence channel for ***/***/ACC (max) and ***/ACC/*** (min). emiRFP670 demonstrated significant fluorescence even when no start codon was present in front of it. We have observed that constructs with no start codons in front of FPs can still produce fluorescence above what was measured for no-color plasmid controls and mCherry-only controls. This suggests unexpected fluorescent spillover is not the cause. Therefore, we hypothesize these results are due to the translation of fluorescent protein products from in-frame, non-canonical start codons or from in-frame, internal TISs.[1] In addition, the expression of emiRFP670 from non-canonical or alternate translation initiation sites appeared to be improved when an actively translating ORF was placed further upstream (i.e. ACC/***/***) compared to the case where the actively translating ORF was placed closer to emiRFP670 (i.e. ***/ACC/***).

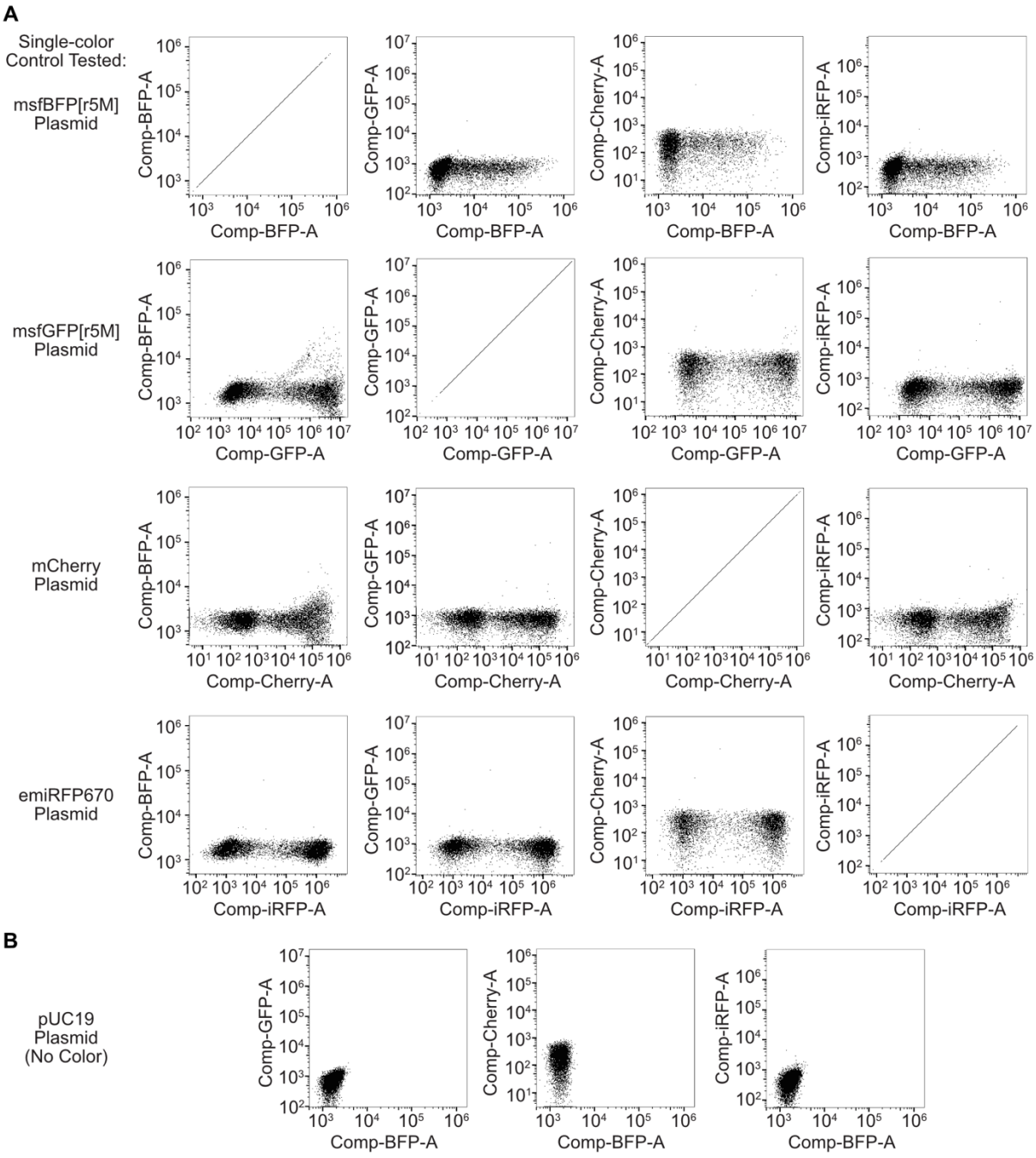

Figure S2-7: HEK293T 3-ORF compensation controls.
A) Flow cytometry plots for single-color control plasmids after transient transfection and compensation. Compensation was conducted using FlowJo. Depicted are cells that have been gated for by size and then doublet-discriminated. Comp-BFP-A, Comp-GFP-A, Comp-iRFP-A, and Comp-Cherry-A represent the channels used for msfBFP[r5M], msfGFP[r5M], emiRFP670 and mCherry respectively. No substantial fluorescence spillover was observed. Measurements were taken using a Cytoflex S cytometer. **B**) Flow cytometry plots for pUC19 plasmid transfections containing no fluorescent marker.

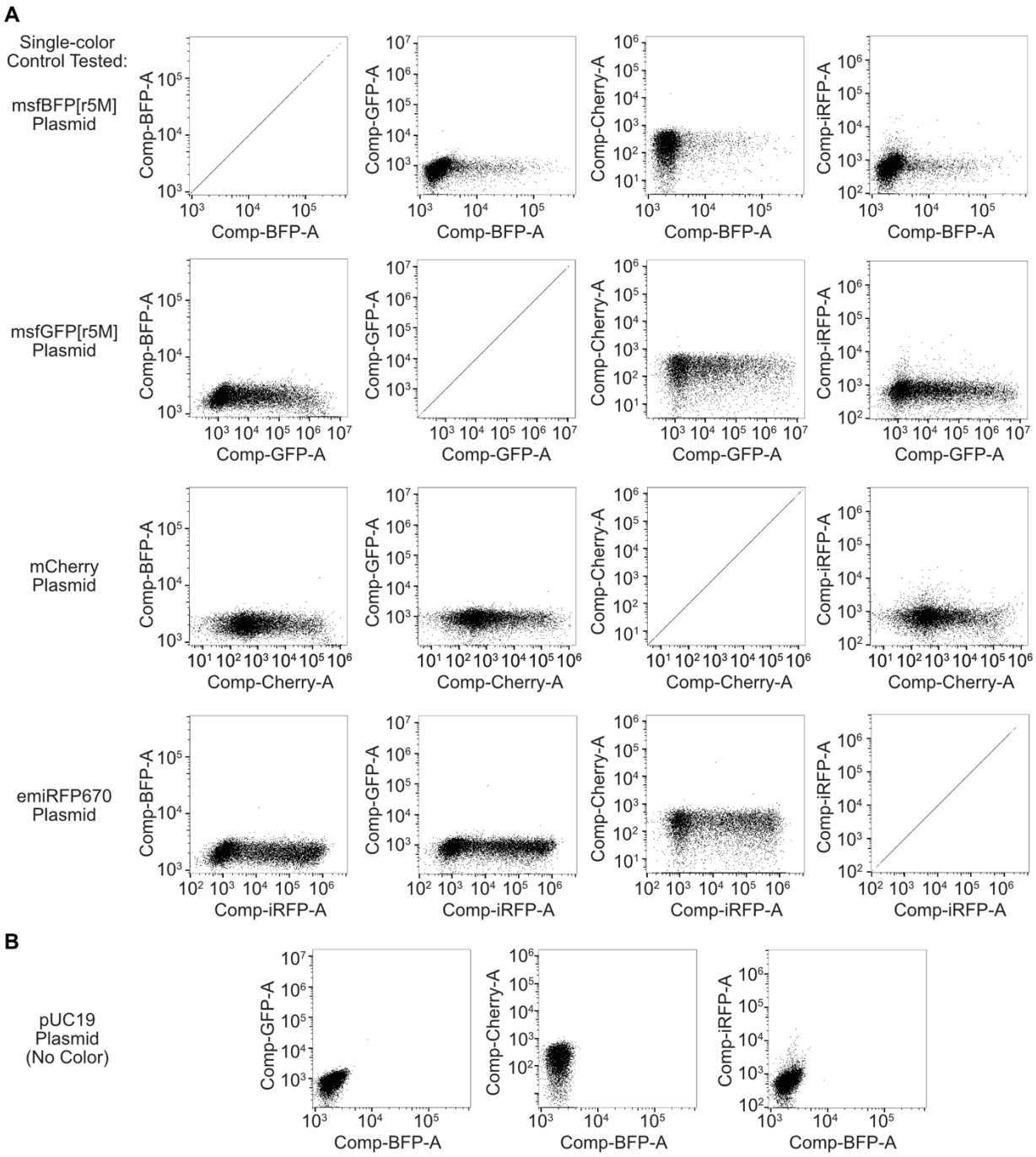

Figure S2-8: CHO-K1 3-ORF Compensation Controls.
A) Flow cytometry plots for single-color control plasmids after transient transfection and compensation. Compensation was conducted using FlowJo. Depicted are cells that have been gated for by size and then doublet-discriminated. Comp-BFP-A, Comp-GFP-A, Comp-iRFP-A, and Comp-Cherry-A represent the channels used for msfBFP[r5M], msfGFP[r5M], emiRFP670 and mCherry respectively. No substantial fluorescence spillover was observed. Measurements were taken using a Cytoflex S cytometer. **B**) Flow cytometry plots for pUC19 plasmid transfections containing no fluorescent marker.

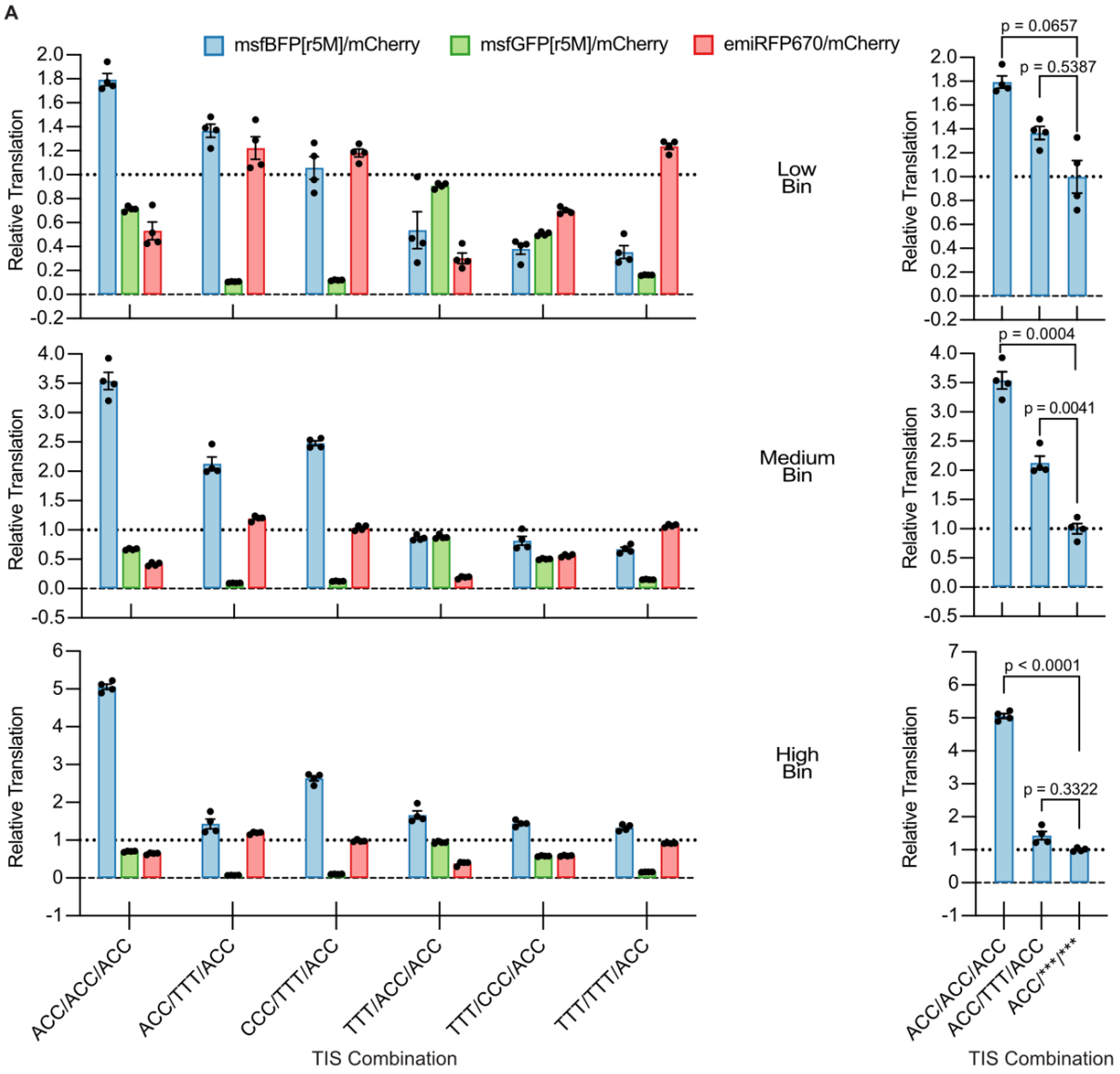

Figure S2-9: 3-ORF SEMPER constructs in CHO-K1 cells.
A) (Left) Min-max normalized relative translation levels for each fluorescent protein. Cells were first clustered into three mCherry fluorescence bins (low, medium, high). Error bars depict SEM (N=4). Welch's ANOVA tests were conducted for blue, green, and red bars independently. The ANOVA tests for all colors across all bins yielded $p < 0.0001$. Post-hoc Dunnett's T3 multiple comparisons tests are provided in **Table S2-5**. (Right) Comparisons of msfBFP[r5M] relative translation levels between the max control and those TIS combinations with multiple ACC driven ORFs. Error bars depict SEM (N=4). Post-hoc Dunnet's T3 comparisons tests demonstrate a significant upregulation in relative translation of msfBFP[r5M] for depicted constructs compared to that observed for ACC/***/*** for some bins. This suggests that the presence of downstream ORFs with strong TISs can lead to boosted expression of the first ORF.

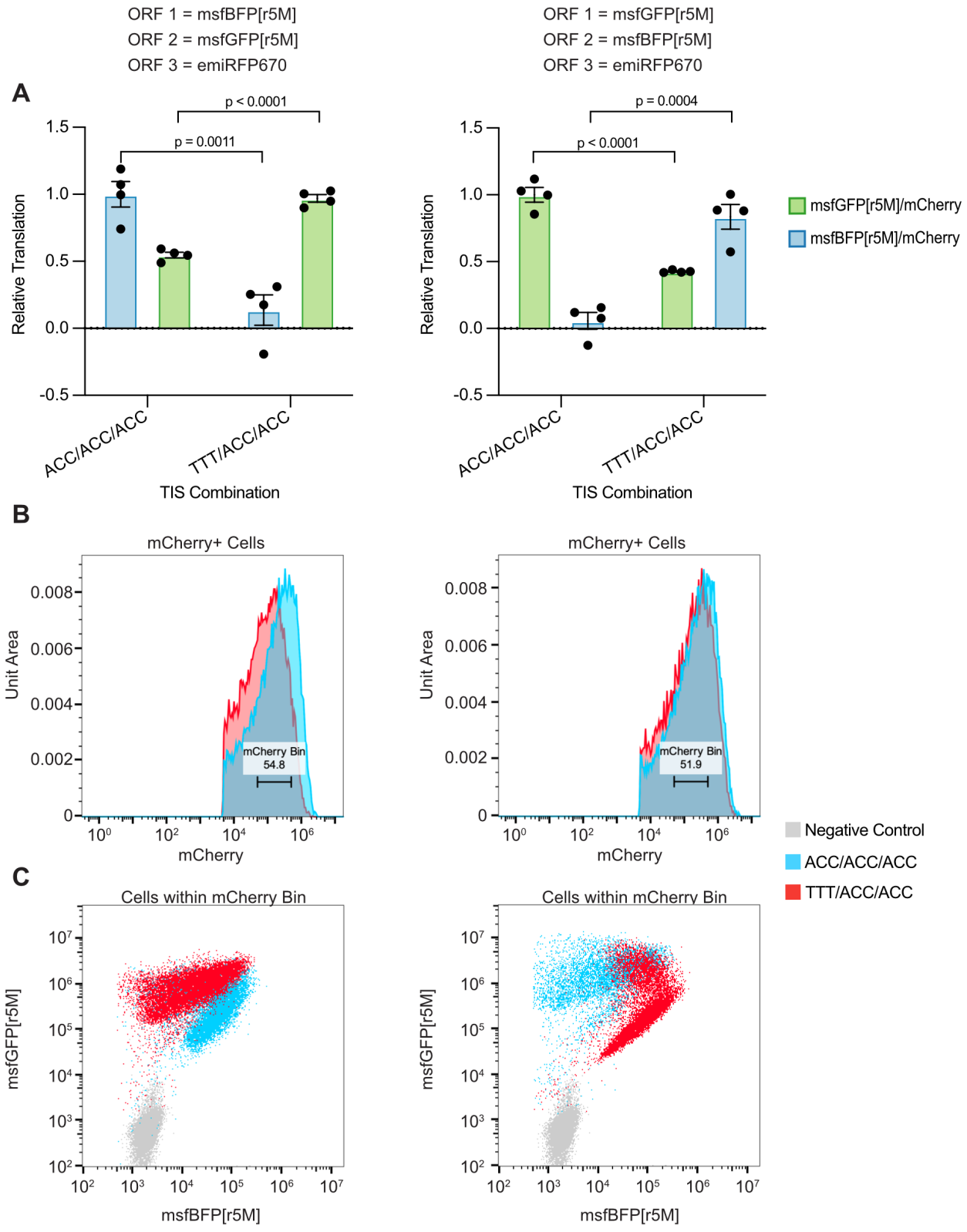

Figure S2-10: SEMPER performance when first two ORFs are interchanged.
As described for **Figure 2-3** of the main text, 3-ORF SEMPER constructs were created with msfBFP[r5M] (a methionineless BFP variant), msfGFP[r5M], and emiRFP670. The first and second ORFs for some of these 3-ORF constructs were interchanged to understand how robust SEMPER is to position effects and ORF identity. The left column of figures depicts constructs that encoded

msfBFP[r5M] first and msfGFP[r5M] second. The right column of figures depicts constructs that encoded msfGFP[r5M] first and msfBFP[r5M] second. emiRFP670 was maintained in the third ORF position under a strong ACC TIS for all constructs. Data was collected on a Cytoflex S flow cytometer and compensated using single-color controls in FlowJo. Cells were then gated by size and then doublet-discriminated. All cells with msfBFP[r5M] fluorescence values less than the negative control were excluded. **A**) Min-max normalized mean values for msfBFP[r5M]/mCherry and msfGFP[r5M]/mCherry within the "mCherry Bin". In the left column, the maximum construct for msfBFP[r5M] was chosen to be ACC/ACC/ACC and the maximum construct for msfGFP[r5M] was chosen to be ***/ACC/ACC (not depicted). In the right column, the maximum construct for msfGFP[r5M] was chosen to be ACC/ACC/ACC and the maximum construct for msfBFP[r5M] was chosen to be ***/ACC/ACC (not depicted). A mCherry single-color control was used as the minimum for all min-max normalization tasks depicted. Error bars depict standard error of the mean (SEM) (N=4). P-values depicted are unpaired two-tailed t-tests. The TTT/ACC/ACC constructs exhibited lower expression of ORF 1 regardless of ORF 1 identity as compared to the expression levels observed for ACC/ACC/ACC. However, the absolute relative translation levels for ORF 1 and ORF 2 did change based on ORF identity, suggesting future researchers will have to conduct some fine-tuning of their gene circuits that utilize the SEMPER approach. Changing or interchanging the ORF sequences can change the secondary structure of the mRNA which may also contribute to deviations in expected expression levels. **B**) mCherry gating strategy. Cells depicted are mCherry positive (mCherry+). As depicted, an "mCherry Bin" was created for these distributions between 4E5 and 5E5 fluorescence units (N=1, representative of four replicates). **C**) Flow cytometry plots for cells in the "mCherry+ Bin". The msfBFP[r5M] variant that was created exhibited dimmer fluorescence compared to msfGFP[r5M], so less separation was observed among conditions in the left column compared to conditions in the right column. Scatter plots in the left column exhibit a "curling" phenotype further from the origin, which we attribute to high transfection efficiency and transgene burden. These curling phenotypes were mitigated in other experiments by reducing transfection efficiency. To produce the bar plots, mCherry normalization was conducted on cells in these flow cytometry plots, followed by min-max normalization of the resultant distribution means. A PUC19 no-color transfection control is depicted in grey (N=1, representative of four replicates).

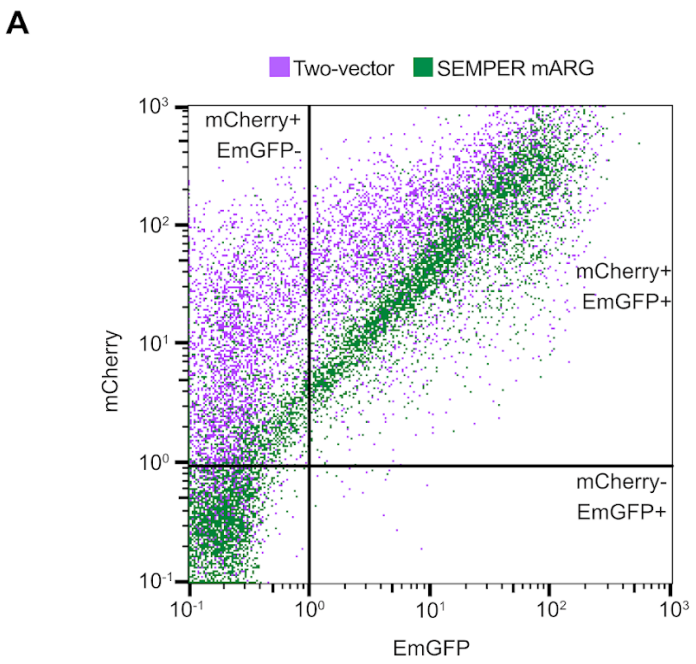

Figure S2-11: SEMPER mARGs are not inhibited by the stochasticity of co-transfection of two plasmids.
A) Comparison of mCherry and EmGFP fluorescence in HEK293T cells transfected with the two-vector mARG expression system or the SEMPER mARG expression system. The gvpA-IRES-mCherry plasmid was transfected in a 4-fold molar excess relative to pgvpNJKFGW-EmGFP, leading to a substantial portion of cells with solely mCherry positivity compared to those transfected with the ACC/ACC SEMPER mARG construct—which contains both fluorescent proteins on a single vector.

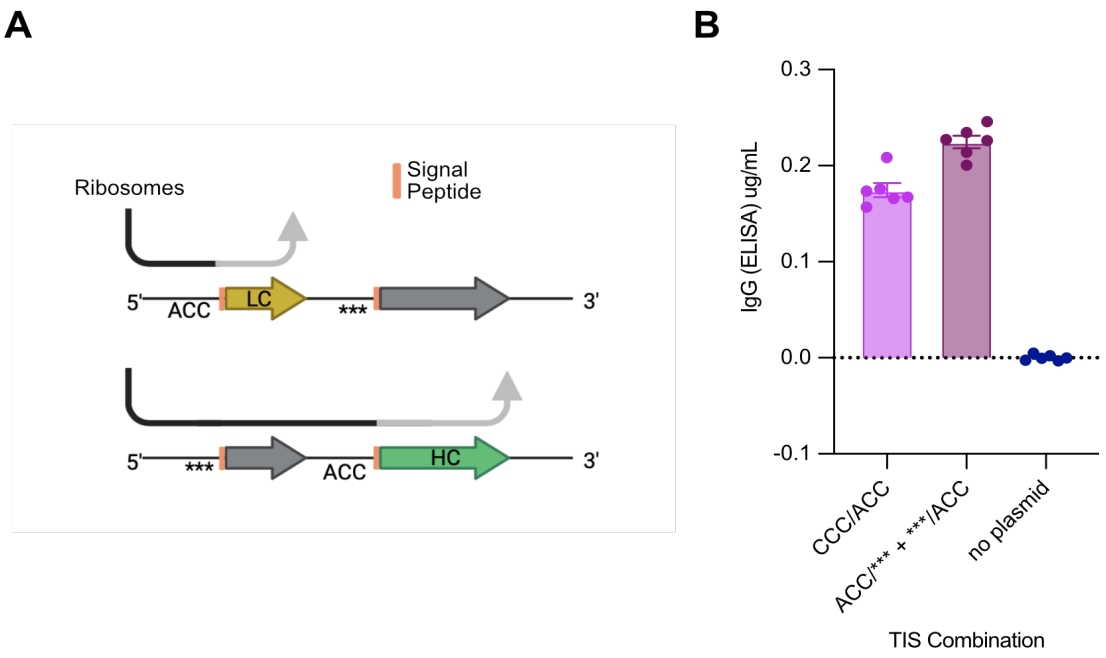

Figure S2-12: Comparison of SEMPER-produced antibody yield with 2-vector control.
A) Architecture of the mRNA transcripts transcribed from transfected SEMPER-b12 plasmid DNA used for positive (ACC/***+***/ACC) and negative (ACC/*** or ***/ACC) controls. **B**) Total human IgG ELISA results following expression and secretion of b12 vectors. HEK293T cells transfected with the 1:1 mixture of the two plasmids while preserving the total amount of transfected DNA (ACC/***+***/ACC) produced similar yield of IgG compared to SEMPER-b12 construct with CCC/ACC TIS combination. Error bars depict SEM (N=6). Statistical analysis was conducted using a one-way ANOVA ($p < 0.0001$) followed by Fisher's LSD post-hoc test with a single pooled variance to compare each treatment group with every other. Pairwise p-values are reported in **Table S2-8.**

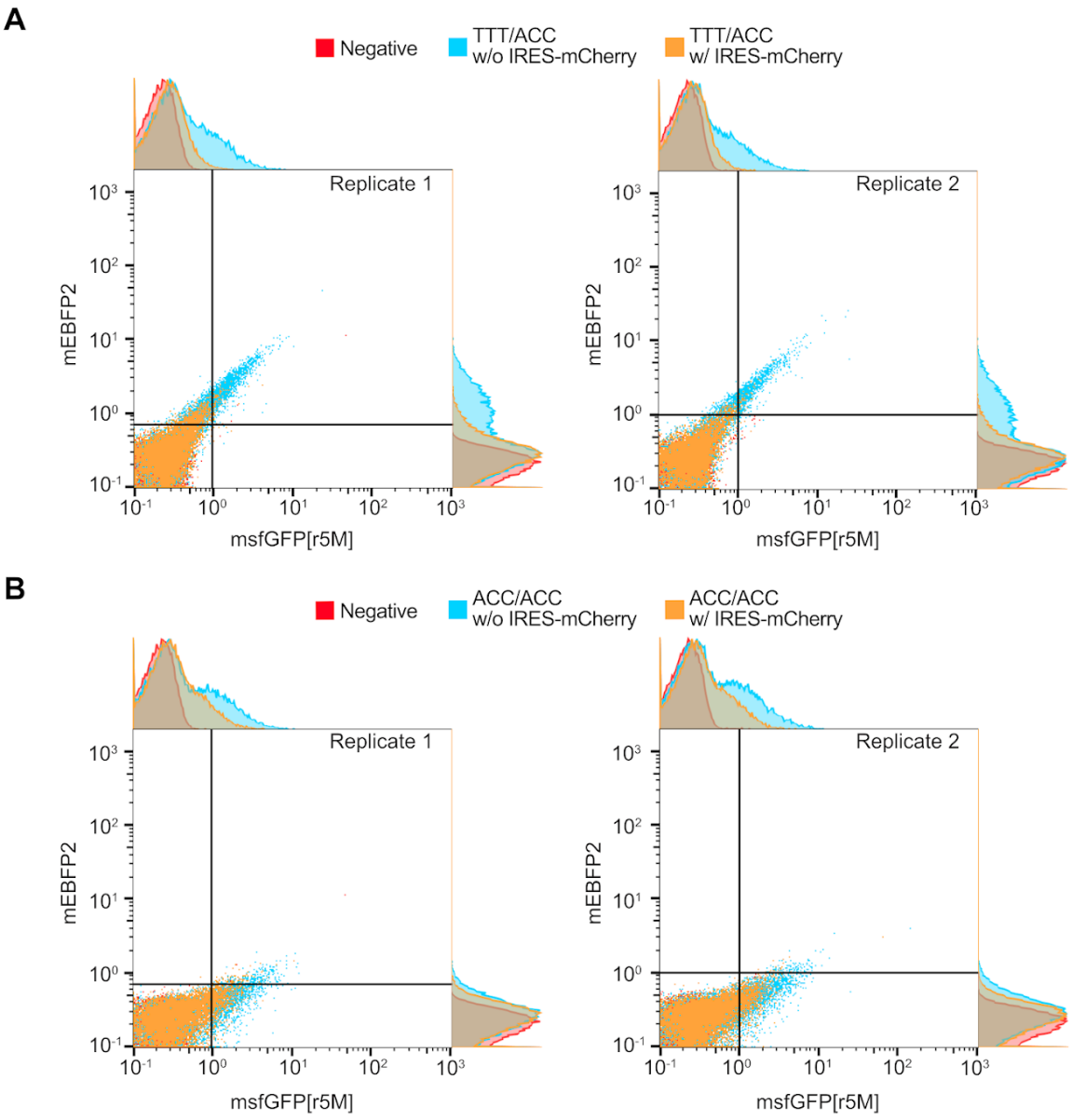

Figure S2-13: 2-ORF IVT mRNA tests with and without IRES-mCherry as measured by flow cytometry.
A) TTT/ACC and **B**) ACC/ACC IVT mRNA constructs were synthesized with and without IRES-mCherry. The IVT mRNA tested was produced with standard UTPs. Negative control cells were transfected solely with Lipofectamine MessengerMAX reagent without mRNA. Cells were first gated by size and then doublet-discriminated. Gates drawn above are based on background fluorescence intensities measured for the negative control cells. For both TIS combinations, the IRES-mCherry mRNA constructs yielded lower mEBFP2 and msfGFP[r5m] expression following transfection into HEK293T cells (N=2). IRES-mCherry was therefore not included in IVT mRNA constructs tested and depicted in the main text. One challenge we saw with 2-ORF IVT mRNA also was generally lower expression of fluorescent proteins as compared to the plasmid DNA transfection case. Improving upon this or even testing a 3-ORF IVT mRNA system will likely require optimization to boost expression levels from IVT mRNA. This may be done by screening for i) 5' UTR sequences and 5' capping reagents that enhance ribosomal loading at the 5' end of transcripts, ii) a 3'UTR sequence and polyA tail track that ensures stability of IVT mRNA, and iii) optimizing transfection conditions to use larger amounts of IVT mRNA. In addition, deviations between plasmid-based expression profiles and IVT mRNA-based expression profiles could be caused by i) poorer capping efficiency of IVT mRNA compared to endogenous mRNA transcribed from plasmid DNA, ii) endogenous post-transcriptional nucleotide modifications that impact expression that cannot be conducted on IVT mRNA, and iii) different mRNA secondary structures between the two cases caused by varying nucleotide modifications or varying transcription environments.[2,3]

Note S2-1: Simulations of SEMPER circuits.

Assumptions of our model:

Our model makes certain simplifications and biological assumptions including:

1. Scanning ribosomes (also referred to below as just "ribosomes") load at the 5'end of an mRNA transcript and will only move across the mRNA from 5' to 3'. This is in direct agreement with the canonical 5' cap-dependent eukaryotic scanning mechanism for translation, which has been heavily researched.[4] We consider the loading of ribosomes at the 5'cap as a fast step and do not directly consider the kinetics of this step in the model.
2. The scanning ribosome (a.k.a. 43S PIC) is a subunit of the full translating ribosome. Upon recognizing a TIS, the 43S PIC interacts with the 60S subunit to form the 80S ribosome which conducts translation. We assume the formation of the 80S ribosome is a kinetically fast step and the rate-limiting step is the interaction of the 43S PIC with the TIS sequence. Therefore, we do not consider a "pooled" model where there may become resource limitations and competition for available 60S subunits. Resource abundances and concentrations in the local environment of the mRNA may also explain boosted first ORF expression observations in **Figure 2-2** of the main text. However, these effects are also not considered in the model.
3. The rate-limiting step in our model is the interaction of the 43S PIC with the TIS sequence and its ability to recognize this TIS as a start site for recruitment of the 60s subunit and subsequent translation. The relative frequency with which a 43S PIC initiates translation at a given TIS are parameters in our model we call p_TIS values. p_TIS takes on a value between 0 and 1. The p_TIS values for ***, TTT, CCC, and ACC are described in the main text and were estimated based on our observed results and the results of Ferreira et al. and Noderer et al.[5,6] As our model is probabilistic in nature, we do not consider any other rate or kinetic parameters. Estimates of these relative p_TIS values can be made for other TIS sequences through analysis of the predicted and measured strengths of TISs made by Noderer et al.[4]
4. We assume there is inherent stochasticity with the number of scanning ribosomes that will traverse an mRNA and the number of mRNA that may exist in a cell. We capture this stochasticity by sampling these values from probabilistic distributions. Specifically, normality was used to describe these processes. However, other distributions such as the Poisson distribution, negative binomial distribution, and gamma distribution could also be considered. The code as written can easily be implemented with these different distributions if a user so desires.
5. As mentioned, our model does not consider "non-linear" effects, such as the increased translation of the first ORF as observed in our screen of the 3-ORF library.
6. Our model does not consider the sequence identity of the ORFs other than the TIS sequence in front of it. The model does not account for ribosomes falling off the transcript during scanning, the length of the transcript, the length of the 5'UTR, local structure of the mRNA, or global structure of the mRNA.

Model architecture:

The pipeline for producing the results **in Figure 2-6B-C** are as follows:

5. *Choosing the number of mRNA to model.* We first estimate the number of unique mRNA simulation results needed to fill the desired number of cells to be modeled with some excess. This is an upper bound estimate to allow the simulation to proceed.
6. *Estimating the number of ribosomes that will traverse each mRNA.* Once the number of mRNA are determined, the number of ribosomes that will traverse each mRNA are sampled from a normal distribution. We chose a normal distribution as it is a commonly used distribution in modeling biological systems. Here, we sample from a normal distribution with mean and standard deviation both equal to 100. The large standard deviation was chosen to

introduce greater variation and randomness into our model. Future users may use our code to easily implement sampling from other distributions if desired. The mean value of 100 was motivated by estimates of the number of observed translation events from single CD147 mRNAs in mammalian cells.[7] The mean value of 100 is likely an underestimate, so further parameter optimization may be considered in future.

7. *Simulating ribosomes traversing an mRNA.* Each mRNA then undergoes a Monte Carlo simulation where R ribosomes will traverse the mRNA. To do this we iterate through a set of Boolean conditions k=1:R times. We initialize variables protein1 and protein2 and set both equal to 0 before this loop begins. We say the k^{th} ribosome has been loaded onto the 5'end of the mRNA, when the k^{th} iteration of this loop begins. On the k^{th} iteration of the loop, the following steps/cases are considered:
    i. We sample a value for a Bernoulli random variable with success probability p_TIS1. If Bernoulli(p_TIS1) equals 1, then the k^{th} ribosome is considered to have succeeded in translating ORF 1, producing one unit of Protein 1. The variable protein1 is incremented up by 1 and the simulation moves to the k+1 iteration. However, if the sampled value from Bernoulli(p_TIS1) equals 0, then the k^{th} ribosome will encounter the next Boolean condition (ii).
    ii. We sample a value for a Bernoulli random variable with success probability p_TIS2. If Bernoulli(p_TIS2) equals 1, then the k^{th} ribosome is considered to have succeeded in translating ORF 2, producing one unit of Protein 2. The variable protein2 is incremented up by 1 and the simulation moves to the k+1 iteration. However, if the sampled value from Bernoulli(p_TIS2) equals 0, then the k^{th} ribosome will encounter the next Boolean condition (iii).
    iii. The final condition is implicit. If the k^{th} ribosome does not initiate translation at either TIS, the ribosome will not create any protein product. The simulation moves onto the k+1 iteration, where the k+1 ribosome is loaded and will encounter these cases (i, ii, iii).
8. *Loading/aggregating mRNA simulation results into cells.* Once all mRNAs have been simulated, we aggregate results of individual mRNA simulations together as a proxy for transient transfection and implicit transcription of plasmid-DNA. Here, the number of individual mRNA simulation results that are aggregated together is sampled from a normal distribution with mean 100 and standard deviation of 100 to capture the stochasticity of transient transfection. This mean value is motivated by observed CD147 mRNA values in single cells. It should be noted the per cell number of CD147 mRNAs measured by Albayrak et al. are not from transient transfection but endogenous expression.[5] Resampling of the mRNA simulations was used in this step as mRNA simulations were computationally expensive to calculate. Specifically, from ~500,000 individual mRNA simulations, 2500 cells were "loaded" with aggregated results. The simulations were then randomized, and another 2500 cells were loaded. This process was conducted four times to yield 10,000 simulated cells. The mean, variance, and choice of distribution to model transient transfection are additional parameters that future researchers can further optimize. The code we have written can easily accommodate the use of different distributions. The quantities of Protein 1 and Protein 2 across all mRNA simulations within a particular group or "cell" are aggregated and then plotted in **Figure 2-6B**. This provides a basis for comparison with other flow cytometry results gathered in this study.

<u>Scaling the model to consider more ORFs or to consider internal methionines</u>

Our model can readily be used to consider systems with more than two ORFs by adding additional p_TIS values in the input parameters. The addition of a p_TIS can be considered as the addition of a desired ORF or an internal methionine within an ORF. The model does not differentiate between a TIS that may exist at the canonical 5' end of an ORF or a TIS that may be encoded in the middle of an ORF.

References Cited in Supplementary Information

1. Andreev, D.E., Loughran, G., Fedorova, A.D., Mikhaylova, M.S., Shatsky, I.N., and Baranov, P.V. (2022). Non-AUG translation initiation in mammals. Genome Biol. *23*, 111. https://doi.org/10.1186/s13059-022-02674-2.
2. Gawroński, P., Pałac, A., and Scharff, L.B. (2020). Secondary Structure of Chloroplast mRNAs In Vivo and In Vitro. Plants *9*, 323. https://doi.org/10.3390/plants9030323.
3. Leamy, K.A., Assmann, S.M., Mathews, D.H., and Bevilacqua, P.C. (2016). Bridging the Gap Between In Vitro and In Vivo RNA Folding. Q. Rev. Biophys. *49*, e10. https://doi.org/10.1017/S003358351600007X.
4. Hinnebusch, A.G. (2014). The Scanning Mechanism of Eukaryotic Translation Initiation. Annu. Rev. Biochem. *83*, 779–812. https://doi.org/10.1146/annurev-biochem-060713-035802.
5. Ferreira, J.P., Overton, K.W., and Wang, C.L. (2013). Tuning gene expression with synthetic upstream open reading frames. Proc. Natl. Acad. Sci. *110*, 11284–11289. https://doi.org/10.1073/pnas.1305590110.
6. Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., and Wang, C.L. (2014). Quantitative analysis of mammalian translation initiation sites by FACS -seq. Mol. Syst. Biol. *10*, 748. https://doi.org/10.15252/msb.20145136.
7. Albayrak, C., Jordi, C.A., Zechner, C., Lin, J., Bichsel, C.A., Khammash, M., and Tay, S. (2016). Digital Quantification of Proteins and mRNA in Single Mammalian Cells. Mol. Cell *61*, 914–924. https://doi.org/10.1016/j.molcel.2016.02.030.

Table S2-1A: Post hoc Games-Howell's multiple comparisons test for 2-ORF HEK293T mCherry distributions. Referenced in **Figure S2-1A.**

Games-Howell's multiple comparisons test	Adjusted P Value	n1	n2
/ACC vs. ACC/	<0.0001	64886	23934
***/ACC vs. ACC/ACC	<0.0001	64886	72256
***/ACC vs. CCC/ACC	0.0044	64886	74879
***/ACC vs. TTT/ACC	<0.0001	64886	81383
ACC/*** vs. ACC/ACC	<0.0001	23934	72256
ACC/*** vs. CCC/ACC	<0.0001	23934	74879
ACC/*** vs. TTT/ACC	<0.0001	23934	81383
ACC/ACC vs. CCC/ACC	<0.0001	72256	74879
ACC/ACC vs. TTT/ACC	<0.0001	72256	81383
CCC/ACC vs. TTT/ACC	<0.0001	74879	81383

Table S2-1B: Post hoc Games-Howell's multiple comparisons test for 2-ORF CHO-K1 mCherry distributions. Referenced in Figure S2-1B.

Games-Howell's multiple comparisons test	Adjusted P Value	n1	n2
/ACC vs. ACC/	<0.0001	32253	12605
***/ACC vs. ACC/ACC	<0.0001	32253	15042
***/ACC vs. CCC/ACC	0.0294	32253	32967
***/ACC vs. TTT/ACC	0.7558	32253	35457
ACC/*** vs. ACC/ACC	<0.0001	12605	15042
ACC/*** vs. CCC/ACC	<0.0001	12605	32967
ACC/*** vs. TTT/ACC	<0.0001	12605	35457
ACC/ACC vs. CCC/ACC	<0.0001	15042	32967
ACC/ACC vs. TTT/ACC	<0.0001	15042	35457
CCC/ACC vs. TTT/ACC	0.0001	32967	35457

Table S2-1C: Post-hoc Games-Howell's multiple comparisons test for 3-ORF HEK293T mCherry distributions. Referenced in **Figure S2-1C**.

Games-Howell's multiple comparisons test	Adjusted P Value	n1	n2
//ACC vs. ***/ACC/***	<0.0001	91946	79348
//ACC vs. ACC/***/***	<0.0001	91946	67093
//ACC vs. ACC/ACC/ACC	<0.0001	91946	73751
//ACC vs. ACC/TTT/ACC	<0.0001	91946	72039
//ACC vs. CCC/TTT/ACC	<0.0001	91946	71607
//ACC vs. TTT/ACC/ACC	<0.0001	91946	81700
//ACC vs. TTT/CCC/ACC	<0.0001	91946	86621
//ACC vs. TTT/TTT/ACC	<0.0001	91946	59893
/ACC/ vs. ACC/***/***	<0.0001	79348	67093
/ACC/ vs. ACC/ACC/ACC	<0.0001	79348	73751
/ACC/ vs. ACC/TTT/ACC	<0.0001	79348	72039
/ACC/ vs. CCC/TTT/ACC	<0.0001	79348	71607
/ACC/ vs. TTT/ACC/ACC	<0.0001	79348	81700
/ACC/ vs. TTT/CCC/ACC	<0.0001	79348	86621
/ACC/ vs. TTT/TTT/ACC	<0.0001	79348	59893
ACC/***/*** vs. ACC/ACC/ACC	0.0016	67093	73751
ACC/***/*** vs. ACC/TTT/ACC	<0.0001	67093	72039
ACC/***/*** vs. CCC/TTT/ACC	<0.0001	67093	71607
ACC/***/*** vs. TTT/ACC/ACC	<0.0001	67093	81700
ACC/***/*** vs. TTT/CCC/ACC	<0.0001	67093	86621
ACC/***/*** vs. TTT/TTT/ACC	<0.0001	67093	59893
ACC/ACC/ACC vs. ACC/TTT/ACC	0.2084	73751	72039
ACC/ACC/ACC vs. CCC/TTT/ACC	<0.0001	73751	71607
ACC/ACC/ACC vs. TTT/ACC/ACC	<0.0001	73751	81700
ACC/ACC/ACC vs. TTT/CCC/ACC	<0.0001	73751	86621
ACC/ACC/ACC vs. TTT/TTT/ACC	<0.0001	73751	59893
ACC/TTT/ACC vs. CCC/TTT/ACC	<0.0001	72039	71607
ACC/TTT/ACC vs. TTT/ACC/ACC	<0.0001	72039	81700
ACC/TTT/ACC vs. TTT/CCC/ACC	<0.0001	72039	86621
ACC/TTT/ACC vs. TTT/TTT/ACC	<0.0001	72039	59893
CCC/TTT/ACC vs. TTT/ACC/ACC	<0.0001	71607	81700
CCC/TTT/ACC vs. TTT/CCC/ACC	0.0006	71607	86621
CCC/TTT/ACC vs. TTT/TTT/ACC	<0.0001	71607	59893
TTT/ACC/ACC vs. TTT/CCC/ACC	<0.0001	81700	86621
TTT/ACC/ACC vs. TTT/TTT/ACC	<0.0001	81700	59893
TTT/CCC/ACC vs. TTT/TTT/ACC	<0.0001	86621	59893

Table S2-1D: Post-hoc Games-Howell's multiple comparisons test for 3-ORF CHO-K1 mCherry distributions. Referenced in **Figure S2-1D.**

Games-Howell's multiple comparisons test	Adjusted P Value	n1	n2
//ACC vs. ***/ACC/***	0.0017	47094	22572
//ACC vs. ACC/***/***	<0.0001	47094	5153
//ACC vs. ACC/ACC/ACC	0.0058	47094	17132
//ACC vs. ACC/TTT/ACC	0.8506	47094	14357
//ACC vs. CCC/TTT/ACC	<0.0001	47094	26154
//ACC vs. TTT/ACC/ACC	<0.0001	47094	23233
//ACC vs. TTT/CCC/ACC	<0.0001	47094	29944
//ACC vs. TTT/TTT/ACC	<0.0001	47094	36965
/ACC/ vs. ACC/***/***	<0.0001	22572	5153
/ACC/ vs. ACC/ACC/ACC	>0.9999	22572	17132
/ACC/ vs. ACC/TTT/ACC	0.0001	22572	14357
/ACC/ vs. CCC/TTT/ACC	0.9914	22572	26154
/ACC/ vs. TTT/ACC/ACC	<0.0001	22572	23233
/ACC/ vs. TTT/CCC/ACC	<0.0001	22572	29944
/ACC/ vs. TTT/TTT/ACC	0.0063	22572	36965
ACC/***/*** vs. ACC/ACC/ACC	<0.0001	5153	17132
ACC/***/*** vs. ACC/TTT/ACC	<0.0001	5153	14357
ACC/***/*** vs. CCC/TTT/ACC	<0.0001	5153	26154
ACC/***/*** vs. TTT/ACC/ACC	<0.0001	5153	23233
ACC/***/*** vs. TTT/CCC/ACC	<0.0001	5153	29944
ACC/***/*** vs. TTT/TTT/ACC	<0.0001	5153	36965
ACC/ACC/ACC vs. ACC/TTT/ACC	0.0005	17132	14357
ACC/ACC/ACC vs. CCC/TTT/ACC	0.9649	17132	26154
ACC/ACC/ACC vs. TTT/ACC/ACC	<0.0001	17132	23233
ACC/ACC/ACC vs. TTT/CCC/ACC	<0.0001	17132	29944
ACC/ACC/ACC vs. TTT/TTT/ACC	0.0027	17132	36965
ACC/TTT/ACC vs. CCC/TTT/ACC	<0.0001	14357	26154
ACC/TTT/ACC vs. TTT/ACC/ACC	<0.0001	14357	23233
ACC/TTT/ACC vs. TTT/CCC/ACC	<0.0001	14357	29944
ACC/TTT/ACC vs. TTT/TTT/ACC	<0.0001	14357	36965
CCC/TTT/ACC vs. TTT/ACC/ACC	<0.0001	26154	23233
CCC/TTT/ACC vs. TTT/CCC/ACC	<0.0001	26154	29944
CCC/TTT/ACC vs. TTT/TTT/ACC	0.1211	26154	36965
TTT/ACC/ACC vs. TTT/CCC/ACC	0.2202	23233	29944
TTT/ACC/ACC vs. TTT/TTT/ACC	<0.0001	23233	36965
TTT/CCC/ACC vs. TTT/TTT/ACC	<0.0001	29944	36965

Table S2-2: Post-hoc Dunnett's T3 multiple comparisons tests for 2-ORF HEK293T constructs.
Referenced in **Figure 2-1D** (N=4 for all conditions).

	Low Bin		Medium Bin		High Bin	
	mEBFP2/mCherry	**msfGFP[r5M]/mCherry**	**mEBFP2/mCherry**	**msfGFP[r5M]/mCherry**	**mEBFP2/mCherry**	**msfGFP[r5M]/mCherry**
One-way Anova P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Dunnett's T3 multiple comparisons test	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**
***/ACC vs. TTT/ACC	0.0047	0.0022	0.248	0.0034	0.0253	0.0022
***/ACC vs. CCC/ACC	<0.0001	0.01	0.0006	0.0065	0.0004	0.0019
***/ACC vs. ACC/ACC	<0.0001	0.0082	0.0017	0.0026	0.0008	0.0004
/ACC vs. ACC/	0.0005	0.0021	0.0014	0.0029	0.0008	0.0004
TTT/ACC vs. CCC/ACC	0.0001	0.0373	0.0019	0.0244	<0.0001	0.002
TTT/ACC vs. ACC/ACC	0.0001	0.016	0.0026	0.0054	0.0005	0.0001
TTT/ACC vs. ACC/***	0.0005	0.0037	0.0019	0.0064	0.0004	0.0001
CCC/ACC vs. ACC/ACC	0.014	0.0594	0.0213	0.0128	0.0024	0.0006
CCC/ACC vs. ACC/***	0.0111	0.0025	0.0099	0.0304	0.0017	0.0015
ACC/ACC vs. ACC/***	0.0369	0.5537	0.0061	0.9652	0.0112	0.6513

Table S2-3: Dunnett's T3 multiple comparisons tests for 2-ORF CHO-K1 constructs.
Referenced in **Figure S2-5** (N=4 for all conditions).

	Low Bin		Medium Bin		High Bin	
	mEBFP2/mCherry	**msfGFP[r5M]/mCherry**	**mEBFP2/mCherry**	**msfGFP[r5M]/mCherry**	**mEBFP2/mCherry**	**msfGFP[r5M]/mCherry**
One-way ANOVA P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Dunnett's T3 multiple comparisons test	Adjusted P Value	Adjusted P Value	Adjusted P Value	Adjusted P Value	Adjusted P Value	Adjusted P Value
***/ACC vs. TTT/ACC	<0.0001	0.0001	0.0022	0.0001	0.0011	<0.0001
***/ACC vs. CCC/ACC	<0.0001	0.0002	0.0004	<0.0001	<0.0001	<0.0001
***/ACC vs. ACC/ACC	<0.0001	0.0002	0.0003	<0.0001	<0.0001	<0.0001
/ACC vs. ACC/	<0.0001	0.0004	0.0001	0.0004	<0.0001	<0.0001
TTT/ACC vs. CCC/ACC	<0.0001	0.0001	<0.0001	<0.0001	<0.0001	<0.0001
TTT/ACC vs. ACC/ACC	<0.0001	0.0005	<0.0001	<0.0001	<0.0001	<0.0001
TTT/ACC vs. ACC/***	<0.0001	0.0017	<0.0001	0.0014	<0.0001	<0.0001
CCC/ACC vs. ACC/ACC	0.0006	0.0008	0.0002	<0.0001	<0.0001	<0.0001
CCC/ACC vs. ACC/***	<0.0001	0.2331	<0.0001	0.0093	<0.0001	0.0002
ACC/ACC vs. ACC/***	<0.0001	0.0033	<0.0001	0.8013	<0.0001	0.0148

Table S2-4: Post-hoc Dunnett's T3 multiple comparisons tests for 3-ORF HEK293T constructs.
Referenced in **Figure 2-2** (N=4 for all conditions).

	Low Bin			Medium Bin			High Bin		
	msfBFP[r5M] /mCherry	**msfGFP[r5M] /mCherry**	**emiRFP670/ mCherry**	**msfBFP[r5M] /mCherry**	**msfGFP[r5M] /mCherry**	**emiRFP670/ mCherry**	**msfBFP[r5M] /mCherry**	**msfGFP[r5M] /mCherry**	**emiRFP670/ mCherry**
One-way ANOVA P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Dunnett's T3 multiple comparisons test	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**
ACC/ACC/ACC vs. ACC/TTT/ACC	0.5759	0.0027	0.0008	0.3098	0.0036	<0.0001	0.0022	0.004	<0.0001
ACC/ACC/ACC vs. CCC/TTT/ACC	<0.0001	0.0075	0.0038	<0.0001	0.008	0.0004	<0.0001	0.0059	0.0055
ACC/ACC/ACC vs. TTT/ACC/ACC	<0.0001	0.0002	0.122	0.0001	<0.0001	0.4009	<0.0001	0.0001	0.0953
ACC/ACC/ACC vs. TTT/CCC/ACC	0.0001	0.0013	0.0018	0.0001	0.0039	0.0018	<0.0001	0.5885	0.0029
ACC/ACC/ACC vs. TTT/TTT/ACC	0.0001	0.0124	0.0014	0.0001	0.0117	0.0001	<0.0001	0.0072	0.0013
ACC/ACC/ACC vs. */***/ACC**	<0.0001	0.0017	0.0108	<0.0001	0.0023	0.0069	<0.0001	0.0027	0.0054
ACC/ACC/ACC vs. */ACC/***	<0.0001	0.0002	0.0125	<0.0001	0.0002	0.0258	<0.0001	<0.0001	0.0033
ACC/ACC/ACC vs. ACC/*/***	0.0063	0.0017	0.0236	0.0062	0.0023	0.0131	<0.0001	0.0029	0.0028
ACC/TTT/ACC vs. CCC/TTT/ACC	0.0001	0.0006	0.0161	<0.0001	0.0005	0.0109	0.0003	0.0035	0.257
ACC/TTT/ACC vs. TTT/ACC/ACC	0.0001	0.0004	0.0021	<0.0001	0.0003	0.0002	<0.0001	0.0003	0.0005
ACC/TTT/ACC vs. TTT/CCC/ACC	0.0002	0.0015	>0.9999	<0.0001	0.0014	0.8692	<0.0001	0.0014	0.0002
ACC/TTT/ACC vs. TTT/TTT/ACC	0.0001	0.0002	0.0012	<0.0001	0.0012	0.0007	<0.0001	0.0007	>0.9999
ACC/TTT/ACC vs. */***/ACC**	<0.0001	<0.0001	0.03	<0.0001	0.0002	0.0214	0.0004	0.0005	0.0882
ACC/TTT/ACC vs. */ACC/***	<0.0001	0.0007	0.0014	<0.0001	0.0005	0.0008	<0.0001	0.0004	0.0004
ACC/TTT/ACC vs. ACC/*/***	0.0366	<0.0001	0.0748	0.0477	<0.0001	>0.9999	<0.0001	0.0001	0.1417
CCC/TTT/ACC vs. TTT/ACC/ACC	0.0012	0.0005	0.0029	0.0008	0.0004	0.0018	0.0007	0.0003	0.0045
CCC/TTT/ACC vs. TTT/CCC/ACC	0.001	0.0022	0.0161	0.0008	0.0021	0.0038	0.0007	0.002	0.0171
CCC/TTT/ACC vs. TTT/TTT/ACC	0.0016	0.0216	0.2065	0.001	0.0242	0.1509	0.0008	0.0489	0.4507

CCC/TTT/ACC vs. */***/ACC**	0.0014	0.0006	0.0552	0.0043	0.0005	0.0341	0.0044	0.0008	0.0216
CCC/TTT/ACC vs. */ACC/***	0.0006	0.0008	0.0023	0.0006	0.0006	0.0014	0.0004	0.0005	0.0036
CCC/TTT/ACC vs. ACC/*/***	0.0014	0.0006	0.0014	0.0021	0.0006	0.0345	0.9979	0.001	0.0173
TTT/ACC/ACC vs. TTT/CCC/ACC	>0.9999	0.0003	0.0044	>0.9999	0.0002	0.0046	>0.9999	0.0002	0.0065
TTT/ACC/ACC vs. TTT/TTT/ACC	>0.9999	0.0005	0.0011	>0.9999	0.0004	0.0006	>0.9999	0.0003	0.001
TTT/ACC/ACC vs. */***/ACC**	0.9215	0.0003	0.0098	0.9227	0.0003	0.0063	0.9393	0.0002	0.0049
TTT/ACC/ACC vs. */ACC/***	0.5496	0.0242	<0.0001	0.5573	0.0168	0.0059	0.4785	0.0097	0.0002
TTT/ACC/ACC vs. ACC/*/***	0.0005	0.0004	0.0266	0.0007	0.0003	0.01	0.0018	0.0002	0.0025
TTT/CCC/ACC vs. TTT/TTT/ACC	>0.9999	0.0027	0.0005	>0.9999	0.0025	0.0002	>0.9999	0.0024	0.0005
TTT/CCC/ACC vs. */***/ACC**	0.8993	0.0012	0.0294	0.9473	0.0011	0.0192	0.951	0.001	0.0052
TTT/CCC/ACC vs. */ACC/***	0.5381	0.0003	0.0028	0.6143	0.0002	0.0028	0.5133	0.0002	0.0043
TTT/CCC/ACC vs. ACC/*/***	0.0007	0.0012	0.1181	0.0007	0.0011	0.9328	0.0019	0.0011	0.0032
TTT/TTT/ACC vs. */***/ACC**	0.8786	0.0004	0.1422	0.9228	0.0003	0.1258	0.9261	0.0004	0.0809
TTT/TTT/ACC vs. */ACC/***	0.4981	0.0009	0.0009	0.5627	0.0007	0.0005	0.4582	0.0005	0.0009
TTT/TTT/ACC vs. ACC/*/***	0.0006	0.0004	0.0002	0.0007	0.0004	0.0047	0.002	0.0005	0.0956
//ACC vs. ***/ACC/***	>0.9999	0.0006	0.0085	>0.9999	0.0005	0.0055	>0.9999	0.0004	0.0043
//ACC vs. ACC/***/***	0.0002	0.4418	0.0199	<0.0001	0.1868	0.0107	0.0205	0.0122	0.9144
/ACC/ vs. ACC/***/***	0.0001	0.0006	0.0141	0.0001	0.0005	0.0068	0.0017	0.0004	0.0022

Table S2-5: Post-hoc Dunnett's T3 multiple comparisons tests for 3-ORF CHO-K1 constructs.
Referenced in **Figure S2-9** (N=4 for all conditions).

	Low Bin			Medium Bin			High Bin		
	msfBFP[r5M] /mCherry	**msfGFP[r5M] /mCherry**	**emiRFP670/ mCherry**	**msfBFP[r5M] /mCherry**	**msfGFP[r5M] /mCherry**	**emiRFP670/ mCherry**	**msfBFP[r5M] /mCherry**	**msfGFP[r5M] /mCherry**	**emiRFP670/ mCherry**
One-way ANOVA P-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Dunnett's T3 multiple comparisons test	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**	**Adjusted P Value**
ACC/ACC/ACC vs. ACC/TTT/ACC	0.023	0.0001	0.022	0.0053	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
ACC/ACC/ACC vs. CCC/TTT/ACC	0.0159	0.0001	0.0172	0.0557	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
ACC/ACC/ACC vs. TTT/ACC/ACC	0.0191	0.0004	0.4576	0.0036	0.0048	0.0011	<0.0001	0.0002	0.0129
ACC/ACC/ACC vs. TTT/CCC/ACC	<0.0001	0.0001	0.6562	0.0011	<0.0001	0.0107	<0.0001	0.001	0.1315
ACC/ACC/ACC vs. TTT/TTT/ACC	<0.0001	0.0002	0.011	0.0031	<0.0001	<0.0001	<0.0001	<0.0001	0.0004
ACC/ACC/ACC vs. */***/ACC**	0.0002	<0.0001	0.0738	0.0004	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
ACC/ACC/ACC vs. */ACC/***	0.003	<0.0001	0.0518	<0.0001	<0.0001	0.0011	<0.0001	<0.0001	<0.0001
ACC/ACC/ACC vs. ACC/*/***	0.0657	<0.0001	0.0637	0.0004	<0.0001	0.0118	<0.0001	<0.0001	0.7929
ACC/TTT/ACC vs. CCC/TTT/ACC	0.3905	0.0507	>0.9999	0.3854	0.0026	0.0328	0.0063	0.0001	0.0012
ACC/TTT/ACC vs. TTT/ACC/ACC	0.0838	<0.0001	0.0115	0.0148	0.0002	<0.0001	0.9562	<0.0001	<0.0001
ACC/TTT/ACC vs. TTT/CCC/ACC	0.0002	0.0002	0.1026	0.0033	<0.0001	<0.0001	>0.9999	<0.0001	<0.0001
ACC/TTT/ACC vs. TTT/TTT/ACC	0.0002	0.0001	>0.9999	0.0032	<0.0001	0.0599	0.9999	<0.0001	0.0046
ACC/TTT/ACC vs. */***/ACC**	0.0009	<0.0001	0.6133	0.0003	<0.0001	0.0131	0.0062	<0.0001	0.0031
ACC/TTT/ACC vs. */ACC/***	0.0095	<0.0001	0.0094	<0.0001	<0.0001	0.0001	0.0048	<0.0001	<0.0001
ACC/TTT/ACC vs. ACC/*/***	0.5387	0.0018	0.6389	0.0041	0.0003	0.253	0.3322	0.0023	0.0643
CCC/TTT/ACC vs. TTT/ACC/ACC	0.379	<0.0001	<0.0001	<0.0001	0.0003	<0.0001	0.0091	<0.0001	0.0006
CCC/TTT/ACC vs. TTT/CCC/ACC	0.036	0.0002	0.0025	0.0001	<0.0001	0.0006	0.0004	<0.0001	0.0002
CCC/TTT/ACC vs. TTT/TTT/ACC	0.0207	<0.0001	0.9791	<0.0001	0.0028	0.9216	0.0002	<0.0001	0.1538

CCC/TTT/ACC vs. */***/ACC**	0.004	<0.0001	0.0844	<0.0001	0.0002	0.9509	<0.0001	<0.0001	0.9939
CCC/TTT/ACC vs. */ACC/***	0.0164	<0.0001	0.0005	0.0001	<0.0001	0.0002	<0.0001	<0.0001	<0.0001
CCC/TTT/ACC vs. ACC/*/***	>0.9999	<0.0001	0.2869	0.0013	<0.0001	0.1044	0.0003	0.0002	0.3346
TTT/ACC/ACC vs. TTT/CCC/ACC	0.9956	<0.0001	0.0126	>0.9999	0.0004	<0.0001	0.7414	0.0001	0.0524
TTT/ACC/ACC vs. TTT/TTT/ACC	0.9911	<0.0001	0.0002	0.0861	0.0003	<0.0001	0.3656	<0.0001	0.0031
TTT/ACC/ACC vs. */***/ACC**	0.3421	<0.0001	0.0056	0.0037	0.0002	<0.0001	0.0019	<0.0001	0.0005
TTT/ACC/ACC vs. */ACC/***	0.6652	0.0293	0.0591	0.0121	0.0441	0.0033	0.0014	0.1314	0.0084
TTT/ACC/ACC vs. ACC/*/***	0.6272	<0.0001	0.0012	0.9005	0.0002	0.0069	0.0752	<0.0001	0.0617
TTT/CCC/ACC vs. TTT/TTT/ACC	>0.9999	0.0003	0.0001	0.8554	<0.0001	<0.0001	0.7539	<0.0001	<0.0001
TTT/CCC/ACC vs. */***/ACC**	0.2133	<0.0001	<0.0001	0.0041	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
TTT/CCC/ACC vs. */ACC/***	0.7103	<0.0001	<0.0001	0.0026	<0.0001	0.0002	<0.0001	<0.0001	<0.0001
TTT/CCC/ACC vs. ACC/*/***	0.1384	<0.0001	0.1201	0.9148	<0.0001	0.0172	0.0039	<0.0001	0.4548
TTT/TTT/ACC vs. */***/ACC**	0.2524	<0.0001	0.0141	0.0122	<0.0001	0.0636	<0.0001	<0.0001	0.0302
TTT/TTT/ACC vs. */ACC/***	0.8134	<0.0001	<0.0001	0.01	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
TTT/TTT/ACC vs. ACC/*/***	0.1304	<0.0001	0.1349	0.2487	<0.0001	0.1214	0.0182	<0.0001	0.5772
*****/***/ACC vs. ***/ACC/***	>0.9999	<0.0001	<0.0001	0.9397	<0.0001	<0.0001	0.9997	<0.0001	<0.0001
*****/***/ACC vs. ACC/***/***	0.0271	0.0157	>0.9999	0.002	0.0309	0.084	0.001	0.001	0.271
*****/ACC/*** vs. ACC/***/***	0.0498	<0.0001	0.0034	0.0014	<0.0001	0.0046	<0.0001	<0.0001	0.0113

Table S2-6: Dunn's multiple comparisons test results of Emerald/mCherry values of SEMPER-mARG.
Referenced in **Figure 2-3B**.

Dunn's multiple comparisons test	Adjusted P Value	n1	n2
Two Vector vs. ACC/ACC	<0.0001	9793	9793
Two Vector vs. CCC/ACC	<0.0001	9793	9793
Two Vector vs. GGG/ACC	<0.0001	9793	9793
Two Vector vs. TTT/ACC	<0.0001	9793	9793
Two Vector vs. */ACC**	<0.0001	9793	9793
Two Vector vs. IRES-mCherry	<0.0001	9793	5278
ACC/ACC vs. CCC/ACC	<0.0001	9793	9793
ACC/ACC vs. GGG/ACC	<0.0001	9793	9793
ACC/ACC vs. TTT/ACC	<0.0001	9793	9793
ACC/ACC vs. */ACC**	<0.0001	9793	9793
ACC/ACC vs. IRES-mCherry	<0.0001	9793	5278
CCC/ACC vs. GGG/ACC	<0.0001	9793	9793
CCC/ACC vs. TTT/ACC	<0.0001	9793	9793
CCC/ACC vs. */ACC**	<0.0001	9793	9793
CCC/ACC vs. IRES-mCherry	<0.0001	9793	5278
GGG/ACC vs. TTT/ACC	<0.0001	9793	9793
GGG/ACC vs. */ACC**	<0.0001	9793	9793
GGG/ACC vs. IRES-mCherry	<0.0001	9793	5278
TTT/ACC vs. */ACC**	<0.0001	9793	9793
TTT/ACC vs. IRES-mCherry	<0.0001	9793	5278
***/ACC vs. IRES-mCherry**	<0.0001	9793	5278

Table S2-7: Multiple comparisons test results of BURST ultrasound signal SEMPER-mARG
Referenced in **Figure 2-3D.**

Uncorrected Fisher's LSD	Individual P Value	n1	n2
Two-vector vs. ACC/ACC	<0.0001	4	4
Two-vector vs. CCC/ACC	<0.0001	4	4
Two-vector vs. TTT/ACC	<0.0001	4	4
Two-vector vs. */ACC**	<0.0001	4	4
ACC/ACC vs. CCC/ACC	0.001	4	4
ACC/ACC vs. TTT/ACC	<0.0001	4	4
ACC/ACC vs. */ACC**	<0.0001	4	4
CCC/ACC vs. TTT/ACC	<0.0001	4	4
CCC/ACC vs. */ACC**	<0.0001	4	4
TTT/ACC vs. */ACC**	<0.0001	4	4

Table S2-8: Multiple comparisons test results of SEMPER-b12 mAb.
Referenced in **Figure 2-4B**.

Uncorrected Fisher's LSD	Individual P Value	n1	n2
ACC/* vs. ACC/ACC**	<0.0001	6	6
ACC/* vs. GGG/ACC**	<0.0001	6	6
ACC/* vs. CCC/ACC**	<0.0001	6	6
ACC/* vs. TTC/ACC**	<0.0001	6	6
ACC/* vs. TTT/ACC**	0.0016	6	6
ACC/* vs. ***/ACC**	0.5476	6	6
ACC/* vs. no plasmid**	0.6651	6	6
ACC/* vs. ACC/*** + ***/ACC**	<0.0001	6	6
ACC/ACC vs. GGG/ACC	<0.0001	6	6
ACC/ACC vs. CCC/ACC	<0.0001	6	6
ACC/ACC vs. TTC/ACC	0.0002	6	6
ACC/ACC vs. TTT/ACC	0.0001	6	6
ACC/ACC vs. */ACC**	<0.0001	6	6
ACC/ACC vs. no plasmid	<0.0001	6	6
ACC/ACC vs. ACC/* + ***/ACC**	<0.0001	6	6
GGG/ACC vs. CCC/ACC	0.0012	6	6
GGG/ACC vs. TTC/ACC	<0.0001	6	6
GGG/ACC vs. TTT/ACC	<0.0001	6	6
GGG/ACC vs. */ACC**	<0.0001	6	6
GGG/ACC vs. no plasmid	<0.0001	6	6
GGG/ACC vs. ACC/* + ***/ACC**	<0.0001	6	6
CCC/ACC vs. TTC/ACC	<0.0001	6	6
CCC/ACC vs. TTT/ACC	<0.0001	6	6
CCC/ACC vs. */ACC**	<0.0001	6	6
CCC/ACC vs. no plasmid	<0.0001	6	6
CCC/ACC vs. ACC/* + ***/ACC**	<0.0001	6	6
TTC/ACC vs. TTT/ACC	<0.0001	6	6
TTC/ACC vs. */ACC**	<0.0001	6	6
TTC/ACC vs. no plasmid	<0.0001	6	6
TTC/ACC vs. ACC/* + ***/ACC**	<0.0001	6	6
TTT/ACC vs. */ACC**	0.0003	6	6
TTT/ACC vs. no plasmid	0.0004	6	6
TTT/ACC vs. ACC/* + ***/ACC**	<0.0001	6	6
***/ACC vs. no plasmid**	0.8656	6	6
/ACC vs. ACC/ + ***/ACC**	<0.0001	6	6
no plasmid vs. ACC/* + ***/ACC**	<0.0001	6	6

Table S2-9: Relevant sequences used to construct plasmids and IVT mRNA.

Please refer to publication to access this table: 10.1016/j.cels.2024.06.001

Supplementary Information 3

This section is temporarily embargoed.

Supplementary Information 4

This section is temporarily embargoed.

Supplementary Information 5

This section is temporarily embargoed.