Expanding Frontiers in Biomedical Imaging and Synthetic Biology: Dynamic Acoustic Reporter Gene Imaging and Ratio-Tuning of Mammalian mRNA Polycistronic Expression

Thesis by
Mengtong (Tom) Duan

In Partial Fulfillment of the Requirements for the Degree
of
Doctor of Philosophy in Bioengineering

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended on March 21st, 2024

© 2024

Mengtong Duan
ORCID: 0000-0002-1601-8876

All rights reserved except where otherwise noted.

Table of Contents

Table of Figures

Acknowledgements

I am deeply grateful to my PhD advisor, Professor Mikhail Shapiro, for his invaluable mentorship and guidance throughout my research journey. His trust in my independence and passion, whether it was directed towards scientific projects or extracurricular activities, has been a cornerstone of my growth and confidence. Even in moments of doubt, his unwavering faith in me was a beacon of support, for which I am profoundly thankful. His role in my life transcends that of an advisor; he has been an older brother, providing both professional guidance and personal support.

I must also extend my heartfelt thanks to my undergraduate advisor, Professor Alex Merz from the University of Washington. Professor Merz was the catalyst for my academic pursuit, instilling in me the curiosity and drive that have shaped my path. Beyond his foundational impact, Professor Merz imbued me with a deep appreciation for skepticism and experimental rigor, qualities that have been crucial in my research. His continued support and mentorship throughout my PhD, akin to that of a father figure, have been indispensable. His guidance has not only laid the foundation for my academic career but has also been a constant source of wisdom and encouragement.

At the University of Washington, in Professor Merz's lab, Rachael Plemel, the lab manager, played a pivotal role in my early research career by teaching me my first laboratory skills. Her patience and expert guidance laid the groundwork for my experimental approach and set a high standard for research excellence.

I am profoundly thankful to Professor Justin Bois for his exceptional teaching and mentorship. Through the numerous classes I took and had the opportunity to serve as a teaching assistant for, I have greatly expanded my knowledge in coding, statistics, and data analysis. His dedication to education and his ability to convey complex concepts with clarity have been pivotal in enhancing my research skills and analytical thinking.

I would also like to express my sincere gratitude to my thesis committee members, Professors Michael Elowitz, Shu-ou Shan, Doug Rees, and Grant Jensen, for their invaluable guidance, insights, and rigorous scrutiny throughout my research journey. Their expertise and thoughtful feedback have significantly contributed to the refinement of my work and my growth as a scientist. Their dedication to fostering a rigorous academic environment has been instrumental in challenging me to achieve my best.

In addition to my advisors, my journey was enriched and made possible by numerous colleagues and friends whose support and collaboration have been invaluable.

I am particularly thankful to Przemek Dutka, a peer who not only taught me the intricacies of electron microscopy but also shared my passion for elucidating the assembly mechanism of gas vesicles. His camaraderie and expertise have been pivotal to my growth.

My profound gratitude goes to Zhiyang Jin, Danny Sawyer, and Claire Rabut, who were instrumental in teaching me everything I know about ultrasound. Their technical support and shared knowledge were foundational to my research.

I would like to express my special thanks to Di Wu, whose exemplary role as a professional model and support has been inspiring at every step of my journey.

Acknowledgement is due to my rotation mentors: Rob Hurt, Arash Farhadi, and Mei Yi You. Arash, for entrusting me with his project, and Mei Yi, for imparting the technical skills necessary at the outset of my research, have both significantly shaped my academic path.

My gratitude extends to Marjorie Buss, Ishaan Dev, and Rob Hurt, my co-first authors on our manuscripts. Their collaboration has been a cornerstone of our shared successes.

Margaret Swift and Pierina Barturen-Larrea deserve my heartfelt thanks for their assistance with animal experiments, without which much of the work would not have been possible.

To my mentees: Pierina Barturen-Larrea, Liz Hughes, Hengyu Li, Ishaan Dev, and Goar Ayrapetyan, witnessing your growth and success has brought me immense satisfaction, surpassing even the joy of publishing papers. Your progress and achievements have been a source of great pride.

Lastly, my journey would not have been the same without the friendship of Bill Ling and Rohit Nayak. Your companionship, both inside and outside the lab, has been a source of comfort, joy, and inspiration.

In terms of funding and institutional support, my research has been generously supported by a variety of sources. I am grateful to the National Institute of General Medical Sciences (NIGMS) of the NIH for funding our Gas Vesicle Ultrasound (GVUS) grant, which has been essential in advancing our work. Additionally, the support from the Center for Environmental Microbial Interactions (CEMI) at Caltech has provided invaluable resources and opportunities for collaboration. As a member of an HHMI lab, I have benefited from the Howard Hughes Medical Institute's commitment to fostering cutting-edge research and innovation. The resources, facilities, and intellectual environment provided by these organizations have been instrumental in the pursuit and achievement of our scientific goals.

To all mentioned and unmentioned who have contributed to this journey, your support has been a beacon of light in the pursuit of knowledge.

Abstract

This thesis presents a comprehensive exploration of the next generation of mammalian Acoustic Reporter Genes (mARGs), unveiling a novel approach for non-invasive, real-time imaging of cellular processes and gene expression within live animals[1]. Building on the foundational work of first-generation ARGs[2,3], which introduced the groundbreaking concept of using gas vesicle (GV) genes as genetically encoded ultrasound contrast agents, this research tackles the inherent limitations of these pioneering systems. The first segment details the development and characterization of the second-generation mARGs which significantly improve upon their predecessors by offering robust expression without the need for monoclonal screening, dynamic non-destructive imaging capabilities, and customizable acoustic properties through gene and protein level modifications. This advancement not only enhances the utility of mARGs in biomedical imaging but also paves the way for their application in novel therapeutic monitoring strategies, as exemplified by real-time tracking of tumor development and ultrasound-guided tumor biopsies that leverage gene expression information.

Further, the thesis delves into the structural, genetic, and biochemical principles underpinning GV assembly, addressing a critical knowledge gap that has persisted despite the utility of GVs in ultrasound imaging. Understanding these assembly mechanisms is crucial for the engineering of improved ARGs.

The exploration then extends into innovative bioengineering methodologies, specifically Stoichiometric Expression of Messenger Polycistrons by Eukaryotic Ribosomes (SEMPER), a synthetic biology breakthrough enabling the expression of multiple proteins at precise stoichiometries from single, compact transcripts.[4] SEMPER represents a strategic advancement in the field, facilitating efficient formation of multi-protein complexes, minimizing cellular toxicity, and broadening the scope of potential applications in genetic engineering, including the creation of enhanced cell lines and circuits for research and therapeutic purposes.

Collectively, this work not only advances our understanding of GV-based ultrasound imaging and gene expression tracking but also introduces versatile genetic tools for the manipulation of cellular machinery. These achievements mark significant strides in the fields of synthetic biology and molecular imaging, setting the stage for future innovations in non-invasive diagnostics, cellular therapy, and cancer monitoring research. Through the integration of improved acoustic reporter genes, insights into gas vesicle assembly, and the SEMPER method for gene expression, this thesis embodies a holistic approach to overcoming current challenges and unlocking new potentials in biomedical engineering and synthetic biology.

Published Content and Contributions

(#Main author, *corresponding author)

- Jin, Z.#, Lakshmanan, A., Zhang, R., Tran, T.A., Rabut, C., Dutka, P., **Duan, M.**, Hurt, R.C., Malounda, D., Yao, Y., Shapiro, M.G.* Ultrasonic reporters of calcium for deep tissue imaging of cellular signals. *BioRxiv preprint,* (2023). doi: https://doi.org/10.1101/2023.11.09.566364

  M.D. conducted some of the experiments.

- **Duan, M.#,** Dev, I.#, Lu, A., Yi, M.Y., Shapiro, M.G.* Stoichiometric expression of messenger polycistrons by eukaryotic ribosomes (SEMPER) for compact, ratio-tunable multi-gene expression from single mRNAs. *Cell Systems,* (2024). doi: 10.1101/2023.05.26.541240

  M.D. conceived the study and contributed to the study design, experiments, data analysis, and writing.

- Wu, D.#, Baresch, D., Cook, C., **Duan, M.**, Malounda, D., Maresca, D., Abundo, M.P., Lee, J., Shivaei, S., Mittelstein, D.R., Qiu, T., Fischer, P., Shapiro, M.G.* Biomolecular actuators for genetically selective acoustic manipulation of cells. *Science Advances*, (2023). doi: 10.1126/sciadv.add9186

  M.D. designed, planned, and conducted the mammalian cell experiments.

- 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., Jin, Z., Abedi, M.H., Farhadi, A., Deshpande, R., Shapiro, M.G.* Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. *Nature Biotechnology,* (2023). doi: 10.1038/s41587-022-01581-y

  M.D. conceived the study and contributed to the study design, experiments, data analysis, and writing. Specifically, M.D. was responsible for all experiments related to mammalian GV expression and imaging.

"Passionate discipline is the alchemy of the mind, where the fervor of passion and the rigor of discipline converge, transforming aspirations into tangible realities. It is not merely the persistence through obstacles, but the joy in the journey, the hunger for knowledge, and the relentless pursuit of excellence. In this state, work transcends obligation, becoming an expression of one's deepest values and desires, a symphony of effort and enthusiasm that elevates both the mundane and the monumental tasks to acts of profound significance."

- *ChatGPT 4*

Chapter 1. Introduction to Thesis

In this introduction, I aim to provide background information to help the reader understand the key concepts of our research, as well as to provide the reader with a brief overview of our field and previous work. We will start with a general overview of gas vesicles (GVs): what they are, where they come from and how they look and work. Next, we will introduce basic concepts of ultrasound imaging and the imaging modalities that we commonly used to visualize GVs. This will be followed by a brief history of our previous attempts at heterologous expression of GVs in bacteria and mammalian cells. Last, we will briefly mention various GV genetic engineering efforts in our lab.

Following the introductory chapter, this thesis unfolds through a series of chapters that encapsulate my doctoral research journey. **Chapter 2** heralds this exploration with a pioneering study, resulting in the manuscript "Genomically Mined Acoustic Reporter Genes for Real-Time *in vivo* Monitoring of Tumors and Tumor-Homing Bacteria," published in Nature Biotechnology.[1] This chapter lays the groundwork for advancing ultrasound imaging through the development of next-generation acoustic reporter genes. Transitioning from this foundational work, **Chapter 3** delves into the genetics and biochemistry underpinning gas vesicle nucleation and assembly, marking a focused exploration into the intricate processes essential for the functionality of acoustic reporter genes. **Chapter 4** takes a significant detour into the realm of synthetic biology, introducing an innovative approach to polycistronic expression in mammalian systems. This exploration is encapsulated in the manuscript titled "Stoichiometric Expression of Messenger Polycistrons by Eukaryotic Ribosomes (SEMPER) for compact, ratio-tunable multi-gene expression from single mRNAs," which is currently under revision for publication in Cell Systems.[4] The concluding **Chapter 5** synthesizes the insights gained from these investigations, weaving together the implications for both synthetic biology and biomedical imaging, and laying out the path forward for future research in these dynamic fields.

1a. Sound vs light

The green fluorescent protein (GFP) has revolutionized biological sciences, allowing for optical visualization of gene expression and molecular and cell dynamics, being used as an output to various conditional systems including molecular sensors and biological circuits at nanometer resolutions. However, the utility of fluorescent proteins and optical imaging methods is limited inside deep tissues of opaque organisms because light scattering in tissue leads to a loss of resolution and penetration.[5] Ultrasound, however, easily penetrates tissues, allowing for deep and noninvasive imaging of biological phenomena at centimeter-scale depths with high spatial (50 – 500 μm) and temporal (sub-millisecond) resolution.[6,7] Ultrasound is also relatively safe and low-cost and is already widely used in clinical and research setting. Whereas contrast agents for ultrasound exist in the form of synthetic microbubbles, a genetically encoded ultrasound contrast agent like the GFP has been elusive until recently. Gas Vesicles (GVs), gas-filled protein nanostructures, have been recently shown to produce ultrasound contrast.[8] This has led to a launch of a new field of genetically encoded ultrasound contrast agents, seeking to improve and utilize GVs for various research, diagnostic and therapeutic applications.

1b. Gas vesicles, the oldest mode of motility ever evolved

Gas vesicles are naturally found across various phyla of predominantly aquatic prokaryotes.[9] Their native function is to regulate buoyancy of these organisms in the water column providing them with a passive form of motility **(Fig. 1a)**.[10] Indeed, gas vesicles represent one of the earliest forms of cellular movement to have evolved.[11] These gas vesicles form a hollow spindle or cylindrical structure on the order of a hundred nanometers in diameter and several hundred nanometers to micrometers in length **(Fig. 1b)**. GVs can exclude liquid from their interior via surface tension. As cells produce GVs, they displace their aqueous cytosol with gas, changing their density and buoyancy. GV-expressing cells also scatter more light, producing an opaque colony morphology and turbid culture suspension. It has been long known that GVs can be collapsed with a sudden increase in hydrostatic pressure **(Fig 1c)**. One of the early experiments involved a flask containing a buoyant and turbid cell culture, which became clear and settled after the flask's cork was struck with a hammer.[12] We now know that GVs can be also collapsed by exposure to strong acoustic pressures emitted by an ultrasound probe.[13] Collapsed GVs do not contain gaseous interior.

a **b** **c**

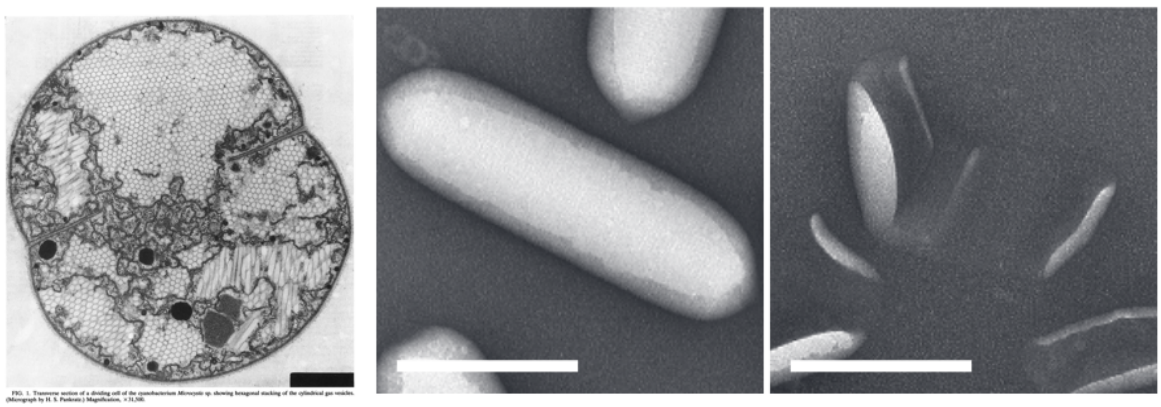


Figure 1 | Gas vesicles. (a) Transverse section of *Microcystis sp.* cell filled with gas vesicles.[9] **(b)** TEM images of *A. flos-aquae* GVs. **(c)** TEM images of collapsed *A. flos-aquae GVs.*[8]

GV gene clusters

Gas vesicle genes are organized within prokaryotic operons located either on the primary chromosome or on episomes. For instance, in Halobacterium salinarum, the p-vac gas vesicle gene cluster is divided into two clusters oriented in opposite directions, each regulated by multiple promoters. This arrangement enables temporal control over the expression of specific gene subsets. Specifically, the operon containing gvpFGHIJKLM is active during the early exponential growth phase, while the gvpACNO operon is expressed later, during the late exponential and stationary phases. (**Fig 2**).[10,14]

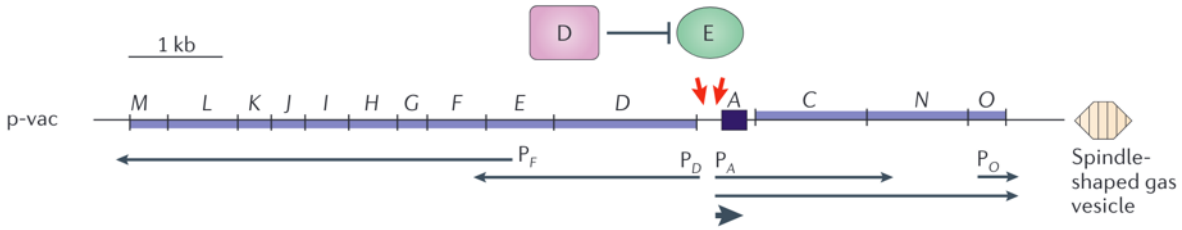


Figure 2 | p-vac GV gene cluster of *H. Salinarum*. P_A, P_D, P_F and P_O are promoters. The arrows represent transcripts, their thickness represents the amount of transcript. P_D and P_F are activated by GvpE, which in turn is inhibited by GvpD.[14]

The sequence homology observed across gas vesicle gene clusters from various prokaryotic species indicates a shared common evolutionary ancestor for all gas vesicles. This widespread distribution among contemporary prokaryotes may have been facilitated by horizontal gene transfer.[9] Gas vesicle genes on GV operons are usually annotated as *gvp* (gas vesicle protein) *or gvr* (gas vesicle regulatory protein), followed by an alphanumeric symbol (e.g., *gvpA1*).[15] Homologous genes are usually annotated with the same letter between species (e.g., *gvpN* in cyanobacterium *Anabaena flos-aquae* is a homolog of *gvpN* in archaeon *Halobacterium Salinarum*), though exceptions exist **(Fig 3)**. Not all genes on a given operon are essential for GV formation, some genes serve a gene-regulatory role, while deletion of others can affect GV shape, acoustic properties and expression levels. Some GV operons have multiple identical or close-to-identical copies of a particular gene that are interchangeable (e.g. Multiple identical copies of *gvpA* in *A. flos-aquae,* three almost-identical copies of *gvpA* in *M. aeruginosa* and *gvpA/B* in *B. megaterium*), in other cases two or more genes on the same operon share high sequence homology but serve distinct functions and cannot substitute for one another (e.g., *gvpJ and gvpS* in *Bacillus megaterium*). Different prokaryotes have evolved different ways to tune in optimal relative expression levels of individual *gvps*. Apart from gene duplication, promoter and ribosome binding site (RBS) strength, some operons employ intracluster promoters and terminators, creating transcripts of various lengths, with some more abundant than others (**Fig 2**).[14]

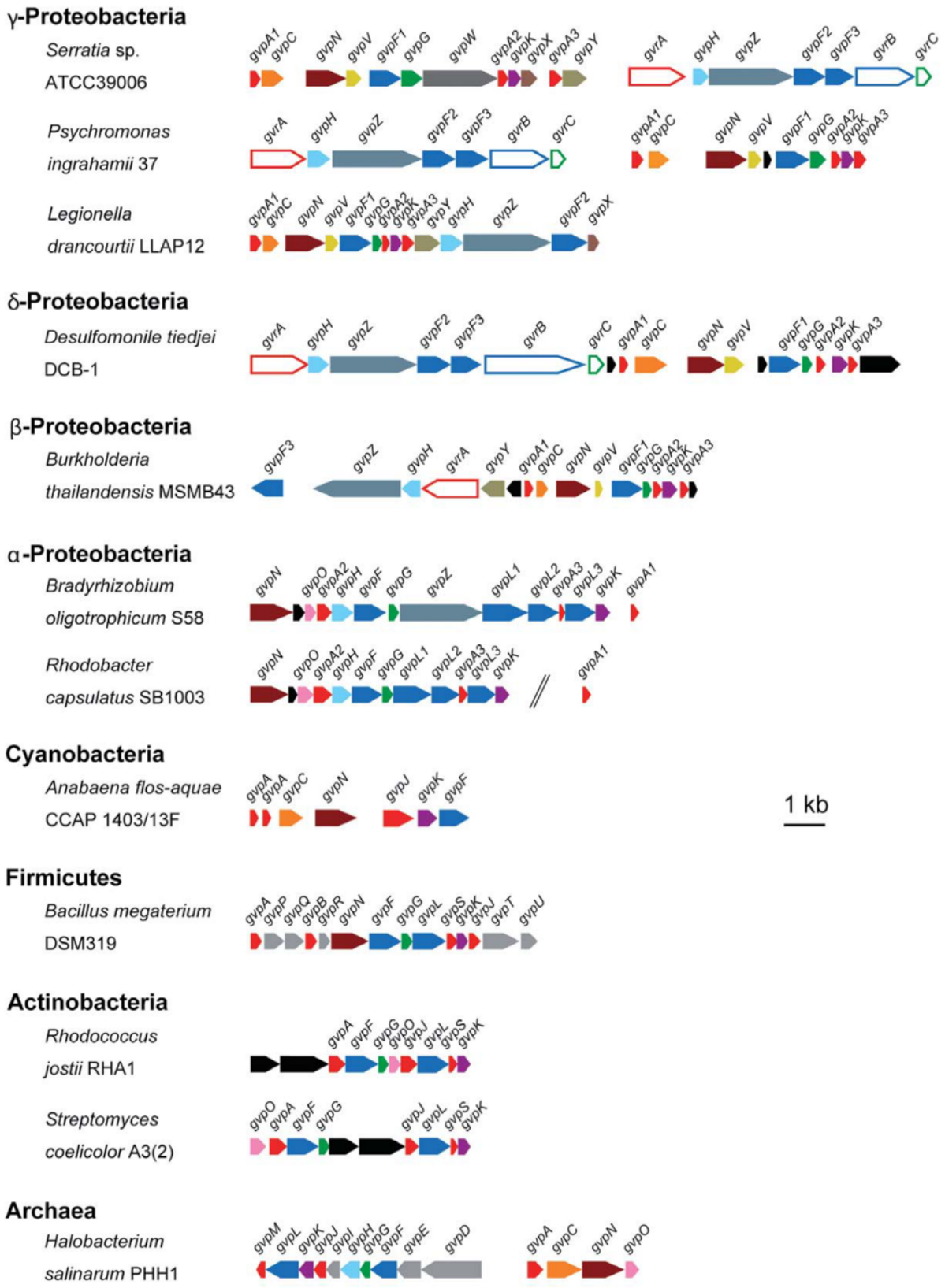


Figure 3 | Gas vesicle gene clusters from prokaryotes of different phyla. Individual genes are annotated as gvp or gvr. Their size and direction is represented by arrows. Homologous genes are colored with the same color.[15]

GV proteins and GV *structure*

GV operons vary in size, number of genes, and complexity but share a number of distinct and conserved features. All gas vesicle producing operons contain a gene that encodes the major structural protein that makes up the majority of the gas vesicle wall **(Fig 4a)**. This protein is usually annotated as *gvpA* but in some species is annotated as *gvpA1 or gvpB*. The major structural protein is a small hydrophobic protein (~8 kDa) with a characteristic alpha-beta-beta-alpha secondary structure, that assembles into a single-start low-pitch (~49 A) left-handed helical polymer (around 200 subunits per turn for GVs with 80 nm diameter), producing the gas vesicle wall with the hydrophobic beta sheets facing the gaseous lumen of the gas vesicle and hydrophilic alpha helices facing the cytosol.[16,17] The helix forms a cylindrical structure for the majority of its length but closes into a conical tip at one end. Two of these helices are joined back-to-back with their cylindrical ends facing each other, creating a cylindrical structure with biconical tips at both ends and with a discontinuity at the junction **(Fig 4b, top)**.[12,16,17]

Some GV-producing species contain a confirmed minor structural protein GvpC, encoded by the *gvpC* gene. This protein consists of variable number of repeats of amphipathic alpha helices **(Fig 4b, bottom)**. Recently, it was discovered that this protein binds to the outside of the GV structure, following the helical path created by GvpA **(Fig 4c)**.[17] This GvpA-GvpC interaction affects structural stability and acoustic properties of GVs, where GVs with GvpC bound collapse at a higher acoustic and hydrostatic pressure and are weaker nonlinear ultrasound scatterers.[13,18,19]

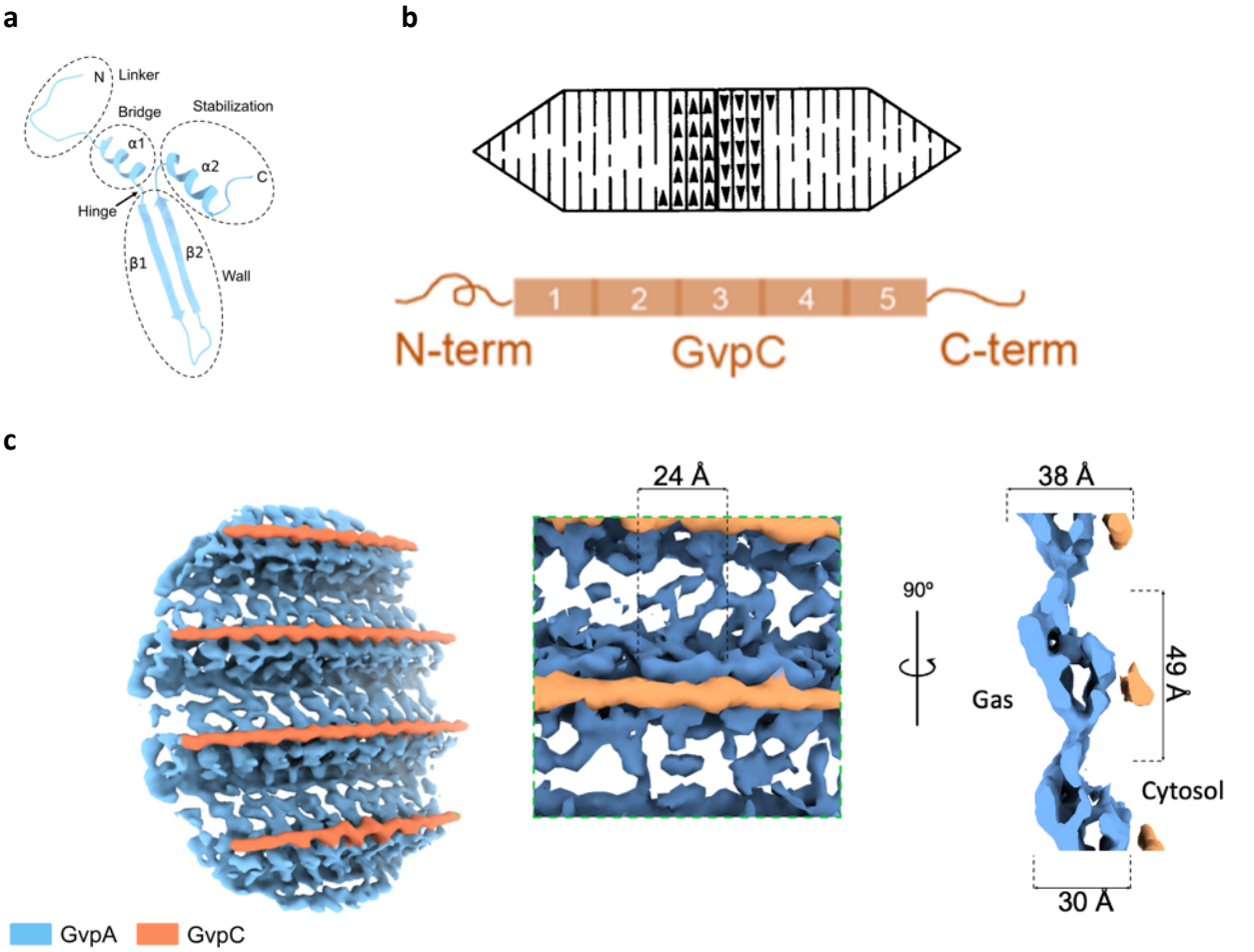


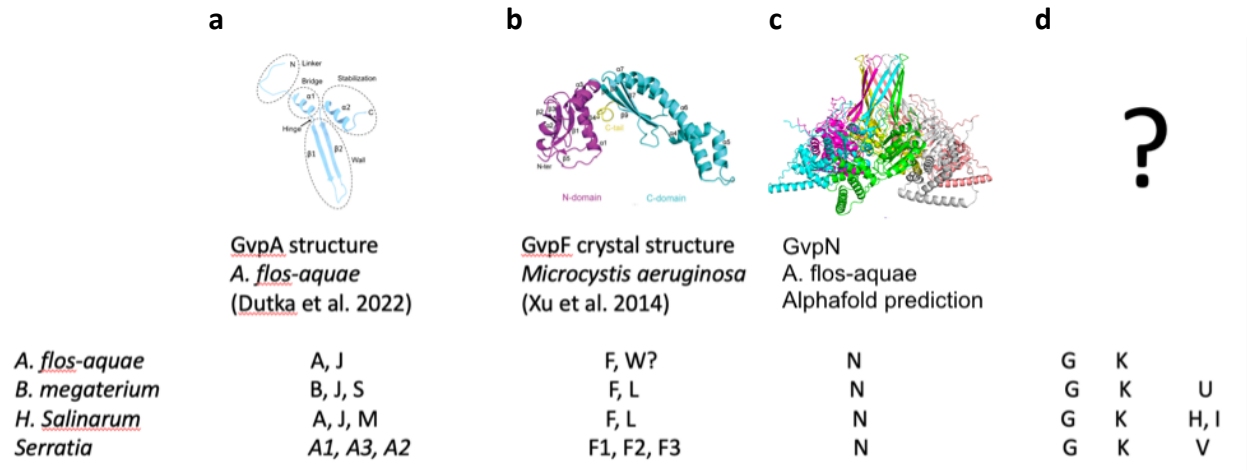
Figure 4 | Structure of gas vesicles.[17] **(a)** cryoET model of a GvpA monomer as it fits into the lattice of the GV wall. **(b)** Schematic of the helical organization of GvpA monomers (top),[12] schematic of GvpC (bottom). **(c)** GV model obtained by sub-tomogram averaging of cryoET tomograms of A. flos-aquae GVs. Blue map represents GvpA, orange map represents GvpC.

Every GV gene cluster that has been observed to produce GVs also contains GvpN, a member of the AAA+ (ATPases Associated with diverse cellular Activities) family of ATPases encoded by the *gvpN* gene **(Fig. 5c).**[20] GvpN and other AAA+ ATPases assemble into homohexamers of a doughnut shape, with a hollow channel. The primary role of AAA+ ATPases is to transform the energy derived from ATP hydrolysis into mechanical work. Specifically for GvpN, its involvement in the elongation process of gas vesicles has been firmly established.[10,15] The AlphaFold prediction for the structure of a GvpN hexamer reveals a highly unique beta-barrel hexamerization domain, which is not observed in other AAA+ ATPases. This distinctive structural feature highlights the specialized nature of GvpN within the broader family of AAA+ ATPases, suggesting a specialized mechanism of action in gas vesicle elongation. **(Fig. 5c).**

Other essential genes that seem to be common in GV clusters include one or more members of the *gvpF/L* family (GvpF is one of the few GV proteins with established structure but its role remains unclear)[21] **(Fig. 5b)** and one or more additional members of *gvpA/J/S/M* family (proteins with sequence and structural homology to the major structural protein but do not appear to be a significant constituent of the GV structure) **(Fig. 5a)**. GvpF has been reported to interact strongly

with the major structural protein GvpA in *H. salinarum*[22], suggesting that GvpF might serve as a molecular chaperone for the major structural Gvp or an intermediary chaperone between the major structural Gvp and another Gvp. Likewise, GvpL has been reported to interact with GvpJ, GvpM and GvpK, suggesting a similar interaction as between GvpA and GvpF.[22] *gvpG* and *gvpK* are essential GV genes that usually occur only once per gene cluster, their structure and function is unknown but a recent study showed a possible interaction of *B. megaterium* GvpG with *B. megaterium* GvpB **(Fig. 5d)**.[23]

GV gene clusters also contain an assortment of less conserved genes, some of which are essential in one organism but could be missing in others. For example, *gvpV* seems to be essential for *Serratia* GV elongation,[15] however, it is absent in *H. salinarum* cluster[14] or is present in *A. flos-aquae* cluster but seems to be non-essential (unpublished result) **(Fig. 5d)**.



A. flos-aquae	A, J	F, W?	N	G	K	
B. megaterium	B, J, S	F, L	N	G	K	U
H. Salinarum	A, J, M	F, L	N	G	K	H, I
Serratia	*A1, A3, A2*	F1, F2, F3	N	G	K	V

Figure 5 | Gvp families. (a) cryoET model of a GvpA monomer as it fits into the lattice of the GV wall (top). Gvps from select species that are homologs of GvpA (bottom). **(b)** Crystal structure of GvpF from *M. aeruginosa* (top).[21] Gvps from select species that are homologs of GvpF (bottom). **(c)** Alphafold prediction of a *A. flos-aquae* GvpN hexamer (top),[24] GvpN is a member of the AAA+ family of ATPases (bottom). **(d)** Gvps with no established structure or function. GvpG and GvpK seem to be present in all clusters.

1c. GVs as ultrasound contrast agents: ultrasound principles & imaging modalities

B-mode (Brightness mode)

Ultrasound imaging operates on the principle of emitting an ultrasound pulse from piezoelectric elements located on an ultrasound transducer. These same piezoelectric elements then detect the backscattered energy (echo) returning from interfaces within the body. The piezo elements convert the amplitude of the detected pressure signal into a voltage signal for processing. The key to ultrasound imaging is the interaction of sound waves with interfaces between two media that have mismatched acoustic impedances, a property defined by the product of the medium's density and the speed of sound within it. A greater mismatch in acoustic impedance results in more sound being scattered back toward the transducer, which is crucial for creating an image.

The distance to a scattering event (and thus, the depth of the object being imaged) is calculated by measuring the time elapsed between the emission of the ultrasound pulse and the detection of the echo. This calculation assumes a known speed of sound through the tissue.

Modern ultrasound systems often use multielement linear array transducers, which sequentially activate elements or groups of elements to emit and receive ultrasound signals in a controlled fashion. This process, scanning "rayline by rayline," allows for the construction of a two-dimensional image from the collected data. In this context, the brightness of a pixel in the resulting image is proportional to the amplitude of the backscattered signal, with higher amplitudes indicating stronger reflections that are typically associated with more significant impedance mismatches.

The term "B-mode" (Brightness mode) imaging refers to this conventional imaging modality, where the intensity of the echo signals is used to generate a grayscale image. B-mode ultrasound is widely used in clinical settings for visualizing anatomical structures inside the body, including the microstructures of cells and the interfaces between different types of tissue. **(Fig 6)**.

Figure 6 | B-mode. An example of a B-mode image of a routine prenatal examination. Image obtained from https://www.medphysics.wisc.edu/~block/Med_Gallery/us_baby.html

Post-collapse signal subtraction

It has been observed that purified GVs and prokaryotic cells that produce GVs can also produce B-mode contrast.[8] This is due to the GV gas interior, which creates a strong acoustic impedance mismatch with the surrounding aqueous environment. However, in applications where GVs are to be imaged against the background of sound-scattering tissue, the increase in B-mode contrast generated by GVs may not be sufficient to distinguish GVs from surrounding tissue. This is

because both GVs and tissue have a linear response of ultrasound scattering to input pressure **(Fig. 7a)**. A novel approach to isolate GV-specific signals in such scenarios involves a two-step imaging process. Initially, a B-mode image is captured to visualize the area of interest. Subsequently, a high-intensity ultrasound pulse is applied to selectively collapse the GVs, after which a second B-mode image is taken. **By subtracting the post-collapse image from the pre-collapse image, it is possible to isolate the contrast specifically attributable to GVs.** This technique effectively differentiates GVs from the background tissue by exploiting the unique, collapsible nature of GVs.

This imaging modality was notably applied by Raymond Bourdeau in a pioneering study, where he successfully visualized E. coli cells, engineered to produce GVs, within the mouse colon. The ability to subtract pre- and post-collapse images to reveal GV-specific contrast represents a significant advancement in the use of ultrasound imaging to track genetically modified prokaryotic cells and potentially other targets labeled with GVs in vivo, offering a unique tool for biomedical research and diagnostics.[2]

Amplitude modulation (AM)

To extend the utility of GV imaging beyond endpoint measurements, a nondestructive and continuous imaging modality with GV-specific contrast was needed. Fortunately, it was discovered that GVs from certain species can produce a non-linear response to input pressure, while GVs from other species can do so when GvpC protein is chemically stripped from the GV surface.[13] Indeed, all GVs have a regime of nonlinear response, however, in most cases, the non-linear regime overlaps or is beyond the collapse pressure of GVs **(Fig. 7a)**. In this case, removal of GvpC was able to shift the non-linear regime of GV scattering to sub-collapse pressures, in a range of pressures where background tissue still responds linearly. **This is thought to be a result of softening of the GV shell by the removal of GvpC making GVs more conducive to reversible bucking under lower pressures.** This buckling creates nonlinear relationship between the backscattered signals and the transmitting acoustic pressure, as well as harmonic echoes (echoes at multiples of the fundamental transmit frequency), which contribute to the nonlinear response **(Fig. 7b)**.[13,18]

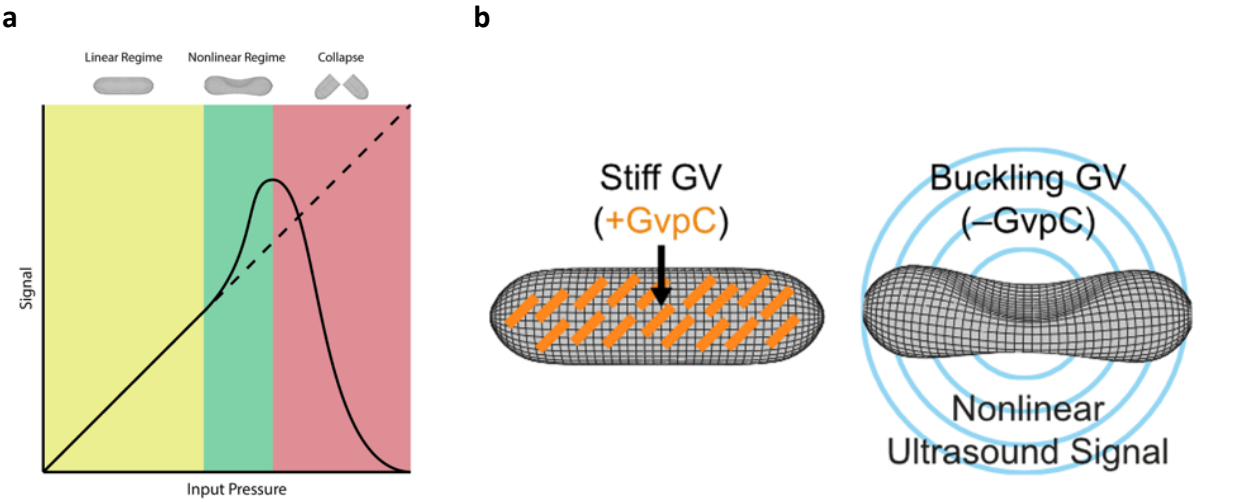


Figure 7 | Nonlinear response of modified GVs. (a) a diagram showing a linear response (dashed line) and nonlinear response (solid curve) to input acoustic pressure. **(b)** a cartoon showing GvpC strengthening of the GV wall (left). A cartoon showing reversible buckling of GVs that are not strengthened by GvpC (right).

David Maresca and colleagues took advantage of this nonlinear response of buckling GVs using an imaging modality that can distinguish GVs under the buckling regime from background tissue. The modality, called parabolic amplitude modulation (pAM),[18] involves sending a focused pulse with parabolic delays at a chosen sub-collapse pressure, followed by two pulses at half the pressure **(Fig. 8a)**. The response from the full pulse is subtracted from the sum of the responses of the two half pulses. **Because the full pulse and half pulses are within the linear response regime of background tissue, the tissue contrast is effectively cancelled, leaving only the nonlinear component of the GV buckling response (Fig 8b)**.

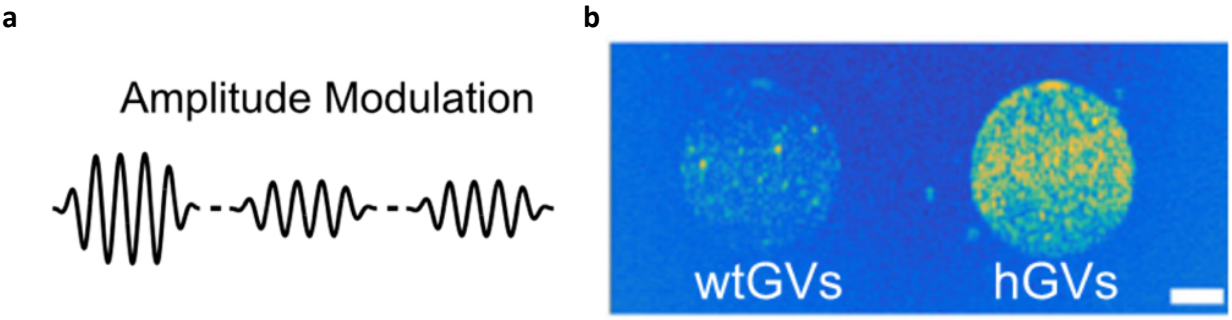


Figure 8 | Nonlinear imaging of GVs. (a) Diagram of amplitude modulation pulse sequence. Y-axis represents amplitude, X-axis represents time. A full-amplitude pulse is followed by two half-amplitude pulses. **(b)** AM image of wild-type GVs strengthened with GvpC (wtGVs) and buckling GVs stripped of GvpC (hGVs).[18] Scalebar is 1mm.

x-wave amplitude modulation (xAM)

However, the imaging depth of field with this modality is rather small as only a thin section around the ultrasound parabolic focus depth can generate GV specific contrast. Additionally, pAM is prone to imaging artifacts below the strong nonlinear scatterers **(Fig. 2b, pAM)**. This modality has since been modified by David Maresca and Danny Sawyer by splitting the aperture in half and sending two angled plane waves that cross each other in the middle of the aperture either individually (for half-amplitude transmit) or together (for full-amplitude transmit) **(Fig. 2a)**.

This modality is called x-wave amplitude modulation (xAM).[19] This change resulted in a number of very important improvements. The imaging window in the z-axis was broadened and is bound at the depth where the two planar waves miss each other, which can be extended by using a transducer with wider aperture. **Because only a thin "rayline" where the two plane waves cross each other can produce pressures that generate nonlinear response, signal from background tissue is further suppressed, resulting in much higher specificity of nonlinear GV buckling signal compared to pAM.** Also, because the wavefront produced by the cross-section moves faster than the speed of sound, xAM is much less prone to artifacts than in pAM **(Fig. 2b)**. This imaging modality is at the center of the dynamic nondestructive imaging of the second-generation ARGs *in vivo*.[1]

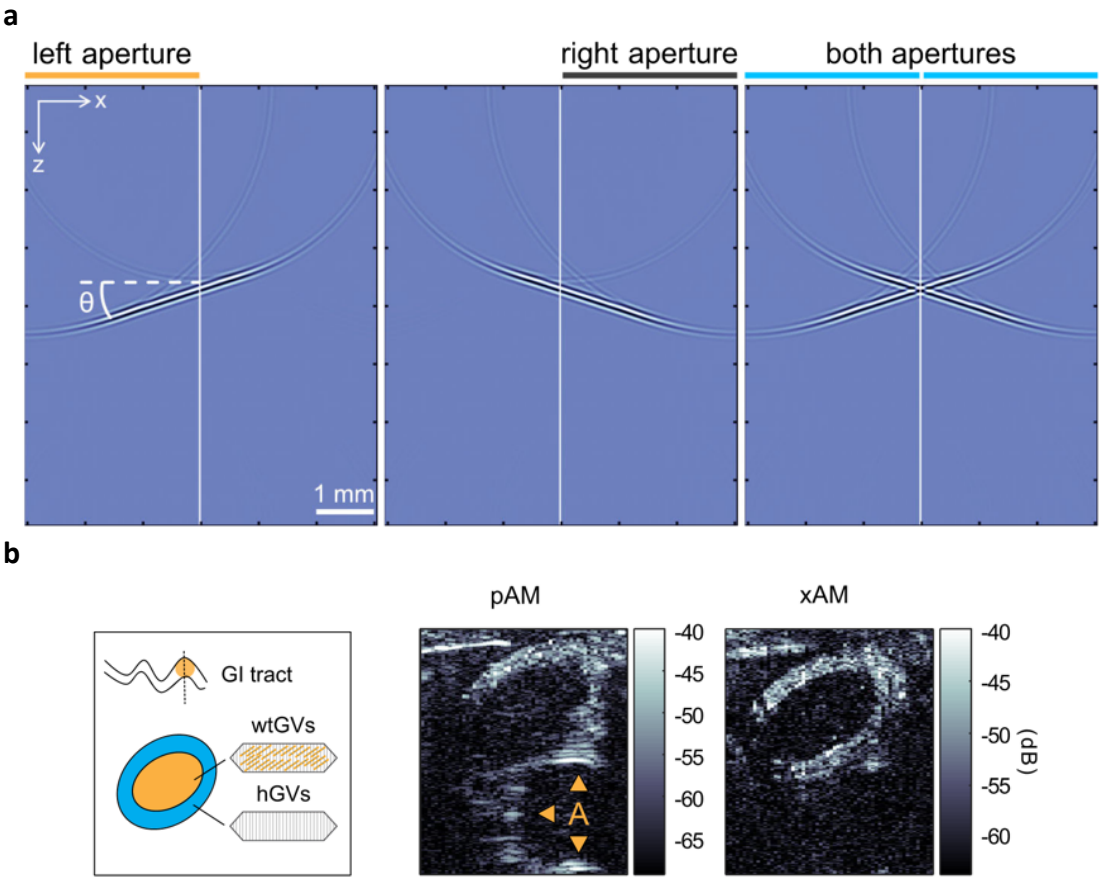


Figure 9 | Nonlinear x-wave imaging of GVs. (a) a simulation of xAM sequence for θ = 18° in water. **(b)** in vivo pAM and xAM imaging of a concentric mixture of hGVs and wtGVs in mouse GI tract. Arrows point to artifact.[19]

Burst ultrasound reconstructed with signal templates (BURST)

In some cases, it is not practical to chemically remove GvpC from GVs, especially when GVs are used as genetic reporters *in vivo*, in other cases GVs do not contain GvpC and are inherently rigid. In yet other cases, even if the properties of produced GVs are conducive to buckling, their expression might be too low. In these situations, amplitude modulation (AM) cannot be effectively used to distinguish GV signal from tissue background. However, the tissue-distinct

signal can still be achieved at the cost of destroying the GVs with high enough acoustic pressure **(Fig. 10a)**. **Danny Sawyer took advantage of this phenomenon to invent the most sensitive (albeit destructive) GV imaging modality to date, capable of single cell detection.**[25] The principle is as follows; ultrasound pressure is stepped up to a high pressure causing GVs to collapse. During the collapse, GVs buckle (and the liberated gas nanobubbles cavitate), which generates a strong linear and nonlinear response, which decays fast while tissue signal remains constant **(Fig. 10b)**. GV signal can then be isolated from the background tissue using pixelwise signal template unmixing or by pixel-wise subtraction of the first and last image frames after the pressure step **(Fig. 10c,d and e)**. This approach can also be combined with AM to filter out the linear components of signal, dramatically reducing the background at the cost of sensitivity. He named this imaging modality **b**urst **u**ltrasound **r**econstructed with **s**ignal **t**emplates (BURST),[25] and this modality was used to image the first generation mARGs *in vivo*.[3] There are a number of drawbacks of this modality; GVs are destroyed during imaging process and, similarly to pre/post-collapse signal subtraction and movement of the specimen during acquisition will generate false signal.

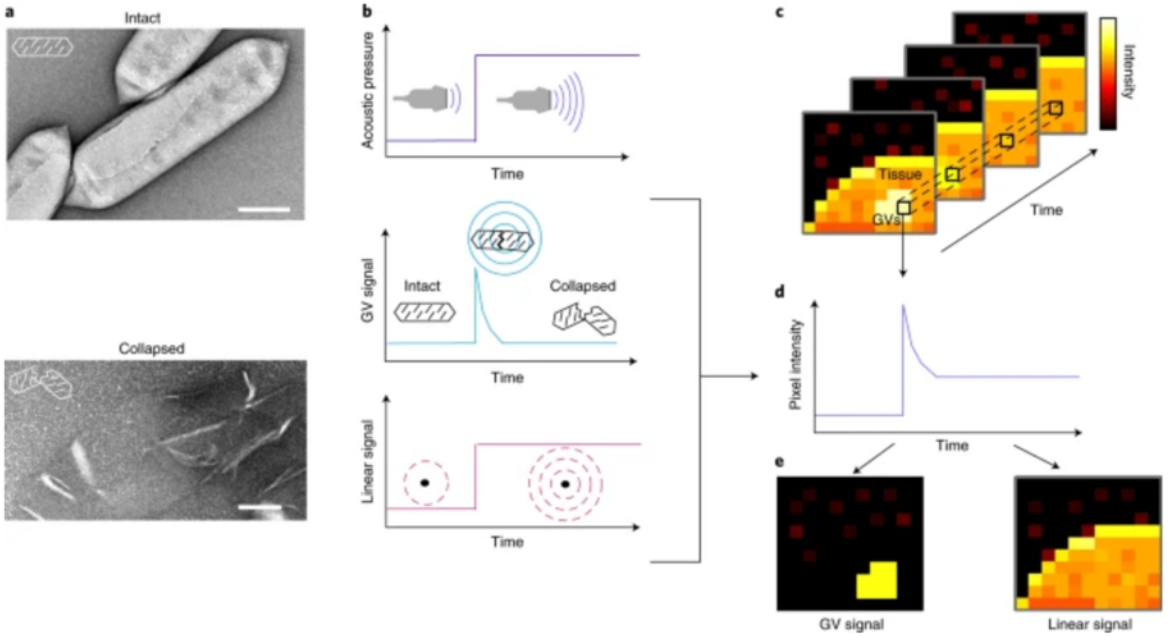


Figure 10 | BURST imaging of GVs. (a) TEM images of intact and collapsed GVs. Scale bars are 100 nm. **(b)** Illustration of the BURST pulse sequence (top) and the decomposed response of GVs (middle) and linear scatterers (bottom). **(c)** Illustrative time series of a BURST sequence response. **(d)** Illustration of the intensity over the course of BURST sequence of a pixel containing both GVs and tissue. **(e)** Illustration of the result of signal template unmixing, which separates the image of GVs from that of tissue.[25]

1d. Heterologous expression of GVs in Bacteria and Mammalian cells

Heterologous expression of GVs in bacteria

As GVs have been demonstrated to produce visible ultrasound contrast, the next obvious step was to express them in heterologous organisms. This was pursued concurrently in two separate efforts in Shapiro lab. The more direct effort was the expression of GVs in laboratory strains of *E.*

coli. The more complicated effort was porting the prokaryotic GV gene clusters into mammalian cells in a compatible way. These efforts were headed by Raymond Bourdeau[2] and Arash Farhadi[3], respectively.

Given the similarities in gene expression regulation among prokaryotes and the ability of prokaryotes to concurrently translate multiple proteins from single polycistronic mRNAs, it was reasonable to assume that GV operons could be moved directly between prokaryotic hosts without significant modifications. Ray Bourdeau and colleagues have therefore attempted to express GVs from GV gene clusters of *A. flos-aquae* (cyanobacterium) and *B. megaterium* (soil bacterium) in *E. coli*. Unfortunately, neither gene cluster could produce sufficient amount of GVs for ultrasound contrast or buoyancy. In the case of *A. flos-aquae* cluster, no GVs were detected by electron microscopy (EM) or ultrasound. **Surprisingly, a hybrid gene cluster consisting of the major structural gene *gvpA* from *A. flos-aquae* combined with the accessory *gvps* from *B. megaterium* produced *E. coli* with visible ultrasound contrast.** Interestingly, addition of the *A. flos-aquae* minor structural gene ***gvpC*** further enhanced expression. They termed this gene cluster ARG1 (now referred to as bARG1 **b**acterial **A**coustic **R**eporter **G**enes gen **1**) **(Fig 11a, b, c and d)**.

The reason for GvpC enhancement of expression is not clear, however it is noteworthy that truncation of GvpC to the first repeat only resulted in GVs with reduced collapse pressure, suggesting that strengthening of the GV shell could have a positive impact on GV expression. Having expressed GVs in laboratory *E.coli*, they set out to image these cells *in vivo*. For this, the authors engineered a probiotic strain of *E. coli* (Nissle 1917) to produce GVs and imaged them inside a mouse GI tract using pre/post-GV collapse B-mode signal subtraction **(Fig 11e)**.

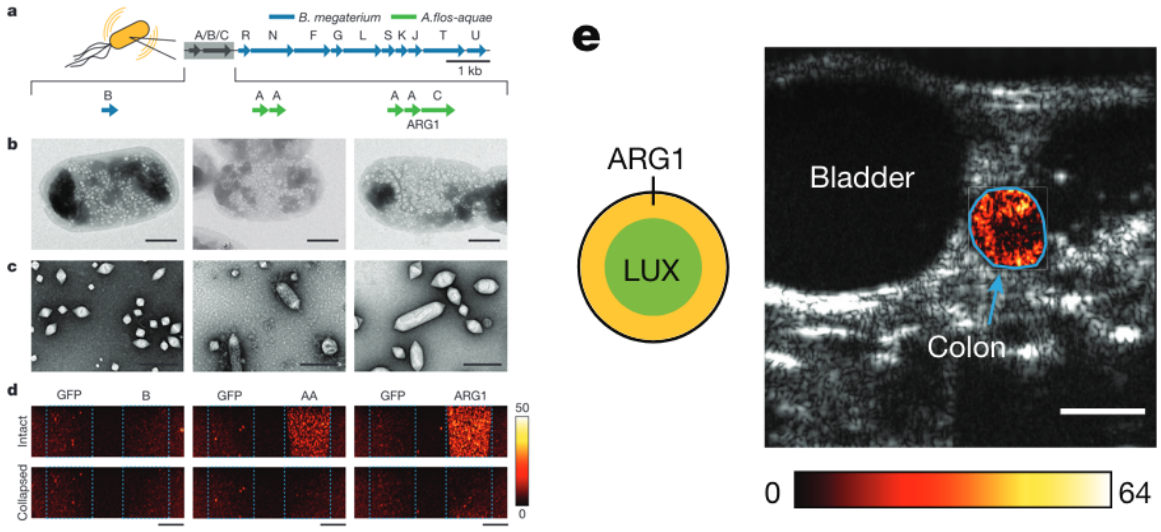


Figure 11 | Heterologous expression and imaging of bARG1. (a) Diagram of genes that were tested for heterologous expression in *E. coli*. **(b)** TEM images of cells expressing each cluster in (a). **(c)** TEM images of purified GVs from each cluster. **(d)** B-mode ultrasound images of *E. coli* expressing either GFP or gene clusters in (a) before and after acoustic collapse. **(e)** Diagram of concentric agarose colon plug with bARG1-expressing *E. coli* on the periphery and luciferase-expressing *E. coli* in the middle (left).

Post-collapse B-mode signal subtraction image (hot) overlaid over B-mode image (gray) of a mouse with the agarose colon plug (right). Scalebar is 2.5 mm.[2]

Heterologous expression of GVs in mammalian cells

The task of expressing gas vesicles (GVs) in mammalian cells presented a formidable challenge, primarily because the mechanisms of gene expression in prokaryotes and mammalian systems are fundamentally distinct. The prokaryotic expression paradigms are orthogonal to those of mammalian cells, which face obstacles such as the inability of mammalian ribosomes to initiate translation from internal open reading frames (ORFs) on mRNA, and the potential mismatch between prokaryotic codons and the preferences of mammalian translation machinery. Directly transferring prokaryotic GV operons into mammalian cells without significant adaptation was thus deemed unfeasible.

Arash Farhadi and colleagues decided to approach these problems systematically. First, to answer the most important question, are mammalian cells able to produce GVs at all, Arash cloned mammalian expression plasmids with individual mammalian codon-optimized GV genes from several prokaryotic species (*A. flos aquae, H. salinarum, B. megaterium and bARG1*) and tested them in equimolar transient cotransfection assay in HEK 293T cells **(Fig. 12a)**. **Unlike in *E. coli*, only the plasmids containing GV genes from *B. megaterium* produced detectable GVs under TEM (Fig. 12b, c)**.

The next challenge was combining all the *B. megaterium* GV genes into a compact mammalian expression cassette. To achieve this, Arash linked the *B. megaterium* GV genes (GvpNFGLSKJU) with self-cleavable viral P2A sequences, which cause the mammalian ribosome to skip a peptide bond formation when each P2A sequence is translated. This results in multiple proteins produced from a singular ribosome flux. Unfortunately, polycistronic expression via P2A has two deleterious consequences. First, since the entire P2A chain is translated using a singular ribosome flux, the relative rate of production of all proteins of the chain is locked at unity. In a system where there might be an optimal ratio of gene products, P2A chains might underproduce some of them. This was indeed confirmed to be the case with the *B. megaterium* cluster. Supplementation of the transfection mixture with an additional P2A chain containing a subset of GV genes (*gvpJFGLK*) was able to enhance expression. Second, since the self-cleavage occurs on the penultimate residue of the P2A sequence, each upstream protein is left with a string of non-native amino acids. It was confirmed that this was an issue for the major structural protein GvpB, which did not tolerate the addition of the non-native amino acids. **Nevertheless, the set of three plasmids (*gvpNFGLSKJU, gvpJFGLK* and *gvpB*) was able to produce GVs detectable by TEM.** This set of three expression cassettes was named mARG (now referred to as $mARG_{Mega}$ **m**ammalian **A**coustic **R**eporter **G**enes from *B. **mega**terium*) **(Fig. 12d, e)**.

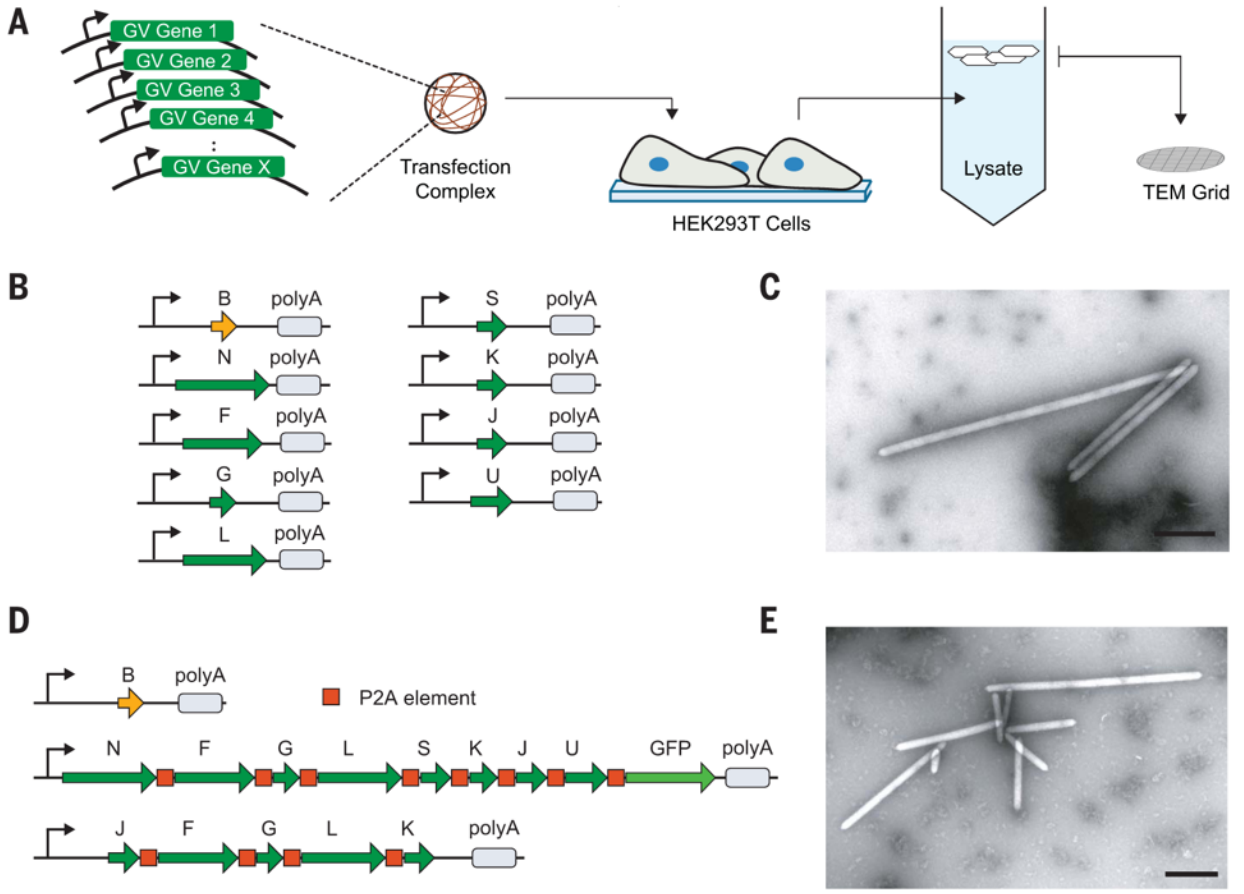


Figure 12 | Heterologous expression and imaging of mARG$_{Mega}$. (a) Diagram of transient co-transfection assay in HEK293T cells. **(a)** Diagram of B. megaterium monocistronic transient co-transfection mixture. **(c)** TEM of HEK293T lysate transfected with mixture in (b). **(d)** Diagram of mARG$_{mega}$ polycistronic transient co-transfection mixture. **(e)** TEM of HEK293T lysate transfected with mixture in (d).[3]

Unfortunately, polycistronic mARG$_{Mega}$ expression of GVs was significantly worse than coexpression of individual monocistronic *gvp* plasmids. The reason for this is unclear, however it is possible that the stochastic nature of transient cotransfection of multiple plasmids can lead to a higher fraction of cells that receive the right stoichiometry of GV genes. Creation of a stable cell line proved to be equally challenging. After PiggyBac genomic integration of mARG$_{Mega}$ into HEK293 TET-ON cells, single cells had to be isolated and screened to find a monoclonal cell line with sufficient GV expression. Gene silencing also proved to be a major problem, as the stable cell line lost expression potential over time. This problem had to be overcome by supplementation with sodium butyrate (a global histone deacetylase inhibitor), to temporarily undo gene silencing effects. Lastly, *B. megaterium* GVs have acoustic properties that are only conducive to destructive ultrasound imaging modalities such as BURST, limiting their utility to endpoint measurements. In spite of that, **Arash demonstrated that mARGs can be imaged *in vivo* even when expressed *in situ* in an engineered tumor xenograft model in mice, opening the door to vast possibilities (Fig. 13)**.

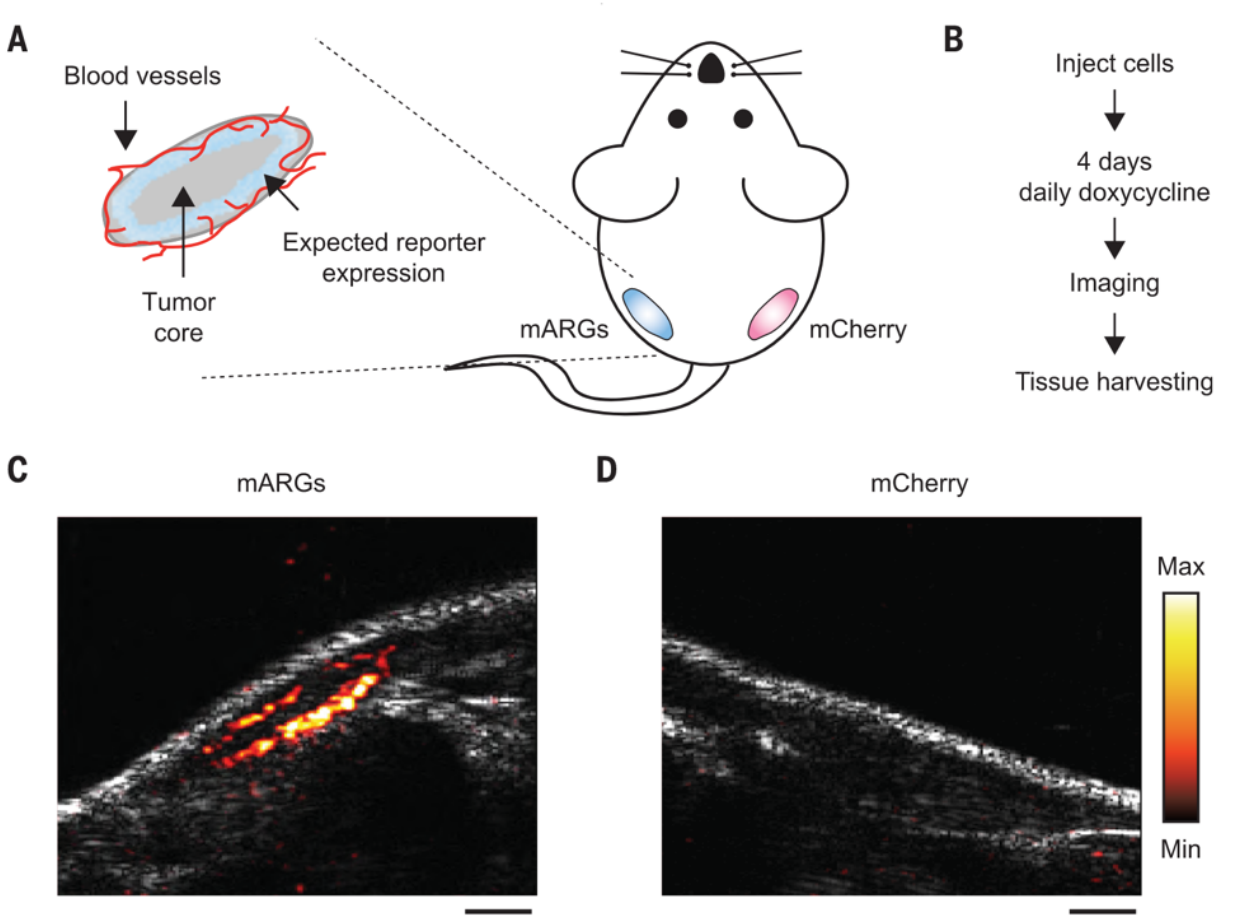


Figure 13 | In vivo imaging of mARG$_{Mega}$. (a) Diagram of a mouse implanted with HEK293T subcutaneous tumors with doxycycline inducible expression of either mARG$_{mega}$ or mCherry. **(b)** diagram of the in situ induction experiment. **(c, d)** BURST images (hot) overlaid over B-mode images (gray) of tumors expressing mARG$_{mega}$ (c) or mCherry (d).

1e. GV-based sensors

The GFP proved to be more than just a genetic reporter. Clever fluorescent protein engineering over the years yielded a vast assortment of molecular sensors capable of responding to molecular and physical phenomena on protein level. One of the examples includes Förster resonance energy transfer (FRET)-based protease sensor, where cleavage of the linker between a FRET pair of fluorophores results in a change in fluorescence.[26] Another example is the genetically encoded calcium indicator (GCaMP), a synthetic fusion of circularly permuted GFP to calmodulin and M13 peptide capable of producing fluorescence in the presence of calcium.[27] Improved variants of fluorescence proteins with greater brightness,[28,29] faster folding and maturation kinetics,[30] and shifted excitation and emission spectra[31] have also been produced in the years following the realization of the GFP potential.

Building on the transformative impact of the Green Fluorescent Protein (GFP) and its derivatives, which have been engineered into a plethora of molecular sensors capable of detecting a wide range of biological activities, our lab has embarked on a parallel journey with gas vesicles (GVs). Just as GFP was reimagined into tools like FRET-based protease sensors and calcium indicators,

GVs have been re-engineered in our latest work to serve as pioneering ultrasonic biomarkers. Our most recent manuscript introduces the first genetically encoded ultrasonic reporter of calcium (URoC), a groundbreaking innovation that leverages gas vesicles to detect calcium ions. URoC is designed to yield an enhanced nonlinear ultrasound signal in response to calcium binding, thereby enabling noninvasive ultrasound imaging of calcium signaling in live organisms. This significant advancement allows for the dynamic imaging of cellular functions in vivo with the superior depth and resolution offered by ultrasound, opening new avenues for the acoustic biosensing of various biological signals.[32]

Chapter 2. The Second Generation of Mammalian Acoustic Reporter Genes

2a. Abstract

Chapter 2 advances mammalian Acoustic Reporter Genes (mARGs) by developing the second generation ($mARG_{Ana}$), aimed at overcoming limitations such as poor expression, gene silencing, and high variability. This new generation leverages improved expression strategies, including the use of the A. flos-aquae gene cluster and tuning of gene expression ratios, resulting in enhanced ultrasound imaging contrast and stability. **Key achievements include the establishment of robust expression in various mammalian cell lines, creation of stable cell lines for tumor modeling, and demonstration of nondestructive, dynamic *in vivo* imaging capabilities.** This work represents a significant step forward in the use of mARGs for biomedical research, offering a powerful tool for real-time, non-invasive imaging of cellular processes.

2b. Introduction

The first generation bARGs and mARGs were paradigm-breaking advancements, breaking the limits of what was thought possible. **However, for ARGs to reach their full potential, number of limitations had to be addressed.** In the case of bARGs, to image them *in vivo*, preinduced cells had to be buoyancy enriched to isolate a cell population visible under ultrasound. mARGs on the other hand suffer from poor expression, gene silencing, high cell to cell variability in expression and a large genetic footprint. In both cases, GVs that were heterologously produced scattered ultrasound linearly, making them hard to distinguish from background tissue during *in vivo* imaging, necessitating the use of destructive imaging modalities, which are not conducive to dynamic imaging.

Our lab has therefore sought to create second-generation ARGs that would overcome as many of these limitations as possible. The search for better bARGs and mARGs was pursued separately but concurrently in our lab and the findings are reported in a manuscript and a research briefing at Nature Biotechnology.[1,33] The following sections pertain to the work on the second-generation mARGs, led by me with the help of colleagues. *Italicized* text and figures have been taken from the manuscript. Figures and text have been rearranged. Each *italicized* section is followed by a section with my personal notes.

2c. Results

Improving total mARG GV expression

Seeking mARGs that are expressed more robustly and produce nonlinear signal, we cloned mammalian versions of the genes contained in each of the three clusters that produced nonlinear signal in E. coli at 37˚C: Serratia, A. flos-aquae, and A. flos-aquae/B. megaterium. Equimolar transient cotransfections of the monocistronic genes derived from each gene cluster into HEK293T cells yielded detectable BURST signal only for the A. flos-aquae gene cluster and the positive control $mARG_{Mega}$ ***(Fig. 14a, b "1-fold excess").***

Given the multiple gvpA copies contained in the native A. flos-aquae GV operon,[9] we hypothesized that expressing the major structural protein GvpA at a higher stoichiometry relative to the other genes in this cluster could improve GV expression. To test this possibility, we titrated the amount

of gvpA plasmid in the A. flos-aquae plasmid set, while keeping the DNA amount corresponding to other genes constant (the total DNA level was also kept constant with a padding vector). ***We found that the BURST signal increased monotonically with increasing gvpA up to 8-fold gvpA excess (Fig. 14b)****. In contrast, the signal peaked at 2-fold excess of the homologous structural protein gvpB when expressing the B. megaterium cluster. To further improve GvpA expression, we stabilized the gvpA transcript with WPRE-hGH poly(A) elements,52 which resulted in peak signal at lower gvpA ratios* ***(Fig. 14 c, d).***

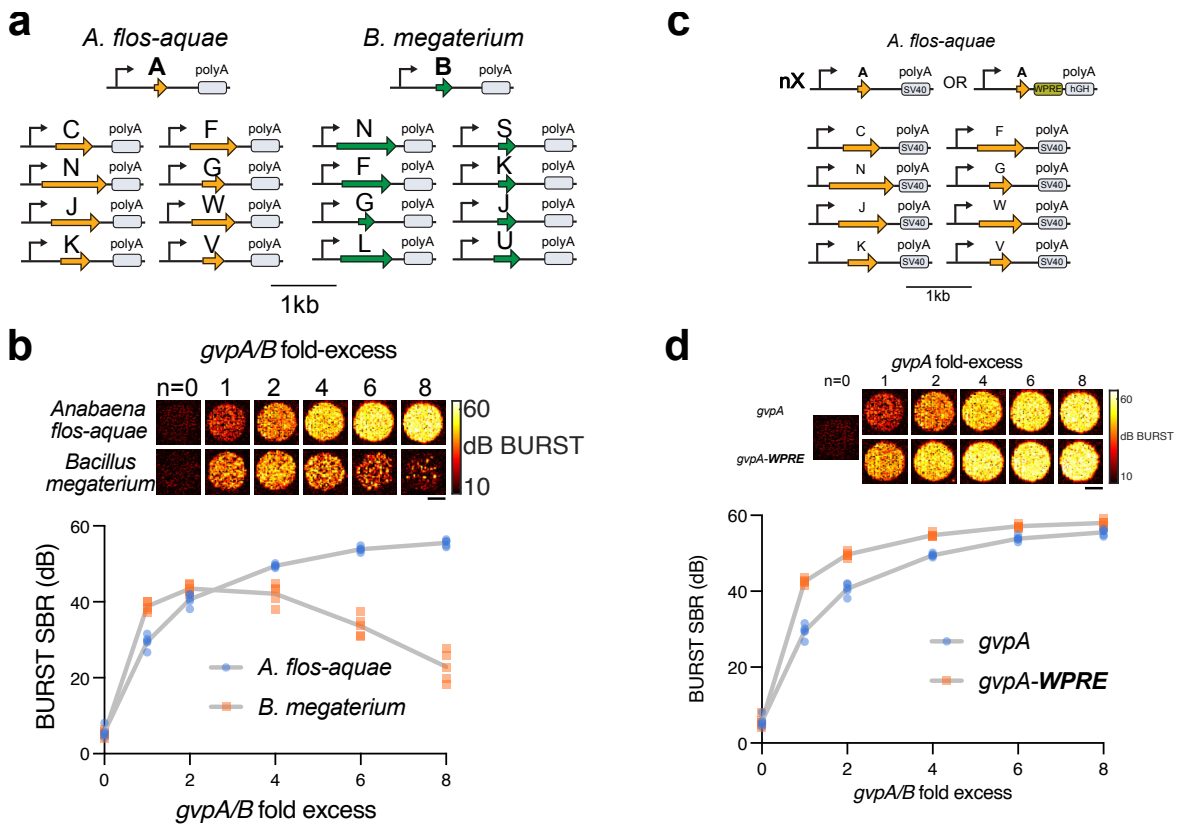


Figure 14 \| Heterologous expression of the A. flos-aquae GV gene cluster in mammalian cells. (a) *Schematic of the codon-optimized monocistronic plasmid sets used in this study.* ***(b)*** *Representative BURST images (top) and SBR quantification (n=5, bottom) of transient GV expression in HEK293T cells 3 days after co-transfection of mixtures from (a) with varying gvpA/B fold excess relative to their respective assembly factor plasmids.* ***(c)*** *Schematic of the codon-optimized A. flos-aquae monocistronic plasmid set with gvpA transcript stabilized with WPRE.* ***(d)*** *Representative BURST images (top) and SBR quantification (n=5, bottom) of transient GV expression in HEK293T cells 3 days after co-transfection of mixturesin (c) with varying gvpA fold excess relative to their respective assembly factor plasmids, with and without WPRE elements on the gvpA DNA. Gray lines connect the means of replicates.*

In this section we sought to find a different GV operon to express in mammalian cells. The *A. flos-aquae* GV operon has always been desirable due to its minimal genetic footprint (only 7 essential genes) and due to its potential as an engineering platform. (cite some papers here). Arash Farhadi has unsuccessfully tried to express the *A. flos-aquae* GV operon in mammalian cells when he was working on the first generation mARGs, however, the only expression readout he had at his disposal was TEM of buoyancy-purified lysates. As we have seen in this section, the initial equimolar expression approach was unlikely to produce enough GVs detectable by TEM. **Fortunately, now we had BURST ultrasound imaging at our disposal, which was able to detect**

the low basal expression of *A. flos-aquae* GVs in intact cells. The key to improving expression, which ultimately exceeded *B. megaterium* expression, was supplying sufficient major structural protein GvpA. It is interesting to note that this approach did not work on *B. megaterium* cluster, as higher excess of the major structural protein GvpB lead to decrease in expression. The reason for this is unclear and the investigation of this phenomenon is subject of later specific aims.

Nonlinear contrast for continuous dynamic nondestructive imaging

*We next looked for nonlinear ultrasound contrast from A. flos-aquae GVs to enable their non-destructive detection versus background. GvpC is a minor structural protein in A. flos-aquae GVs that binds to and mechanically reinforces the GV shell **(Fig. 15a)**.*[13] *in vitro chemical removal of GvpC from purified GVs was previously shown to enhance nonlinear ultrasound scattering by allowing the GVs to deform more strongly and nonlinearly in response to acoustic pressure.*[13,34,35] ***When we omitted gvpC from our mammalian cotransfection mixture, we observed a dramatic enhancement of nonlinear signal in xAM imaging, with the peak signal produced at around 0.6 MPa (Fig. 15b).** By comparison, transfections including gvpC produced a much weaker xAM signal, while B. megaterium plasmids and GFP-expressing cells did not produce appreciable nonlinear contrast at any pressure. The omission of gvpC did not appreciably alter BURST contrast **(Fig. 15c)**. These results indicated that the mammalian GVs derived from A. flos-aquae can provide strong, nondestructive, nonlinear ultrasound contrast.*

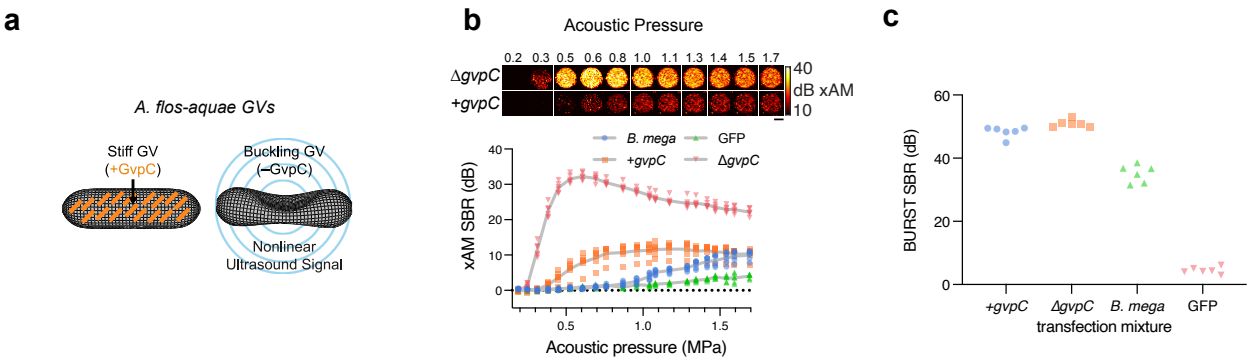


***Figure 15 | Nonlinear contrast in mammalian cells.* (a)** *Diagram of GV structure with GvpC highlighted in orange. When GvpC is absent, GVs buckle reversibly.* **(b)** *Representative xAM ultrasound images (top) and SBR quantification (n=6, bottom) of transient co-transfection experiments of A. flos-aquae GV plasmids (4-fold gvpA excess) with and without gvpC at varying acoustic pressures. B. megaterium GV (at 2-fold gvpB excess) and GFP expression is included for quantitative comparison.* **(c)** *BURST SBR quantification (n=6) of 293T cells expressing constructs tested in* (b)

Polycistronic expression of mARG2.0

*To create a convenient vector for mammalian expression of A. flos-aquae GVs, we constructed a polycistronic plasmid linking the assembly factor genes gvpNJKFGWV through P2A co-translational self-cleavage elements. gvpA was supplied on a separate plasmid to enable stoichiometric tuning. The gvpA and gvpNJKFGWV plasmids were labeled with IRES-BFP and P2A-GFP, respectively, to allow for fluorescent analysis and sorting. Both plasmids were driven by CMV promoters, and their transcripts were stabilized with WPRE-hGH poly(A) elements. We termed this pair of plasmids mARG$_{Ana}$ – **m**ammalian **ARG**s adapted from **Ana**baena flos-aquae **(Fig. 16a)**.*

mARGAna produced robust GV expression and ultrasound contrast in HEK293T cells transiently co-transfected with a 1- to 6-fold excess of gvpA (Fig. 16b).

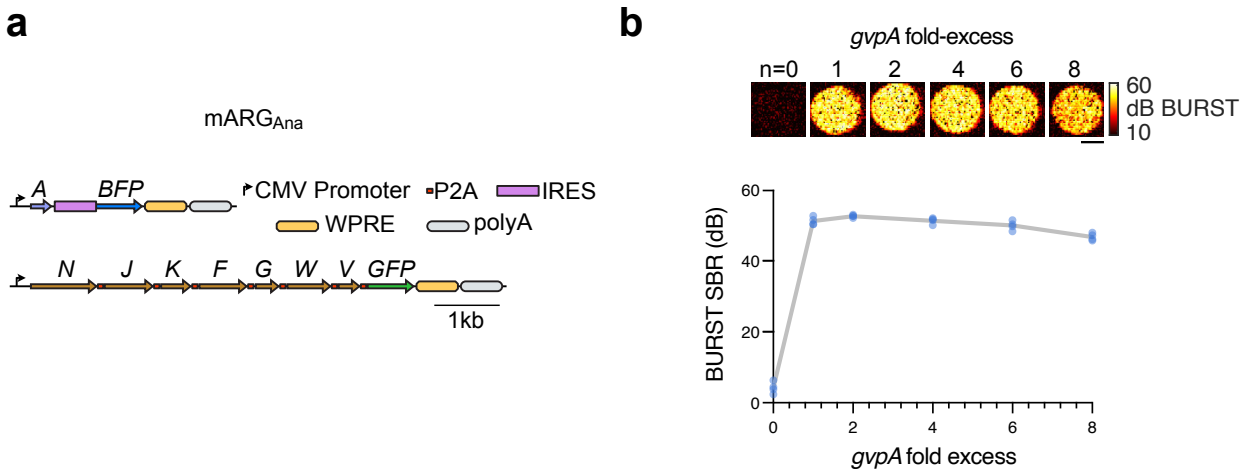


Figure 16 | Design and imaging of polycistronic mARG$_{Ana}$ in vitro. ***(a)*** *Schematic of the mARGAna polycistronic plasmids.* ***(b)*** *Representative BURST images (top) and SBR quantification (n=4, bottom) of transient GV expression in HEK293T cells 3 days after co-transfection of mARGAna mixtures with varying gvpA-fold excess relative to the assembly factor plasmid.* Gray lines connect the means of replicates.

Omission of *gvpC*, in hindsight, seems like an obvious way to generate nonlinear contrast. However, it should be noted that *gvpC* omission in bARG significantly abrogated GV expression.[2] The reason why *gvpC* omission in mammalian cells does not reduce expression is unclear. One of the theories speculates that because bacterial cells have a cell wall, the increased turgor pressure necessitates GvpC for structural reinforcement, whereas mammalian cells do not have turgor pressure. Further studies are needed to conclusively establish the model for this effect.

Second thing to note is that the mARG$_{Ana}$ polycistronic plasmid set produces levels of GV expression comparable to monocistronic cotransfection, which was not the case for mARG$_{Mega}$, which saw significant reduction in expression when expressed from polycistronic cassettes. The mARG$_{Ana}$ polycistronic plasmid set also does not require a booster cassette, significantly reducing its genetic footprint. Interestingly, the peak expression in polycistronic expression was achieved at 2-fold *gvpA* excess, after which expression seemed to plateau with a slight downward trend, mimicking the effect seen in *gvpB* titration experiment in the *B. megaterium* cluster. This could potentially signify that one or more accessory proteins in the polycistronic mARG$_{Ana}$ (and in *B. megaterium* cluster) could be limiting, causing *gvpA* expression to exceed the "capacity" of the assembly factors beyond 2-fold excess. **Unpublished results indicate that expression at higher (>2-fold) *gvpA* molar amounts can be improved by increasing the molar amount of the *pNJKFGWV* assembly factor plasmid.** Identification of the specific subset of limiting assembly factors could provide insight into the mechanism of GV assembly and inspire improved mARG$_{Ana}$ genetic design.

Creation of stable mammalian cell lines capable of conditional expression of mARG$_{Ana}$

One of the most common applications of mammalian reporter genes is the visualization of tumor growth in animal models of cancer, which are a critical platform for basic oncology research and

the development of new treatments. To produce a stable cancer cell line expressing mARG$_{Ana}$, we cloned our polycistronic constructs into PiggyBac integration plasmids under a doxycycline-inducible TRE promoter.[36,37] *As a clinically relevant cancer model, we chose the human breast cancer cell line MDA-MB-231, which is widely used in xenograft studies. We engineered these cells to constitutively express the rtTA transactivator and Antares optical reporter,*[38] *then electrically transduced them with a mixture of mARG$_{Ana}$ and PiggyBac transposase plasmids at a 2:1 (gvpA:gvpNJKFGWV) molar ratio, and fluorescently sorted for co-expression of Antares, GFP and BFP* ***(Fig. 17a)****.* ***The resulting polyclonal MDA-MB-231-mARG$_{Ana}$ cell line showed xAM contrast after a single day of doxycycline induction, which increased substantially through day 6 (Fig. 17b)****. We confirmed the expression of GVs in these cells by electron microscopy* ***(Fig. 17c).***

The ultrasound signal increased steeply with increasing doxycycline doses up to 1 µg/mL, above which the signal saturated ***(Fig. 17d)****. xAM was detected from induced cells starting from an acoustic pressure of 0.31 MPa, whereas uninduced control cells did not produce signal at any pressure* ***(Fig. 17e).*** *At pressures above 0.42 kPa, the xAM signal gradually decreased over time, indicating the partial collapse of GVs. We chose 0.42 MPa as the xAM imaging pressure for subsequent experiments, providing the optimal balance of signal stability and signal strength.*

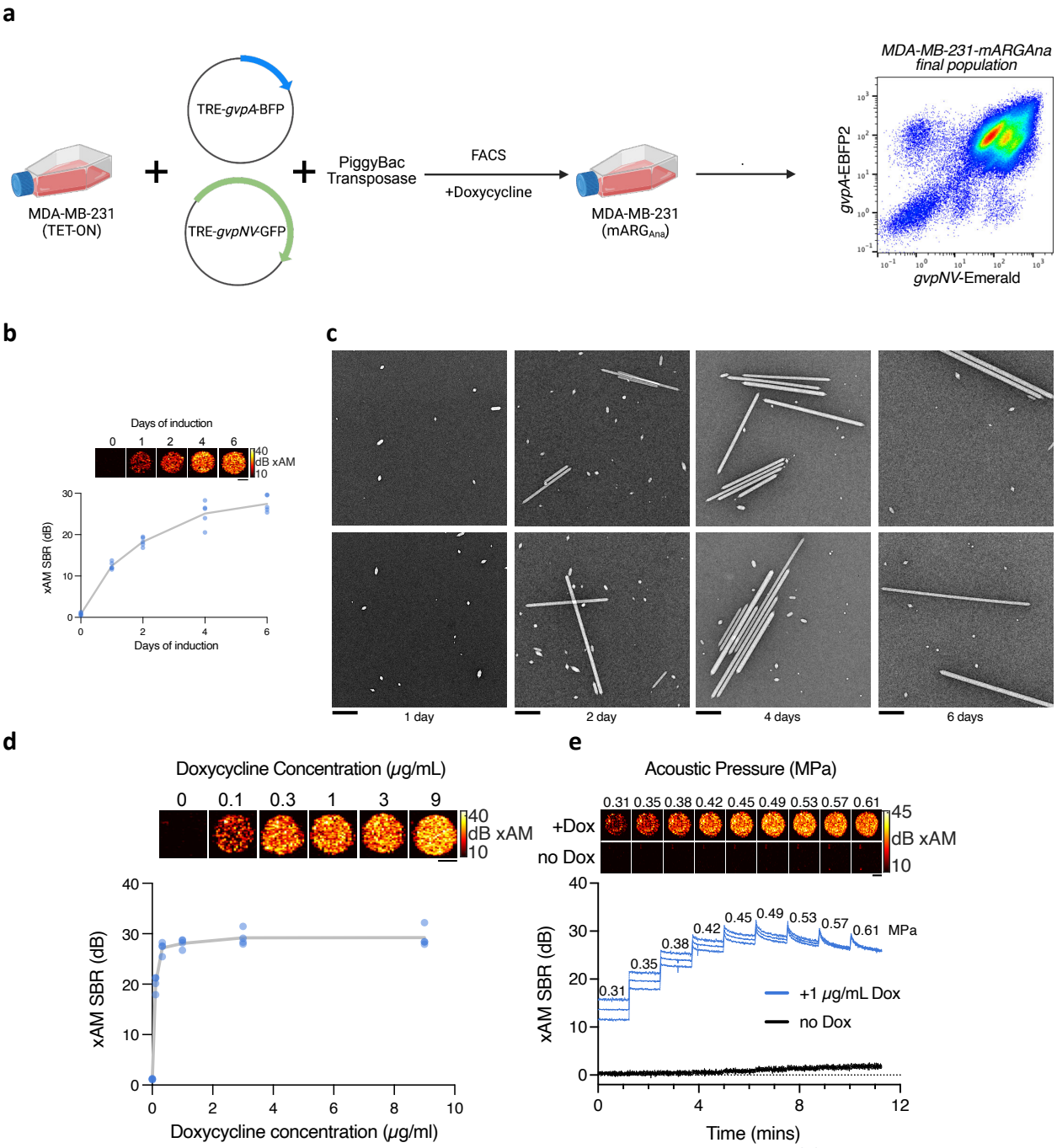


Figure 17 | Creation of mARG$_{Ana}$-expressing stable cell line and their characterization* in vitro.** ***(a) *Schematic of MDA-MB-231-mARGAna engineering (created with BioRender.com and FlowJo). The final population was ~95% double positive for gvpA and gvpNJKFGWV expression.* ***(b)*** *Representative xAM images (top) and SBR quantification (n=5, bottom) of MDA-MB-231-mARGAna cells at 0.54 MPa after 1, 2, 4 and 6 days of 1 μg/ml doxycycline induction.* ***(c)*** *TEM images of GVs purified from MDA-MB-231-mARG$_{Ana}$ detergent lysates* from (b). *Scalebars represent 0.5 μm.* ***(d)*** *Representative xAM images (top) and SBR quantification (n=4, bottom) of MDA-MB-231-mARGAna cells at 0.42 MPa as a function of doxycycline concentration after 4 days of expression.* ***(e)*** *Representative xAM images (top) and SBR quantification (n=4, bottom) of induced and uninduced MDA-MB-231-mARGAna cells as a function of time under varying acoustic pressures.*

The major advantage of ultrasound imaging over optical methods is the deep penetration depth of ultrasound waves. To ensure that mARG$_{Ana}$ expression can be detected in deep tissue, we imaged pre-induced MDA-MB-231-mARG$_{Ana}$ cells using xAM under >1cm of beef liver tissue or under PBS ***(Fig. 18a)****. The results show that xAM imaging of mARG$_{Ana}$ is highly sensitive, even as liver tissue attenuates acoustic pressure, resulting in a right-shifted pressure response curve. Moreover, the signal is specific given the low contrast generated by the liver tissue compared to MDA-MB-231-mARG$_{Ana}$ cells, even when imaged under deep tissue* ***(Fig. 18b)****.*

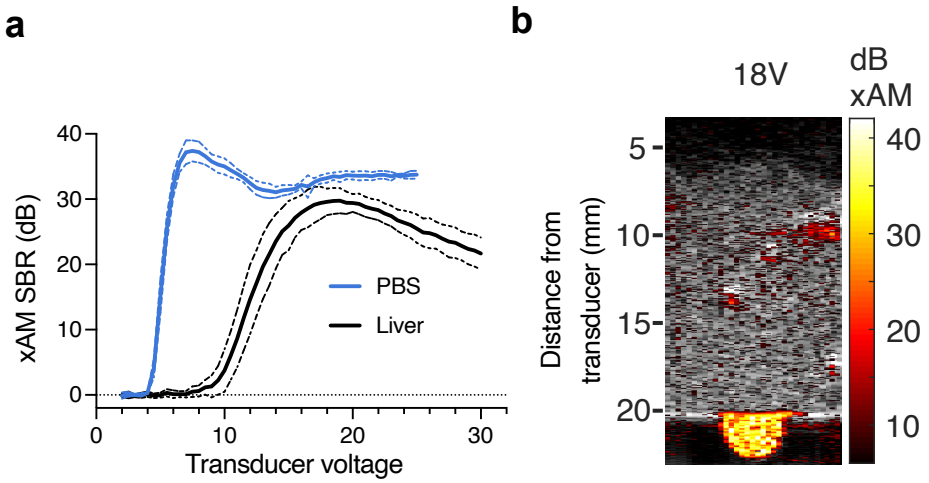


Figure 18 | Imaging of mARG$_{Ana}$-expressing stable cell line under deep tissue. **(a)** *Quantification of xAM SBR of MDA-MB-231-mARGAna cells imaged under >1 cm of beef liver tissue (n=7) or under PBS (n=4) as a function of transducer voltage. Thick lines represent the mean of replicates and thin lines represent ± standard deviation.* **(b)** *Representative xAM/B-mode overlay of MDA-MB-231-mARGAna cells imaged under liver tissue using 18V transducer voltage for xAM.*

To ensure that mARG$_{Ana}$ expression is not limited to HEK293T and MDA-MB-231 cells, we also engineered 3T3 mouse fibroblast cells and HuH-7 human hepatocarcinoma cells to produce GVs. We then decided to determine the sensitivity of nondestructive mARG$_{Ana}$ imaging by determining the lowest number of cells that can be imaged using xAM with each cell line ***(Fig. 19a)****. Nondestructive xAM imaging sensitivity of both mARG$_{ana}$ cell lines (30K cells/mL) surpassed that of the lowest reported cell concentration of mARG$_{mega}$ cells (2.5M cells/mL) imaged using destructive BURST imaging.[3] However, BURST imaging still proved to be the more sensitive imaging method, detecting cell concentrations down to 300 cells/mL* ***(Fig. 19b)****.*

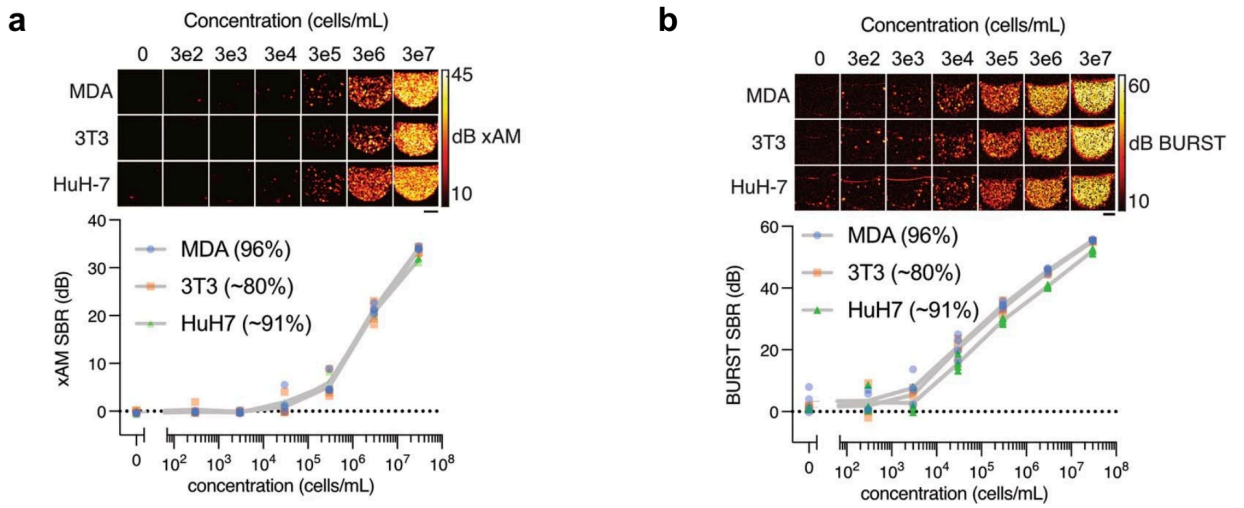


Figure 19 | Additional mARG$_{Ana}$-expressing stable cell line and limits of US detection. **(a)** *Representative xAM images (top) and SBR quantification (n=4, bottom) of induced MDA-MB-231-mARG$_{Ana}$, 3T3-mARG$_{Ana}$ and HuH-7-mARG$_{Ana}$ cells at 0.61 MPa as a function of cell concentration.* **(b)** *Representative BURST images (top) and SBR quantification (n=4, bottom) of induced MDA-MB-231-mARG$_{Ana}$, 3T3-mARG$_{Ana}$ and HuH-7-mARG$_{Ana}$ cells as a function of cell concentration. For* (a) *and* (b)*, gray lines connect the means of the replicates.*

In this section we saw that mARG$_{Ana}$ expression is robust across various cell lines as we have demonstrated strong expression in 4 unrelated cell lines. Second, mARG$_{Ana}$ expression also has a lower cell-to-cell expression variability as bulk FACS gating strategy, as opposed to single cell sorting, was able to produce polyclonal cell lines that are robust GV producers, contrasting mARG$_{Mega}$, which required single cell sorting to isolate a monoclonal cell line that was able to produce GVs robustly. Third, mARG$_{Ana}$ is more genomically stable as the expression in neither case necessitated supplementation with sodium butyrate.

The gene silencing problem

mARG$_{Ana}$ expression is not limited to these 4 cell lines. Other members of our lab have since been able to produce their own cell lines capable of robust mARG$_{Ana}$ expression including THP-1 human monocytic leukemia, U87 human glioblastoma or N2a mouse neuroblastoma. It is important to also note that many attempts at engineering cell lines to produce mARG$_{Ana}$, including INS-1, RAW 264.7, did not yield robust mARG$_{Ana}$ expressing cell lines. In many cases the major problem is gene silencing, where expression is lost over time after FACS or never attains high expression. In some of these cases, expression can be boosted by the addition of epigenetic agents such as sodium butyrate or 5-aza-2'-deoxycytidine. Some of the variability in engineering success could be caused by variations in engineering approaches (PiggyBac vs viral transduction, differences in the mARG$_{Ana}$ cluster design, different FACS gating strategies, etc.). Genomic or metabolic differences between different cell lines are also suspected as a potential contributor to this variability. However, general principles behind this variability remain unclear. **A systematic study to explain and address this variability is absolutely essential if mARGs are to be used as a general tool and for reliable delivery of mARGs to primary cells and systemically *in vivo*.**

As part of the investigation into gene silencing, we created a HEK293T cell line that constitutively expresses mARG$_{Ana}$ driven by dual CMV promoters. As perhaps expected, GV expression was dramatically boosted by the addition of sodium butyrate as evidenced by BURST

(Fig. 20a). **What was surprising to see is that sodium butyrate (NaB) did not have symmetrical effect on the two mARG cassettes**, *gvpA*-mCherry and *gvpNJKFGW*-GFP, as the expression of the former remained relatively unchanged while the expression of the latter improved 40-fold after NaB treatment **(Fig. 20b)**. A similar effect was observed when this experiment was performed with the HEK293-mARG$_{Mega}$ cell line. NaB had a much more dramatic effect on the *gvpJFGLK* booster cassette expression than on the other two cassettes **(Fig. 20c)**. However, it should be noted that the fluorescence reporter for *gvpJFGLK* was driven by a bidirectional promoter in the opposite direction of the *gvp* genes **(Fig. 20d)**.

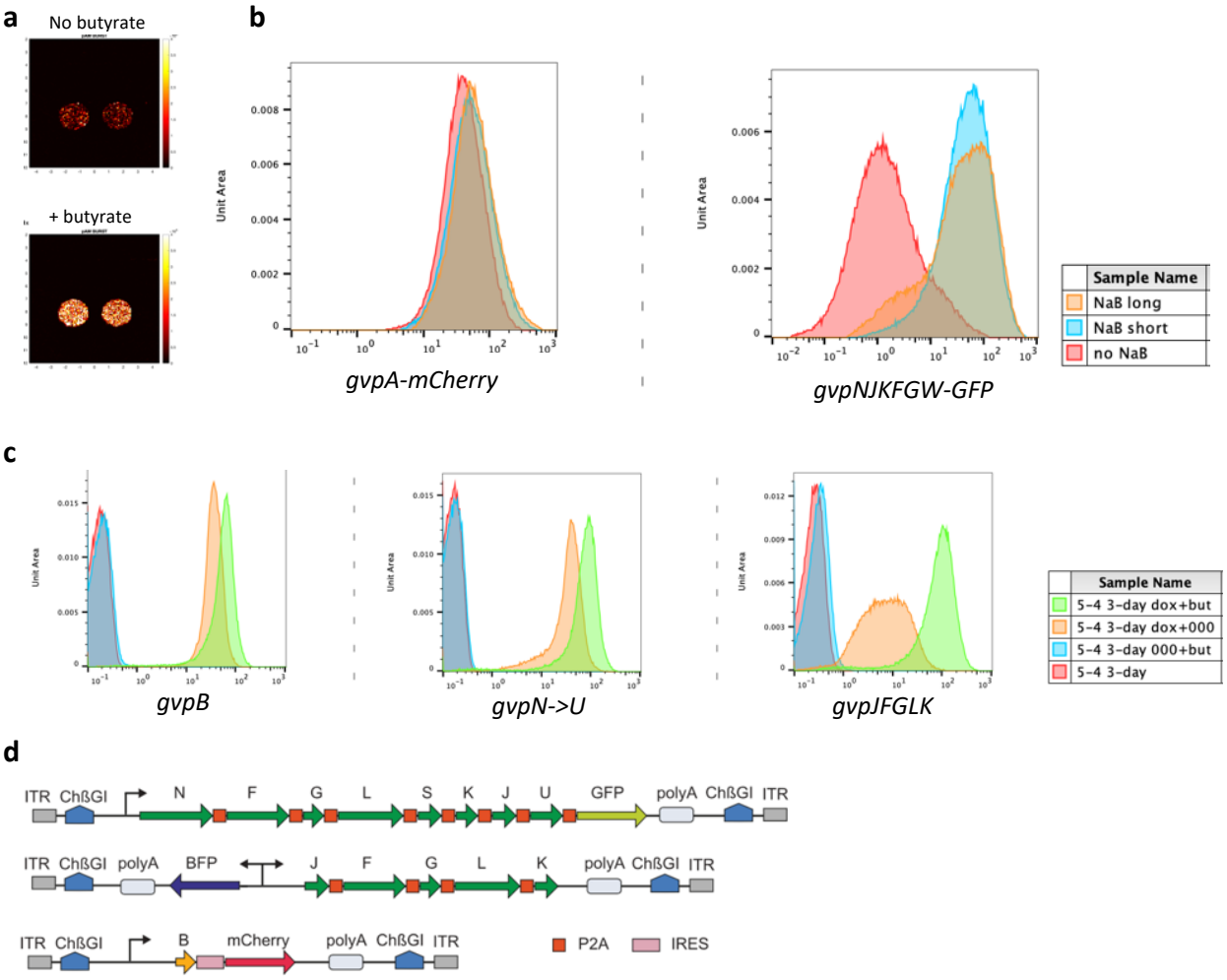


Figure 20 | Sodium butyrate boosts genome-integrated expression cassettes assymetrically. (a) BURST images of HEK293T cells constitutively expressing mARG$_{Ana}$ in the absence (top) or presence (bottom) of NaB. **(b)** Flow cytometry analysis of cells in (a) treated with either no NaB, 3-day NaB (NaB short) or 5-day NaB (NaB long). **(c)** Flow cytometry analysis of HEK293TET-ON mARG$_{mega}$ expressing cells treated for three days with either doxycycline only (dox+000), NaB only (000+but), both (dox+but) or neither (000+000). **(d)** diagram of the mARG$_{mega}$ expression cassettes of cells in (c).

In vivo nondestructive and dynamic imaging of mARGs

We next tested the ability of mARGAna to image the spatial distribution of gene expression in tumor xenografts in living mice. We formed orthotopic tumors by injecting MDA-MB-231-mARG$_{Ana}$ cells bilaterally in the fourth mammary fat pads of female immunocompromised mice.

The mice were then split into doxycycline-induced and uninduced groups. We acquired ultrasound images of the tumors as they grew, with 3 imaging sessions distributed over 12 days ***(Fig. 21a)****.* ***All induced tumors produced bright and specific xAM contrast starting from the first timepoint (day 4), whereas the uninduced tumors did not (Fig. 21b)****. The acquisition of adjacent planes allowed 3D visualization of expression patterns* ***(Supplementary Video 1)****. The nonlinear xAM signal was highly specific to the viable tumor cells, being absent outside the anatomically visible tumor boundaries and within the tumor cores of the larger tumors imaged at 8 and 12 days.*

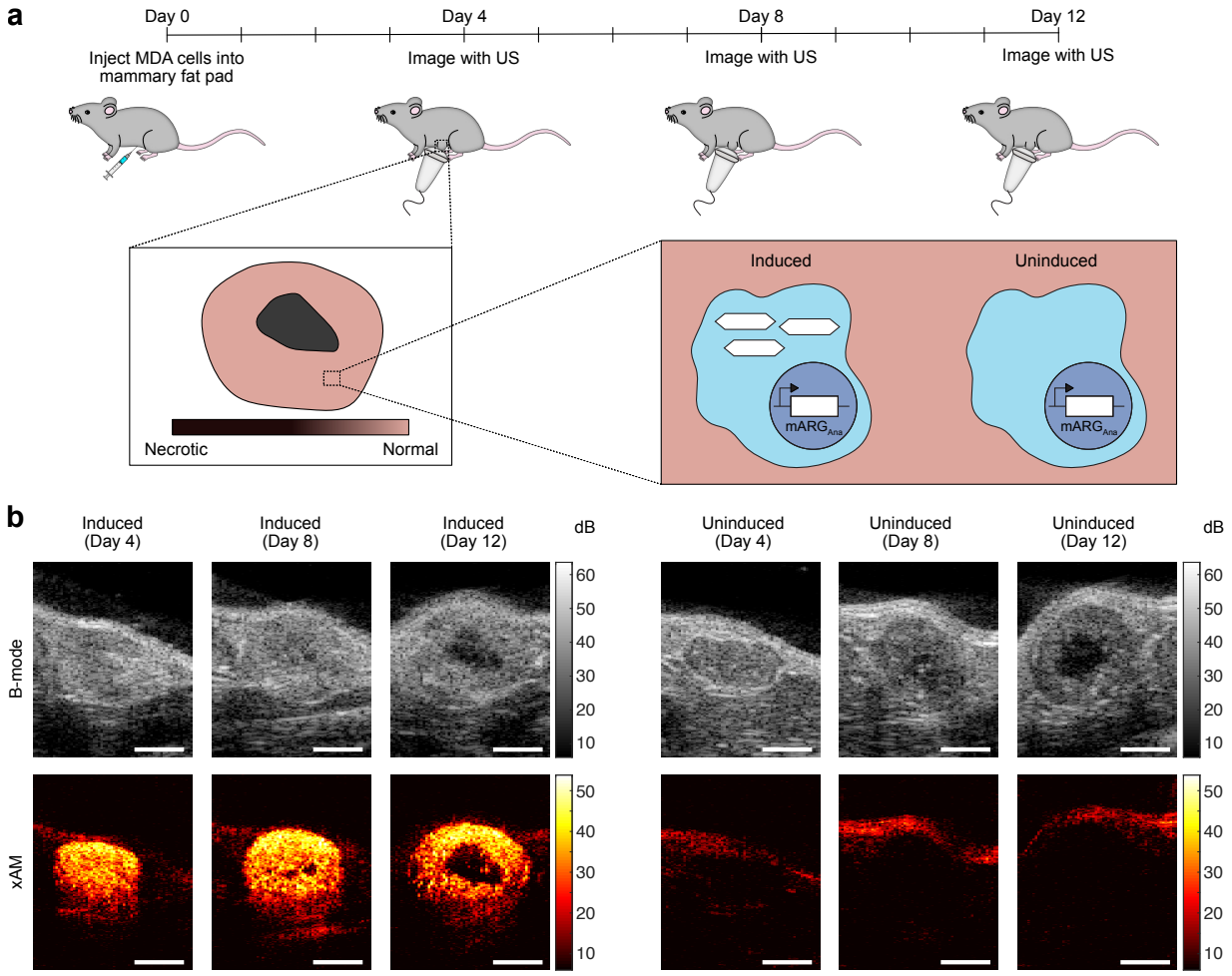


Figure 21 | In situ mARG$_{Ana}$ expression enables nondestructive ultrasound imaging of orthotopic tumors. (a) *Diagram of the in vivo protocol for assessing in situ mARG$_{Ana}$ expression in orthotopic tumors. Mice were injected bilaterally in 4th mammary fat pads with engineered MDA-MB-231-mARGAna human breast adenocarcinoma cells on day 0. mARG$_{Ana}$ expression was induced by regular intraperitoneal (IP) doxycycline injections starting from the day of tumor injections. Tumors were imaged with ultrasound after 4, 8 and 12 days of expression.* ***(b)*** *Representative middle sections of B-mode and xAM ultrasound tomograms of MDA-MB-231 mARG$_{Ana}$ tumors induced with doxycycline (left) and uninduced control (right) imaged on day 4, 8 and 12. Scalebars are 2 mm.*

The observed spatial pattern of gene expression activity in these tumors was corroborated by fluorescence microscopy of formalin-fixed tumor sections obtained from euthanized mice on day 12 ***(Fig. 22a, b)****, confirming the ability of mARG$_{Ana}$ to report microscale patterns of gene expression noninvasively in intact animals. In contrast, in vivo fluorescence images of the mice*

lacked information about the spatial distribution of gene expression within the tumor ***(Fig. 22c)****. We quantified GV expression over time as a volumetric sum of xAM signal over all acquired image planes* ***(Fig. 22d)****. The induced tumors had significantly higher total signal than the uninduced controls at all time points.* ***These experiments demonstrate the ability of mARG$_{Ana}$ to serve as a highly effective reporter gene for the noninvasive monitoring of tumor growth and gene expression activity.***

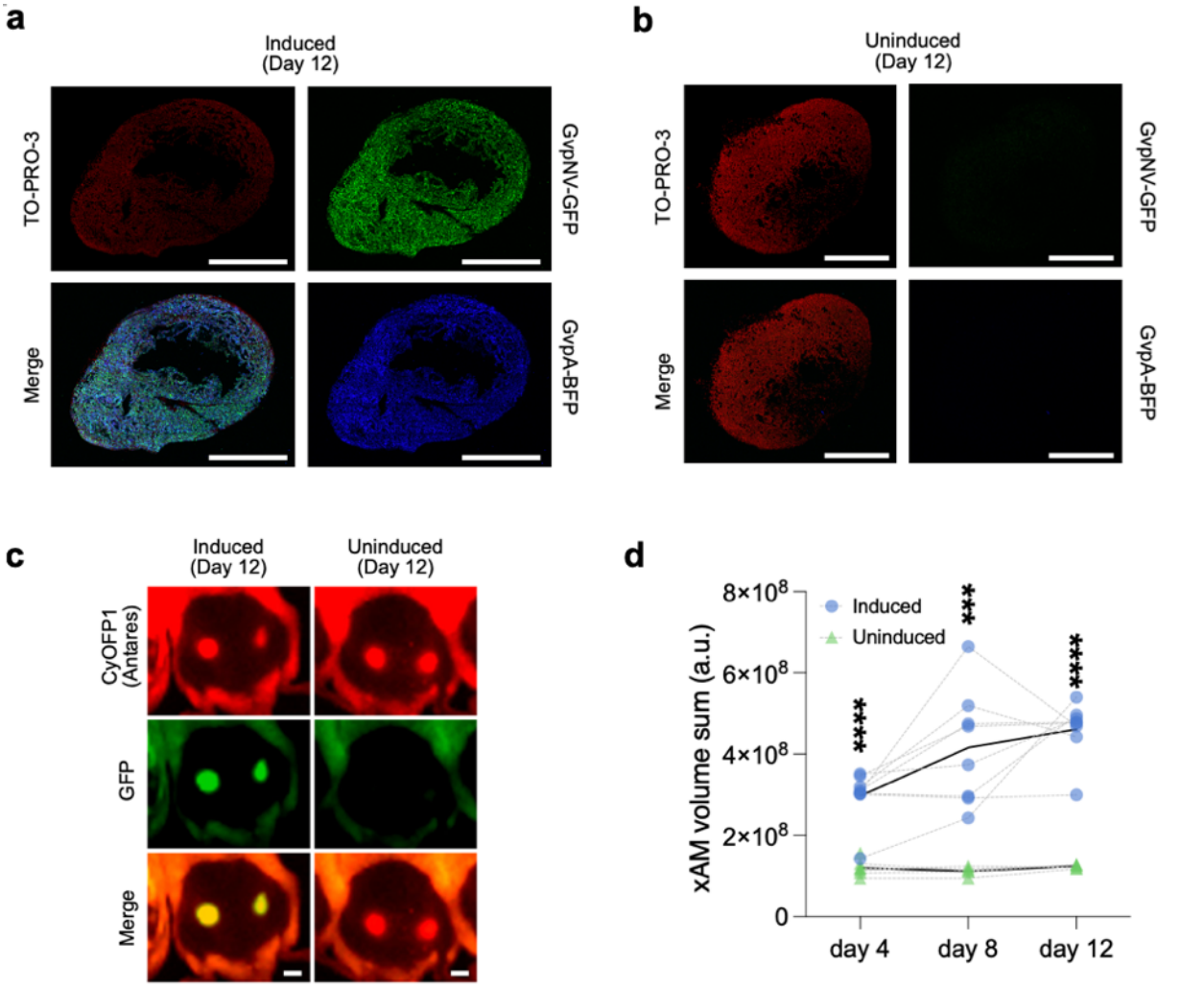


Figure 22 |* in situ mARG$_{Ana}$ expression corroborated with optical imaging.** *Fluorescence micrographs of 100 nm thin* induced **(a)** or uninduced **(b)** *tumor section. Green color shows GFP fluorescence, blue color shows BFP fluorescence, and red color shows TO-PRO-3 nuclear stain.* **(c)** *Whole-animal fluorescence imaging of induced (left) and uninduced (right) mouse after 12 days of expression. All tumors are constitutively expressing CyOFP1 (Antares, red) whereas mARG$_{Ana}$ expression is linked to expression of GFP (green). Scalebars represent 5 mm.* **(d)** *Three-dimensional sum of xAM signal from ultrasound tomograms of induced (n=8) and uninduced (n=7 on day 4, n=5 on days 8 and 12) tumors from all three imaging sessions plotted on a linear scale in arbitrary units. Black curves connect the means and grey curves connect points for each mouse. Asterisks represent statistical significance by two-tailed, unpaired Student's t-tests between induced and uninduced conditions (* = p<0.0001, *** = p<0.001).*

The first gene expression informed ultrasound-guided biopsy

To showcase the utility of dynamic xAM imaging of mARG$_{Ana}$, we decided to perform ultrasound-guided biopsy procedure of chimeric tumors. We created chimeric tumors using two proximal

subcutaneous injections of Antares-labelled MDA-MB-231 cells and MDA-MB-231-mARG$_{Ana}$ cells, resulting in uniform tumors with two regions of differential gene expression ***(Fig. 23a)****. We validated the chimeric nature of the tumors after 5 days of doxycycline induction by performing xAM/B-mode tomograms of entire tumors and reconstructing them in 3D* ***(Supplementary video 2, 3)****. We then performed fine needle aspiration biopsy targeting either the xAM+ or the xAM- regions of each tumor* ***(Fig. 23b, Supplementary video 4, 5)****. Flow cytometry analysis of dissociated biopsy samples showed that xAM imaging of mARGAna expression can be used as a reliable real-time spatial guide given the high correlation between mARG$_{Ana}$ positive (GFP+) cells that were sampled from xAM+ regions and vice versa* ***(Fig. 23c)****.* ***This is to our knowledge the first demonstration of ultrasound-guided biopsy where gene expression information has been used as a spatiotemporal guide as opposed to anatomical information alone.*** *In addition to research applications, this approach could be used in a clinical setting where tissues could be marked with sentinel cells that would produce ultrasound contrast upon confirmation of cancer tissue.*

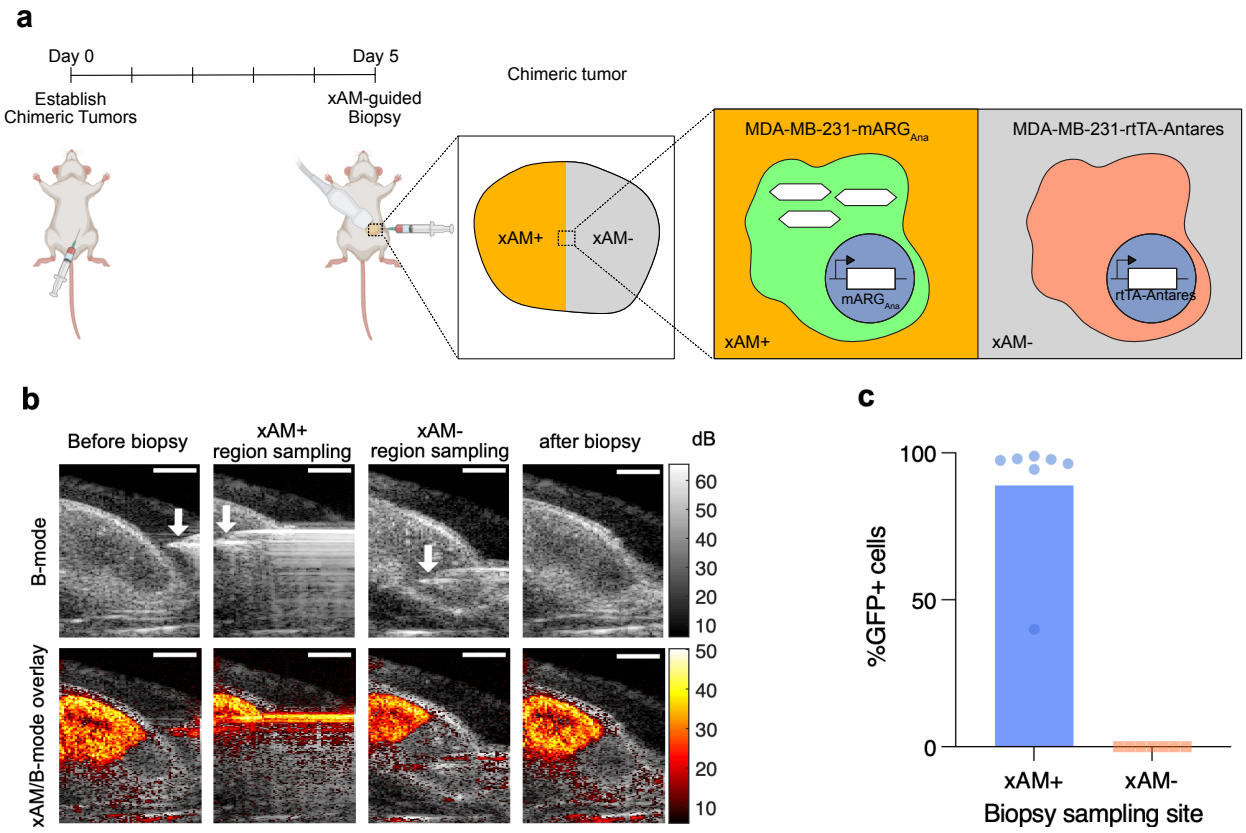


Figure 23 | xAM imaging of mARG$_{Ana}$ enables ultrasound-guided tumor biopsy. ***(a)*** *Diagram of the in vivo protocol for establishing chimeric tumors, in situ expression of GVs and tumor biopsy. Chimeric tumors were established on day 0. mARG$_{Ana}$ expression was induced by regular intraperitoneal (IP) doxycycline injections starting from the day of tumor injections. Tumors were imaged with ultrasound and biopsied after 5 days of expression.* ***(b)*** *Representative frames of B-mode/xAM ultrasound-guided biopsy procedure. Scalebars represent 2 mm. See supplementary movies 2,3,4 and 5 for the full ultrasound tomogram of the induced chimeric tumor, 3D reconstruction of the chimeric tumor as well as the videos of the full procedure.* ***(c)*** *Results of flow cytometric analysis of biopsied samples from xAM-positive and xAM-negative regions of chimeric tumors (n=7 tumors).*

Loss of mARG expression in old tumors

While we demonstrated that mARG$_{Ana}$ can be used to nondestructively and dynamically image gene expression *in vivo*, we found that **mARG$_{Ana}$ expressing tumors lose their ability to produce**

strong contrast after about three weeks of growth *in vivo*, even when mARG expression was not induced until 4 days before imaging **(Fig. 24a,b). Interestingly, cells harvested and recultured from day 30 tumors regained strong GV expression *in vitro* regardless of whether they were induced *in vivo* (Fig. 24c)**. Flow cytometry analysis of the recultured/reinduced tumors show that **the polyclonal MDA-MB-231-mARG$_{Ana}$ cell line was subjected to selection *in vivo* of a specific subpopulation that was high in *gvpNJKFGWV-GFP* expression**, regardless of whether the cells have been induced *in vivo* or not **(Fig. 24d)**. This selection was not observed when the polyclonal MDA-MB-231-mARG$_{Ana}$ cells were cultured *in vitro* (not shown). Reason for this loss of expression is unclear, however, it is possible that as tumors grow, tumor density becomes high enough to limit the diffusion of doxycycline, nutrients, oxygen, or all of the above. This theory is however inconsistent with the low expression even at the tumor periphery **(Fig. 24a day 30)**. It is also possible that the cells are dying over the 30-day growth period. However, tumor size grows continuously during that period. The phenomenon of subpopulation selection *in vivo* is even more confusing given that the selection appears to be stronger when the expression of mARGs is never induced *in vivo* **(Fig. 24d)**. In any case, whatever changes that caused attenuation of expression during the 30-day tumor growth were undone once the cells were returned to culture. It would be interesting to see in a future experiment if these 30-day recultured tumors retain strong expression when returned to mouse tumors.

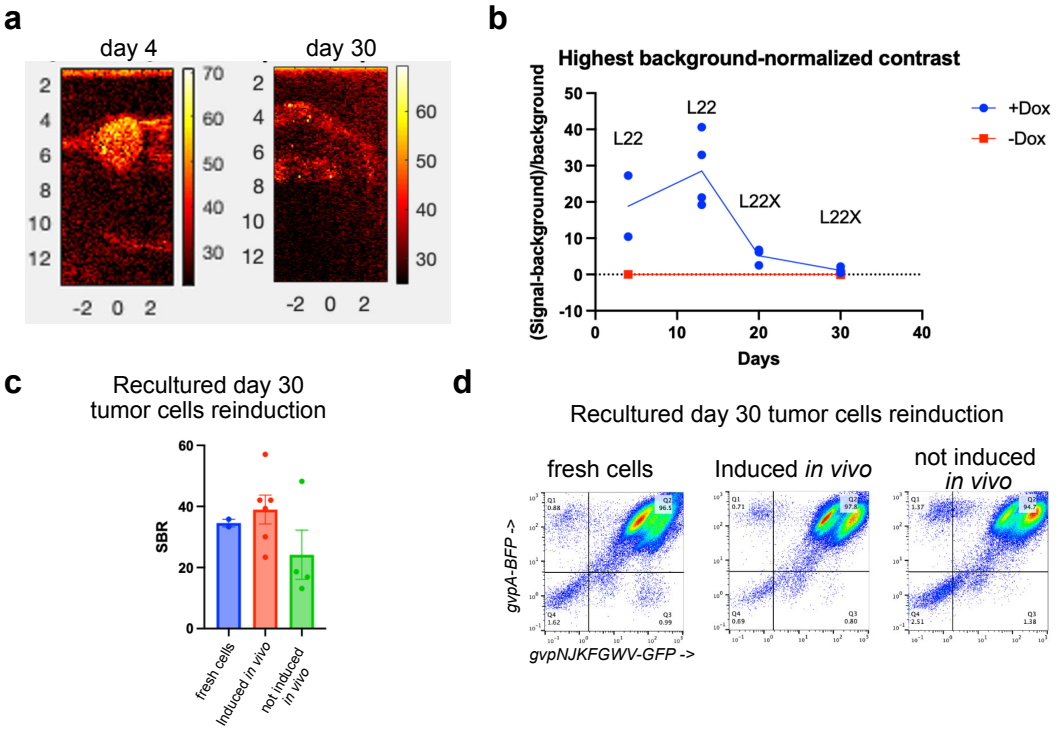


Figure 24 | Loss of mARG expression in old tumors. (a) xAM images of MDA-MB-231-mARG$_{Ana}$ tumors induced for 4 days prior to imaging. Tumors were imaged either 4 or 30 days after tumor establishment. **(b)** Quantification of the xAM images at the highest contrast voltage of the tumors over the course of 30 days. +Dox tumors were induced 4-days before imaging. **(c)** quantification of xAM images of tumor cells that were harvested from mice and returned to culture and reinduced *in vitro*. Fresh cells that have been taken out of cyostock were used as control. **(d)** Flow cytometry analysis of cells from (c).

2d. Discussion

The discussion section emphasizes the groundbreaking potential of acoustic reporter genes (ARGs) in non-invasive biomedical imaging. The $mARG_{Ana}$ construct, derived from Anabaena flos-aquae, represents a significant advancement in the field of molecular imaging by providing a highly sensitive and specific method for real-time ultrasound imaging of gene expression in mammalian cells. This development enables the detailed observation of cellular processes *in vivo* without the need for invasive procedures, offering a window into the dynamic states of disease progression and therapeutic response.

The utility of $mARG_{Ana}$ extends beyond basic research, with implications for clinical applications such as targeted therapy and regenerative medicine. By allowing for the non-destructive imaging of cells deep within the body, $mARG_{Ana}$ has the potential to facilitate the precise localization and monitoring of therapeutic cells, including T cells and stem cells, offering a new dimension to cell-based therapies. The non-linear ultrasound contrast produced by $mARG_{Ana}$ enables the visualization of these cells in real-time, providing critical insights into their behavior, viability, and therapeutic effectiveness in a way that was previously unattainable.

Moreover, the molecular engineering of gas vesicles from A. flos-aquae opens up new possibilities for customizing ARGs for specific applications. Through targeted modifications, it is conceivable to enhance the stability, sensitivity, and specificity of mARGana, tailoring it to various biomedical needs. This could lead to the development of acoustic biosensors capable of detecting cellular signals such as enzyme activity, further expanding the scope of ARGs in biomedical research and diagnostics.

In summary, the $mARG_{Ana}$ construct represents a paradigm shift in the field of non-invasive imaging, offering unparalleled sensitivity and specificity for the real-time observation of mammalian gene expression. Its development not only underscores the potential of ARGs in advancing our understanding of complex biological processes but also paves the way for novel diagnostic and therapeutic strategies that leverage the unique capabilities of ultrasound imaging. As this technology matures, it promises to provide critical insights into the molecular underpinnings of disease, enabling more precise and effective treatments that are grounded in the real-time monitoring of cellular function.

Chapter 3. Insights into Gas Vesicle nucleation and assembly

Chapter 3 delves into the intricate processes of gas vesicle (GV) nucleation and assembly, pinpointing these as critical areas for advancing the utility of Acoustic Reporter Genes (ARGs). The development of future ARG generations faces several formidable challenges. One such challenge is the significant delay between the induction of GV expression and the emergence of detectable ultrasound contrast, which spans hours, while reaching a steady state of expression can take days in the case of mammalian ARGs (mARGs). Enhancing our understanding of GV assembly could provide valuable insights into accelerating the nucleation and assembly processes of GVs, thereby reducing this time lag.

Moreover, the expression of ARGs, whether in bacterial or mammalian cells, often imposes stress on the host cells, potentially resulting in extended doubling times or even cell death. Investigating the underlying mechanisms of cell toxicity associated with GV expression is crucial for developing safer and more efficient ARG systems.

Another area for improvement is the genetic footprint of ARGs. Despite advancements, there remains a need to further minimize the size of ARG constructs. At this juncture, straightforward strategies for size reduction appear to be exhausted, suggesting that future efforts will likely require innovative engineering approaches that simplify the assembly process by reducing the number of necessary proteins.

This chapter proposes that unlocking further advancements in ARG technology hinges on a deeper understanding of GV nucleation and assembly principles, which are currently inadequately understood. To this end, a two-pronged approach is recommended:

1. **Exploring the Dynamics of GV Expression:** This involves a thorough investigation of how GVs are expressed and re-expressed at both individual cellular levels and across cell populations, aiming to identify factors that can expedite or enhance the process.
2. **Understanding Biochemical Interactions:** Delving into the biochemical interactions and mechanisms of action of GV assembly factors during the nucleation and assembly phases. This includes identifying the roles of specific proteins and how their interactions contribute to the efficient formation of GVs.

Model of GV assembly

The current model of GV assembly starts with a nucleation event where putative nucleation factors bring enough GV subunits to start forming a bicone. The bicone increases in length and diameter independent of GvpN. At some point the diameter increase ceases and the bicone begins a phase of cylindrical elongation, a process involving GvpN **(Fig. 25)**. The roles of individual Gvps is unclear. Some of the accessory Gvps could potentially form protein complexes that work together to either nucleate or elongate the GV. It is also unclear if any of the accessory Gvps are consumables in the GV production process. Careful genetic, biochemical and structural analyses of the accessory Gvps is needed to get the full picture of GV assembly.

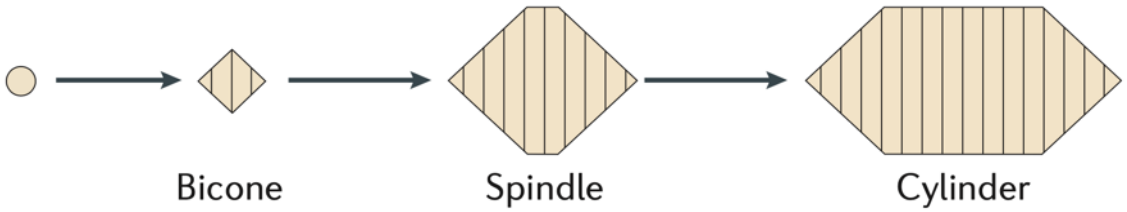


Figure 25 | Model of GV nucleation and elongation.[14]

3a. Optical microscopy of GVs and Gvps

GVs under phase contrast microscopy

GVs have been known to be observable under phase contrast microscopy in intact cells.[39] This is especially true for GVs expressed in mammalian cells, as mammalian GVs can reach lengths of several microns. It is generally thought that individual GVs are not visible under phase contrast but bundles of GVs are. Using this approach, it is possible to monitor GV bundle formation and elongation in live mammalian cells over the course of hours of GV assembly. Upon closer inspection of GV-producing MDA-MB-231-mARG$_{Ana}$ cells under phase contrast microscopy, it becomes evident that at any given time, **the cell-to-cell variability of GV expression is high, with some cells containing many large GV bundles, some containing small bundles while others do not contain any (Fig. 26a,b)**. The principles that govern this variability are unclear and should be the focus of future research. It also becomes apparent that GVs bundles are dynamic, able to split into multiple independent bundles **(Fig. 26c)** or for multiple bundles to coalesce into single bundles **(Fig. 26d)**. It is also possible to monitor the birth of new bundles **(Fig. 26e)**, which is consistent with the observation in bacteria that GV assembly occurs in bundles.

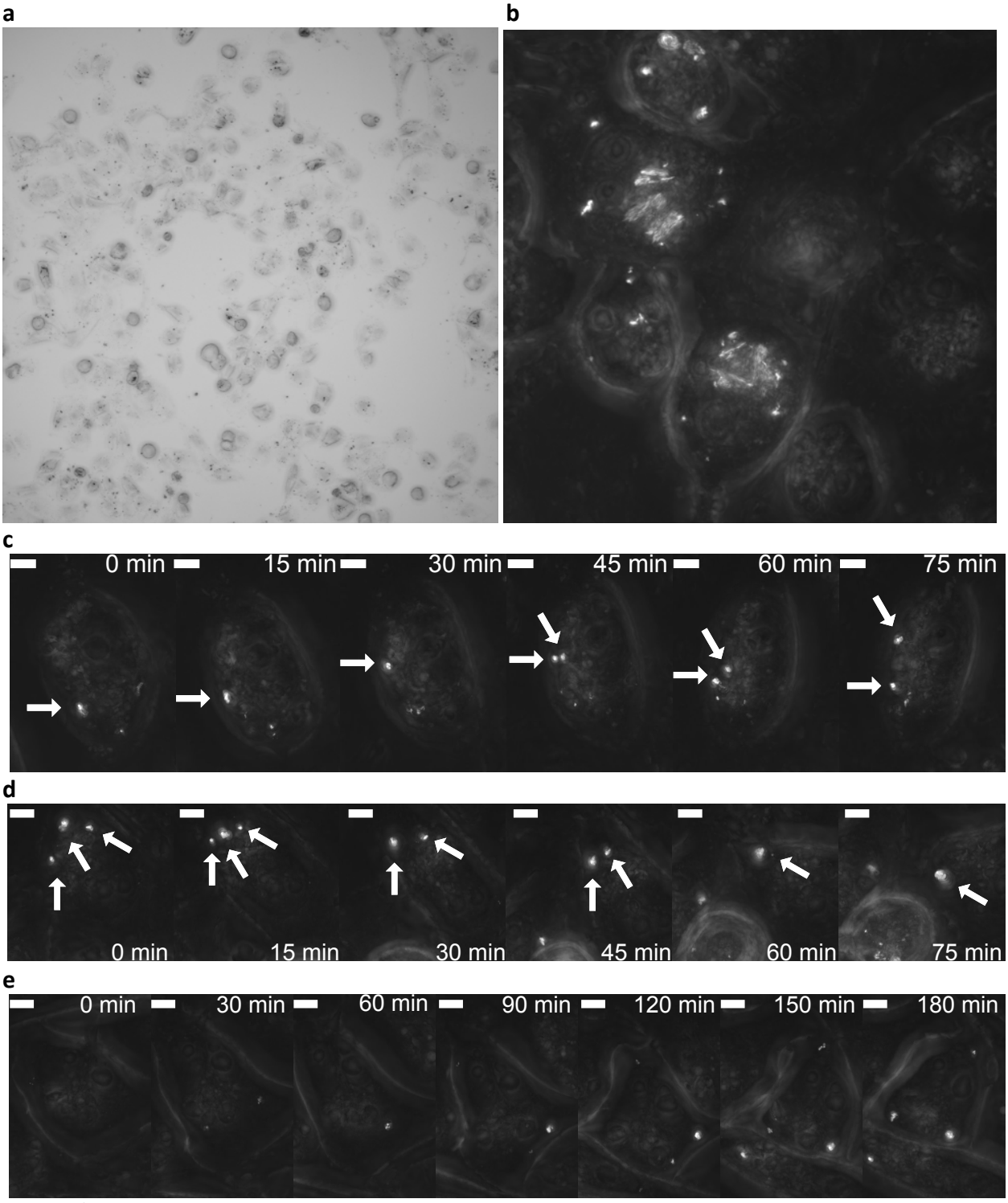


Figure 26 | Phase contrast microscopy of live mARG$_{Ana}$ expressing MDA-MB-231 cells. Z-stacks were collected for all images and time points and either minimum (20x) or maximum (100x) z-projection is shown in this figure. **(a)** 20x minimum z-projection of cells after expressing GVs (dark). **(b)** 100x maximum z-projection of cells after expressing GVs (light). **(c,d and e)** Time series of 100x maximum z-projections of cells expressing GVs. (c) Shows a single GV bungle split into two over the course of 75 minutes.

(d) Shows multiple GV bundles coalesce into a single bundle over 75 minutes. (e) Shows appearance of newly formed GV bundles. Scalebars are 5 µm.

GV bundling is an interesting phenomenon. It is unclear whether GV nucleation occurs in clusters, from which bundles arise as GVs elongate, or if GV nucleation of individual GVs occurs separately, which then coalesce into bundles. **During our efforts to express mARGs and to solve the GV structure, we have noticed that bicones (immature, early stage of GV assembly) often co-purify together in clusters (Fig 27a)**. What is holding them together is unclear, but Przemek Dutka noticed that purified bicones are often embedded in a matrix of impurities of unknown nature when imaged with cryoET **(Fig 27b)**. It is possible that these clusters of bicones are the birthplace of GVs. It would be interesting to see how these seemingly randomly oriented bicones mature into long and parallel GV bundles.

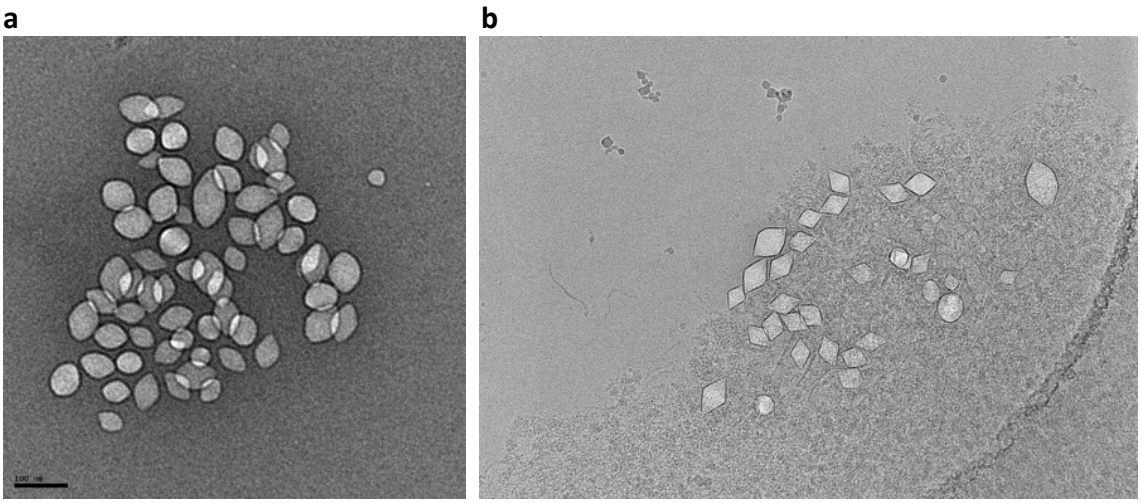


Figure 27 | Electron microscopy of bicones. (a) Unstained TEM of buoyancy purified bicones produced in HEK293T transient expression with B. megaterium monocistronic transfection mixture without gvpN. **(b)** CryoET of bicones found in a A. flos-aque GV prep (credit Przemek Dutka).

Optical imaging of GV re-expression

One of the key questions concerning GV re-expression is how does a cell's former ability to produce GV bundles correlate with their ability to re-express them after collapse? At the population level, if at any given moment, a specific subset of cells contains visible GV bundles, will that same subset of cells contain visible bundles after the GVs have been collapsed? To answer this question, we need to build a specialized live microscopy setup which can collapse GVs at any point during live imaging to enable individual cell tracking before and after collapse. Di Wu and colleagues are building such setup and once complete, we will be able to perform such experiments. Meanwhile, we have performed a proof-of-principle experiment where we pre-expressed GVs in mammalian cells, imaged them, used hydrostatic collapse chamber to collapse GVs and monitored their re-expression on the population level. **We saw that visible GV clusters returned as soon as 1 day after hydrostatic collapse, which is faster than it takes to see them starting from naïve induction (Fig. 28a)**. It should be noted that hydrostatic collapse is not a perfect collapse method as evidenced by collapsed cells still producing visible BURST signal, even when no bundles were observed under phase contrast. Hydrostatic collapse was able to almost completely abrogate AM signal from treated cells **(Fig. 28b,c)**. It is possible that the visibility of

the bundles is a prerequisite for AM signal, as it has been shown that GV bundling enhances ultrasound contrast.[8] It is also possible that the leftover GVs that contribute to the BURST signal after hydrostatic collapse serve as nuclei for GV bundling and elongation, giving those cells a head start, which could explain the fast return of GV bundles under phase contrast microscopy.

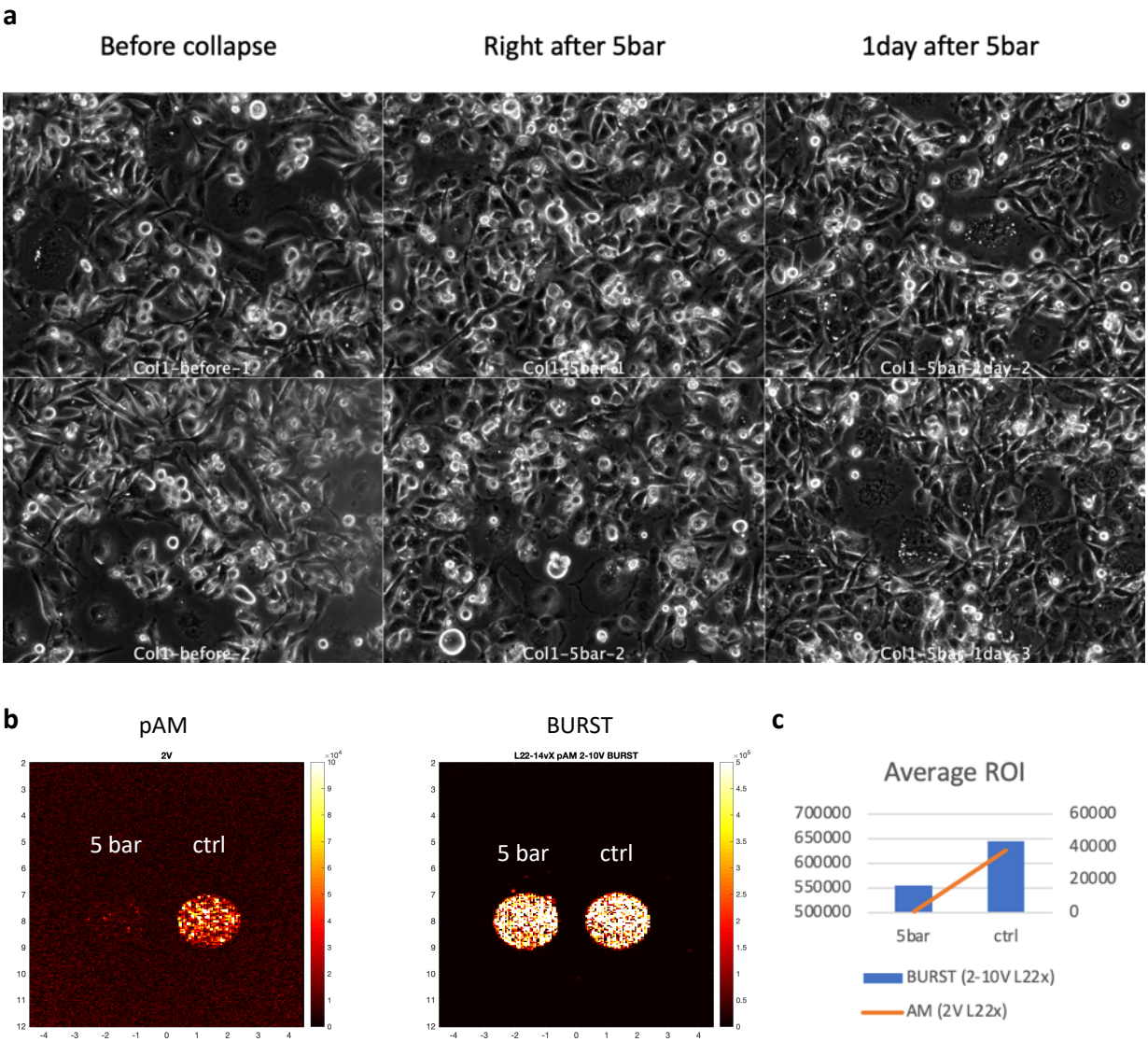


Figure 28 | Mammalian GV reexpression after hydrostatic collapse. (a) Phase contrast microscopy of MDA-MB-231-mARG$_{Ana}$ cells right before, right after and one day after hydrostatic collapse in a 5 bar chamber. **(b)** pAM and BURST images of cells exposed to 5 bar collapse chamber or nonexposed control. **(c)** pAM and BURST quantification of ROI in (b).

Fluorescent labeling of Gvps and GVs for live cell imaging

What has not yet been observed is whether heterogously expressed GVs can be removed or disassembled inside cells, whether via cellular degradation (e.g., autophagy) or whether GV assembly factors have an intrinsic ability to disassemble mature GVs into monomers. We would also like to image colocalization of accessory Gvps relative to growing GV bundles. Likewise, it would benefit our understanding of GV assembly if we could measure the kinetics of elongation of individual GVs or GV bundles. Such experiments would be well suited for confocal fluorescence

microscopy. Unfortunately, while some Gvps can tolerate fusion of fluorescent proteins, GvpA is not one of them. Until recently, there was not a practical way to fluorescently label individual GV subunits in live cells. We recently found that the C-terminus of GvpA, which is exposed to the cytosol, tolerates point mutations**. We were able to introduce a tetra-cysteine tag to the C-terminus of GvpA and still produce full GVs. A cell-permeable non-fluorescent compound fluorescein arsenical hairpin binder-ethanedithiol (FlAsH-EDT$_2$) can selectively and covalently react with the tetra-cysteine tag, which renders it fluorescent (green) (Fig. 29a)**.[40] This labeling approach, combined with confocal fluorescence microscopy, could be used to study the rates of elongation, GV assembly and degradation in live cells. One potential experiment could determine where new GvpA subunits are added on the growing GV. After cells produce visible long GV bundles, a nonfluorescent cysteine-reactive compound (e.g., N-Ethylmaleimide) could be used to block all reactive sites on established GVs (pulse). A later addition of FlAsH-EDT$_2$ (chase) will then label only newly synthesized GV regions. The location of the label will therefore confirm the site of new subunit addition on the growing GV.

Another approach could leverage the use of 3-(6-acetylnaphthalen-2-ylamino)-2-aminopropanoic acid (Anap)[41] or L-Acridonylalanine (Acd) **(Fig. 29b)**,[42,43] **unnatural amino acids with a blue-fluorescent sidechain with can be directly incorporated into proteins using orthogonal amber stop codon suppression.** Using the two above orthogonal approaches we could label GVs and individual Gvps with multiple colors to study how they colocalize relative to growing GV bundles and to each other.

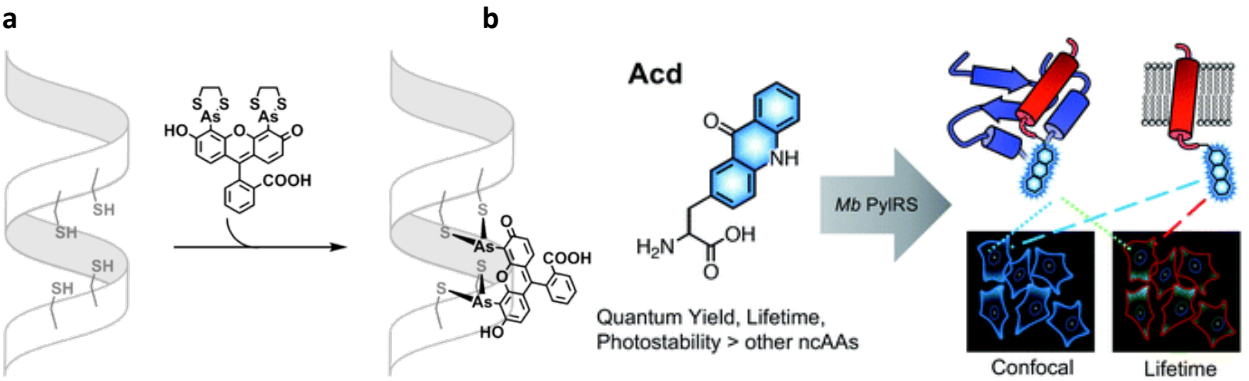


Figure 29 | Fluorescent GV labeling strategies for live cell imaging. (a) *fluorescin arsenical hairpin binder-ethanedithiol (FlAsH-EDT2) reacting with tetracysteine tag on engineered proteins. Model by Dkim118, distributed under a CC BY-SA 3.0 license.* ***(b)*** *L-Acridonylalanine (Acd), an unnatural amino acid can be incorporated into engineered proteins via orthogonal amber stop codon suppression.*[43]

Mass photometry and Interferometric scattering microscopy (iSCAT)

Apart from directly labeling GVs and Gvps for fluorescent imaging, there are other labe-free approaches to optically characterize nanoscale objects and single molecules and complexes. Mass photometry is a label-free method for single molecule measurement in solution with the ability to measure protein-protein interactions and their mass.[44,45] **Mass photometry could be well suited to study Gvp complexes and Gvp interactions (Fig. 30a)**.

Interferometric scattering microscopy (iSCAT) could in principle be used to track single GVs and image and measure their assembly dynamics in a label-free manner.[46] iSCAT has been used to measure the length and motion of individual microtubules **(Fig. 30b,c)**

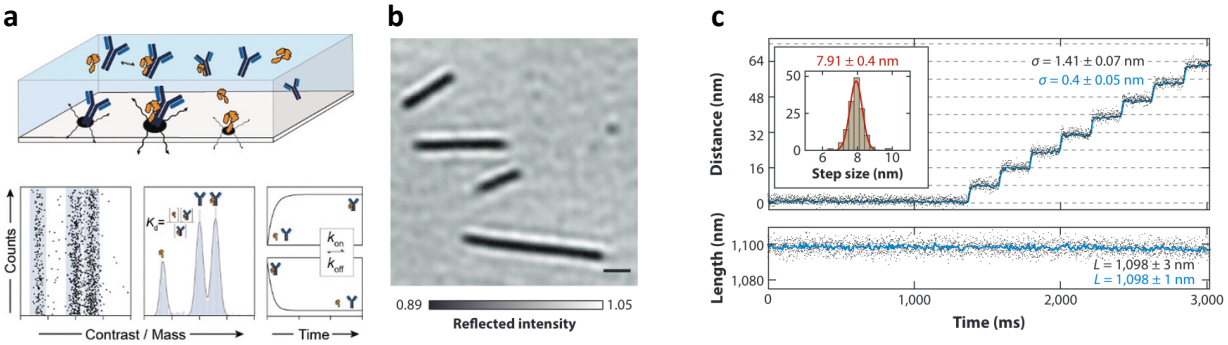


Figure 30 | Mass photometry and iSCAT. (a) Mass photometry. Protein-protein interactions can be quantified based on light scattering of molecules upon binding to the glass/water interface. It enables quantification of relative abundances of proteins and their complexes as well as kinetics of protein-protein interactions.[44] **(b)** iSCAT image of individual microtubules **(c)** tracking of length and displacement of microtubules undergoing nanometer-scale artificial movement generated by nanometric stage.[46]

3b. Genetic analysis of *gvps*

While optimizing the $mARG_{Ana}$ cluster, we realized that the mammalian transient cotransfection of gvp genes can serve as a platform for genetic analysis of *gvps*; *A. flos*-aquae contains perhaps the most minimal set of essential *gvp* genes (*gvpANJKFGW*), individual genes can be modified on single plasmids, stoichiometry of genes can be easily tuned by changing their ratios in the plasmid mixture, genes from different prokaryotic species and their mutants can be easily mixed and matched. **Most importantly, mammalian expression provides us with several phenotypic readouts; phase contrast microscopy and ultrasound can be used to quantify expressed GV quantity and their acoustic properties, cell toxicity assays allow for assessment of burden of GV assembly in different genetic conditions. EM of purified lysates can provide insight into the shapes and shape distributions of resulting GVs.** What started as a side project has turned into a systematic study aimed at deciphering interactions of individual Gvps and their roles in GV assembly.

Essential gvp identification and gvpN-independent assembly

One of the early experiments we performed when designing the $mARG_{Ana}$ system was to identify essential *gvps* by individual gene omission from the plasmid cotransfection mixture **(Fig. 31a)**. As mentioned earlier, we found that omission of *gvpC* and *gvpV* was still able to produce robust BURST ultrasound contrast. All other *gvps* seemed to be essential. *ΔgvpN* (AAA+ ATPase implicated in GV elongation) produced a very weak BURST ultrasound contrast. As seen in Serratia,[15] bARG1 and $mARG_{Mega}$, deletion of *gvpN* results in immature GVs arrested at the bicone stage of assembly, which is consistent with a weak BURST signal.

Surprisingly, when we repeated the *ΔgvpN* experiment in the absence of *gvpC*, we observed a stronger BURST signal **(Fig. 31b, top)** and even a weak AM signal (not shown). Heterologously expressed bicones have never been observed to produce AM signal, which prompted us to image buoyancy purified lysate of HEK cells cotransfected with *ΔgvpNΔgvpC* mixture using TEM. What we found was shocking, **while the double omission of *gvpN and gvpC* resulted in most of the GVs arrested at the bicone stage, there were numerous microns-long full-sized GVs in the lysate**, which were the most likely explanation for the weak AM signal **(Fig.**

31b, bottom). This is to our knowledge the first evidence that GvpN, while greatly stimulating GV assembly, might not be necessary in the "essential sense" of the GV assembly model.

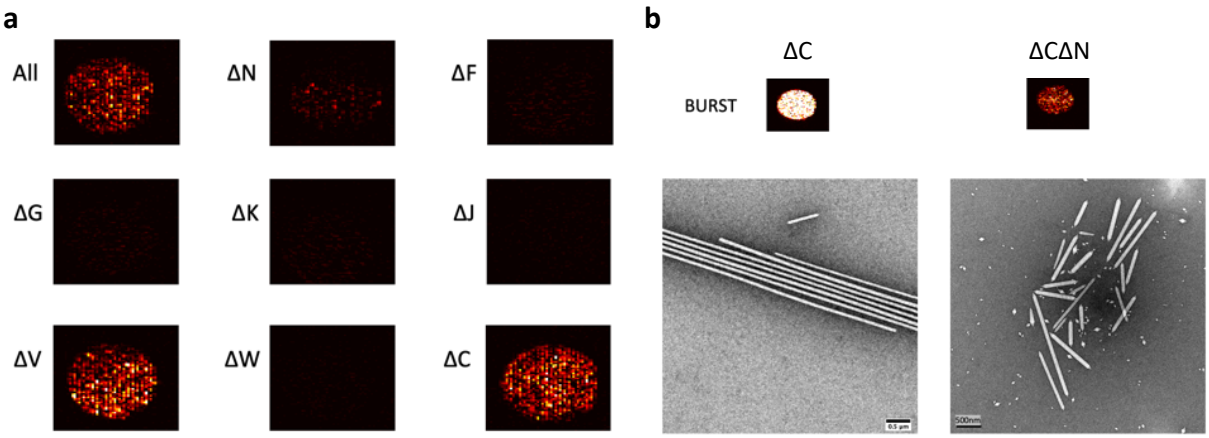


Figure 31 | gvp omission phenotype analysis. (a) BURST images of HEK293T cells transiently cotransfected with *A. flos-aquae* monocistronic gvp plasmids (A, N, F, G, K, J, V, W, C) with individual gvps omitted "Δ". **(b)** BURST images (top) and TEM images of lysates (bottom) of HEK293T cells transiently cotransfected with *A. flos-aquae* monocistronic gvp plasmids without gvpC (A, N, F, G, K, J, V, W) with or without gvpN. Scalebars are 500 nm.

The inconsistency of this result with the ones observed in bacteria puzzled us. Individual deletion of GvpN in Serratia and bARG1 only produced bicones, however, in both cases GvpC was present. We hypothesized that GvpC might be inhibiting this apparent GvpN-independent elongation. We decided to test that by adding or omitting *gvpC* in the presence or absence of *gvpN*. Consistent with our hypothesis, **GvpC was able to completely inhibit GvpN-independent GV elongation** as no full-sized GVs or AM signal were observed when *gvpN* was omitted in the presence of excess *gvpC*. **When *gvpN* was added to the mixture containing *gvpC* elongation was restored (Fig. 32)**. One theory to explain this result is that as GvpC binds to the growing GV wall, it might have to be removed by GvpN for new GvpA subunits to access the site of assembly.

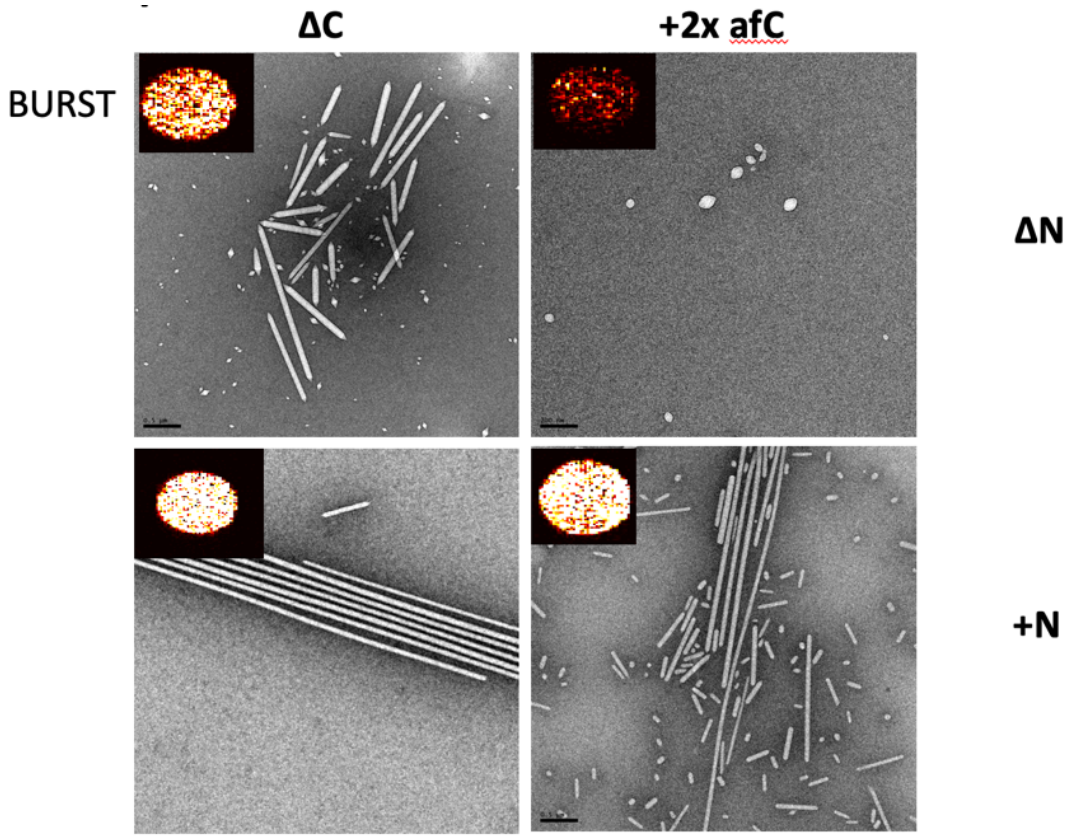


Figure 32 | gvpC inhibits gvpN independent elongation. BURST images of HEK293T cells and TEM images of their lysates. Cells were either transfected with gvpC and with gvpN (lower right), with gvpC but without gvpN (upper right), without gvpC but with gvpN (lower left), or without gvpC and without gvpN (upper left).

This theory is however inconsistent with the $mARG_{Mega}$ result, where omission of *gvpN* results in bicones **(Fig. 33a,b)**, especially since $mARG_{Mega}$ does not contain *gvpC*. The question then became: can we observe GvpN-independent assembly with a hybrid cluster in mammalian cells? Unfortunately, bARG1-like hybrid cluster configuration (*gvpA* form *A. flos-aquae* with *B. megaterium gvpNFGLSKJU*) did not produce any GVs in HEK cells, not even bicones **(Fig. 33c)**. **However, when we swapped in *gvpN* from *A. flos-aquae*, we were able to observe full GVs and strong BURST contrast (Fig. 33d)**. Interestingly, despite afGvpA also being the structural protein in the hybrid gene set, this gene configuration produced GVs with much lower AM contrast relative to BURST than GVs made with full *A. flos-aquae* gene set, mirroring the acoustic properties of *B. megaterium* GVs (not shown). This result shows that acoustic properties of GVs are not just a function of the structural protein, but also of the assembly factors that assemble them. Similar to the hybrid cluster with *B. megaterium gvpN*, omission of *gvpN* from the hybrid cluster also did not produce detectable GVs or bicones **(Fig. 33e)**. Neither was *A. flos-aquae gvpN* able to substitute for *B. megaterium gvpN* **(Fig. 33f)**. **This apparent genetic specificity between the structural proteins and their respective *gvpN* strongly suggests that the structural protein is likely one of the substrates of GvpN**. While, these results do not explain the mystery of *A. flos-aquae gvpN*-independent elongation, they suggest that this behavior is specific to the *A. flos-aquae* assembly factors. Future genetic and biochemical analyses are necessary to fully understand this behavior.

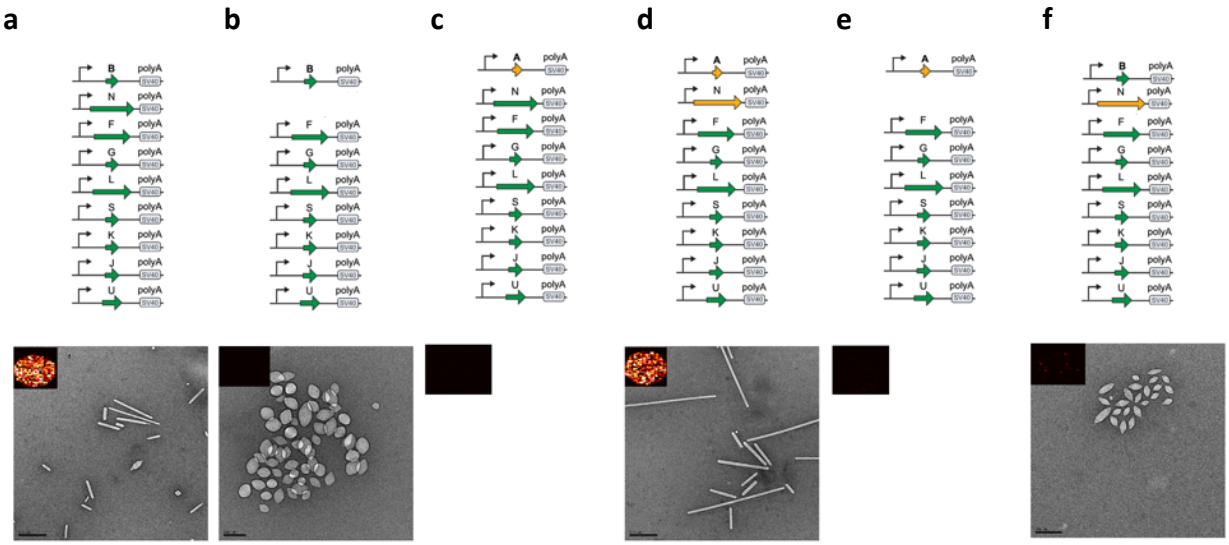


Figure 33 | *gvpN*-independent assembly is specific to *A. flos-aquae* GV gene cluster. B. megaterium *gvps* are labeled with green arrows, A. flos-aquae *gvps* are labeled with orange arrows. **(a-f)** Each column represents the plasmid mixture used (top), BURST image of cells expressing plasmid mixtures on top, and TEM image of their buoyancy-purified lysates (Bottom). (c) and (e) do not contain TEM images because nothing identifiable could be imaged.

GvpW is GvpL (probably)

Völkner and colleagues from Felicitas Pfeifer lab have recently published a systematic analysis of Gvp protein-protein interactions in *H. salinarum* cluster using split-GFP assay in vivo.[22] **Using the results, they proposed a model of Gvp interactions, suggesting that a subset of conserved Gvps form a protein complex, perhaps as a part of elongation or nucleation machinery (Fig. 34)**. There are several differences between the *H. salinarum* gene set and the *A. flos-aquae* gene set, with the *A. flos-aquae* having a smaller set. Specifically, while *H. salinarum* contains two members of the *gvpA/J/S/M* family (*gvpJ* and *gvpM), A. flos-aquae* only contains *gvpJ*. Additionally, *A. flos-aquae* does not contain *gvpH, I or L*, making this model confusing as hsGvpL seems to have a central role in this Gvp-interaction map. *A. flos-aquae*, however, contains a gene annotated as *gvpW*, which is not present in the *H. salinarum* gene set. Is it possible that *A. flos-aquae* GvpF or GvpW could serve the role of GvpL in the putative interaction map?

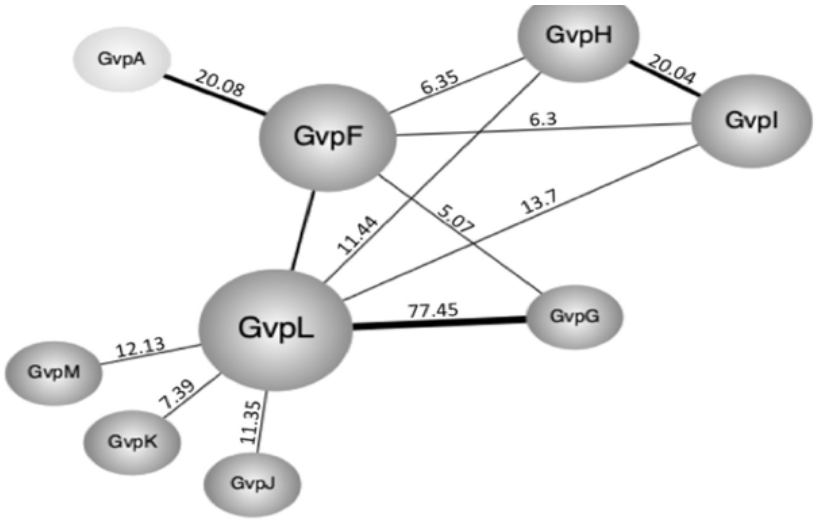


Figure 34 | Model interaction network of H. salinarum Gvps based on Split-GFP experiments.[22]

To test this theory, we returned to the hybrid cluster experiments (afGvpA with bmGvpJSFLGKU) as *B. megaterium* gene set contains both F and L, as well as two members of the GvpA/J/S/M family (bmGvpJ and bmGvpS). Substitution of afGvpJ for bmGvpJ,S in the hybrid cluster abrogated GV expression **(Fig. 35a,b)**. We then supplemented this transfection mixture with individual or sets of afGvps to see if any combination of afGvps can rescue GV expression **(Fig. 35c-g)**. **Surprisingly, supplementation of afGvpW alone (Fig. 35d) or simultaneous supplementation of afGvpF and afGvpG (Fig. 35e,f) was able to partially rescue the phenotype.** Simultaneous supplementation of all three genes dramatically improved expression, beyond the expression of the hybrid cluster **(Fig. 35a and g)**. **These apparent genetic interactions between afGvpJ with afGvpW and between afGvpW with afGvpF and G suggests that afGvpW might occupy the GvpL biochemical niche in the *A. flos-aquae* cluster.** This conclusion is further supported with GvpW in *M. aeruginosa*, a closely related cyanobacterium, having a length and pI that are similar to halobacterial GvpL. Additionally, the closest BLAST search match of *M. aeruginosa* GvpW, which is highly similar to *A. flos aquae* GvpW, is the *B. megaterium* GvpL **(Fig. 35h)**.[47]

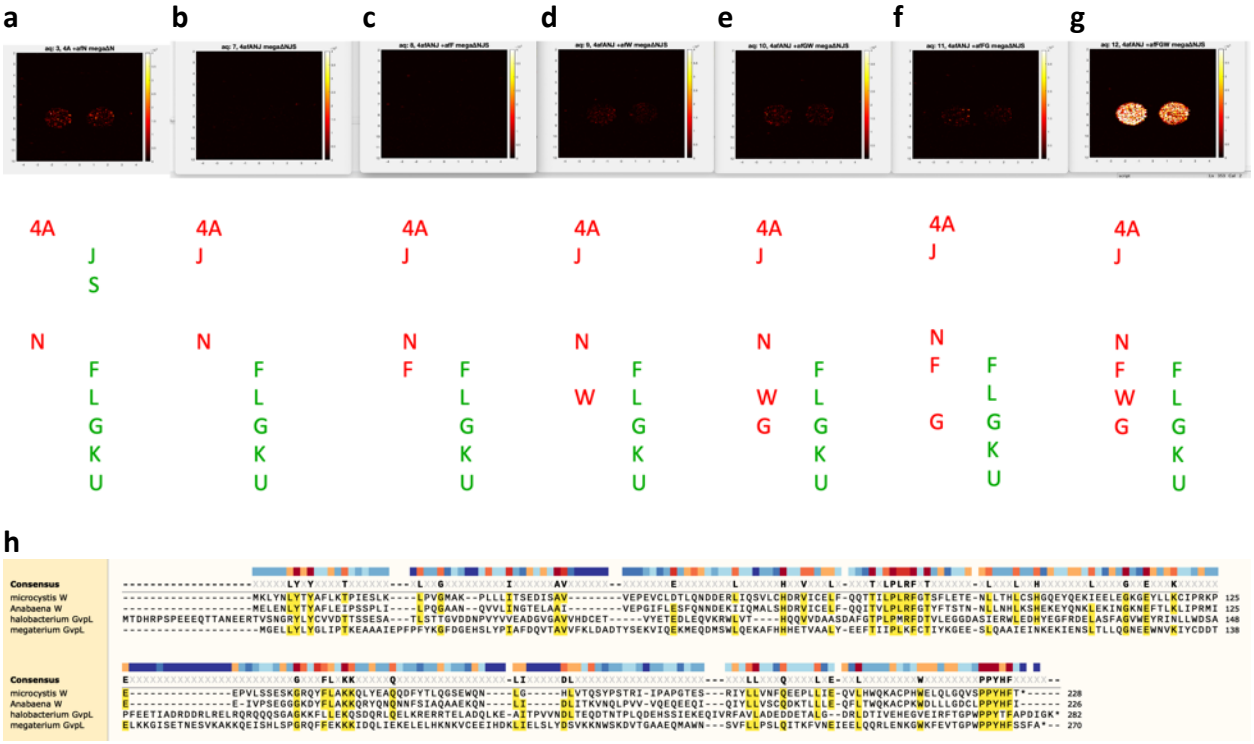


Figure 35 | A. flos-aquae GvpW is closest to GvpL of other GV custers. (a-g) BURST ultrasound images (top) of HEK293T cells transiently transfected with hybrid transfection mixtures (bottom). Genes in red are from *A. flos-aquae*, genes in green are from *B. megaterium*. **(h)** Multiple sequence alignment of maGvpW, afGvpW with hsGvpL and bmGvpL.

GvpJ is a potential GV nucleation factor

Why does *A. flos-aquae* only have one additional member (afGvpJ) of the GvpA/J/S/M family? It is common for GV gene clusters to contain multiple members of GvpA/J/S/M in addition to the major structural protein GvpA. As we have seen, *B. megaterium* contains *gvpJ* and *gvpS*, *H. salinarum* contains *gvpJ* and *gvpM*, *Serratia* contains *gvpA2 and gvpA3*. In all these cases, the "accessory" GvpA/J/S/M sequences are similar in size to their respective structural protein GvpA. However, **afGvpJ protein sequence is 3.4-times longer than its respective afGvpA**. We decided to test what would happen if we truncated afGvpJ to a size more similar to afGvpA, which is 72 amino acids long **(Fig. 36a)**. We split afGvpJ into two parts, 1-108 and 109-249. The 1-108 fragment of afGvpJ was able to support GV expression but the expression was lower than that of the full afGvpJ fragment. **EM analysis of GVs produced with the 1-108 fragment of GvpJ still showed full-sized GVs.** The 109-249 fragment alone was not able to produce any detectable GVs. Cotransfection of both fragments did not enhance expression compared to transfection of the 1-108 fragment alone, suggesting that the 109-249 fragment needs to be attached to the 1-108 to perform its function **(Fig. 36b)**.

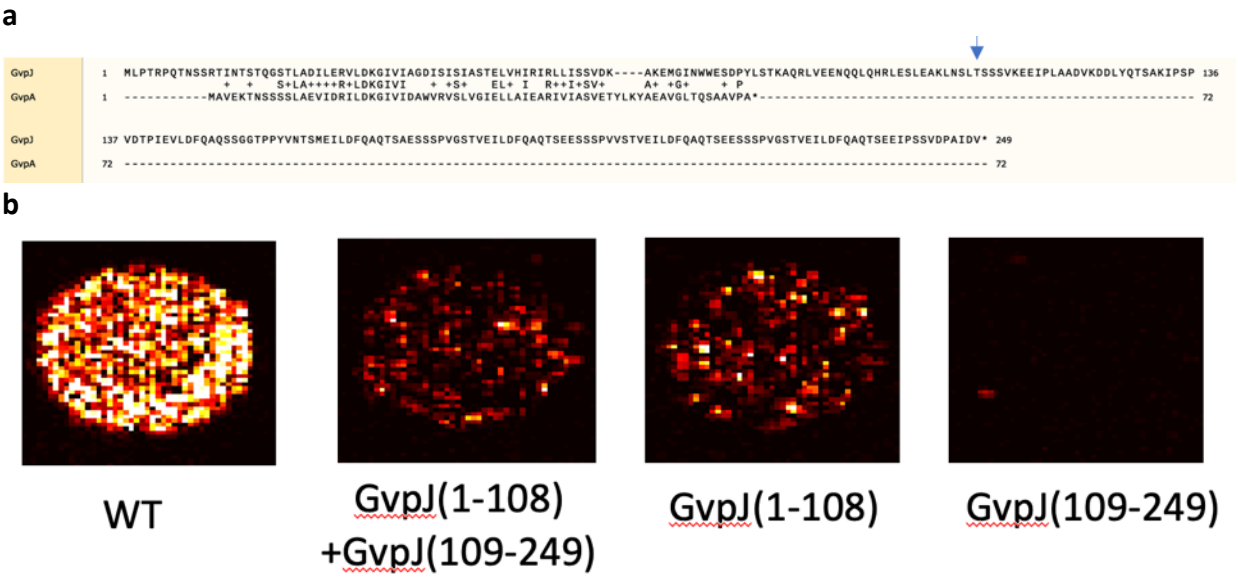


Figure 36 | *A. flos-aquae* GvpJ split experiment. (a) *A. flos-aquae* GvpJ and GvpA protein sequence alignment. afGvpA shares sequence homology with the N-terminal portion of afGvpJ. "+" indicates similar residues. Blue arrow points between residues 108/109. **(b)** BURST images of transiently transfected HEK293T cells with different mutants of afGvpJ.

Given that the halobacterial GvpJ is located on the operon that is expressed earlier than the GvpACNO operon, is suggesting that GvpJ is involved in the early stages of GV assembly, maybe as a part of the GV nucleation complex. It is unclear what the function of the C-terminal fragment of afGvpJ is since most other species do not have it. Interestingly, the *M. aeruginosa* GvpJ (maGvpJ), a related cyanobacterium to *A. flos-aquae* with a very similar Gvp gene cluster organization and almost identical GvpA, is only 136 amino acids long, lacking most of the C-terminal fragment that is present in afGvpJ. However, the last 14 residues of maGvpJ ORF overlap with a putative ORF annotated as GvpX (102 aa), which is not present in *A. flos-aquae*, in an out-of-frame fashion. The putative role of maGvpX is not known, however the extra sequence of afGvpJ following the end of maGvpJ/afGvpJ homology region (aas 138-249) shares sequence homology with the putative maGvpX (17.86% identity, 32.14% similarity) **(Fig. 37a)**. In fact, the N-terminal half of afGvpJ shares homology to the C-terminal half of afGvpJ (21.19% identity, 34.44% similarity) **(Fig. 37b)**. Similarly, maGvpX shares homology with maGvpJ (32.39% identity, 42.96% similarity) (**Fig. 37c**),[47] suggesting both **maGvpX and the C-terminal fragment of afGvpJ may have arisen through gene duplication of part of GvpJ.**[47] We do not know whether the fusion of afGvpJ or separation of maGvpJ/X occurred first but phylogenetic analysis of related GV gene clusters could answer that question.

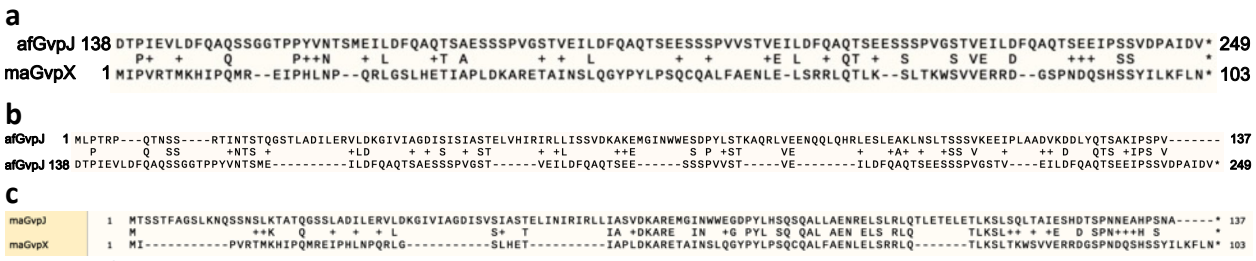


Figure 37 | maGvpX and the C-terminal part of afGvpJ could have arisen via gene duplication. Needleman-Wunsch global alignments. **(a)** C-terminal fragment of afGvpJ shares homology with maGvpX. **(b)** N-terminal fragment of afGvpJ shares homology to C-terminal fragment of afGvpJ. **(c)** maGvpJ shares homology with maGvpX.

While the maGvpJ and the N-terminal fragment of afGvpJ are predicted to form the characteristic alpha-beta-beta-alpha secondary structure according to Jpred secondary structure prediction algorithm, neither maGvpX nor the C-terminal fragment of afGvpJ have this predicted structure **(Fig. 38a)**. Given the similarity of C-terminal fragment of afGvpJ with maGvpX, it is not clear why these fragments evolved to work fused in *A. flos aquae* but separated in *M. aeruginosa*. It is interesting to note that the calculated translation rate of maGvpX RBS (which resides in the ma-*gvpJ* ORF) is 10-times higher than that of maGvpJ.[48] However, in our mammalian expression experiments, the omission of maGvpX from the maGvp transfection mixture did not lead to a noticeable difference in GV expression, suggesting that maGvpX has lost its function **(Fig. 38b)**. Interestingly the maGV WT cluster produced fewer GVs compared to afGV cluster **(Fig. 38b)**. We suspected that the reason for this difference lies in GvpJ. We performed a series of maGvpJ substitutions with either afGvpJ, truncated afGvpJ (afJ136) and fusion of the C-terminal fragment of afGvpJ onto maGvpJ (maJafJ). Strikingly, while the truncated afGvpJ produced similar amount of GVs compared to maGvpJ, both the wild type afGvpJ and the maJafJ fusion produced GVs comparable to the anabaena GV cluster, suggesting that the enhancement in expression we see lies solely in the C-terminal fragment of afGvpJ. **Upon closer inspection of the C-terminal tail of afGvpJ sequence, we noticed a highly repetitive unstructured 21-residue sequence with consensus TS**N**E**NN**SPV**NNNN**EILDFQAQ (Fig. 38a,c). We termed this region the QAQ-domain.** Interestingly, the maGvpX does not contain such repeating motif. Even more interestingly afGvpC also contains a similar repeating motif of 33-amino acids that also ends with QAQ.

Next, we tested if the QAQ-domain of afGvpJ can enhance GV expression of a distant GV cluster from *B. megaterium (megaGV)* in mammalian expression. megaGV cluster contains two homologues of afGvpJ termed bmGvpJ and bmGvpS. As it was not clear which protein would benefit from the QAQ-domain fusion, we decided to test both bmJ-QAQ and bmS-QAQ separately and together in the context of varying the amount of the bmGvpB **(Fig. 38d). Strikingly, fusion of the QAQ-domain onto bmGvpS or S and J reduced expression at all levels of bmGvpB and fusion of the QAQ-domain to bmGvpJ enhanced expression drastically at 4-fold excess of bmGvpB, suggesting that the QAQ-domain increases the capacity of GV assembly factors to handle excess structural protein and that the mechanism of this enhancement is species-agnostic.** To get a better idea of the QAQ-domain's mechanism of function, we inserted the full coding sequence of GFP between the N-terminal fragment of afJ and the QAQ domain (afJ-GFP-QAQ) and also created a version with two extra QAQ repeats (afJ-superQAQ) and tested them in a mammalian anaGV assembly assay with varying amounts of GvpJ **(Fig. 38e)**. **afGvpJ was able to tolerate GFP insertion, suggesting that the function of the QAQ domain may be mainly biophysical rather than mechanistic, perhaps aiding in phase separation of assembly factors.** It also seems that additional QAQ-repeats neither harm nor improve GV expression.

a

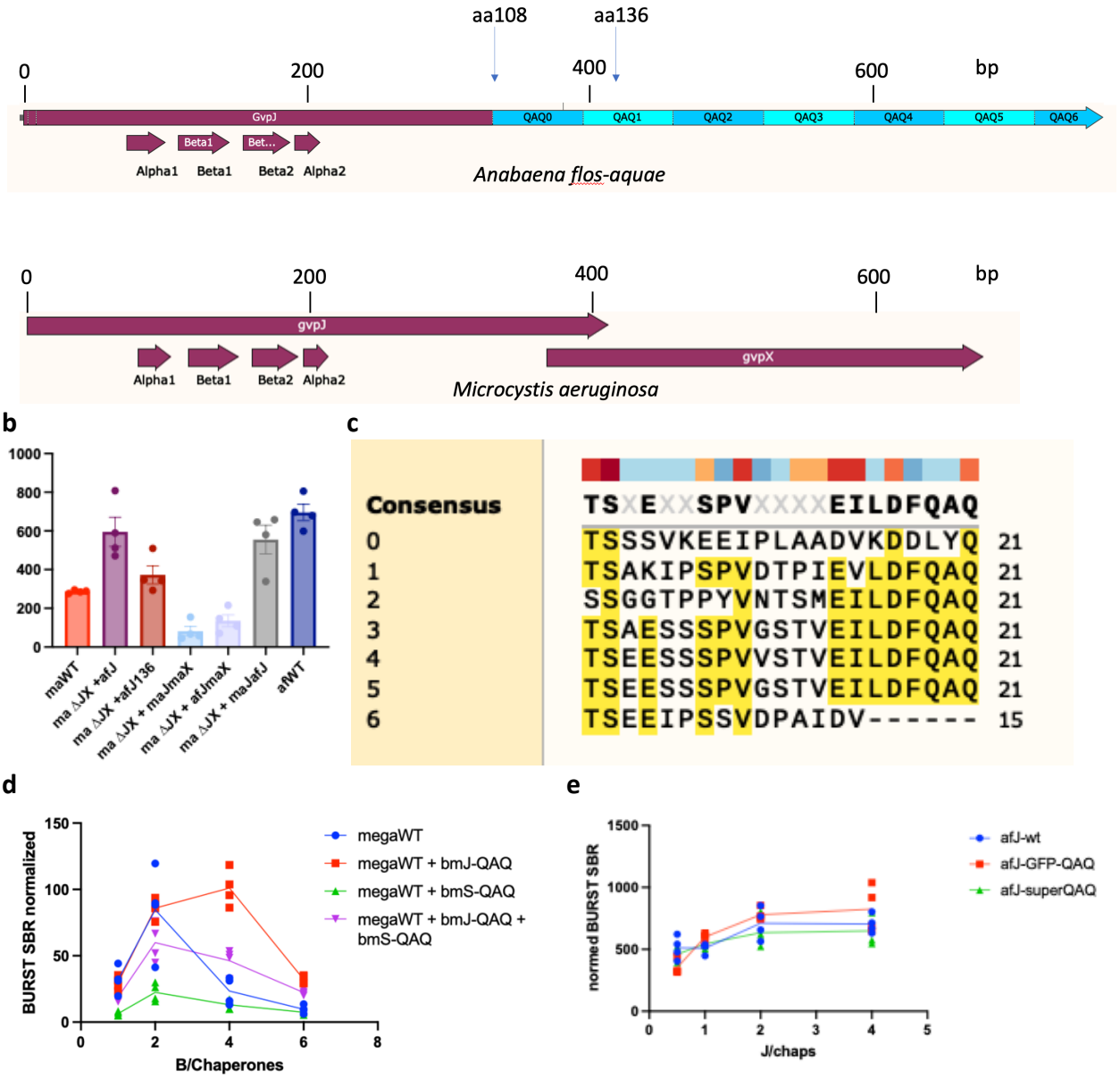


Figure 38 | C-terminal fragment of GvpJ. **(a)** C-terminal fragment of afGvpJ consists of repeats of a 21-aa motif not present in other GvpJs. **(b)** maGvpJ-QAQ graft enhances GV expression in the maGV assembly assay **(c)** alignment of the individual repeats from afGvpJ C-terminal fragment. **(d)** QAQ-domain enhances expression in megaGV assembly assay. **(e)** mutants of afGvpJ including GFP insertion and additional QAQ-repeats do not affect GV expression in anaGV assembly assay.

3c. Biochemical & structural analysis of gvps

GvpN

GvpN, the highly conserved AAA+ ATPase implicated in GV elongation, would be a prime target for structural study using single particle cryoEM with the added intention of identifying its substrates. It has been recently shown that GvpN can be purified in a form that retains its enzymatic activity.[49] Other AAA+ ATPases have been purified with substrate bound **(Fig. 39a,b)**,[50] or the ATPase can be arrested with the substrate bound using EDTA or nonhydrolyzable ATP analogs.[51] In either case, the identity of the specific substrate can be identified by biochemical or

structural analysis, such as specific substrate elution with Mg-ATP.[51] So far, we have observed genetic interactions of GvpN with GvpA, and of GvpN with GvpC. **We plan to co-express epitope-tagged GvpN with all other Gvps in hopes of capturing and identifying the substrates and solving the GvpN structure.**

Additionally, establishing a procedure for purifying high quantities of active GvpN hexamer protein will allow us to use it in GV assembly, enzymatic and Gvp interaction assays. In the case of cell free GV assembly, an alternative approach to transcribing and translating GvpN, along with other Gvps, during assembly is to purify it and supply it as an active protein. This method has been used with AAA+ ATPase NSF/Sec18 in numerous cell-free and chemically defined SNARE-mediated fusion assays.[26,52] In addition to identifying GvpN substrate via structural or biochemical methods, **we can also quantitatively measure GvpN substrate preference and processing kinetics using *in vitro* phosphate release assays**, where we can measure the rate of ATP hydrolysis in the presence or absence of various Gvp substrates or during GV assembly assay.[53] Lastly, **Alphafold prediction of afGvpN shows a highly unique beta-barrel hexamerization domain, unobserved in other AAA+ ATPases (Fig. 39c).**

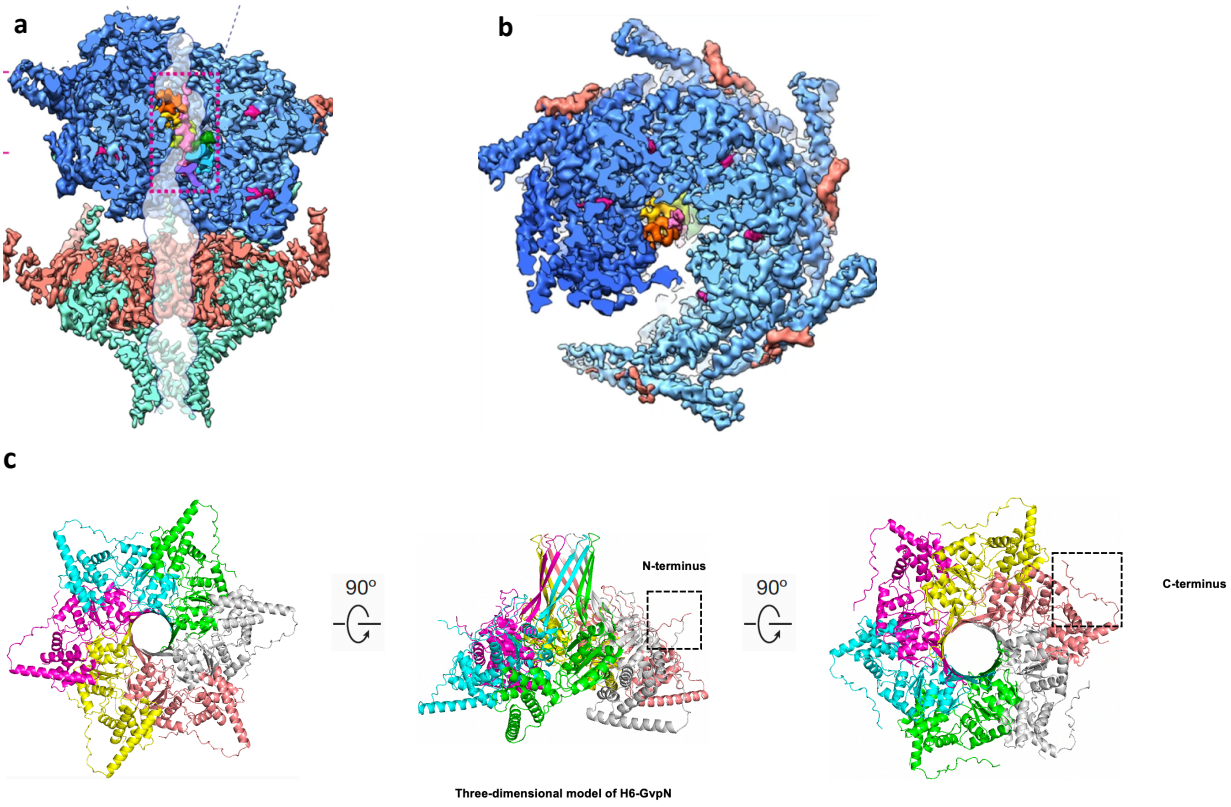


Figure 39 | AAA+ ATPases. Cryo-EM map of bisected side view **(a)** or cross-sectional top view **(b)** of Plasmodium translocon of exported proteins (PTEX) complex containing AAA+ ATPase HSP101 (Blue hexamer) engaged with its substrate (pink rectangle)[50] **(c)** Alphafold 2 prediction of afGvpN hexamer.

Gvp complexes

The genetic and biochemical interaction models suggest that a subset of Gvps may form a protein complex machinery, whose function is to either initiate GV nucleation or catalyze GV elongation.

We hope to identify if such complexes exist by individually tagging gvps and hopefully co-purifying them with their interaction partners. One such potential target could involve the hypothetical initiation complex, consisting of all or a subset of GvpFGJKW/L, as these proteins are expressed early in halobacteria. One approach would be to coexpress and purify these proteins recombinantly in the presence or absence of GvpA, or express individual components separately and assembling them *in vitro*. Identifying the components and solving the structure of such complexes would greatly enhance our understanding of GV assembly.

In vitro GV assembly

To gain a truly deep mechanistic understanding of GV assembly and to create a platform for GV engineering and prototyping, **it would be the ultimate goal to reconstitute GV assembly in a cell-free system or a chemically defined system**. Previous attempts involving the TXTL transcription-translation system have not succeeded. An alternative approach could involve purification of sufficient amounts of stable GV monomers and adding them to pre-nucleated GVs in the presence of either purified or co-expressed assembly factors to produce assembly *in vitro*. Purification of GvpA monomers is challenging, given the hydrophobic nature of GvpA and its tendency to aggregate and produce cell toxicity as evidenced by the bright fluorescent puncta from cells expressing GvpA only **(Fig. 40a)**. However, when GvpA is co-expressed with its assembly factors, this toxicity is rescued **(Fig. 40b)**.

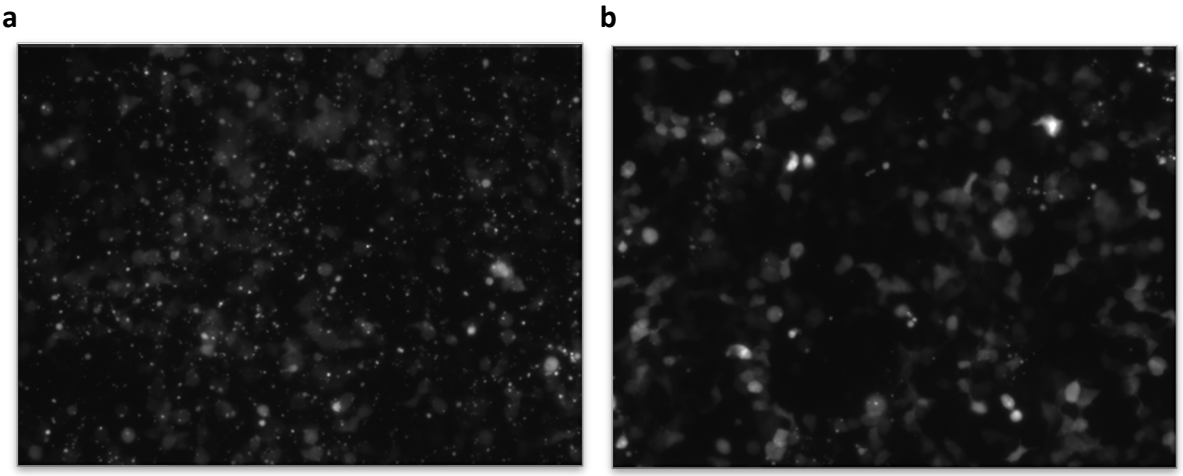


Figure 40 | GvpA expression without accessory Gvps causes cell toxicity. (a) fluorescence micrograph of HEK293T cells transiently transfected with *gvpA-IRES-mCherry* plasmid only. **(b)** fluorescence micrograph of HEK293T cells transiently transfected with *gvpA-gvpNJKFGWV-IRES-mCherry* plasmid.

The next step therefore is identification of the minimal subset of accessory Gvps that would rescue GvpA toxicity. This hypothetical subset of Gvps could theoretically be involved in chaperoning GvpA subunits and preventing them from aggregating and causing toxicity by shielding their hydrophobic domains. Some candidates for this hypothetical chaperone are GvpF due to its strong biochemical interaction with GvpA, or GvpG, which rescues the toxicity phenotype of bmGvpB expression in bacteria. Interestingly, promoter and ribosome binding site (RBS) strength prediction algorithms[48,54] predict that after GvpA, GvpG is the second most abundantly expressed Gvp in both *A. flos aquae* and *M. aeruginosa*, even surpassing GvpC

structural protein. **It might be possible to co-purify GvpA monomers complexed with these chaperones in high quantities, which could be subsequently used for *in vitro* GV assembly.**

Chapter 4. Stoichiometric Expression of Messenger Polycistrons by Eukaryotic Ribosomes (SEMPER)

4a. Abstract

Applications of mammalian synthetic biology increasingly require the ability to express multiple proteins at user-determined stoichiometries from single, compactly encoded transcripts. Here we present an approach for expressing multiple open reading frames (ORFs) from a single transcript, taking advantage of the leaky scanning model of translation initiation. In this method, adjacent ORFs are translated from a single messenger RNA at tunable ratios determined by their order in the sequence and the strength of their translation initiation sites. We call this approach Stoichiometric Expression of Messenger Polycistrons by Eukaryotic Ribosomes (SEMPER). We demonstrate the principles of this approach by expressing up to two *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, showing that enforcing the optimal ratio in every cell enables efficient formation of the multi-protein complex while minimizing cellular toxicity. SEMPER will enable a broad range of applications requiring tunable expression from compact eukaryotic constructs.*

4c. Introduction

In the realm of genetic engineering, achieving precise gene stoichiometry and polycistronic expression has emerged as a cornerstone for the efficient and robust recombinant expression of gas vesicles (GVs) in mammalian cells. This is particularly vital in the engineering of cell lines and circuits where GV expression is a critical input or output. Traditional methods partially address these challenges; for instance, viral elements like 2A or IRES sequences facilitate polycistronic expression. However, fine-tuning the expression ratios from a single mRNA transcript presents persistent difficulties. This necessitates the separation of the major structural protein GvpA onto a distinct cassette to ensure its adequate overexpression relative to assembly factors.

The breakthrough came somewhat unexpectedly, leading to the development of a novel, synthetic approach to polycistronic expression that also enables precise ratio tuning. This was inspired by research on short upstream open reading frames (uORFs), which demonstrated that the initiation of translation at any given ORF is probabilistic and can be modulated by altering the nucleotide sequence surrounding the start codon. A pivotal insight was realizing that ribosomes scanning an mRNA could bypass a translation initiation signal (TIS) and initiate translation at a downstream ORF based on a distinct probability of initiation.[55,56] By designing a single mRNA linking the ORFs for GvpA and assembly factors, we adjusted the translation initiation probability so that GvpA's initiation was significantly higher than that of the downstream assembly factors. This strategy enabled us to achieve the desired stoichiometry and produced GV amounts comparable to those of the multi-cassette system, signifying a major advancement in our approach.

Realizing the broader implications of this discovery, we named the method Stoichiometric Expression of Messenger Polycistrons by Eukaryotic Ribosomes (SEMPER). We embarked on an extensive characterization of SEMPER, exploring its myriad potential applications, from toxicity

rescue of misfolded proteins to mRNA-based antibody therapies. Our findings have been detailed in a manuscript currently under revision for publication and are the subject of multiple patent applications.[4]

The ensuing sections delve into the work I spearheaded, with invaluable assistance from my colleagues. *Italics* are used to denote text and figures sourced directly from our manuscript, with adjustments made to their arrangement for coherence in this context. Sections about three-ORF expression, computational simulation and polycistronic expression from IVT mRNA are omitted in this thesis as these efforts have been spearheaded by my colleague and co-author Ishaan Dev. However, these sections are available in the referenced pre-print and subsequent publication.

4b. Results

The SEMPER mechanism

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.[57,58] *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.*[59] *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).*[60] *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.*[61]

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.[62] *Recently, Ferreira and colleagues demonstrated that it is possible to use synthetic uORFs to regulate the expression of a downstream recombinant GOI.*[63] *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. 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.*

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 41A**). 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 following TIS sequence (NNNAUGG) to initiate translation of our ORFs, where NNN represents one of the following sequences in order of decreasing translation initiation efficiency: ACC, CCC, TTT.[63] *We engineered all our FPs to contain a valine residue (GTG) following the N-terminal methionine to ensure changes in translation initiation were due to the trinucleotide preceding the start codon. 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.*[64,65] *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 41B**). 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,*[66,67] *the mCherry allowed us to normalize single-cell fluorescence measurements for msfGFP[r5M] and mEBFP2, a strategy common to LRS-focused studies.*[63,68] *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.*

*After cloning various combinations of TISs in front of our two ORFs, we transfected these "2-ORF SEMPER" constructs into 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. As hypothesized, our cell lines produced both msfGFP[r5M] and mEBFP2 from single transcripts **(Figure 41C, left)**. Strikingly, 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 **(Figure 41C, right)**. As we increased the TIS strength in front of msfGFP[r5M], we observed increases in msfGFP[r5M] relative translation levels and decreases in mEBFP2 relative 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 **(Figure 41C, right)**.*

To confirm generalizability across species, 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 cell line for the production of biologics (Figure S1A-B).26 Upon comparing the distributions of log10(msfGFP[r5M]/mEBFP2) values between cell types (Figure 1E), we

determined that the three TIS combinations tested maintained the same rank order in both HEK293T and CHO-K1 cell lines.

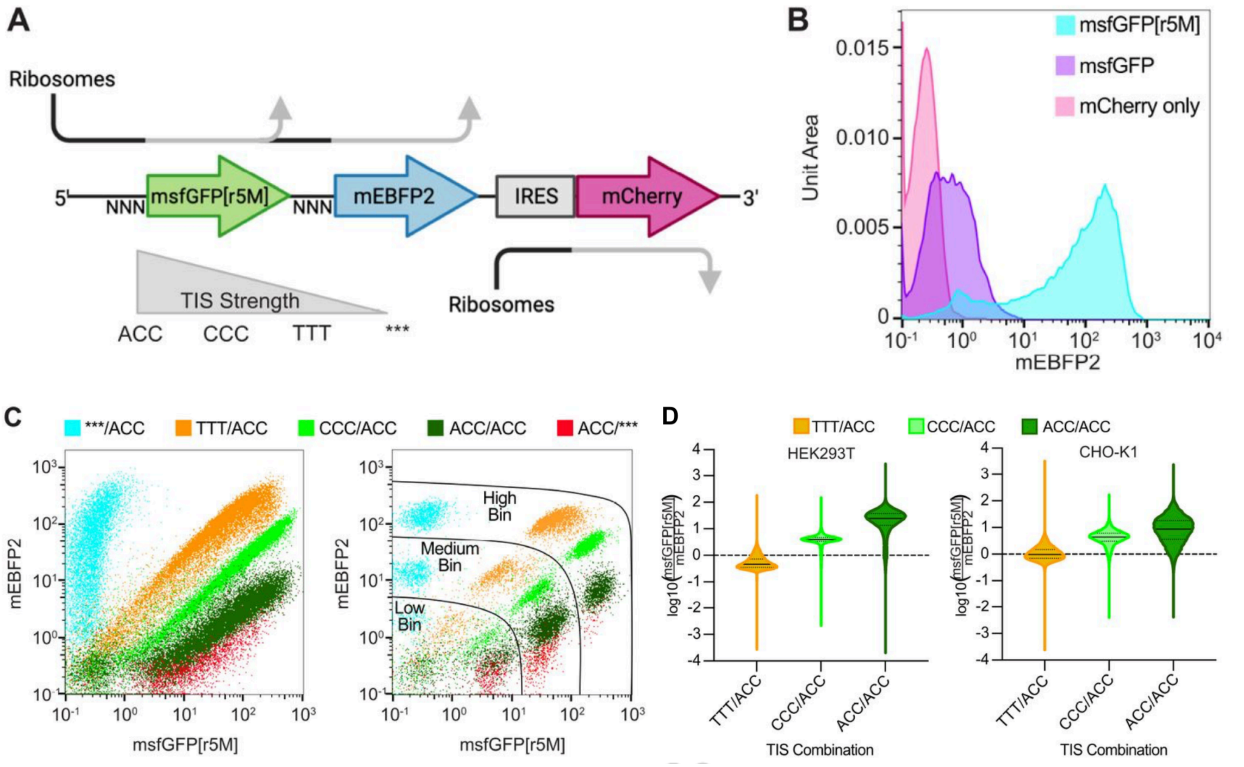


Figure 41 | 2-ORF SEMPER constructs demonstrate tunable, bicistronic expression. (A) *Architecture of an mRNA transcript produced by transfected 2-ORF SEMPER plasmid DNA. Cap-dependent ribosomes translate ORF 1, msfGFP[r5M], or ORF 2, mEBFP2, with frequencies dependent on the trinucleotide (NNN) upstream of each ORF. The relative strengths of the trinucleotides and TISs they represent are depicted. *** (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 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 replicates). (Right) Binning cells into three mCherry fluorescence ranges produces unique clusters of cells (N=1, representative of four replicates).* ***(D)*** *Violin plots of log10(msfGFP[r5M]/mEBFP2) values for all mCherry positive cells for four combined replicates of TTT/ACC, CCC/ACC, and ACC/ACC plasmids transfected into HEK293T and CHO-K1 cell lines. The median and quartiles of the distribution are represented by the solid and dotted lines respectively.*

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).[69] *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.*[70,71] *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.[72,73] *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). Additionally, we have found that co-transfecting the pgvpA-IRES-mCherry in excess of pgvpNJKFGW-EmGFP improves acoustic contrast.*[69] *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.*[74]

*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 42A**). We compared these SEMPER constructs to our published two-vector expression system, 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 42B**), as expected from our 2-ORF SEMPER FP experiments. As measured by BURST ultrasound imaging,*[75] *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 42C,D**).*

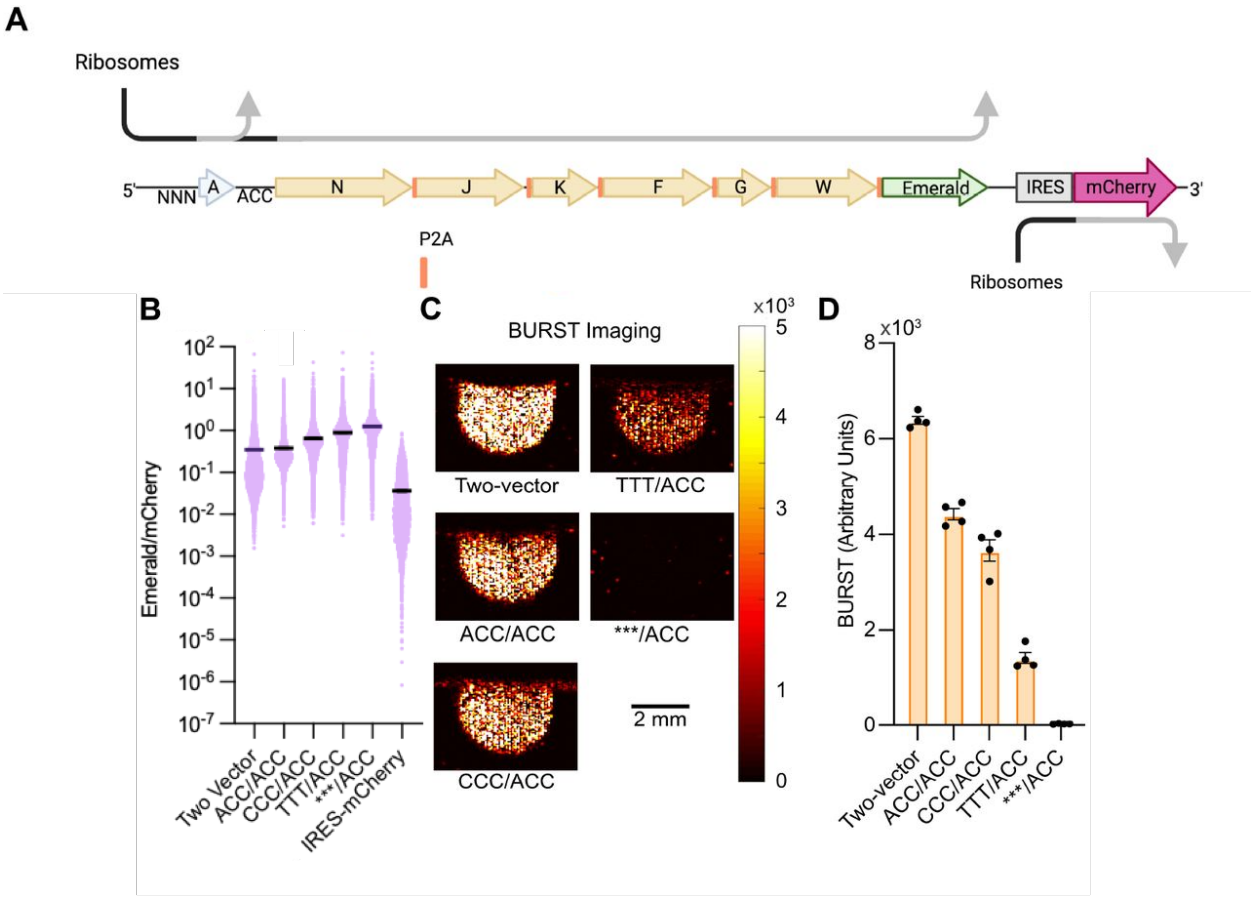


Figure 42 | Utilizing 2-ORF SEMPER constructs to express gas vesicles in mammalian cells. (A) *Architecture of an mRNA transcript transcribed from transfected SEMPER mARG plasmid DNA. The first ORF, gvpA, encodes the main structural protein, while the second ORF, gvpNJKFGW-EmGFP, encodes all necessary 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 line depicts the mean of the distribution. Error bars depict standard error of the mean (SEM) (N=1, representative of four replicates).* ***(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 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 depict SEM (N=4).*

Relative ribosome translation initiation affects Gas Vesicle expression and acoustic properties

We characterized produced gas vesicles using two ultrasound modalities, namely BURST imaging[3,25], which derives contrast from the collapse of GVs when a high acoustic pressure is applied, and amplitude modulation imaging (AM)[18], which derives contrast from the nonlinear scattering of ultrasound signal by buckling GVs with reduced shell stiffness. For this experiment, transiently transfected HEK293T cells were harvested after 3 days of expression and embedded in agarose phantoms for ultrasound imaging. ***Our results show that GVs produced with the strongest GvpA TIS produced the highest AM signal which dropped significantly with decreasing TIS strength (Figure 43A). Interestingly, the BURST signal did not start to drop until the fourth strongest TIS (TTC), implying that redistribution of the ribosome flux between the two ORFs, and also their relative expression, can modulate not only the gas vesicle expression level but also acoustomechanic properties of GVs (Figure 43B)****. Lastly, we lysed the cells and*

purified GVs for TEM imaging. All TIS strengths produced visible Gas Vesicle structures characterized by their inherent TEM contrast in the absence of staining. The formation of GVs requires the expression of GvpA and the accessory proteins, confirming that both the upstream and downstream ORFs are translated in all our constructs that contain start codons on both ORFs ***(Figure 43C)****.*

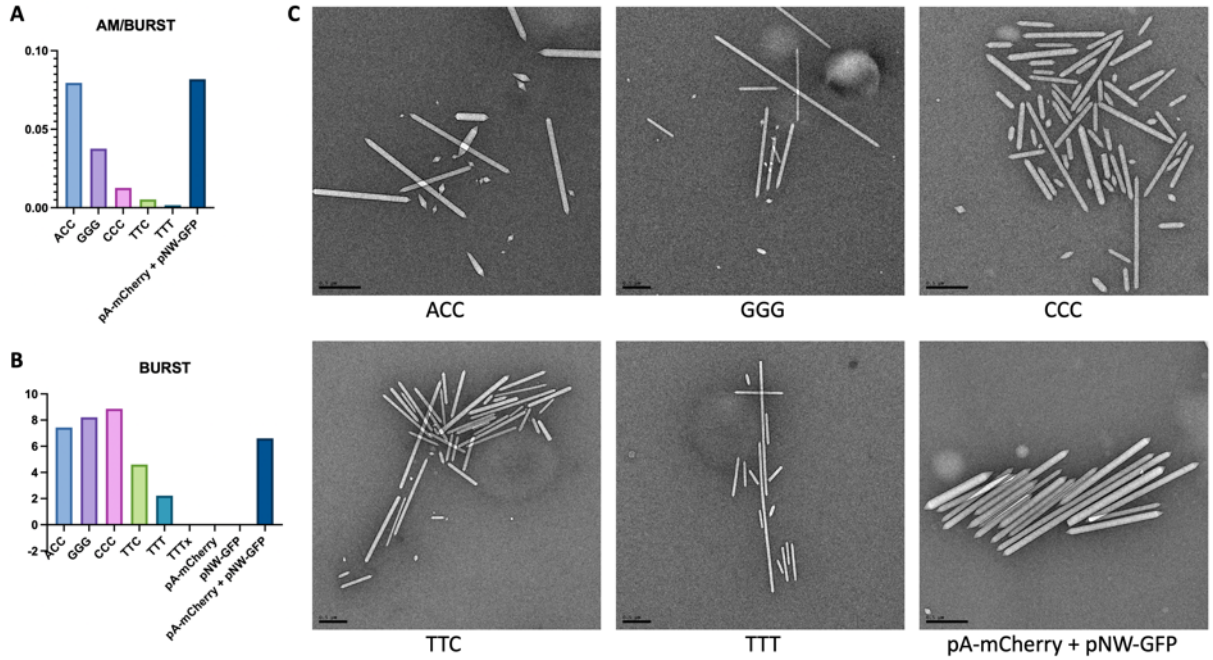


Figure 43 | Ratio tuning of SEMPER mARG affects GV expression and GV acoustic properties. (A) *BURST normalized AM signal of mARG expressing HEK293T cells. Three-letter labels indicate the three bases upstream of GvpA start codon on each SEMPER mARG plasmid. Plasmids were transfected using Polyethyleneimine. pNW-GFP was supplied at 1/4 of the amount of pA-mCherry, which was supplied at the same amount as SEMPER mARG plasmids.* ***(B)*** *Background normalized BURST signal. Annotations as in A.* ***(C)*** *TEM micrographs of purified gas vesicles. Cells were lysed with Solulyse-M and Benzonase, lysates were overlayed with 10mM HEPES pH7.5, centrifuged overnight at 300xg, 4˚C and the top fraction was collected. Scalebar is 0.5 microns*

SEMPER mARG expression system reduces cell toxicity

*A significant 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 44A****). This is not surprising due to the inherent stochasticity of transient co-transfection. As GvpA subunits have been speculated to nonspecifically aggregate when expressed without GV assembly factors,*[76,77] *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 44b)**. 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 significantly 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.*

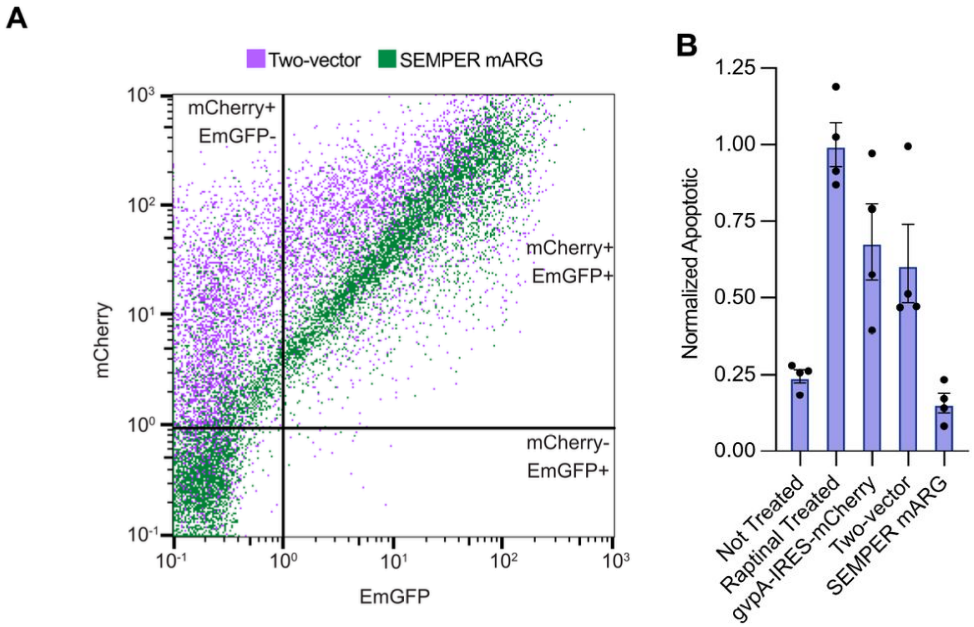


***Figure 44 | SEMPER mARGs are not inhibited by the stochasticity of co-transfection of two plasmids. (A)** Comparison of mCherry and EmGFP fluorescence in HEK293T cells transfected with the two-vector mARG expression system or the SEMPER mARG expression system. The gvpA-IRES-mCherry plasmid was transfected in a 4-fold molar excess relative to pgvpNJKFGW-EmGFP, leading to a significant 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. **(B)** 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 depict SEM (N=4).*

Recombinant b12 IgG expression in HEK293T cells using SEMPER

*Next, we successfully expressed and secreted recombinant b12 IgG in HEK293T cells using the SEMPER system. The SEMPER-b12 plasmid was designed to express both heavy and light chains of the b12 IgG from a single mRNA transcript with tunable HC to LC stoichiometry **(Figure 45a).** For comparison to the traditional approach, we employed a two-plasmid system, expressing separate heavy (HC) and light (LC) chain plasmids **(Figure 45b).** For this, we used modified versions of the SEMPER-b12 plasmid, where the start codon was omitted from either the heavy chain or*

the light chain, effectively creating single-chain expressing control of the same plasmid size as the SEMPER-b12 plasmids to reduce size-dependent variability in transfection efficiency. This served as a control to evaluate the efficacy of SEMPER in coordinating the expression of IgG chains from a single transcript. IgG production was quantified from a sample of media supernatant using the Human Total IgG Coated ELISA Kit.

Our results indicated that the SEMPER-b12 plasmid achieved comparable levels of IgG expression compared to the two-plasmid control. Specifically, IgG production was achieved by the optimal SEMPER-b12 system at approximately 0.2 ug/mL, a level statistically lower than the two-plasmid approach as shown in Figure 45c, although the magnitude of this difference is likely not biologically significant. This finding underlines the capability of SEMPER to streamline antibody production by simplifying the genetic engineering required for bi-functional protein expression. In addition to yield, the stoichiometry of heavy and light chain expression is critical for functional antibody assembly. The SEMPER system's ability to control expression stoichiometry was evident, as demonstrated by varying levels of IgG production across different expression ratios of the heavy and light chain ***(Figure 45c)****. This aspect is particularly advantageous over traditional methods that often require fine-tuning of individual plasmid transfections to achieve balanced expression.*

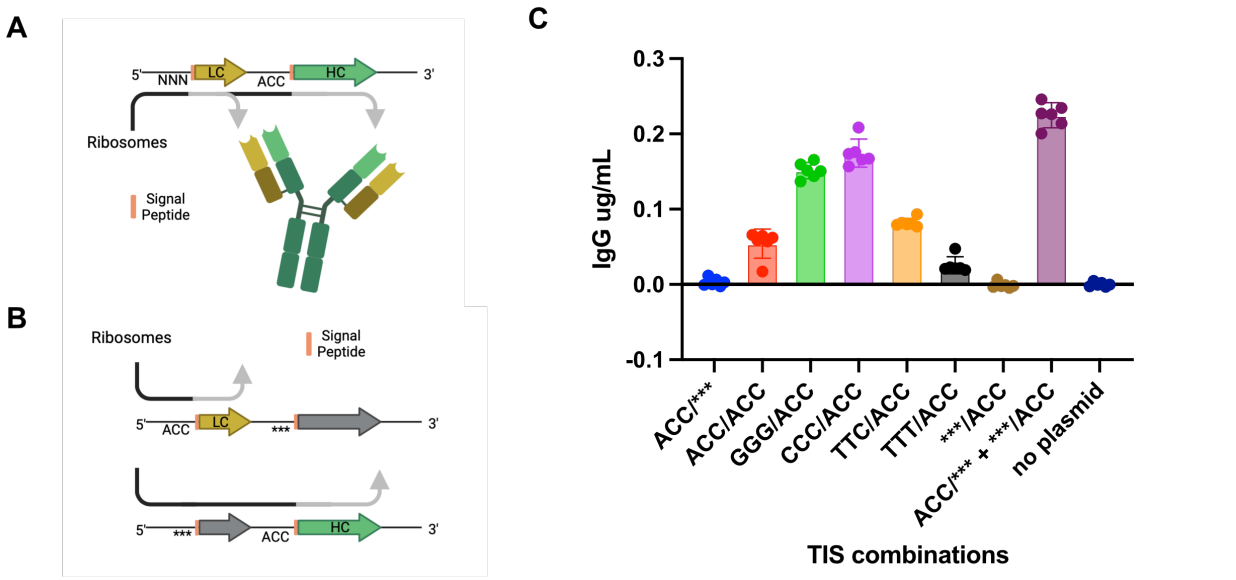


Figure 45 | Recombinant expression of SEMPER mAb. (A) *Architecture of an mRNA transcript transcribed from transfected SEMPER-b12 plasmid DNA. The first ORF, LC, encodes the b12 IgG light chain, while the second ORF, HC, encodes the b12 IgG heavy chain. Both polypeptides contain a signal peptide for secretion. The TIS of the first ORF, NNN, is varied while the TIS of the second ORF, ACC, remains constant in different ratio constructs. B) Architecture of the mRNA transcripts transcribed from transfected SEMPER-b12 plasmid DNA used for positive (ACC/***+***/ACC) and negative (ACC/*** and ***/ACC) controls C) Total human IgG ELISA results following expression of SEMPER-b12 vectors. HEK293T cells transfected with either the light chain (ACC/***) or heavy chain (***/ACC) only produced no detectable IgG similar to the mock transfection (no plasmid). However, the 1:1 mixture of the two plasmids while preserving the total amount of transfected DNA (ACC/***+***/ACC) produced the highest measured IgG yield. SEMPER-b12 constructs with varying expression stoichiometry produced intermediate amounts of IgG with CCC/ACC producing the highest yield, comparable to the two-plasmid controls. Error bars depict 95% CI (N=6).*

These findings validate the utility of SEMPER for the production of complex, bi-functional and secreted proteins like IgGs in mammalian cells. The efficiency and simplicity of this system, requiring only a single transcript for dual peptide expression and secretion, present a significant advancement in biotechnological applications, particularly in therapeutic antibody production.

4b. 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 variable strength TISs. The SEMPER paradigm is applicable across cells from multiple species and can yield a range of user-tunable expression stoichiometries for at least three GOIs. 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. This is particularly valuable in scenarios demanding high toxin/antitoxin ratios for optimal production, where traditional cotransfection methods fall short, especially in in vivo settings where the DNA/mRNA-to-cell ratios are orders of magnitude lower. Additionally, the application of SEMPER for monoclonal antibody (mAb) expression not only showcases its potential in synthetic biology but also opens up significant opportunities in therapeutic applications. Finally, we show 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 new avenues to further tune protein stoichiometries with synthetic mRNA.

In the development of the SEMPER system, we utilized TIS sequences from Ferreira et al.[55] *as an established and accessible reference to demonstrate our system's capabilities. However, the SEMPER framework is designed to be independent of the specific methodology for determining TIS strength, and is also agnostic to the underlying biochemical mechanisms underlying the variability in TIS strengths. This flexibility highlights the potential of SEMPER as a foundational tool in synthetic biology.*

Our approach positions the SEMPER system as a versatile 'knob' for gene expression control. We demonstrate its functionality using established TIS strengths, focusing on providing a tool for expression modulation. The determination of specific TIS strengths, and the exploration of their biochemical mechanisms, are left for other studies, such as those employing predictive models like position weight matrix.[56]

Moreover, we acknowledge that the exact ratios from specific TIS combinations may vary across different sets of ORFs due to complex gene expression dynamics, influenced by factors like ORF lengths, codon optimization or resource utilization. Nevertheless, the rank order of expression ratios is maintained across any set of TIS combinations. This consistency is a key aspect of our system's utility, allowing for predictable modulation of gene expression within a defined range. This principle is supported by our experiments with diverse ORFs like GV genes and recombinant b12 IgG, showcasing the relative consistency of TIS combinations on rank order of expression stoichiometries.

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 important 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.

We believe SEMPER represents a significant advancement in the field of recombinant multi-protein expression and synthetic biology, enabling 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.[78] *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 novel strategy to tackle protein misfolding precisely. SEMPER's compact framework and tunability make it a critical step towards increasingly complex engineering efforts and finer control of mammalian systems. 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.[79,80] *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.*[81–83] *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.*

Chapter 5. Conclusions

This thesis has charted a significant journey through the development and optimization of second-generation mammalian Acoustic Reporter Genes (mARGs), the exploration of gas vesicle nucleation and assembly, and the introduction of Stoichiometric Expression of Messenger Polycistrons by Eukaryotic Ribosomes (SEMPER). Through these interconnected studies, we have advanced the capabilities of non-invasive, real-time imaging of cellular processes within live

organisms, enhanced our understanding of the structural, genetic, and biochemical principles underlying gas vesicle assembly, and introduced a versatile tool for the precise control of gene expression.

Regarding the development of $mARG_{Ana}$, we achieved significant improvements in the expression, stability, and imaging capabilities of mARGs, overcoming limitations related to gene silencing, expression variability, and the need for destructive imaging modalities. The introduction of $mARG_{Ana}$, adapted from the Anabaena flos-aquae GV operon, stands as a highlight, enabling robust and tunable gene expression conducive to non-destructive, nonlinear ultrasound imaging.

Our investigations into the mechanisms of GV nucleation and assembly have identified several key protein interactions essential for GV formation, providing a foundation for future work to fully elucidate their function.

The SEMPER framework represents a significant leap forward, allowing for the expression of multiple proteins at precise stoichiometries from single, compact transcripts. This methodology not only facilitates the formation of multi-protein complexes but also minimizes cellular toxicity, broadening the applications of genetic engineering in research and therapeutic contexts.

References

1. Hurt, R. C. *et al.* Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. *Nat Biotechnol* **41**, 919–931 (2023).
2. Bourdeau, R. W. *et al.* Acoustic reporter genes for noninvasive imaging of microorganisms in mammalian hosts. *Nature* **553**, 86–90 (2018).
3. Farhadi, A., Ho, G. H., Sawyer, D. P., Bourdeau, R. W. & Shapiro, M. G. Ultrasound imaging of gene expression in mammalian cells. *Science* **365**, 1469–1475 (2019).
4. Duan, M., Dev, I., Lu, A., You, M. Y. & Shapiro, M. G. Stoichiometric expression of messenger polycistrons by eukaryotic ribosomes (SEMPER) for compact, ratio-tunable multi-gene expression from single mRNAs. 2023.05.26.541240 Preprint at https://doi.org/10.1101/2023.05.26.541240 (2023).
5. Piraner, D. I. *et al.* Going Deeper: Biomolecular Tools for Acoustic and Magnetic Imaging and Control of Cellular Function. *Biochemistry* **56**, 5202–5209 (2017).
6. Foster, F. S., Pavlin, C. J., Harasiewicz, K. A., Christopher, D. A. & Turnbull, D. H. Advances in ultrasound biomicroscopy. *Ultrasound in Medicine and Biology* **26**, 1–27 (2000).
7. Maresca, D. *et al.* Biomolecular Ultrasound and Sonogenetics. *Annu. Rev. Chem. Biomol. Eng.* **9**, 229–252 (2018).
8. Shapiro, M. G. *et al.* Biogenic gas nanostructures as ultrasonic molecular reporters. *Nature Nanotech* **9**, 311–316 (2014).
9. Walsby, A. E. Gas vesicles. *Microbiol. Rev.* **58**, 94–144 (1994).
10. Pfeifer, F. Gas Vesicles of Archaea and Bacteria. in *Complex Intracellular Structures in Prokaryotes* (ed. Shively, J. M.) vol. 2 115–140 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).
11. Staley, J. T. The gas vacuole: An early organelle of prokaryote motility? *Origins Life Evol Biosphere* **10**, 111–116 (1980).
12. Walsby, A. E. Structure and function of gas vacuoles. *Bacteriological Reviews* **36**, 1–32 (1972).
13. Lakshmanan, A. *et al.* Molecular Engineering of Acoustic Protein Nanostructures. *ACS Nano* **10**, 7314–7322 (2016).
14. Pfeifer, F. Distribution, formation and regulation of gas vesicles. *Nat Rev Microbiol* **10**, 705–715 (2012).
15. Tashiro, Y., Monson, R. E., Ramsay, J. P. & Salmond, G. P. C. Molecular genetic and physical analysis of gas vesicles in buoyant enterobacteria. *Environ Microbiol* **18**, 1264–1276 (2016).
16. Huber, S. T., Terwiel, D., Evers, W. H., Maresca, D. & Jakobi, A. J. Cryo-EM structure of gas vesicles for buoyancy-controlled motility. *Cell* **186**, 975-986.e13 (2023).
17. Dutka, P. *et al.* Structure of *Anabaena flos-aquae* gas vesicles revealed by cryo-ET. *Structure* **31**, 518-528.e6 (2023).
18. Maresca, D. *et al.* Nonlinear ultrasound imaging of nanoscale acoustic biomolecules. *Appl. Phys. Lett.* **110**, 073704 (2017).
19. Maresca, D., Sawyer, D. P., Renaud, G., Lee-Gosselin, A. & Shapiro, M. G. Nonlinear X-Wave Ultrasound Imaging of Acoustic Biomolecules. *Phys. Rev. X* **8**, 041002 (2018).
20. Smits, M., Oyserman, B. & Medema, M. Genomic analysis of microbial gas vesicle gene clusters. 16.

21. Xu, B.-Y. *et al.* Structure of the gas vesicle protein GvpF from the cyanobacterium *Microcystis aeruginosa*. *Acta Crystallogr D Biol Crystallogr* **70**, 3013–3022 (2014).
22. Völkner, K., Jost, A. & Pfeifer, F. Accessory Gvp Proteins Form a Complex During Gas Vesicle Formation of Haloarchaea. *Front. Microbiol.* **11**, 610179 (2020).
23. Iburg, M. *et al.* Elucidating the Assembly of Gas Vesicles by Systematic Protein-Protein Interaction Analysis. 2023.07.23.550228 Preprint at https://doi.org/10.1101/2023.07.23.550228 (2023).
24. Davies, J. M., Brunger, A. T. & Weis, W. I. Improved Structures of Full-Length p97, an AAA ATPase: Implications for Mechanisms of Nucleotide-Dependent Conformational Change. *Structure* **16**, 715–726 (2008).
25. Sawyer, D. P. *et al.* Ultrasensitive ultrasound imaging of gene expression with signal unmixing. *Nat Methods* **18**, 945–952 (2021).
26. Duan, M. *et al.* Gatekeeper helix activates Golgi SM protein Sly1 and directly mediates close-range vesicle tethering. 2020.01.16.906719 Preprint at https://doi.org/10.1101/2020.01.16.906719 (2020).
27. Nakai, J., Ohkura, M. & Imoto, K. A high signal-to-noise Ca2+ probe composed of a single green fluorescent protein. *Nat Biotechnol* **19**, 137–141 (2001).
28. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38 (1996).
29. Heim, R., Cubitt, A. B. & Tsien, R. Y. Improved green fluorescence. *Nature* **373**, 663–664 (1995).
30. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79–88 (2006).
31. Heim, R., Prasher, D. C. & Tsien, R. Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proceedings of the National Academy of Sciences* **91**, 12501–12504 (1994).
32. Jin, Z. *et al.* Ultrasonic reporters of calcium for deep tissue imaging of cellular signals. 2023.11.09.566364 Preprint at https://doi.org/10.1101/2023.11.09.566364 (2023).
33. Real-time monitoring of tumor-homing bacteria and tumor cells in vivo using ultrasound. *Nat Biotechnol* **41**, 913–914 (2023).
34. Maresca, D. *et al.* Nonlinear ultrasound imaging of nanoscale acoustic biomolecules. *Appl Phys Lett* **110**, (2017).
35. Maresca, D., Sawyer, D. P., Renaud, G., Lee-Gosselin, A. & Shapiro, M. G. Nonlinear X-Wave Ultrasound Imaging of Acoustic Biomolecules. *Phys. Rev. X* **8**, 041002 (2018).
36. Zhou, X., Vink, M., Klaver, B., Berkhout, B. & Das, A. T. Optimization of the Tet-On system for regulated gene expression through viral evolution. *Gene Ther* **13**, 1382–1390 (2006).
37. Loew, R., Heinz, N., Hampf, M., Bujard, H. & Gossen, M. Improved Tet-responsive promoters with minimized background expression. *BMC Biotechnology* **10**, 81 (2010).
38. Chu, J. *et al.* A bright cyan-excitable orange fluorescent protein facilitates dual-emission microscopy and enhances bioluminescence imaging in vivo. *Nat Biotechnol* **34**, 760–767 (2016).
39. Farhadi, A. *et al.* Genetically encoded phase contrast agents for digital holographic microscopy. *Nano Lett* **20**, 8127–8134 (2020).
40. Schrunk, E., Dutka, P., Hurt, R. C., Wu, D. & Shapiro, M. G. Bioorthogonal labeling enables in situ fluorescence imaging of expressed gas vesicle nanostructures.

41. Chatterjee, A., Guo, J., Lee, H. S. & Schultz, P. G. A Genetically Encoded Fluorescent Probe in Mammalian Cells. *J. Am. Chem. Soc.* **135**, 12540–12543 (2013).
42. Zagotta, W. N. *et al.* An improved fluorescent noncanonical amino acid for measuring conformational distributions using time-resolved transition metal ion FRET. *eLife* **10**, e70236 (2021).
43. Jones, C. M. *et al.* Genetic encoding of a highly photostable, long lifetime fluorescent amino acid for imaging in mammalian cells. *Chem. Sci.* **12**, 11955–11964 (2021).
44. Soltermann, F. *et al.* Quantifying Protein–Protein Interactions by Molecular Counting with Mass Photometry. *Angewandte Chemie International Edition* **59**, 10774–10779 (2020).
45. Foley, E. D. B., Kushwah, M. S., Young, G. & Kukura, P. Mass photometry enables label-free tracking and mass measurement of single proteins on lipid bilayers. *Nat Methods* **18**, 1247–1252 (2021).
46. Young, G. & Kukura, P. Interferometric Scattering Microscopy. *Annu. Rev. Phys. Chem.* **70**, 301–322 (2019).
47. Mlouka, A., Comte, K., Castets, A.-M., Bouchier, C. & Tandeau de Marsac, N. The Gas Vesicle Gene Cluster from Microcystis aeruginosa and DNA Rearrangements That Lead to Loss of Cell Buoyancy. *JB* **186**, 2355–2365 (2004).
48. Reis, A. C. & Salis, H. M. An Automated Model Test System for Systematic Development and Improvement of Gene Expression Models. *ACS Synth. Biol.* **9**, 3145–3156 (2020).
49. Cai, K. *et al.* The model cyanobacteria Anabaena sp. PCC 7120 possess an intact but partially degenerated gene cluster encoding gas vesicles. *BMC Microbiol* **20**, 110 (2020).
50. Ho, C.-M. *et al.* Malaria parasite translocon structure and mechanism of effector export. *Nature* **561**, 70–75 (2018).
51. Söllner, T. *et al.* SNAP receptors implicated in vesicle targeting and fusion. *Nature* **362**, 318–324 (1993).
52. Song, H., Orr, A., Duan, M., Merz, A. J. & Wickner, W. Sec17/Sec18 act twice, enhancing membrane fusion and then disassembling cis-SNARE complexes. *eLife* **6**, e26646 (2017).
53. Snoberger, A., Anderson, R. T. & Smith, D. M. The Proteasomal ATPases Use a Slow but Highly Processive Strategy to Unfold Proteins. *Front Mol Biosci* **4**, 18 (2017).
54. Fleur, T. L., Hossain, A. & Salis, H. M. Automated Model-Predictive Design of Synthetic Promoters to Control Transcriptional Profiles in Bacteria. 2021.09.01.458561 Preprint at https://doi.org/10.1101/2021.09.01.458561 (2021).
55. Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. *Proceedings of the National Academy of Sciences* **110**, 11284–11289 (2013).
56. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology* **10**, 748 (2014).
57. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**, 113–127 (2010).
58. Sonenberg, N. & Hinnebusch, A. G. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell* **136**, 731–745 (2009).
59. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences* **109**, E2424–E2432 (2012).

60. Kozak, M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299**, 1–34 (2002).
61. Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292 (1986).
62. Barbosa, C., Peixeiro, I. & Romão, L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLOS Genetics* **9**, e1003529 (2013).
63. Ferreira, J. P., Overton, K. W. & Wang, C. L. Tuning gene expression with synthetic upstream open reading frames. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11284–11289 (2013).
64. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* **24**, 79–88 (2006).
65. Soundrarajan, N. *et al.* Conjugation of Proteins by Installing BIO-Orthogonally Reactive Groups at Their N-Termini. *PLoS ONE* **7**, e46741 (2012).
66. Kaminski, A., Howell, M. T. & Jackson, R. J. Initiation of encephalomyocarditis virus RNA translation: the authentic initiation site is not selected by a scanning mechanism. *EMBO J* **9**, 3753–3759 (1990).
67. Jackson, R. J., Howell, M. T. & Kaminski, A. The novel mechanism of initiation of picornavirus RNA translation. *Trends in Biochemical Sciences* **15**, 477–483 (1990).
68. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS -seq. *Mol Syst Biol* **10**, 748 (2014).
69. Hurt, R. C. *et al.* Genomically mined acoustic reporter genes for real-time in vivo monitoring of tumors and tumor-homing bacteria. *Nat Biotechnol* 1–13 (2023) doi:10.1038/s41587-022-01581-y.
70. Farhadi, A., Ho, G. H., Sawyer, D. P., Bourdeau, R. W. & Shapiro, M. G. Ultrasound Imaging of Gene Expression in Mammalian Cells. *Science* **365**, 1469–1475 (2019).
71. Bourdeau, R. W. *et al.* Acoustic reporter genes for noninvasive imaging of microbes in mammalian hosts. *Nature* **553**, 86–90 (2018).
72. Huber, S. T., Terwiel, D., Evers, W. H., Maresca, D. & Jakobi, A. J. Cryo-EM structure of gas vesicles for buoyancy-controlled motility. *Cell* **186**, 975-986.e13 (2023).
73. Dutka, P. *et al.* Structure of Anabaena flos-aquae gas vesicles revealed by cryo-ET. *Structure* **31**, 518-528.e6 (2023).
74. Wang, H. *et al.* Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium Anabaena sp. strain 90. *BMC Genomics* **13**, 613 (2012).
75. Sawyer, D. P. *et al.* Ultrasensitive ultrasound imaging of gene expression with signal unmixing. *Nat Methods* **18**, 945–952 (2021).
76. Jost, A., Knitsch, R., Völkner, K. & Pfeifer, F. Effect of Mutations in GvpJ and GvpM on Gas Vesicle Formation of Halobacterium salinarum. *Frontiers in Microbiology* **12**, (2021).
77. Pfeifer, F. Recent Advances in the Study of Gas Vesicle Proteins and Application of Gas Vesicles in Biomedical Research. *Life (Basel)* **12**, 1455 (2022).
78. Gao, X. J., Chong, L. S., Kim, M. S. & Elowitz, M. B. Programmable protein circuits in living cells. *Science* **361**, 1252–1258 (2018).
79. Hoffmann, M. A. G. *et al.* ESCRT recruitment to SARS-CoV-2 spike induces virus-like particles that improve mRNA vaccines. *Cell* (2023) doi:10.1016/j.cell.2023.04.024.
80. Cohen, A. A. *et al.* Mosaic nanoparticles elicit cross-reactive immune responses to zoonotic coronaviruses in mice. *Science* **371**, 735–741 (2021).

81. Van Hoecke, L. & Roose, K. How mRNA therapeutics are entering the monoclonal antibody field. *Journal of Translational Medicine* **17**, 54 (2019).
82. Pardi, N. *et al.* Administration of nucleoside-modified mRNA encoding broadly neutralizing antibody protects humanized mice from HIV-1 challenge. *Nat Commun* **8**, 14630 (2017).
83. Stadler, C. R. *et al.* Elimination of large tumors in mice by mRNA-encoded bispecific antibodies. *Nat Med* **23**, 815–817 (2017).