Engineering novel rhodopsins for neuroscience

Thesis by

Claire N. Bedbrook

In Partial Fulfillment of the Requirements for

the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2018

(Defended May 10, 2018)

Claire Nicole Bedbrook
ORCID: 0000-0003-3973-598X

ACKNOWLEDGEMENTS

I love Caltech. Caltech is an intense scientific oasis full of outstanding minds with a different standard for quality. Things I will always remember about Caltech: steep learning curves, the expectation of scientific excellence, walks through the beautiful campus, and writing in Kerckhoff's Morgan library under the watchful eye of Caltech biology giants, Thomas Hunt Morgan, Max Delbrück, and Ed Lewis. Caltech has had a key role in shaping me into my best scientific self and I leave here with great pride for what I have accomplished.

I would like to thank my two advisors, Professor Frances Arnold and Professor Viviana Gradinaru. I started my PhD with a rotation in Frances' lab and I never left. Frances has fostered an amazing group that is always changing and challenging what is possible with proteins. Frances' high expectations pushed me to do my best work and I will forever strive to meet or exceed these expectations. Somehow, Frances holds herself to even higher expectations. Frances always asks the most important questions and has provided me with key advice throughout my PhD that I will continue to use for the rest of my career in science.

I was Viviana's second graduate student. The year I started my PhD, she started as an assistant professor at Caltech. Something that's truly special about meetings and interactions with Viviana is that they are filled with laughter. This is not to belittle the intensity and quality of the scientific discussion, but rather an example of how having fun enhances science. Viviana's excitement about projects and results is contagious and her thoughtful and caring nature inspires everyone in her lab to do better and be better. I am so grateful to have Viviana as a mentor.

I would also like to thank my committee, which in addition to my two advisors includes David Anderson and Mikhail Shapiro. I have the utmost respect for both David and Mikhail and am honored that they agreed to be on my committee.

Dr. Phil Patton was a significant contributor to my scientific development prior to Caltech. I think of Phil as my third PhD advisor. I worked at Pioneer hybrid for a year before coming to Caltech and Phil was my mentor. As a Caltech alumnus, Phil was a big reason why I chose to come to Caltech. Phil taught me how to present data effectively, to think creatively about scientific problems from the most tedious incremental detail to the big-picture goal, and to always feel comfortable asking questions.

I would like to thank my dad, Dr. John Bedbrook, my fourth PhD advisor. Dr. John Bedbrook is in the acknowledgment of every paper I have published during my time at Caltech because he is a constant contributor of advice, edits, and support through all my projects. He knows every detail of my PhD thesis. I call my dad on my walk home from work almost every day and he will ask about what data I collected that day, or ask when I will have a draft for the next paper, or tell me I should relax and take a weekend off. I am so thankful to have his support and scientific input, and every day I try to be more like him.

My sister, Jessie Bedbrook, has been a constant source of light and love throughout my PhD. She inspires me with her incredible work ethic, exceptional creativity, and never-ending ability to care for others.

I would like to thank my partner, Ravi Nath, who has been my companion on this Caltech PhD adventure. Ravi's positivity about life and science never ceases to amaze me. Ravi has wonderful and fearless scientific questions and thoughts. When we work together, I know we can change the world.

I dedicate my PhD thesis to my mum, Dr. Pamela Bedbrook. I will forever aspire to be like you. I know you would be proud of me.

ABSTRACT

The overarching goal of my PhD research has been engineering proteins capable of controlling and reading out neural activity to advance neuroscience research. I engineered light-gated microbial rhodopsins, primarily focusing on the algal derived, light-gated channel, channelrhodopsin (ChR), which can be used to modulate neuronal activity with light. This work has required overcoming three major challenges. First, rhodopsins are trans-membrane proteins, which are inherently difficult to engineer because the sequence and structural determinants of membrane protein expression and plasma membrane localization are highly constrained and poorly understood (Chapter 3-5). Second, protein properties of interest for neuroscience applications are assayed using very low throughput patch-clamp electrophysiology preventing the use of high-throughput assays required for directed evolution experiments (Chapter 2, 5-6). And third, *in vivo* application of these improved tools require either retention or optimization of multiple protein properties in a single protein tool; for example, we must optimize expression and localization of these algal membrane proteins in mammalian cells while at the same time optimizing kinetic and functional properties (Chapter 5-6). These challenges restricted the field to low-throughput, conservative methods for discovery of improved ChRs, e.g., structure-guided mutagenesis and testing of natural ChR variants. I used an alternative approach: data-driven machine learning to model the fitness landscape of ChRs for different properties of interest and applying these models to select ChR sequences with optimal combinations of properties (Chapters 5-6). ChR variants identified from this work have unprecedented conductance properties and light sensitivity that could enable non-invasive activation of populations of cells throughout the nervous system. These ChRs have the potential to change how optogenetics experiments are done. This work is a convincing demonstration of the power of machine learning guided protein engineering for a class of proteins that present multiple engineering challenges. A component of the novel application of these new ChR tools relies on recent advances in gene delivery throughout the nervous system facilitated by engineered AAVs (Chapter 7). And finally, I developed a behavioral tracking system to monitor behavior and demonstrate sleep behavior in the jellyfish *Cassiopea*, the most primitive organism to have this behavior formally characterized (Chapter 8).

PUBLISHED CONTENT

McIsaac RS, Bedbrook CN, & Arnold FH (2015) Recent advances in engineering microbial rhodopsins for optogenetics. Current opinion in structural biology 33:8-15. doi: 10.1016/j.sbi.2015.05.001
url: http://www.ncbi.nlm.nih.gov/pubmed/26038227
C.N.B. made figures and wrote the manuscript.

Flytzanis NC, Bedbrook CN, et al. (2014) Archaerhodopsin variants with enhanced voltage-sensitive fluorescence in mammalian and Caenorhabditis elegans neurons. Nature communications 5:4894. doi: 10.1038/ncomms5894
url: http://www.ncbi.nlm.nih.gov/pubmed/25222271
C.N.B. designed the experiments, performed the experiments, analyzed all of the data, and wrote the manuscript.

Bedbrook CN, et al. (2015) Genetically Encoded Spy Peptide Fusion System to Detect Plasma Membrane-Localized Proteins In Vivo. Cell Chemistry & biology 22(8):1108-1121. doi: 10.1016/j.chembiol.2015.06.020
url: http://www.ncbi.nlm.nih.gov/pubmed/26211362
C.N.B. conceived the project, designed the experiments, performed the experiments, analyzed all of the data, and wrote the manuscript.

Bedbrook CN, et al. (2017) Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. Proceedings of the National Academy of Sciences of the United States of America 114(13):E2624-E2633. doi: 10.1073/pnas.1700269114
url: http://www.ncbi.nlm.nih.gov/pubmed/28283661
C.N.B. conceived the project, designed research, performed research, analyzed data, and wrote the manuscript.

Bedbrook CN, Yang KK, Rice AJ, Gradinaru V, & Arnold FH (2017) Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. PLoS computational biology 13(10):e1005786. doi: 10.1371/journal.pcbi.1005786
url: http://www.ncbi.nlm.nih.gov/pubmed/29059183
C.N.B. conceived the project, designed the experiments, performed the experiments, analyzed data, and wrote the manuscript.

Bedbrook CN, Deverman, B.E., and Gradinaru V. (2018). Viral Strategies for Targeting the Central and Peripheral Nervous Systems. Annual Reviews Neuroscience 41, 323–48. (*in press*)
C.N.B. made figures and wrote the manuscript.

Nath RD, Bedbrook CN, et al. (2017) The Jellyfish Cassiopea Exhibits a Sleep-like State. Current biology: CB 27(19):2984-2990 e2983. doi: 10.1016/j.cub.2017.08.014
url: http://www.ncbi.nlm.nih.gov/pubmed/28943083
C.N.B. conceived the project, performed experiments and data analysis, and wrote the manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTARY FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION: ENGINEERING MICROBIAL RHODOPSINS FOR OPTOGENETICS

Adapted from (1)

1.1 Introduction

Optogenetics refers to the ability to control or monitor cellular activities with light ('opto') using genetically encoded machinery ('genetics'). For nearly a decade, a major focus of optogenetics has been neuroscience. Light-activated microbial rhodopsins can be transgenically expressed in neurons to reversibly control and sense neural activity with relevant speed and precision (2). Coupling targeted perturbations stimulated by light to specific readouts (e.g., behavioral phenotypes or electrical recordings) enables the functional dissection of neural circuits (3). Certain rhodopsins can also function as fluorescent voltage indicators providing optical detection of neuronal activity (and perhaps other electrically active cell types) (4, 5).

Rhodopsins are a family of light-activated integral membrane proteins that adopt a seven trans-membrane α-helical fold referred to as the G protein-coupled receptor fold. The polyene chromophore retinal is covalently attached to the ε-amino group of a conserved lysine residue on the seventh α-helix through a protonated Schiff base (PSB) linkage (6). In microbes, rhodopsins can act as receptors that change conformation in response to light to trigger intracellular signaling, as pumps that drive protons or chloride ions across the cell membrane, or as non-specific cation channels (7).

Microbial rhodopsin pumps and channels are widely used for optogenetic applications. Light-triggered isomerization of retinal from all-*trans* to 13-*cis* initiates the rhodopsin photocycle and ultimately results in the movement of ions across the membrane (6). When transgenically expressed in neurons, channelrhodopsins (ChRs) mediate light-dependent transport of cations into the cell, causing depolarization and stimulation of action potentials

(2, 8-11). In contrast to the excitatory ChRs, both proton and chloride-pumping rhodopsins can be used to selectively hyperpolarize the cell and inhibit action potentials through either pumping protons out or pumping chloride into the cell (2, 12). Collectively, these tools facilitate genetically targeted, fast, reversible loss and gain of function experiments *in vivo*. Since these proteins allow light- dependent 'actuation' of neuronal activity, we refer to them as actuators (**Figure 1.1**, **Table 1.1**).

Over the past few years, several proton-pumping rhodopsins have been identified that exhibit weak fluorescence that is sensitive to changes in the local electronic environment (e.g., changes in pH and trans-membrane voltage) (4, 5, 13). One proton pumping rhodopsin, Archaerhodopsin-3 (Arch) from *Halorubrum sodomense*, has been extensively characterized in mammalian neurons for both light-activated proton pumping and voltage sensitive fluorescence (5, 14-16). Wild-type Arch transports protons in response to light used to excite opsin fluorescence (635–655 nm). This activity can be attenuated or eliminated by introducing mutations at residues known to be critical for pumping (5, 14-16), thereby creating a tool for voltage sensing independent of hyperpolarization. We refer to these rhodopsin variants as sensors (**Figure 1.1**).

Rhodopsins have evolved to convert sunlight into a more useful currency for their microbial host. Since rhodopsins have been optimized with sunlight as the main substrate for activation, they have broad activation spectra in the visible range (400-650 nm) and require high intensity light for activation (~1 mW mm^{-2}, equivalent to the average intensity of sun light on the earth's surface) (17). The rhodopsins are also naturally low-conductance channels [single channel conductance of ChR2 < 1 pS (8, 18)] Rhodopsins' natural, broad activation spectra makes multiplexed control of cells with various light colors challenging due to spectral overlap, their poor sensitivity to low light levels necessitates delivery of high intensity light deep into brain tissue for neuronal activation (19), and their low conductance necessitates very high transgenic expression levels in neurons to produce sufficient photocurrents for neuronal activation (20). Further, the voltage-sensitive fluorescence detected from some rhodopsins is a byproduct of their natural chromophore mediated function (light-gated proton pumping) and thus has not been optimized through

evolution. As a result, current fluorescent variants are extremely dim, limiting the scope of potential ‘all-optical electrophysiology’ (14, 15, 21, 22).

Improving rhodopsin-based actuators and sensors has and will continue to require various elements of protein engineering. There is a diversity of available rhodopsin tools from both protein engineering and also from discovery of rhodopsin sequences found in different natural hosts. Within this collection of tools, each rhodopsin has different properties optimized for specific neuroscience applications. However, there are still gaps in the available optogenetic tool kit.

1.2 Rhodopsin engineering

Because of the limitations in screenability of the ChRs, it has not been possible to take full advantage of directed evolution techniques for optimization of different properties. While fluorescence and spectral properties of the rhodopsin sensors are screenable in high-throughput, other important properties like on/off kinetics and voltage sensitivity are not screenable in high throughput. Despite this limitation, there has been progress in rhodopsin engineering, the approaches used include recombination based methods where the positive properties of two or more rhodopsins are recombined to make a more optimal opsin, structure-guided directed mutagenesis, low-throughput screening of natural rhodopsin sequences, and more recently with our own work, machine learning guided protein engineering. We will highlight specific examples here and also expand on different approaches to this engineering challenge.

1.3 Spectral tuning of microbial rhodopsins

Microbial rhodospsin actuators from nature are optimally activated by light in the range of 450–570 nm. The absorption maximum of rhodopsin is determined by the energy gap between the resting state (S0) and excited state (S1) of the retinal chromophore. Narrowing or increasing the S0–S1 energy gap results in red or blue shifts, respectively. Stabilization of these states is governed by interactions between the protein and retinal, which itself is surrounded by a hydrophobic binding pocket with five conserved aromatic residues in

trans-membrane helix 3, 5, and 6 (23). Experimental and theoretical work suggests that the amino acids surrounding the retinal affect the S0-S1 energy gap by altering the polarity of the retinal binding cavity (24, 25) and the distance between the Schiff base linkage to retinal and its counter-ion (26). Empirically, it has been observed that red-shifted rhodopsin variants produce currents with blue light, while rhodopsins that peak with blue light often have no currents with red light. It is possible that it is difficult to generate a protein environment for the retinal that specifically promotes activation with lower energy light but not high-energy light.

Identifying variants with well-separated activation spectra is of great interest to neuroscience since it would enable multiplexed optogenetic control of excitation and inhibition using different colors of light in a single cell or in a mixed population of cells. Lin *et al.* reported a variant called ReaChR that is optimally excited by ~590 nm light and can be significantly excited by orange-red light in the range 590–630 nm (27). ReaChR is an engineered chimeric variant of VChR1, a cation-conducting ChR from *Volvo carteri* that is maximally excited at 535 nm (27, 28). ReaChR has helix 6 replaced with that of VChR2 (also from *V. carteri*), which improves protein expression, and has the sequence of ChR1 from *Chlamydamonas reinhardtii* at the N-terminus, which further improves plasma membrane localization. To further improve the chimera's spectral properties, a number of single amino acid mutations were tested based on mutations that had previously been shown to alter ChR light activation properties. One such single amino acid mutation (L171I) increased the amplitude of the photo response at 610 nm and 630 nm (27). The L171 position was previously mutated in the ChR chimera ChIEF (29) and was targeted because of its position proximal to the retinal-binding pocket. ReaChR demonstrates that transferring mutations or even parts of domains between variants can confer desired properties (i.e., improved photostability and membrane localization). More broadly, chimeragenesis has proven to be a good engineering strategy to achieve spectral shifts in ChRs: in an earlier study from Prigge *et al.*, helix swapping between ChR1, ChR2, VChR1, and VChR2 resulted in variants with red and blue-shifted spectra, though none as red-shifted as ReaChR (30).

Spectral tuning of ChRs using higher throughput approaches has remained a challenge in part due to limited ChR expression in *Escherichia coli*, a common host for directed evolution (31, 32). The presence of predicted N-glycosylation sites in several rhodopsins suggests that glycosylation, which *E. coli* does not naturally perform, is required for functional ChR expression (32). If the lack of glycosylation is limiting expression, then expressing ChRs in *E. coli* with a reconstituted eukaryotic glycosylation pathway (which was recently reported in (33)) may be possible. ChRs can be expressed in *Pichia pastoris* (18), suggesting that directed evolution should be possible in this system or in other laboratory yeasts such as *Saccharomyces cerevisiae*.

In contrast to ChRs, proton-pumping rhodopsins (PPRs) can typically be expressed in *E. coli*. Recently, spectral tuning of a PPR from *Gloeobacter violaceus* called GR, expressed in E coli, was performed using directed evolution in high-throughput (2000 variants/round of screening) (25). Site-saturation mutagenesis at 19 positions around the retinal chromophore followed by recombination of beneficial mutations and further site-saturation mutagenesis generated large spectral shifts in absorption spectra relative to wildtype GR. Collectively, variants with shifts of ~80 nm compared to wildtype GR were achieved. The large shifts, however, came at the cost of proton pumping capacity (25). Further characterization of evolved variants revealed that blue-tuning mutations modulate the polarity along the retinal chromophore. Blue-tuning mutations near the PSB generally increased polarity relative to the native residues, while blue-tuning mutations near the beta-ionone ring decreased polarity (25), consistent with recent theoretical predictions (34). In contrast, red-tuning mutations occurred near the PSB linkage to retinal and probably disrupted its interaction with the negatively charged counter-ion (25). While directed evolution is clearly an effective strategy for spectral tuning, identifying variants with large shifts in absorbance and wildtype activity levels remains a challenge that the screening methods used to date have not been able to address.

1.4 Engineering rhodopsin ion selectivity

Currently, inward-pumping chloride-transporting rhodopsins and outward-pumping proton-transporting rhodopsins are widely used for inhibiting neurons (2). Rhodopsin channels (ChRs) can transport many ions for every photon of absorbed light, while pumps can only move a single ion per photon. Increased efficiency of ion translocation enables targeted perturbations with less light (often advantageous for optogenetics applications) but comes at the cost of transient perturbations of membrane conductance. Engineering potassium and chloride-selective ChRs would enable selective inhibition in a way that better mimics natural neuronal physiology, with decreased photon flux.

Recently, two groups independently engineered ChR chloride channels that can silence neurons (35, 36) with the aid of the dark state crystal structure of the ChR variant, C1C2 (a chimera of ChR1 and ChR2) (23) (**Figure 1.2**). Berndt *et al.* speculated that since the ion-selectivity pore in C1C2 is less ordered than that in potassium-selective channels (37), natural cation-specific activity is driven by the electrostatic potential surrounding the C1C2 pore and vestibule (35). By identifying single amino acid mutations in this region that modified the channel reversal potential and combining the single mutations into a variant called inhibitory C1C2 (iC1C2), they created a chloride-specific channel that can silence action potentials in response to light (35). Wietek *et al.* took a different approach: using molecular dynamics simulations, they identified five residues that form a hydrophobic barrier in darkness to prevent water from entering the protein vestibule (36). One of these residues, E90, when mutated to lysine or arginine, decreased ChR2's reversal potential and turned ChR2 into a light-activated chloride channel at membrane holding potentials above about -40 mV. Introduction of the T159C mutation improved membrane targeting of the protein in mammalian cells (36). The resulting variant, ChloC, required two mutations to transform ChR2 into an effective tool for silencing action potentials in neurons in the presence of light (36).

Ideally, inhibitory channels would have a decelerated channel closure, which would enable a prolonged ion-conducting state with a brief light stimulation. This has been achieved for the excitatory channel, ChR2, by introduction of a mutation at C128 which significantly decreased the time for channel closure once light is turned off (off kinetics, t_{off}) (38). The

C128 mutation was introduced into ChR2 by analogy to previous work on bacteriorhodopsin (bR), a light-driven proton pump, showing that the equivalent position in bR, when mutated, affects kinetics of the photocycle and lifetimes of intermediates (39). The C128 residue is within 4 Å of the 12th carbon of retinal and, based on the C1C2 crystal structure (23), the thiol group is associated with the p-electron system in the retinal molecule (23). Berndt *et al.* applied the equivalent mutation in iC1C2, which resulted in the construction of an inhibitory channel with slower channel closure that was named $SwiChR_{CT}$ (35). Wietek *et al.* engineered a slow-closing version of the inhibitory channel ChloC with mutations at position D156, a residue thought to interact with C128 (36).

1.5 Exploring natural variants for new rhodopsin actuators

Combining protein engineering with environmental sample mining via de novo transcriptome sequencing has led to the identification of dozens of new rhodopsins (40, 41). Two recently identified, valuable, ChRs, Chronos (activated with low intensity blue light) and Chrimson (activated with red light), together enable wavelength specific multiplexed perturbations of neurons (41).

A single mutation, K176R (which was previously shown to enhance photocurrents at the equivalent position in ChR2 (42)), was introduced into Chrimson to improve its slow kinetics to generate ChrimsonR (41).

Screening members of the cruxhalorhodopsin family led to identification of Halo57 from *H. salinarum* (40). Introducing two single mutations into Halo57 to boost photocurrents and appending trafficking sequences from (43) resulted in an optimized variant called Jaws, a red-shifted inhibitor of neuronal activity (40).

A major limitation in synchronous sensing and perturbing of neuronal activity for all-optical electrophysiology is that the light used to activate the actuator can perturb the fluorescence readout of the sensor. A highly light-sensitive, blue-shifted channelrhodopsin variant (sdChR, (41)) identified in a screen of plant genomes was further engineered for

faster kinetics and improved membrane localization to produce CheRiff to enable subcellular excitation (14) (**Figure 1.3**).

1.6 Machine-learning guided protein engineering of ChRs

To overcome the limits of functional screening throughput of interesting rhodopsin properties, we have used machine learning guided protein engineering enabling the data of a small set of variants to predict properties of a larger set of variants. Focusing on ChRs we built recombination libraries of ChRs and selected ~100 diverse sequence variants and measured their expression, localization, and photocurrent properties. We also collected much of the published ChR sequence/function data. We used all these data to train machine-learning models to approximate what we call the 'protein fitness landscape' of ChRs for different properties of interest. The protein fitness landscape simply means how protein fitness (as defined by the experimenter, e.g. photocurrent strength) changes with sequence. i.e., how protein sequence maps to protein function. The mapping of sequence to function is a complicated task. By measuring the properties of a small number of ChR sequences we are sparsely sampling the ChR fitness landscape. We can use machine learning along with our measured data to model this fitness landscape, then essentially interpolate and extrapolate from our measured data points to predict the 'fitness' of new, untested ChRs.

Using this strategy, we have built models of ChR expression and membrane localization trained with our empirical measurements (44, 45). With these models, we are able to very accurately predict whether or not a ChR sequence will express and localize in mammalian cells (45). We have also built models, trained with empirical measurements, for four different photocurrent properties: peak photocurrent, steady state photocurrent, off kinetics, and spectral properties. We then predicted which ChR sequences would express, localize, and have a desired combination of photocurrent properties (e.g. fast kinetics, red-shifted with strong currents). Despite the fact that we used empirical data of only ~100 variants to build the models we were able to accurately predict the photocurrent properties of highly diverse, untested ChRs.

Using this machine learning method we have made many highly functional ChR variants. Several variants stand out as exceptional for neuroscience applications. Three of our variants have very strong photocurrents with low light levels. Further, the ChR variants developed have diverse spectral properties, one has a narrow activation spectra (more narrow than ChR2) that peaks at 480 nm light, while the other highly light sensitive ChR has a very broad activation spectra. This broad spectra peaks at 480 nm light, but the ChR is still very strong with violet light, green light, and also yellow light.

1.7 Engineering of rhodopsin voltage indicators

Adam Cohen and colleagues recently discovered that rhodopsins can be used as genetically encoded voltage indicators (GEVIs); however, the natural proteins suffer from extremely low quantum efficiencies (~10^{-4}) (4). Eliminating pumping activity while retaining fast kinetics also presents an engineering challenge since the relationship between pumping, fluorescence, and kinetics is not completely understood. The photocycle of Arch, a leading candidate for GEVI development, is thought to proceed as follows: absorption of a photon initiates the photocycle (g → M), leading to an equilibrium between the M state (protonated counter-ion) and N state (protonated Schiff base) (46). Following conversion of N → Q (through absorption of photon at 540 nm) and excitation of the Q-state (absorption of photon at 570 nm), a photon at 710 is emitted as fluorescence as Arch returns to the N intermediate (46). Retinal thermally isomerizes back to all-*trans* (N → O) and a proton is released at the extracellular side (O → g). On the basis of this model, mutants with a longer-lived Q-state should exhibit increased fluorescence.

Directed evolution is an effective strategy for enhancing the brightness of Arch (14, 47). For example, introduction of mutations near the lysine that forms the covalent Schiff base linkage to retinal and screening for fluorescence enabled identification of two variants of Arch, one a double mutant, D95E/T99C (Archer) and another containing five mutations (referred to as QuasAr1). Both Archer and QuasAr1 show enhanced voltage sensitive fluorescence with emission in the far-red (maximal emission >680 nm) (14, 15, 47). Both of these engineered variants have improved brightness and dynamic range compared to two

previously published variants, Arch EEQ and Arch EEN (16). Directed evolution of Archer revealed two fluorescence-enhancing mutations, V59A and I129T (47), that were independently identified at the homologous positions in bR (V49A and I119T) and shown to stabilize the Q-state intermediate (48). Many mutations at P60 (<5 Å from retinal) also increase Arch fluorescence (47); similarly, many mutations at the homologous bR position (P50) stabilized the Q state (48). These observations are consistent with the Q state being the fluorescent state in the Arch photocycle (46).

Since their absorbance is sensitive to changes in electric potential (49), rhodopsins can also potentially be used in FRET sensors, assuming the absorbance overlaps with the emission of a bright fluorescent protein. Recently, a FRET-opsin sensor (a fusion between L. maculans [Mac] rhodopsin and mOrange2) was developed, achieving a response time of ~5 ms following a step change in membrane voltage and successful detection of sub-threshold events (5). However, current Mac-mOrange2 derivatives have a lower dynamic range (defined as voltage-dependent changes with respect to the probe's baseline fluorescence) than recently engineered Arch variants (14, 15, 47). Using a vector that can drive expression in both prokaryotic and eukaryotic cells, Zou *et al.* developed a screening strategy in which brighter Arch-mOrange2 variants can be identified in *E. coli* and subsequently transfected into HEK293 cells to measure their voltage sensitivity (50). This engineering strategy accelerates the speed at which brighter, multi-colored, and voltage-sensitive rhodopsins can be identified and has resulted in FRET sensors with rise times in the range 1–7 ms (50).

Engineered rhodopsin-based sensors are still quite dim, with quantum yields of <1%. Alternative voltage sensors have been engineered by fusing the *Ciona intestinalis* voltage-sensor containing domain (Ci-VSD), a non-rhodopsin protein that undergoes a voltage-dependent conformational change, to a fluorescent protein (51). The issue of slow kinetics of these non-rhodopsin sensors (52) has been largely overcome (53), but they exhibit non-linear voltage sensitivity, which may limit their capacity for detecting sub-threshold events (53). Despite being fused to bright fluorescent proteins and increased basal fluorescence over rhodopsins, the spectral overlap between Ci-VSD-based sensors and rhodopsins limits

their compatibility for all-optical electrophysiology (**Figure 1.3**); furthermore, rhodopsins appear to be less susceptible to photo-bleaching (15).

1.8 Conclusion

Rhodopsins are powerful tools for brain research. Identifying actuators with shifted and narrowed spectra would improve the ability to multiplex perturbations with different colors of light, whereas enhancing ion specificity will enable more physiological studies within and beyond neuroscience. Rhodopsins with increased light sensitivity and increased conductance could enable less invasive optogenetic experiments and improve the efficiency of optogenetic experiments enabling activation of large tissue volumes, activation of the entire brain nuclei or large volumes of diffuse circuits throughout the body (e.g. the enteric nervous system). Brighter rhodopsin sensors have been engineered, but further improved brightness would facilitate imaging populations of neurons (and perhaps other electrically active cell-types such as cardiomyocytes) with wide-field microscopy. The development of opsin-FRET sensors could also enable monitoring different cell types with different colors of light, a potentially powerful application of all-optical electrophysiology. Rational design and machine-learning guided engineering have been useful to overcome the key engineering limitation, low throughput screening, and enabled important advances in rhodopsin properties. Future work would greatly benefit from an understanding of how characterized mutations impact the photocycle and the elements of protein structure that lead to desired properties found in engineered rhodopsins. Chimeragenesis, structure-guided mutagenesis, and directed evolution have and will continue to play central roles in the development of improved rhodopsins for optogenetics.

1.9 Figures and tables

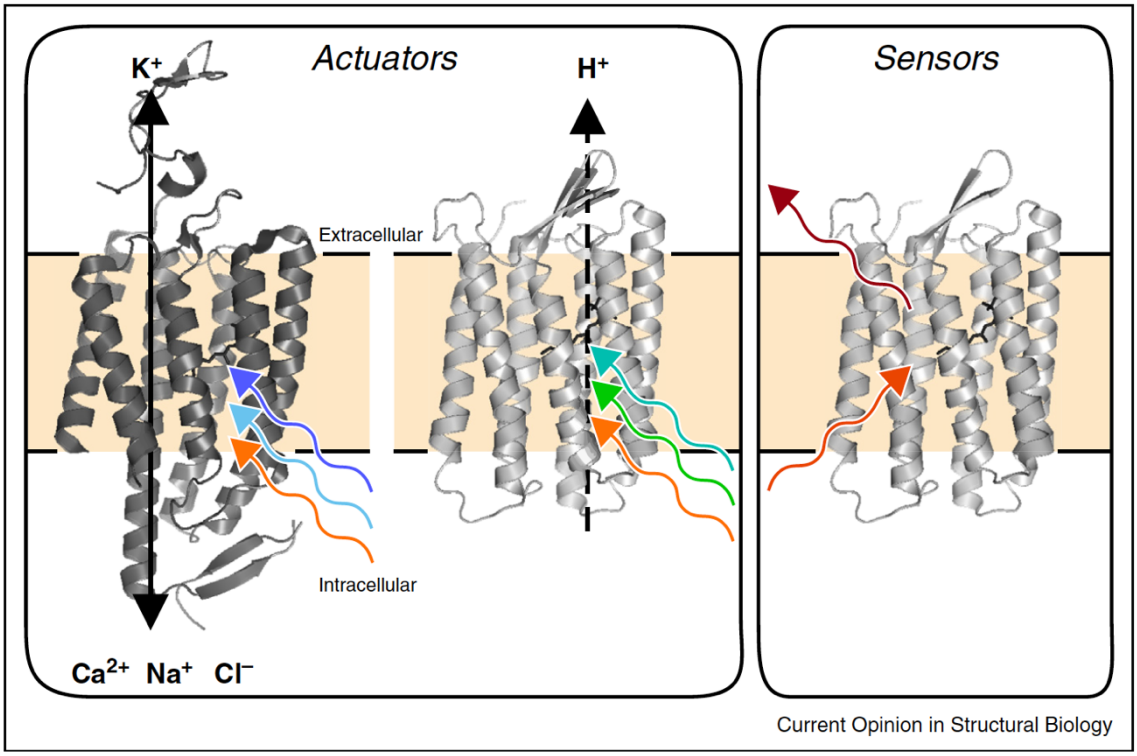


Figure 1.1. Rhodopsins can be used as actuators and sensors in optogenetics. Actuators transport ions across the membrane to activate or repress neuronal activity. ChRs transport positively charged ions into the cell, while proton-pumping rhodopsins (PPRs) move protons out of the cell. In the ideal case, engineered rhodopsin sensors emit light as fluorescence in the far-red in a voltage-dependent fashion.

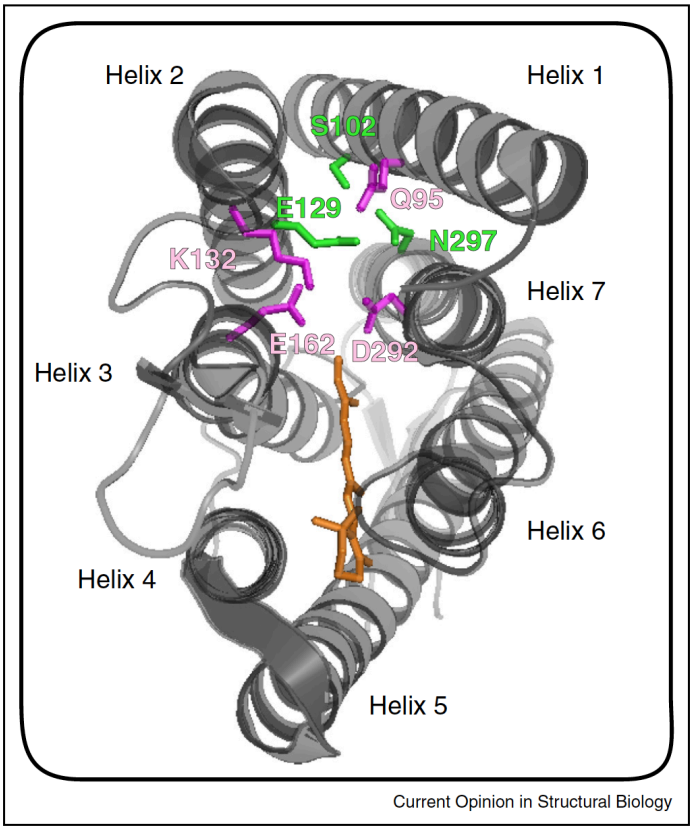


Figure 1.2. Residues that affect ion selectivity in the channelrhodopsin C1C2. The illustration shows crystal structure of C1C2, with putative ion gating residues S102, E129 and N297 highlighted in green. Mutation of the gating residue N297 to D results in a significant increase in selectivity for Ca^{2+}, while mutation of E129 to Q or A results in a significant decrease in the channel's Ca^{2+} selectivity (23). Mutating the highly conserved gating residue E129 has significant effects on the channel's selectivity for Cl^- in both the C1C2 backbone and the ChR2 backbone (position E90 in the ChR2 backbone) (35, 36). Mutation of E90 in ChR2 to R or K increases the reversal potential as a result of increased Cl^- selectivity to generate a light activated inhibitory channel (36). Residues outside of the putative ion gate also influence channel selectivity (residues highlighted in purple). Mutations Q95A, E162A, and D292A have all been shown to enhance H^+ selectivity. Mutants K132A and Q95A display increased K^+ permeability in the C1C2 backbone (23).

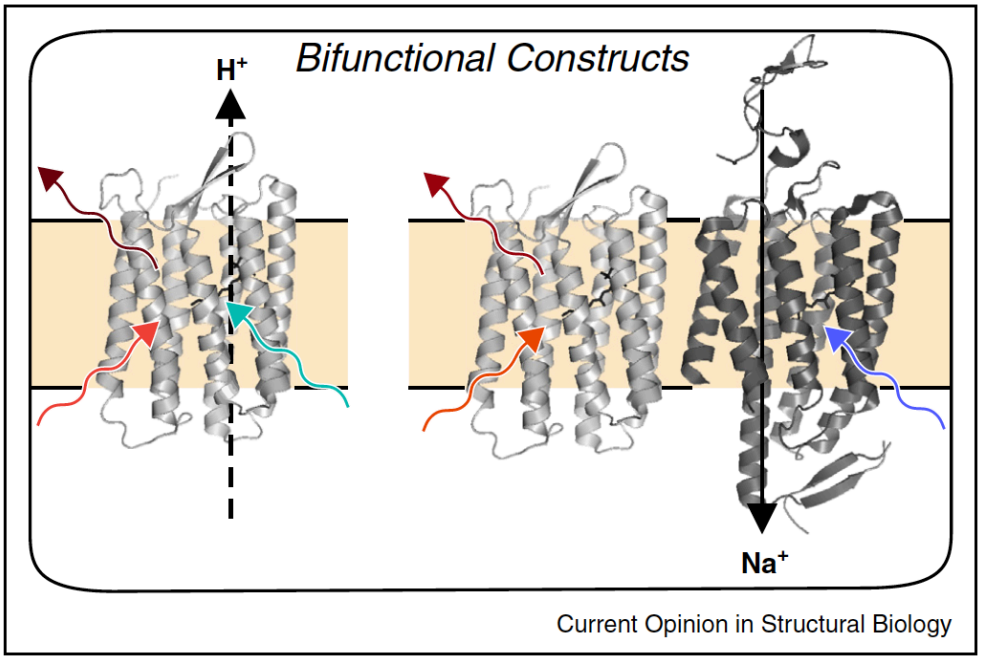


Figure 1.3. Bifunctional constructs for all-optical electrophysiology. Archer, an engineered Archaerhodopsin-3 variant, enables optical monitoring of voltage with red light, and perturbation of membrane potential with blue light (left) (15). Alternatively, one rhodopsin can be used for sensing with red light, while an engineered ChR can be used for perturbing the membrane with blue light (right) (14).

Table 1

Comparison of engineered rhodopsin actuators for a number of relevant characteristics and engineering methods

Rhodopsin	Function	λ_{max} (nm)	τ_{off} (ms)	Reversal potential (mV)	Distinguishing properties	Engineering approach
ChR2	Excitatory	480	~10	~0	Most commonly used ChR.	Native.
ReaChR	Excitatory	545–590	137 ± 7	7 ± 4	Red-shifted ChR.	Chimera of ChR1, VChR1 and VChR2 with single amino acid substitution.
Chronos	Excitatory	~500	3.6 ± 0.2	NR	Fast kinetics and sensitive to low intensity light.	Native: de novo transcriptome sequencing.
ChrimsonR	Excitatory	590	15.8 ± 0.4	NR	Red-shifted ChR.	Native: de novo transcriptome sequencing with single amino acid substitution to improve kinetics.
CheRiff	Excitatory	460	16.0 ± 0.8	NR	Very sensitive to low intensity light.	Native: de novo transcriptome sequencing with single amino acid substitution to improve kinetics and trafficking sequence to improve membrane localization.
iC1C2	Inhibitory	480	24	−64	C1C2 with decreased reversal potential and increased Cl^- ion specificity.	Structure based site-directed mutagenesis of C1C2.
$SwiChR_{CT}$	Inhibitory	480	7300	−61	Slow channel closure version of iC1C2.	iC1C2 with single amino acid substitution.
ChloC	Inhibitory	~480	NR	−61.8 ± 1.0	ChR with decreased reversal potential and increased Cl^- ion specificity.	Mutagenesis of ChR2 based off of molecular dynamics simulations based off of C1C2 structure.
Slow ChloC	Inhibitory	~480	10500 ± 300	−68.2 ± 0.8	Slow channel closure version of ChloC.	ChloC with single amino acid substitution.
Jaws	Inhibitory	600	~7	NR	Red-shifted inhibitory chloride pump.	Native with amino acid substitutions for improved photocurrents and added trafficking sequences for improved localization.

Table 1.1. Comparison of engineered rhodopsin actuators for a number of relevant characteristics and engineering methods. Rhodopsin molecules are functionally classified as either ‘excitatory’ or ‘inhibitory’. The rhodopsin actuators are compared for optimal wavelength for photocurrent excitation (λ_{max}), kinetic off rate (τ_{off}) indicating how quickly the molecule closes once light stimulation is turned off, and reversal potential. The engineering approach is briefly described.

Chapter 2

ARCHAERHODOPSIN VARIANTS WITH ENHANCED VOLTAGE SENSITIVE FLUORESCENCE IN MAMMALIAN AND CAENORHABDITIS ELEGANS NEURONS

A version of this chapter has been published as (15).

2.1 Abstract

Probing the neural circuit dynamics underlying behavior would benefit greatly from a genetically encoded voltage indicator capable of optically monitoring the activity of large populations of neurons simultaneously. The proton pump Archaerhodopsin-3 (Arch), an optogenetic tool commonly used for neuronal inhibition, has been shown to emit voltage sensitive fluorescence. Here we report two Arch variants that in response to 655 nm light have 3-5 times increased fluorescence and 55-99 times reduced photocurrents compared to Arch WT. The most fluorescent variant, Archer1, has 25-40% fluorescence change in response to action potentials while using 9 times lower light intensity compared to other Arch-based voltage sensors. Archer1 is capable of wavelength specific functionality as a voltage sensor under red-light and as an inhibitory actuator under green-light. As a proof-of-concept for the application of Arch-based sensors *in vivo*, we show an example of fluorescence voltage sensing in behaving *C. elegans*. Archer1's characteristics contribute to the goal of all-optical detection and modulation of activity in neuronal networks in vivo.

2.2 Introduction

The study of brain circuitry encompasses three frames of reference: neuron-level spiking activity, circuit-level connectivity, and systems-level behavioral output. A pervasive goal in neuroscience is the ability to examine all three frames concurrently. Fluorescent sensors, which enable measurements of simultaneous changes in activity of specific populations of neurons, are envisioned to provide a solution (54-58). Successful detection of both high frequency trains of action potentials and sub-threshold events in neuronal populations *in*

vivo requires a genetically encoded voltage indicator (GEVI) (51) with fast kinetics, high sensitivity and high baseline fluorescence. Recent developments of genetically encoded calcium (56) and voltage sensors (5, 53, 59) have yielded progress towards achieving this goal. The calcium sensor family GCaMP has been used to monitor populations of neurons in intact behaving organisms (57). However, the detection of fast spiking activity, sub-threshold voltage changes, and hyperpolarization is difficult with GCaMP due to its relatively slow kinetics and reliance on calcium, a secondary messenger, flux into the cell (56, 60, 61). Newer iterations of voltage-sensitive fluorescent proteins (VSFPs) based on fusions with circularly permuted GFP (cpGFP), e.g. ASAP1 (53), improve upon both the speed and sensitivity of previous sensors, e.g. Arclight (52), but are still limited by the ability to be combined with optogenetic actuators (27, 41, 62). This spectral overlap prohibits the combined use of these sensors with opsins for all-optical electrophysiology. No currently available sensor is able to meet all of the needs for optical imaging of activity *in vivo*, calling for continued efforts to evolve GEVIs.

Archaerhodopsin-3 (Arch) (62, 63), a microbial rhodopsin proton pump that has recently been introduced as a fluorescent voltage sensor (64), is fast and sensitive but suffers from low baseline fluorescence and strong inhibitory photocurrents. Previous optimizations of Arch successfully reduced photocurrents, e.g. Arch D95N (64) and Arch EEQ (16), and increased sensitivity and speed, e.g. QuasArs (59), but have still to enable its use *in vivo*. All previous *in vivo* voltage sensing has been accomplished using lower power of fluorescence excitation light than is possible with reported Arch variants to date (5, 55, 56). For example, Arch WT (64) uses 3,600x higher intensity illumination than ASAP1 (53). The high laser power used to excite Arch fluorescence causes significant autofluorescence in intact tissue (51) and limits its accessibility for widespread use.

Here we report two Arch mutants (Archers: Arch with enhanced radiance), Archer1 (D95E and T99C) and Archer2 (D95E, T99C, and A225M) with improved properties for voltage sensing. These mutants exhibit high baseline fluorescence (3-5x over Arch WT), large dynamic range of sensitivity (85% $\Delta F/F$ and 60% $\Delta F/F$ per 100 mV for Archer1 and Archer2 respectively) that is stable over long illumination times, and fast kinetics, when

imaged at 9x lower light intensity (880 mW mm^{-2} at 655 nm) than the most recently reported Arch variants (59) [and 20.5x lower than Arch WT (64)]. We demonstrate that Archer1's improved characteristics enable its use to monitor rapid changes in membrane voltage throughout a single neuron and throughout a population of neurons *in vitro*. Though Archer1 has minimal pumping at wavelengths used for fluorescence excitation (655 nm) it maintains strong proton pumping currents at lower wavelengths (560 nm). We show that this single protein, Archer1, is a bi-functional tool that provides both voltage sensing with red light and inhibitory capabilities with green light. Finally, we demonstrate that Archer1 is capable of detecting small voltage changes in response to sensory stimulus in the context of intact multicellular organisms such as *C. elegans*.

2.3 Results

The combination of D95E, T99C, and A225M mutations was first identified in a site-saturation mutagenesis library of the proton pump *Gloeobacter violaceus* rhodopsin (GR) designed to evolve for spectral shifts (65). Far-red shifted mutants of the GR library were then screened for fluorescence intensity in *E. coli*, which revealed numerous hits with higher fluorescence than GR WT (65). The corresponding mutations found in the most intensely fluorescent variants can be transferred to the homologous residues of Arch WT (**Supplementary Figure 2.1**) and greatly improve its quantum efficiency and absolute brightness (66). The selected mutants were expressed in neurons to test if their improved characteristics were maintained in a mammalian system.

2.3.1 Characterization of two new mutant Arch voltage sensors

Arch variants designed with TS and ER export domains for enhanced membrane localization (43) (**Figure 2.1a** and **Supplementary Figure 2.1b**) were screened in neurons for enhanced baseline fluorescence, decreased photocurrents at imaging wavelengths, increased voltage sensitivity, and fast fluorescence kinetics, and compared with previously reported variant Arch EEQ (16). Of the Arch variants screened, Archer1 and Archer2 exhibited ~5x and ~3x increased fluorescence, respectively, over Arch WT (**Figure 2.1a**). Archer1 and Archer2 also have 55x and 99x reduced photocurrents in response to 655 nm

laser illumination, respectively, when compared to Arch WT (**Figure 2.1b** and **Supplementary Figure 2.2a**). Archer1 exhibits a peak current upon initial laser exposure, which then reaches a residual average steady state of 5.6 pA, while Archer2 produces no peak current, and an average steady state of 3.1 pA (**Figure 2.1b** and **Supplementary Figure 2.2b,c**). Voltage sensitivity was measured as a fluorescence response to steps in membrane potential ranging from -100 mV to +50 mV. Due to Arch EEQ's low baseline fluorescence, its single cell fluorescence traces show considerably more noise than those for Archer1 and Archer2 (**Figure 2.1c**). Archer1 shows the highest voltage sensitive fluorescence, as depicted by single cell sensitivity measurements (**Figure 2.1c**), and by the averaged traces (**Figure 2.1d**, **Supplementary Figure 2.3**). Facilitated by Archer1's increased baseline fluorescence, imaging can be done with short 1 ms exposure times and at lower laser intensities (880 mW mm^{-2}) than previously published Arch-based sensors (16, 59, 64). To characterize the stability of Archer1's fluorescence, sensitivity was measured before and after prolonged laser illumination. Archer1 showed no reduction in voltage sensitivity over the 10-15 minute timeframe measured (**Supplementary Figure 2.4**).

2.3.2 Sensitivity Kinetics enables comparison across sensors

The choice of a specific voltage sensor for a given experimental application depends on whether the sensor will yield a significant fluorescence change in response to a given voltage change within the time frame of interest. Traditionally, sensitivity is quantified by measuring the steady-state fluorescence change for a step in voltage (5, 16, 52, 53, 59, 64), but the steady-state value does not provide information about the initial dynamics of the fluorescence response (sensor kinetics). The methods for kinetic analysis vary with different types of sensors. Following a previously used method for Arch-based sensor kinetics (5, 16) we compared Archer1 to Arch WT by normalizing the fluorescent responses of each sensor during a 1 s voltage step (-70 mV to +30 mV) to the steps maximum fluorescence. These results indicate very similar kinetics between the two (**Figure 2.1e**), without addressing Archer1's 35x larger change in fluorescence. The large timescale of these voltage steps is not relevant for neuronal applications. However,

normalizing over a shorter timescale produces variable results depending on the timepoint used for normalization (**Figure 2.2b**). A method that takes into account the sensitivity of a sensor on the timescale relevant to an action potential is necessary.

Our proposed method for analysis, Sensitivity Kinetics (*SKi*), expands upon the traditional method by providing %ΔF/F for any given voltage change over time (**Figure 2.2a**). With this method, both the sensitivity and kinetics can be compared simultaneously amongst sensors. *SKi* is calculated by evaluating the slope of the fluorescence response to steps in voltage for each time point after the step's initiation. The sensitivity-slopes are then plotted over time (**Figure 2.2a,c**). Characterization of the sensitivity kinetics for Arch variants reveals that Archer1 produces the largest changes in fluorescence of the sensors we tested (**Figure 2.2d**), within any timeframe.

2.3.3 Tracking action potentials in primary neuronal cultures

Action potentials were evoked in cultured rat hippocampal neurons expressing Archer1 through current injection. Archer1 fluorescence is capable of tracking action potentials in both individual processes and the cell body (**Figure 2.3a,b**). In addition, the magnitude and shape of dendritic fluorescence changes closely mimics that of the cell body in response to the same event. As predicted by the sensitivity kinetics, Archer1 fluorescence, with a > 6x increase in signal-to-noise ratio (SNR), more closely follows the electrical recording of action potentials than Arch EEQ at similar frequencies (**Figure 2.3c,d**). Archer1 exhibits a large percentage change in fluorescence in response to action potentials (25-40% ΔF/F), and can track 40 Hz firing rate as well as simulated changes in membrane voltage occurring at 100 and 150 Hz (> 50% ΔF/F) (**Figure 2.3e,f**). The ability to follow action potential throughout neurons by imaging with significantly lower laser intensity (880 mW mm^{-2}) is enabling for monitoring voltage sensitive fluorescence *in vivo*.

2.3.4 Archer1 functions as a voltage sensor and inhibitory actuator

All-optical electrophysiology requires an optical method for both sensing and perturbing cells. Recent work (59) presented a construct with dual capabilities: voltage sensing and

neuronal activation at distinct wavelengths through co-expression of a sensor and a light-gated channel. Archer1 also provides two useful functionalities, but in a single protein. While minimally active with high intensity 655 nm laser illumination (880 mW mm^{-2}), Archer1 is significantly more active at low intensity 560 nm LED illumination (3 mW mm^{-2}) (51x at peak and 35x at steady state) (**Figure 2.4a,b**). The hyperpolarizing photocurrents generated by Archer1 in response to green light successfully inhibit action potentials, while red light does not (**Figure 2.4c,d**). Archer1 is capable of inducing inhibitory currents with green light and simultaneously sensing activity with red, without crossover.

2.3.5 Optical monitoring of cultured neuronal networks

Fluorescent voltage sensors should enable the detection of spiking activity across all neurons in a population. Original Arch variants require the use of high optical magnification combined with binning and heavy pixel weighing (16) to detect modest changes in fluorescence, due to low baseline. Until recently (59) these stringent imaging requirements had prevented microbial rhodopsin-based voltage sensors from being used to monitor multiple cells simultaneously. Archer1, similar to QuasAr (59), by virtue of its increased fluorescence and higher sensitivity kinetics, allows simultaneous imaging of activity for a population of cells while perturbing only one of them through current injection (**Figure 2.5a**, schematic). Within the same optical field, we tracked the fluorescence of three cells with different behaviors: one showed a step change (due to an induced voltage step), one had spontaneous spikes that increased concurrently with the step, and one remained unchanged (**Figure 2.5a**, traces).

2.3.6 Optical monitoring of sensory neurons in behaving Caenorhabditis elegans

A major application for voltage sensors is all-optical neuronal activity monitoring in model organisms in which electrophysiological recordings are inherently difficult, e.g. *C. elegans.* The aforementioned improved fluorescence and sensitivity kinetics of Archer1 have enabled us to extend its use from cultured cells to live, behaving nematodes. To test whether Archer1 will work in *C. elegans*, we examined the olfactory neuron AWC-ON (WormBase cell WBbt:0005832), one of the pair of C type Amphid Wing cells. Previously,

sensory-evoked Ca^{2+} transients that were monitored using GCaMP show a fluorescence increase upon odor removal, which peaks within 10 s and gradually decreases over minutes post-stimulation (67). To monitor the small voltage changes underlying this effect, we expressed Archer1 in AWC-ON, and observed fluorescence changes in response to turning off the odorant stimulus (isoamyl alcohol; IAA) in anesthetized and non-anesthetized animals. According to Kato S. *et al*., the chemosensory responses in AWC neurons are not affected by the application of cholinergic agonist (68). As shown in **Figure 2.5b-d,** Archer1's fluorescence indicates that voltage transients peak within 2 s, and end 10 s after turning off stimulus (**Figure 2.5c** and **Supplementary Figure 2.5**). These observed fluorescence changes, which correspond to small reported changes in AWC membrane voltage (69), validate the sensor's *in vivo* utility. A combination of results from Archer1 and GCaMP experiments can be used to better understand the dynamics of *C. elegans* voltage-gated calcium channels.

2.4 Discussion

Replacing electrophysiology with all-optical methods for *in vivo* recording will require a genetically encoded voltage indicator with fast kinetics, high sensitivity, high baseline fluorescence, and compatibility with optical methods for controlling neuronal activity. Here we report an Arch mutant, Archer1, in which these combined improvements enable the accurate tracking of action potentials at high speed, the detection of simultaneous activity within populations of neurons, wavelength specific inhibition of neuronal activity, and the real-time observation of voltage changes in response to a stimulus in live nematodes. Fluorescence measurements of Archer1 and Archer2 were achieved at a lower intensity of laser illumination than has been possible in experiments using previously reported Arch variants (16, 59, 64). Reduction in excitation light intensity required for fluorescent measurements increases the accessibility of Arch-based voltage sensors and their potential use *in vivo*.

Archer1 is an enhanced voltage sensor under red light and it also enables inhibition of action potentials under green light. Recent work has been done to generate an all-optical

system for neuronal excitation and voltage sensing (Optopatch (59)). Archer1, on the other hand, provides the first example of a combination of wavelength specific sensing and hyperpolarization with a single protein. This wavelength specific bi-functionality can enable all-optical dissection of a neural network through targeted inhibition and global fluorescence monitoring. Tools like Archer1 and Optopatch could be used for all-optical loss and gain of function circuit analysis, respectively.

Voltage sensors can also provide insights into neuronal response to stimuli in organisms in which electrophysiology is challenging, such as *Caenorhabditis elegans* and *Drosophila melanogaster*. Archer1 represents the first genetically encoded voltage sensor that has been used in live, behaving nematodes. This work provides a foundation for more detailed characterization of cell types with unknown voltage dynamics as well as fast-spiking muscle cells in *C. elegans (70)*. Additional applications of this tool likely include other transparent organisms, i.e. fly larvae and zebrafish, where a fluorescent voltage sensor could be used to dissect neural circuitry.

Until recently, due to their low baseline fluorescence (51), Arch-based sensors were not compatible with *in vivo* applications. This work on Archer1, as well as recent work on QuasArs (59), demonstrates that Arch-based sensors are not fundamentally limited, but can be used for a variety of neuronal applications, including *in vivo*. Our data shows that variants of Arch are capable of increased fluorescence, enabling practical detection, while retaining their superior speed and dynamic range (71). Even though this work uses the lowest excitation intensity for an Arch-based sensor (<5% original illumination intensity of Arch WT (64), ~60% of Arch EEQ (16) and 11% of QuasArs (59)), it is still ~200 times higher than that for XFP-based sensors. Further enhancements of baseline fluorescence while maintaining fast kinetics and high sensitivity of Arch-based sensors could result in a GEVI capable of detecting both high frequency trains of action potentials and sub-threshold events in mammalian neuronal populations *in vivo*.

2.5 Methods

Ethics statement

All experiments using animals in this study were approved by Institutional Animal Care and Use Committee (IACUC) at the California Institute of Technology.

Sensor constructs

Arch variant constructs were generated by first amplifying EGFP from FCK-Arch-GFP (**Supplementary Table 2.1**) and adding the ER export domain using GFPfwd_overlapTSend and FCK-GFPrev_ERexport primers to make EGFP-ER. Arch-TS was then amplified from pLenti-CaMKIIa-eArch3.0-EYFP (**Supplementary Table 2.1**) using Archfwd and TSrev_into_GFPstart primers, assembled with EGFP-ER using Archfwd and ERrev primers, and subsequent cloned back into pLenti-CaMKIIa-eArch3.0-EYFP cut with *BamHI* and *EcoRI* restriction enzymes, to make pLenti-CaMKIIa-eArch3.0-EGFP. To make pLenti-CaMKIIa-Archer1-EGFP and pLenti-CaMKIIa-Archer2-EGFP, the D95E, T99C, and A225M mutations were introduced in the pLenti-CaMKIIa-eArch3.0-EGFP vector through overlap assembly PCR using Archfwd, ERrev, Arch3.0_D95E_T99C_fwd, Arch3.0_D95E_T99C_rev, Arch3.0_A225M_fwd, and Arch3.0_A225M_rev primers and subsequent cloning back into the backbone via *BamHI* and *EcoRI* sites. pLenti-Arch-EEQ (**Supplementary Table 2.1**), an EYFP fusion, was used as a comparison.

To make *Pstr-2::Archer1eGFP::unc-54 3'UTR*, Archer1 was amplified from pLenti-CaMKIIa-Archer1-EGFP using Arch-NheI-AAA-F and Arch-EcoRI-R primers and inserted into the pSM vector using *NheI* and *EcoRI* sites. The *C. elegans* Kozak sequence AAA, and the restriction enzyme sites mentioned above were engineered into the primers (72). The AWC specific promoter, which is a 2kb sequence 5' to the start codon of *str-2*, was amplified from genomic DNA using str-2p-SphI-F2(2K) and str-2p-AscI-R2 primers and cloned into the vector via *SphI* and *AscI* sites.

Primary neuronal cultures

Rat hippocampal cells were dissected from Wistar pups (postnatal days 0-1, Charles-River Labs), and cultured at 37^{o}C, 5% CO_2 in Neurobasal media supplemented with B27,

glutamine, and 2.5% FBS. 3 days after plating, glial growth was inhibited by addition of FUDR. Cells were transfected 4-5 days after plating with Arch WT and variants using calcium chloride. Neurons were imaged 3-5 days after transfection.

Fluorescence Imaging

Imaging was performed concurrently with electrophysiology recordings of voltage and current clamped cultured rat hippocampal neurons. For both cultured neurons and *in vivo C. elegans* experiments, a Zeiss Axio Examiner.D1 microscope with a 20x 1.0 NA water immersion objective (Zeiss W Plan Apochromat 20x/1.0 DIC D=0.17 M27 75mm) was used. A diode laser (MRL-III-FS-655-1.3W; CNI) with a 650/13 nm excitation filter, 685 nm dichroic mirror and 664 nm long-pass emission filter (all SEMROCK) was used for rhodopsin fluorescence excitation throughout. For cultured neuron experiments Arch WT, Archer1, and Archer2 fluorescence was excited with 880 mW mm^{-2} illumination intensity at the specimen plane, while for Arch EEQ, 1,500 mW mm^{-2} illumination intensity was used. Higher illumination intensity was used for Arch EEQ compared to other Arch variants due to its lower baseline fluorescence with our imaging setup. For *C. elegans* experiments, 880 mW mm^{-2} illumination intensity was used to visualize Archer1 fluorescence. For all experiments, fused EGFP fluorescence was imaged with 485±25 nm LED light using a Lumencor SPECTRAX light engine with quad band 387/485/559/649 nm excitation filter, quad band 410/504/582/669 nm dichroic mirror and quad band 440/521/607/700 nm emission filter (all SEMROCK) at 0.05 mW mm^{-2}.

All fluorescence traces were recorded using an Andor Neo 5.5 sCMOS camera cooled to -30 ^{o}C at 500 or 1,000 Hz. Pixels were binned up to 0.54 mm x 0.54 mm to achieve the image acquisition speeds. All recordings were taken using Andor's Solis software.

Electrophysiology

Conventional whole-cell patch-clamp recordings were done in cultured rat hippocampal neurons at > 2 days post transfection. Cells were continuously perfused with extracellular solution at room temperature (in mM: 140 NaCl, 5 KCl, 10 HEPES, 2 $MgCl_2$, 2 $CaCl_2$, 10

glucose; pH 7.35) while mounted on the microscope stage. Patch pipettes were fabricated from borosilicate capillary glass tubing (1B150-4; World Precision Instruments, Inc., Sarasota. FL) using a model P-2000 laser puller (Sutter Instruments) to resistances of 2-5 MΩ. Pipettes were filled with intracellular solution (in mM): 134 K gluconate, 5 EGTA, 10 HEPES, 2 $MgCl_2$, 0.5 $CaCl_2$, 3 ATP, 0.2 GTP. Whole-cell patch-clamp recordings were made using a Multiclamp 700B amplifier (Molecular Devices, Sunnyvale, CA), a Digidata 1440 digitizer (Molecular Devices), and a PC running pClamp (version 10.4) software (Molecular Devices) to generate current injection waveforms and to record voltage and current traces.

Patch recordings were done simultaneously with imaging for measurements of voltage sensitive fluorescence. For sensitivity measurements cells were recorded in voltage-clamp with a holding potential of -70 mV for 0.5 s and then 1 s voltage steps were applied ranging from -100 mV to +50 mV in 10 mV increments. Action potentials were generated in current clamp by current injection in either a long step (10-200 pA; 0.8s) or in short pulses (100-500 pA; 2-10 ms).

Patch-clamp recordings were done with short light pulses to measure photocurrents. Photocurrents induced by the excitation wavelength used for voltage sensing were measured using a 655 nm laser at 880 mW mm^{-2}. Photocurrents induced by green light were measured using 560±25 nm LED at 3 mW mm^{-2}. Photocurrents were recorded from cells in voltage clamp held at -50 mV with 3-10 light pulse trains (0.5 s each pulse; 2 s apart). Voltage changes induced by 655 nm laser at 880 mW mm^{-2} were measured in a current clamp mode with three 0.5 s light pulses separated by 2 s and zero current injection.

To test for inhibitory capabilities of Arch mutants, pulses (300 ms) of illumination with either red laser (655 nm at 880 mW mm^{-2}) or green LED (560±25 nm at 3 mW mm^{-2}) were applied to cells during a 900 ms train of induced action potentials (generated in current clamp by current injections from 30-100 pA).

Action spectra measurements were performed for the following wavelengths: 386±23 nm, 438±24 nm, 485±20 nm, 513±17 nm, 560±25 nm, and 650±13 nm with light intensity

matched across all experiments at 0.08 mW mm^{-2}. Each light pulse was delivered for 0.6 s with 10 s breaks between light pulses. All wavelengths were produced using LED illumination from a SPECTRAX light engine (Lumencor). Cell health was monitored through holding current and input resistance.

Microinjection and germ line transformation in Caenorhabditis elegans

The transgenic line used in this work is PS6666 *N2; syEx1328[Pstr-2(2k)::Archer1eGFP(*75 ng ml^{-1}*); Pofm-1::RFP(*25 ng ml^{-1}*)]*. *Pstr-2::Archer1eGFP::unc-54 3'UTR* was co-injected with a *Pofm-1::RFP* marker into Bristol N2 using the method described by Melo, *et al.* (73). The two plasmids were diluted to the desired concentration in water to make a 5 mL injection mix. The injection mix was spun down at 14,000 rpm for 15 min and transferred to a new tube prior to injection to prevent needle clogging. Late L4 hermaphrodites were transferred to a newly seeded plate and maintained at 22 °C one day before injection. The microinjection was performed the next morning when the worms had become young adults. Worms were glued on a 2% agarose pad and covered with Halocarbon Oil (Halocarbon Products Corporation, HC-700) before injection. 0.8 mL of the injection mix was loaded into the injection needle. For generating this particular transgenic line, 32 hermaphrodites (P_0S) were injected for both arms of the gonad. 27 F_1 were identified 3 days after injection based on *Pofm-1::RFP* expression in coelomocytes. Among them, 5 eventually became stable lines. The best line used in this study was determined by the highest transmission rate and the strongest expression level of Archer1eGFP.

Caenorhabditis elegans in vivo stimulation experiments

Late L4 transgenic worms were transferred to a plate seeded with the mixture of OP50 and all-trans-Retinal (ATR) (Sigma-Aldrich, USA), and maintained at 22°C in the dark 18 hours before imaging. The final concentration of ATR in the mixture was 100 mm (diluted from 100mM stock: 100 mg ATR powder dissolved in 3.52 ml 100% ethanol) using fresh OP50. Five times higher concentration of ATR was previously used for wild type Arch activity in worms (74). The microfluidic device is adapted for *in vivo* imaging (75, 76). The

PDMS chip contains four buffer inlets, one worm loading channel, and one suction channel connected to house vacuum. Two buffer inlets in the middle are the 'buffer' and the 'stimulus' channels, which are loaded with the default solution S Basal medium and 1:1,000 isoamyl alcohol (IAA) (Sigma-Aldrich), respectively. S Basal medium containing 0.15% phenol red (Sigma-Aldrich) is loaded in the side channels for detecting the laminar flow. An ATR-fed worm was first transferred to an empty NGM plate and washed in a drop of S Basal. It was then loaded in the microfluidic chip, where its nose was presented with either the buffer or the stimulus streams. The switch between buffer and stimulus stream was accomplished by changing the flow pressure from the side channels, which was regulated via an external valve controlled using a LabView script (National Instruments). The worm was exposed to the stimulus stream for 5 minutes (stimulus on), to the buffer stream for 30 seconds (stimulus off), and to the stimulus stream again. For performing the control experiments on the same worm, the flow switch remained the same but the stimulus channel was loaded with S Basal. Imaging of Archer1 fluorescence began 5 seconds before stimulus was switched off and lasted for 40 seconds. For anesthetized experiments only, 0.1% levamisole (Sigma-Aldrich) was added to the worm loading channel to minimize movement artifacts.

Data analysis

Unless otherwise noted all fluorescence analysis was done with raw measurements of cell fluorescence background subtracted. Cells and background regions were selected manually in ImageJ and fluorescence measurements were recorded for each region of interest (ROI) and background fluorescence was subtracted from cell fluorescence.

Sensitivity analysis was performed using background subtracted fluorescence recordings. Baseline fluorescence (mean fluorescence of the cell 20 ms prior to voltage step) and step fluorescence (fluorescence over whole 1 s voltage step) were used to generate %ΔF/F traces for each voltage step. The mean %ΔF/F over the entire 1 s step was calculated for each voltage step and then plotted (%ΔF/F vs. voltage step).

On & off kinetics analysis was performed on fluorescence traces in response to a 100 mV step (-70 mV to +30 mV). Percentage change in fluorescence %ΔF/F for each time point is normalized to the maximum step response (%ΔF/F averaged over the whole step).

Sensitivity kinetics analysis was performed using time-locked, average %ΔF/F traces (voltage steps ranging from -100 mV to 50 mV in 10 mV increments) for all cells. At each time point throughout a voltage step (t = 0 at time of voltage step trigger), %ΔF/F was plotted vs. the respective voltage step. A linear best fit was then performed for the %ΔF/F vs. voltage step for each time point. The slope of the best fit for each time point was then plotted over time (%ΔF/F / voltage step vs. time).

Signal-to-noise ratio analysis for action potentials tracked by Archer1 and Arch EEQ fluorescence was performed. SNR was computed as SNR = abs(s-n)/s, where s = peak fluorescence during action potential, n = average of pre-action potential noise and s = standard deviation of the pre-action potential noise (55).

Worm AWC cell and background regions were selected manually in ImageJ, fluorescence measurements were recorded for each ROI and background fluorescence was subtracted from cell fluorescence. The ROI for the fluorescent cell was drawn to contain the cell soma for all time points of the experiment. ΔF is reported instead of %ΔF/F due to low detected baseline fluorescence. Calculating %ΔF/F would result in amplified signal, as well as amplified noise.

Worm movement analysis was performed on the worm fluorescence traces, which were first thresholded so that the only pixels above a certain threshold are considered pixels of the cell. The cell location was then determined by averaging coordinates of pixels above the set threshold for the first frame in the 10,000 frame experiment to get the coordinates at the center of the cell. A 70x70 pixel region around the center of the cell was then set as the ROI. The center of the cell was corrected by again taking the averaging coordinates of pixels above the set threshold within the 70x70 pixel region to eliminate any influence of pixel noise within the full frame. The corrected cell center ($x_{c,1}$; $y_{c,1}$) was then calculated for every frame of the 10,000 frame experiment ($x_{c,1}$ - $x_{c,10000}$; $y_{c,1}$ - $y_{c,10000}$). The x and y

displacement (x_d; y_d) were calculate for each frame as the difference from $x_{c,1}$ and $y_{c,1}$. The x_d and y_d were then plotted over time.

Statistical methods

Paired and unpaired student's *t*-tests were performed using GraphPad Prism (version 6.04 for Windows, GraphPad Software, San Diego California USA, www.graphpad.com).

2.6 Figures

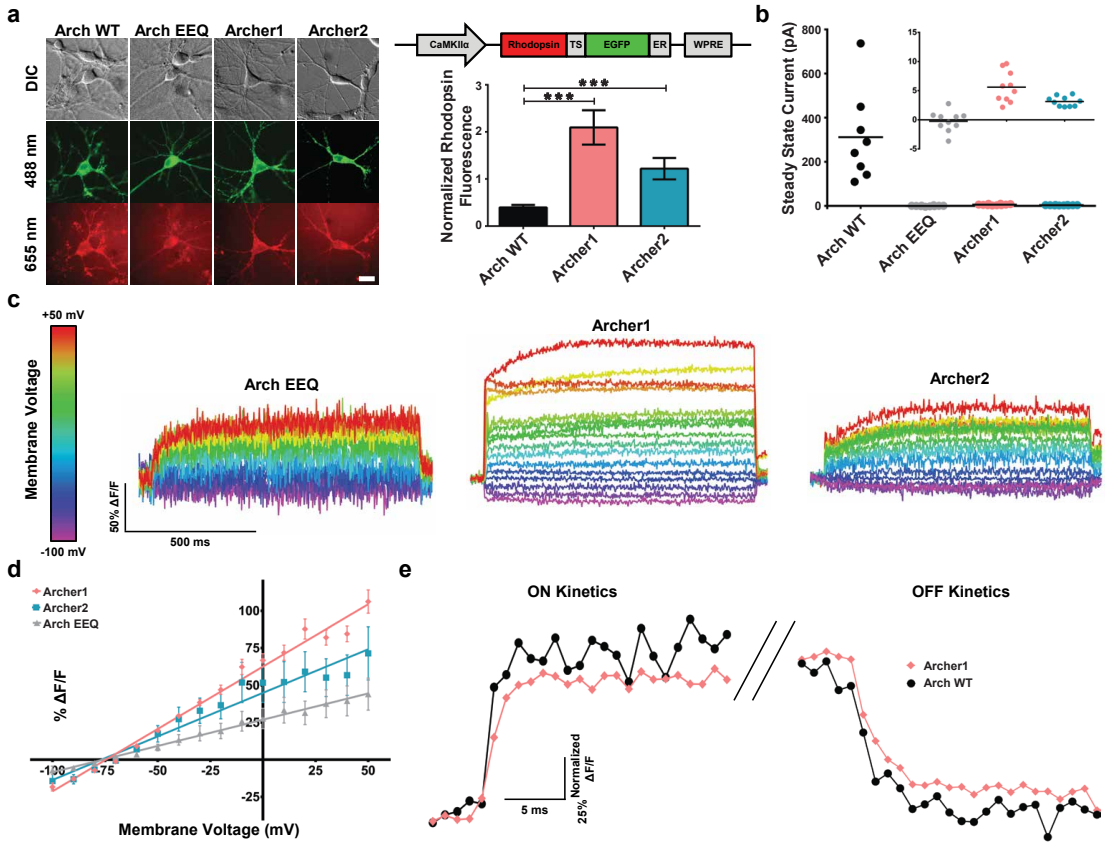


Figure 2.1. Characterization of Arch variants in mammalian neurons

(**a**) Quantification of Archer1 (n = 12) and Archer2 (n = 11) fluorescence compared to Arch WT (n = 13). Left – representative images of rhodopsin and fusion protein fluorescence; the published Arch EEQ-EYFP fusion is used, while all other sensors are fused to EGFP. Right graph – summary data. Baseline rhodopsin fluorescence normalized to EGFP fluorescence. Arch EEQ not included in comparison as it has a different fluorescent protein fusion. Right construct – Arch-EGFP fusion vector design. Scale bar, 10 mm. (**b**) Average steady-state photocurrents generated by Arch WT (n = 10) and different variants (n = 9, 10 and 9 respectively for Arch EEQ, Archer1, and Archer2) in neurons voltage clamped at V = -50 mV. Inset shows low levels of photocurrents expanded to indicate differences between variants. (**c**) Fluorescent responses (imaged at 500 Hz) of single neurons expressing Arch EEQ, Archer1, and Archer2 to voltage clamped steps in

membrane potential. Neurons are held at -70 mV and stepped to voltages ranging from -100 mV to +50 mV in 10 mV increments. (**d**) Sensitivity of Arch variants measured as the functional dependence of fluorescence to change in voltage. Fluorescence changes are averaged over 1,000 ms voltage steps and plotted against voltage. Results exhibit linear dependence with R^2 values of 0.98, 0.95, and 0.99 for Archer1 (n = 10), Archer2 (n = 3), and Arch EEQ (n = 5) respectively. (**e**) On/Off kinetics in response to a 100 mV step (-70 mV to +30 mV) for Archer1 (n = 10) compared to Arch WT (n = 6). %ΔF/F for each time point is normalized to the maximum step response (%ΔF/F averaged over the whole step) (imaged at 1,000 Hz). Laser illumination for Arch WT, Archer1, and Archer2 (λ = 655 nm; I = 880 mW mm^{-2}) is lower than that used for Arch EEQ (λ = 655 nm; I = 1,500 mW mm^{-2}). Error bars represent standard error of the mean (s.e.m.). *** $P < 0.001$, ns $P > 0.05$, un-paired student's t-test.

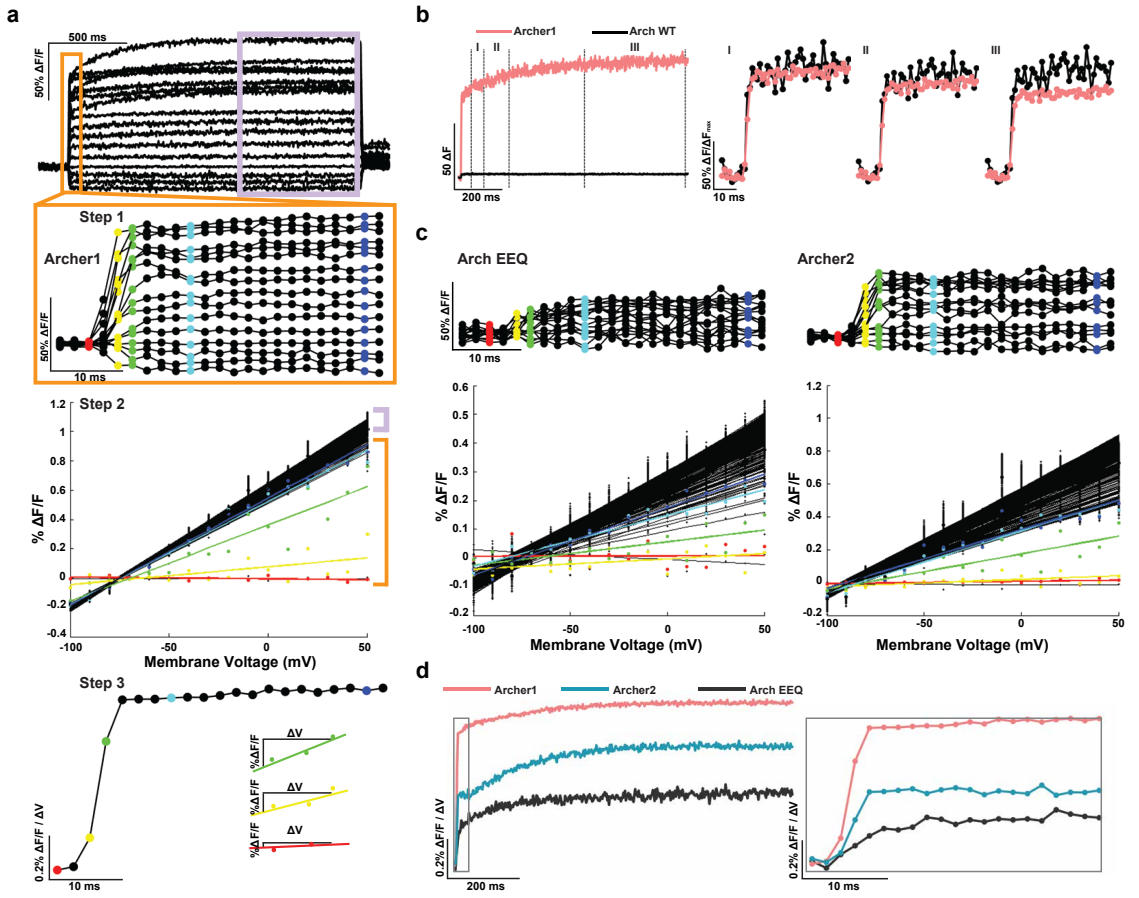


Figure 2.2. **A method for comparing different voltage sensors**

(**a**) Overview of the method used to quantify sensitivity kinetics. Step 1: averaged fluorescence responses (imaged at 500 Hz) of neurons expressing Archer1 (*n* = 10) to voltage clamped steps in membrane potential. Neurons are held at -70 mV and then stepped to voltages ranging from -100 mV to +50 mV in increments of 10 mV. Step 2: voltage sensitivity of fluorescence is plotted for each time point and a linear fit is calculated. This step assumes a linear dependence of fluorescence on voltage. Step 3: the slope for each linear fit is plotted over time. This measure allows one to calculate %ΔF/F for a desired voltage change over any timescale. (**b**) Averaged change in fluorescence due to a 100 mV step (-70 mV to +30 mV) of Archer1 (*n* = 10) compared to Arch WT (*n* = 6) shows significant differences in response magnitude (25-30x). To compare the kinetics of the two sensors, normalization across the step is necessary. The maximum value within

three different regions (I, II, and III) is used as a normalization factor, resulting in different apparent kinetics and prompting the need for a different method for kinetic analysis. (**c**) Plotting the voltage sensitivity for each time point with linear best fits for Arch EEQ (n = 5) and Archer2 (n = 3) shows a slower rise to the steady state value than Archer1 (n = 10). (**d**) Summarizing the sensitivity kinetics comparison of Archer1, Arch EEQ, and Archer2. Inset expands the first 40 ms. Laser illumination for Arch WT, Archer1, and Archer2 (λ = 655 nm; I = 880 mW mm^{-2}), and for Arch EEQ (λ = 655 nm; I = 1,500 mW mm^{-2}).

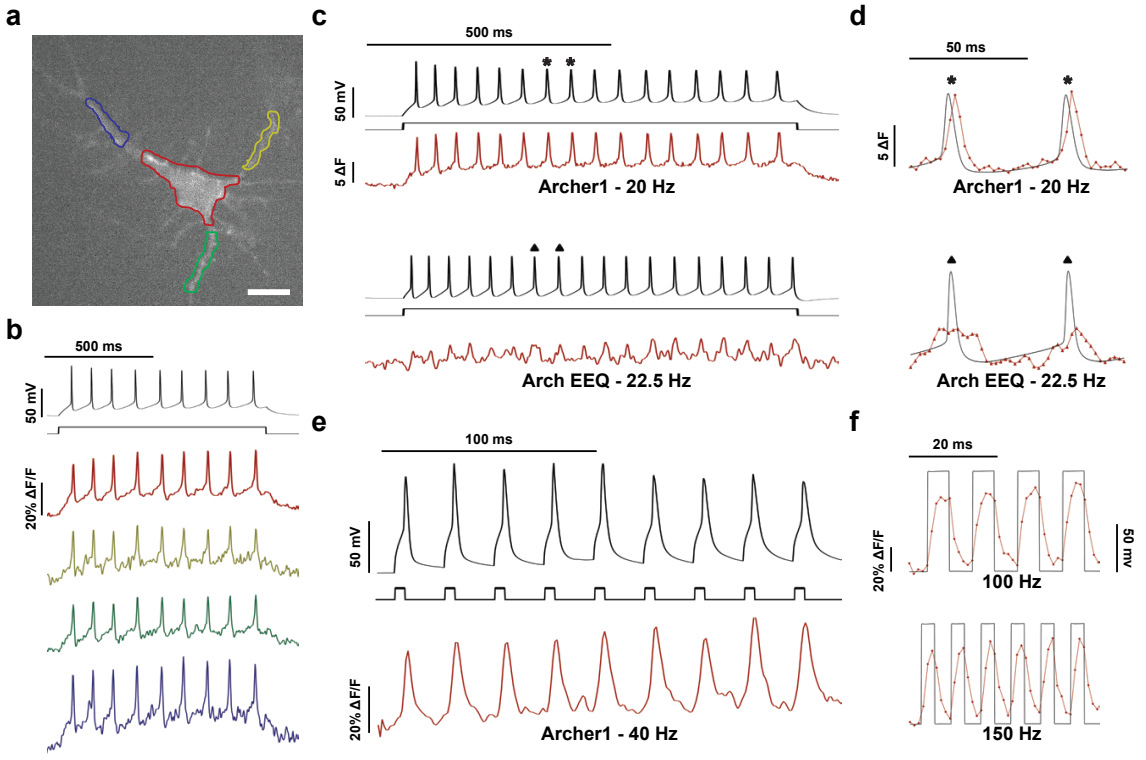


Figure 2.3. Archer1 fluorescence tracks action potentials in cultured neurons

(**a**) Fluorescence of Archer1 expressing rat hippocampal neuron. Cell body and individual processes are outlined. Scale bar, 10 mm. (**b**) Fluorescence (imaged at 500 Hz) from single-trial optical and electrophysiological recordings of action potentials induced by a step current injection (800 ms, 50 pA) analyzed for the color-matched somatic and dendritic areas outlined in (**a**). (**c**) Fluorescence (imaged at 500 Hz) from single trial recordings of action potentials in neurons expressing Archer1 and Arch EEQ. Firing of 20 and 22.5 Hz respectively is generated through a step current injection (800 ms, 50 pA) in current-clamped cells. Fluorescence change is measured in absolute terms, as opposed to a percentage change, due to the lower baseline fluorescence of Arch EEQ. (**d**) Expanded regions of action potentials from (**c**). Archer1 shows ~2x higher change in fluorescence and > 6x increase in SNR (24.03 vs. 3.75) when compared to Arch EEQ, allowing it to better track action potential waveforms. Each fluorescent point is 2 ms apart. (**e**) Archer1 fluorescence (imaged at 1000 Hz) successfully tracks action potentials in cultured rat

hippocampal neurons at 40 Hz: higher limit for such cultures, generated through a succession of brief, large amplitude current pulses (5 ms, 500 pA). Individual action potentials at 40 Hz show ~40% change in ΔF/F. (**f**) Single-trial recording of high frequency (100 Hz and 150 Hz) voltage steps (-70 mV to +30 mV) are generated in neurons to test Archer1's ability to detect fast trains of depolarization and hyperpolarization. Fluorescence changes (imaged at 1,000 Hz) exhibited by Archer1 are > 50% ΔF/F for both frequencies and return near baseline between each pulse. Each fluorescent point is 1 ms apart. Laser illumination for Archer1 (λ = 655 nm; I = 880 mW mm^{-2}) and Arch EEQ (λ = 655 nm; I = 1,500 mW mm^{-2}). Fluorescence traces in (**b**)-(**e**) have undergone background subtraction and Gaussian averaging.

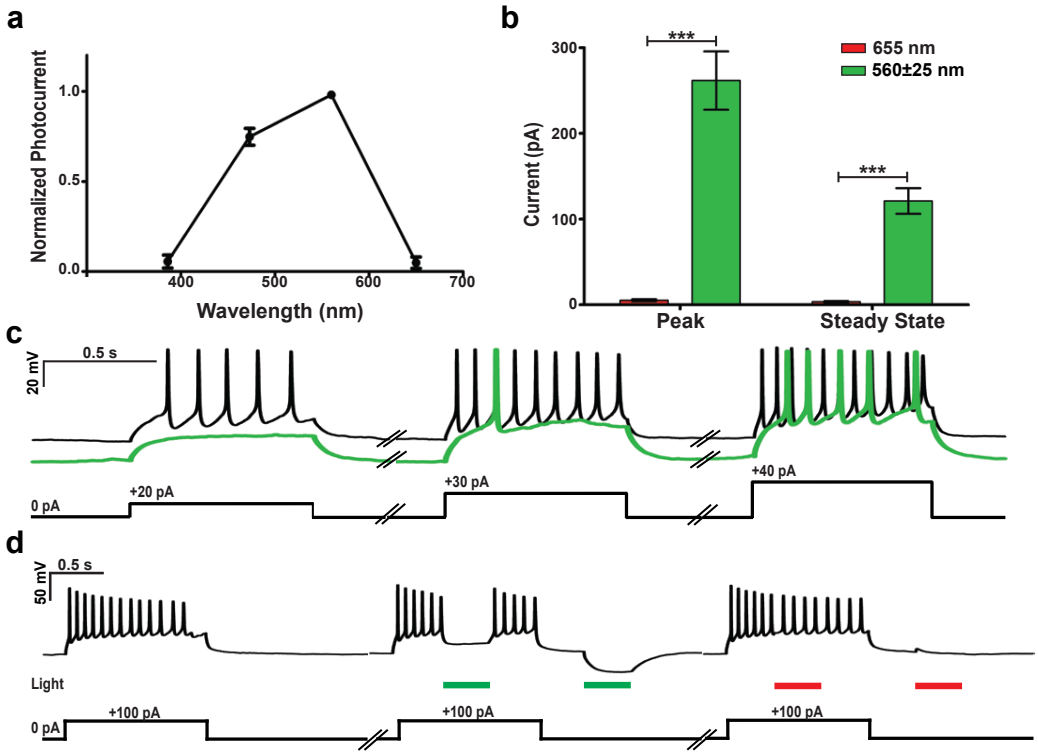


Figure 2.4. Archer1 acts as either a sensor or actuator at separate wavelengths

(**a**) Normalized steady-state activation spectrum of Archer1 spanning wavelengths between 386 – 650 nm (n = 11). (**b**) Currents induced by low intensity green LED illumination (n = 8, λ = 560±25 nm; I = 3 mW mm^{-2}) are significantly larger than those induced by high intensity red laser illumination (n = 16, λ = 655 nm; I = 880 mW mm^{-2}). (**c**) Archer1 exposed to green light successfully inhibits action potentials induced by step current injections (at 20, 30, and 40 pA) when compared to non-illuminated current injections in the same cell. (**d**) Action potentials induced by a 100 pA current injection (900 ms) are inhibited by a pulse of green light (300 ms; I = 3 mW mm^{-2}), while no inhibition of action potentials is observed with a pulse of red laser at the power used to excite fluorescence (300 ms; I = 880 mW mm^{-2}). Additionally, with no current injection, hyperpolarization is observed with exposure to green, but not red light. Error bars represent standard error of the mean (s.e.m.). *** $P < 0.0001$, unpaired student's t-test.

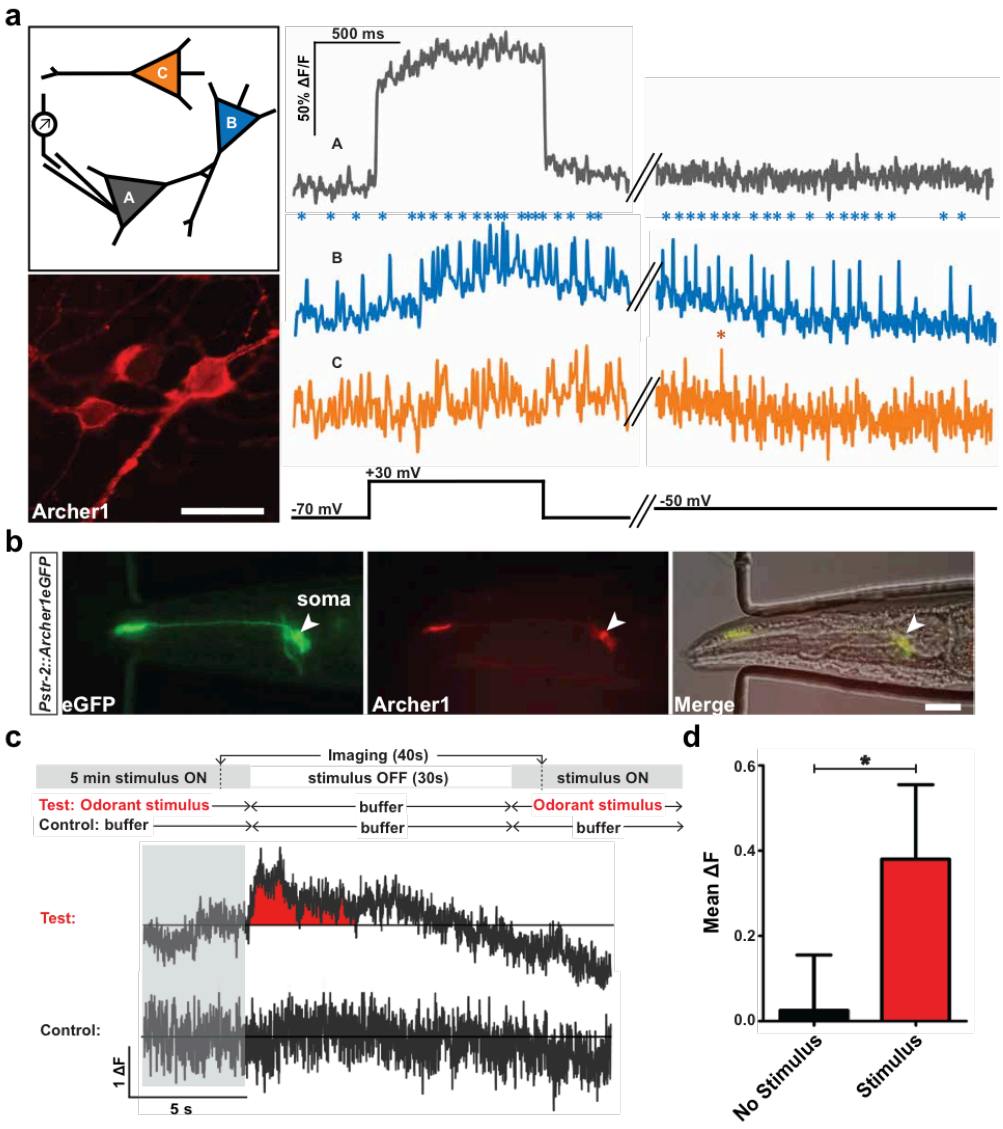
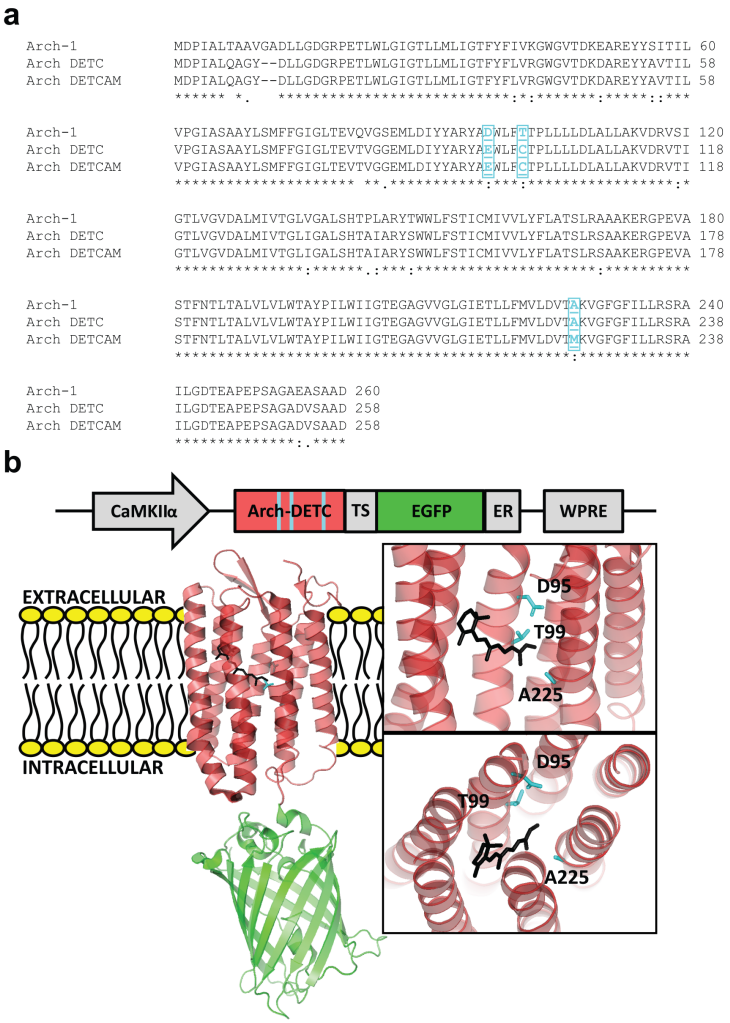


Figure 2.5. **Archer1 tracks activity in populations of cultured neurons and behaving worms**

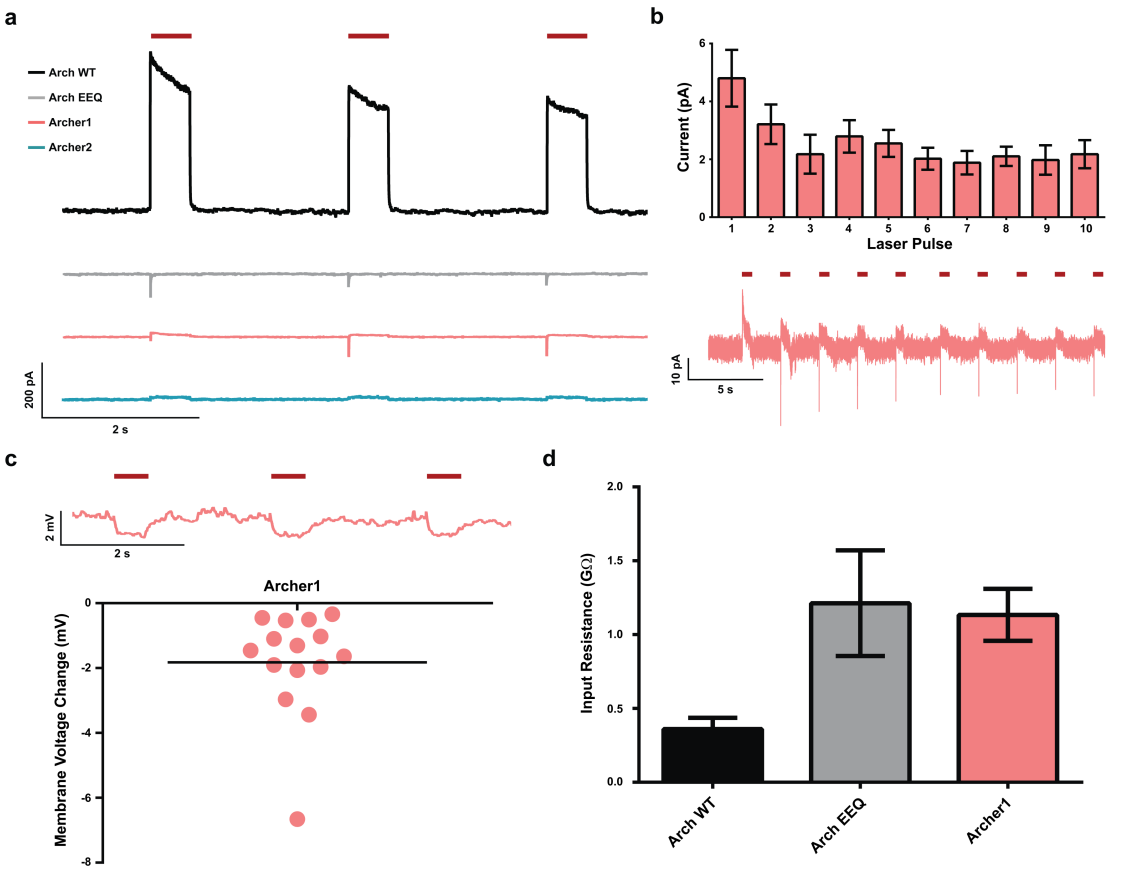
(**a**) Monitoring fluorescence in three Archer1 expressing cultured neurons with electrical stimulation of one cell. Cell A undergoes a voltage clamped 100 mV step and fluorescence changes in the population are measured simultaneously. Cell A exhibits a step-like increase in fluorescence corresponding to the voltage step. Cell B, whose fluorescence indicates spontaneous firing previous to the step, shows an increase in firing rate concurrent with the voltage step in Cell A, with continued firing after the step is completed. Fluorescence of

Cell C appears not responsive to the voltage step in Cell A. Asterisks indicate action potential-like changes in fluorescence (~35-40% ΔF/F increase within 10 ms). Scale bar, 20 mm. (**b**) *C. elegans* expressing Archer1 in one AWC neuron shows opsin fluorescence (λ = 655 nm; I = 880 mW mm^{-2}, 100 ms exposure) co-localizing with fused EGFP fluorescence (λ = 485±20 nm; I = 0.05 mW mm^{-2}, 100 ms exposure). Scale bar, 20 mm. (**c**) Top: behavioral paradigm: worms are stimulated with odorant (Isoamyl alcohol, IAA) for 5 minutes, flow is switched to buffer (S-Basal) for 30 seconds, and then odorant flow is restored. On the same worm, a control is performed where odorant is replaced with buffer. Bottom traces: imaging of Archer1 fluorescence (250 Hz) is performed continuously for 40 seconds, starting 5 seconds prior to flow switch. Averaged ΔF traces for two worms are shown. (**d**) Mean fluorescence of the 4 second time window after switch shows a significant increase with stimulus compared to no-stimulus controls (n = 4 worms). Fluorescence traces imaged at λ = 655 nm; I = 880 mW mm^{-2}. Fluorescence traces in (**a**) and (**b**) have undergone background subtraction and Gaussian averaging. Error bars represent standard error of the mean (s.e.m.). * $P < 0.05$, paired student's t-test.

2.7 Supplementary figures and tables

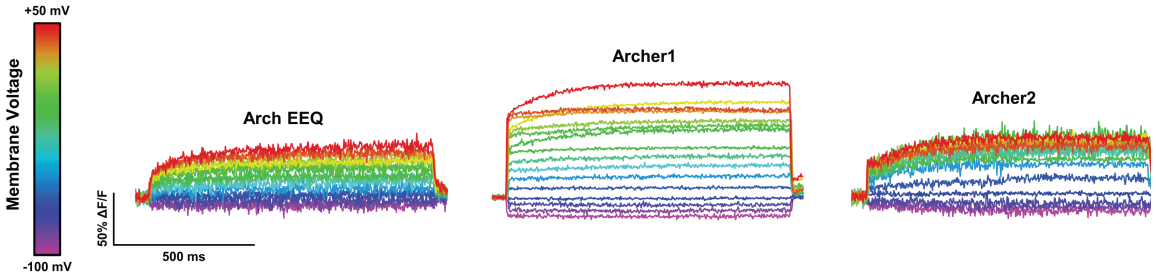


Supplementary Figure 2.1. Structural alignment of Arch variants with Arch-1. (**a**) Sequence alignment via ClustalW2. Arch-1 (77) (Uniprot P69051), Archer1, and Archer2 share 93% amino acid identity. The alignment shows the D95E, T99C and A225M mutations of Archer1 and Archer2 from Arch WT boxed in blue. (**b**) Archer1 construct design and schematic of location of opsin-fluorescent protein fusion in membrane. Locations of the mutated residues (D95, T99, and A225) are shown in blue and their relative positions to the retinal chromophore in black.

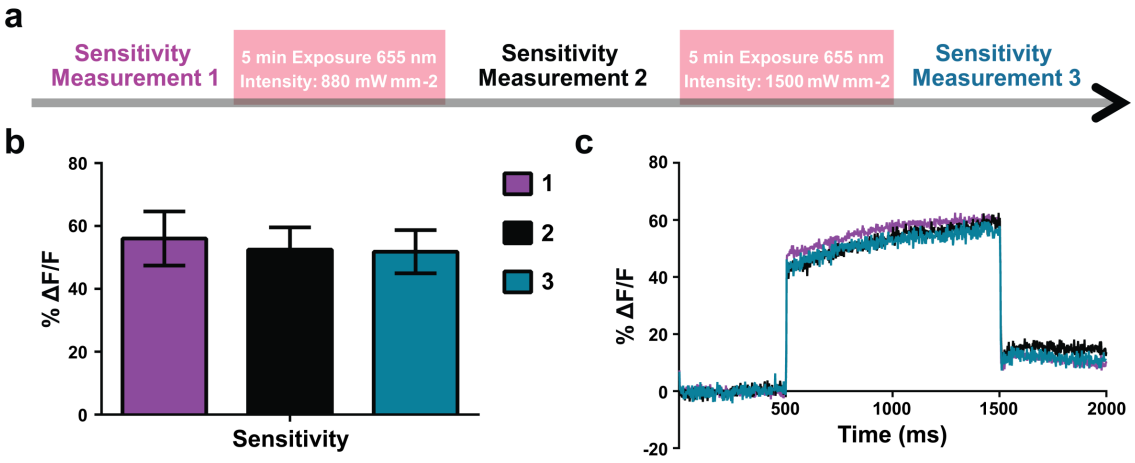


Supplementary Figure 2.2. Residual photocurrents of Arch variants and effect on membrane potential. (**a**) Single trace voltage-clamp recordings of photocurrents in neurons expressing Arch WT and variants in response to three consecutive pulses of laser illumination at the intensity used for fluorescence imaging. Arch EEQ, as previously reported (16), shows no steady-state photocurrent in response to laser illumination, while Archer1 and Archer2 exhibit small steady-state currents. Arch EEQ and Archer1 both respond to laser illumination with a brief peak of depolarizing photocurrent before reaching steady state. This has been observed with microbial rhodopsin-based voltage sensors as previously reported for Mac (5). (**b**) Archer1 photocurrent characteristics are measured in response to 10 consecutive laser pulses (n = 10). An initial peak current is generated in naïve cells exposed to laser illumination for the first time. Subsequent pulses reach a lower steady state without a peak. (**c**) Current clamp recordings of changes in membrane voltage of neurons expressing Archer1 (n = 15) induced by pulses of laser illumination. (**d**) Input resistance of patched cells expressing Arch WT (n = 8), Arch EEQ (n = 10), and Archer1 (n = 10) recorded as a measure of quality of the seal break. Laser illumination for Arch

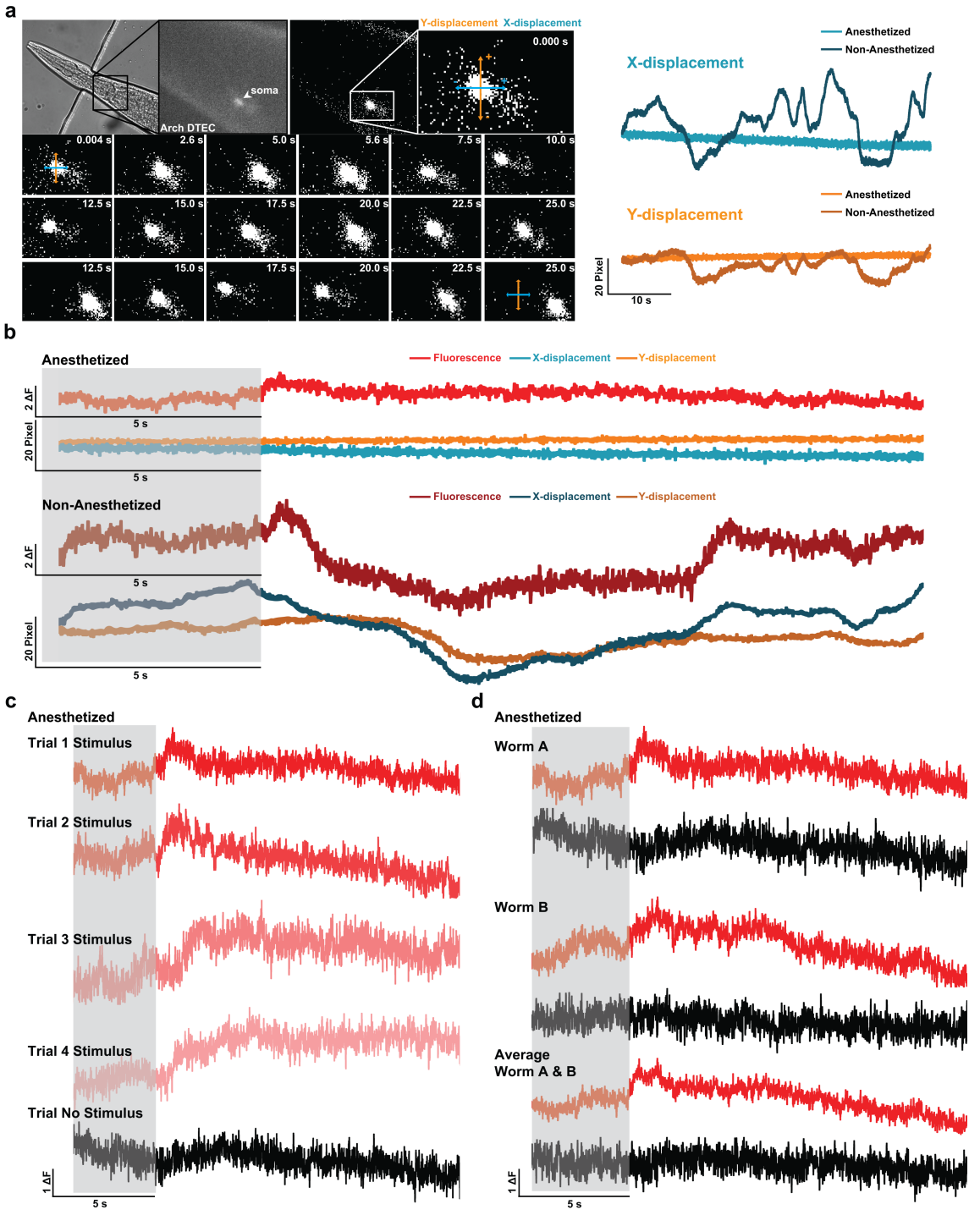
WT, Archer1 and Archer2 (λ = 655 nm; I = 880 mW mm^{-2}), and Arch EEQ (λ = 655 nm; I = 1,500 mW mm^{-2}). Error bars represent standard error of the mean (s.e.m.).



Supplementary Figure 2.3. Averaged fluorescence sensitivity of Arch variants. Averaged fluorescence responses (imaged at 500 Hz) of neurons expressing Arch EEQ (n = 5), Archer1 (n = 10) and Archer2 (n = 3) to voltage clamped steps in membrane potential. Neurons are held at -70 mV and then stepped to voltages ranging from -100 mV to +50 mV in increments of 10 mV. Laser illumination for Archer1 and Archer2 (λ = 655 nm; I = 880 mW mm^{-2}), and Arch EEQ (λ = 655 nm; I = 1,500 mW mm^{-2}).



Supplementary Figure 2.4. Archer1 fluorescence sensitivity is stable with prolonged illumination. (**a**) Laser exposure and sensitivity measurement paradigm consists of detecting the sensitivity of fluorescence response to 100 mV voltage step in three consecutive measurements separated by 5 minutes of continuous laser exposures, with the first exposure at 880 mW mm^{-2} and the second at 1,500 mW mm^{-2}. (**b**) The average percentage change in fluorescence in response to 100 mV step in voltage does not significantly change after the first (n = 8) or second (n = 6) prolonged laser exposure. (**c**) Average fluorescence waveforms for the sensitivity measurements described in (**a**, **b**) show no change in the characteristics of fluorescence response. Laser illumination for Archer1 (λ = 655 nm; I = 880 mW mm^{-2}). Error bars represent standard error of the mean (s.e.m.).



Supplementary Figure 2.5. Worm movement and fluorescence in anesthetized vs. non-anesthetized worms. (**a**) Tracking fluorescence of an AWC cell throughout a stimulus paradigm. Cell location is determined by averaging coordinates of fluorescent pixels above a set threshold and monitoring their position on an x-y coordinate plane over time. Non-

anesthetized worms show significant movement in both x (blue) and y (red) direction throughout the stimulation protocol compared to anesthetized worms. (**b**) Changes in fluorescence in response to cessation of exposure to odorant stimulus (IAA) are time-locked to respective cell movement for anesthetized vs. non-anesthetized worms. Non-anesthetized worms show frequent changes in fluorescence correlated with movement, not apparent in anesthetized worms. (**c**) Fluorescence traces of repeated trials of stimulation (red) within the same worm compared to control (black). (**d**) Single trial fluorescence response to stimulus and control paradigms for two worms (A and B) and the average fluorescence trace of the two. Fluorescence traces imaged at λ = 655 nm; I = 880 mW mm^{-2}. Fluorescence traces (ΔF) in (**b**)-(**d**) have undergone background subtraction and Gaussian averaging.

Supplementary Table 2.1. Accession codes

Construct	**Addgene #**
pLenti-CaMKIIa-eArch3.0-EYFP	35514
FCK-Arch-GFP	22217
pLenti-Arch-EEQ	45188

Chapter 3

GENETICALLY ENCODED SPY PEPTIDE FUSION SYSTEM TO DETECT PLASMA MEMBRANE-LOCALIZED PROTEINS *IN VIVO*

A version of this chapter has been published as (78).

3.1 Summary

Membrane proteins are the main gatekeepers of cellular state especially in neurons, serving either to maintain homeostasis or to instruct response to synaptic input or other external signals. Visualization of membrane protein localization and trafficking in live cells facilitates understanding the molecular basis of cellular dynamics. We describe here a method for specifically labeling the plasma membrane-localized fraction of heterologous membrane protein expression using channelrhodopsins as a case study. We show that the genetically encoded, covalent binding SpyTag and SpyCatcher pair from the *Streptococcus pyogenes* fibronectin-binding protein FbaB can selectively label membrane-localized proteins in living cells in culture and *in vivo* in *Caenorhabditis elegans*. The SpyTag/SpyCatcher covalent labeling method is highly specific, modular, and stable in living cells. We have used the binding pair to develop a channelrhodopsin membrane localization assay that is amenable to high-throughput screening for opsin discovery and engineering.

3.2 Introduction

Real-time visualization of biochemical processes in living cells is aided by methods for specific protein labeling, including genetically encoded fluorescent proteins and synthetic probes. Since their first application as markers for transgenic protein expression and localization in live cells (79), genetically encoded fluorescent proteins (FP) have been engineered (80) to offer a palette of colors with enhanced brightness (80, 81) and various useful properties such as reversible or irreversible photoswitching (82-85) to aid in tracking protein dynamics (86). Synthetic fluorescent probes that covalently label proteins have

facilitated live cell imaging (87-91) due to their irreversible, highly specific binding. These bright, cell permeable, spectrally diverse, fluorescent probes are ideal for microscopy of cells in culture (92). However synthetic probes must be applied exogenously, making real-time *in vivo* protein tracking difficult. Methods for specific covalent labeling using synthetic fluorescent probes also requires protein tag fusions to the protein of interest: SNAP-tag, 181 amino acids (90, 91, 93); CLIP-tag, 181 amino acids (88); or Halo tag, 295 amino acids (87). The large size of these tags presents the risk that the assay system itself disturbs the natural compartmentalization and localization of the targeted protein.

Here we report a general method for post-translational, covalent labeling of cell surface exposed transgenic proteins using all genetically encoded components. This method specifically and quantitatively labels membrane proteins in living cells without impacting cell viability and therefore enables further experimentation with the labeled cells (e.g. electrophysiology or imaging of protein dynamics). The method uses the covalent SpyTag-SpyCatcher peptide-protein system first described by Zakeri *et al.* (94) which was structurally characterized and optimized by Li *et al.* (53). We show that the short peptide tag (SpyTag, 13 amino acids) fused to a membrane protein of interest can form a covalent bond with an exogenously added or expressed SpyCatcher-XFP labeling protein (SpyCatcher, 139 amino acids). This short tag system is ideal for visualizing membrane protein localization since its small size will likely minimize the effect on protein folding and membrane localization relative to the larger tag methods previously described. Here we demonstrate that the inexpensive and scalable SpyTag/SpyCatcher system can be used to 1) label membrane-localized proteins used for optogenetics (channelrhodopsins C1C2 (23) and ReaChR (27)) and receptors (TrkB) transfected in HEK cells and primary neuronal cultures; 2) aid in membrane protein engineering via an assay for membrane localization in a 96-well plate format platform; and 3) identify membrane protein localization in whole living organisms in an all-genetically encoded fashion.

3.3 Results

The SpyTag/SpyCatcher pair labels membrane-localized channelrhodopsins in live cultures. We used the SpyTag/SpyCatcher system to label membrane-localized, light-activated ion channels, channelrhodopsins (ChRs), in live cells. Since the SpyCatcher-XFP is too large to passively cross the membrane, specific labeling of membrane-localized protein requires the SpyTag be fused to a portion of the protein displayed on the extracellular surface. To limit potential disruption to the three-dimensional membrane protein structure we chose to target the SpyTag to the N-terminal region of the channelrhodopsin C1C2, a variant with a known crystal structure (23) (**Figure 1A**), immediately C terminal to the proposed post-translationally cleaved, signal peptide sequence (residues 1-23) (23) (**Figure 1A**). Though previous work on the SpyTag/SpyCatcher system has shown that it is not limited to N- or C-terminal application (95), for our application N-terminal application was optimal. The fluorescent protein mCherry was fused to the C-terminus of the opsin as a marker of total protein expression (Tag-C1C2-mCherry) (**Figure 1A**). The SpyCatcher binding partner was produced separately for exogenous labeling by expression in *E. coli* with an elastin-like protein (ELP) inserted between SpyCatcher and its GFP fluorescent label (Catcher-GFP), in an attempt to minimize steric interference between the fluorescent protein and the cell membrane. A 6xHis tag was inserted at the N-terminus of the SpyCatcher for purification purposes (**Figure 1A**). Catcher-GFP was expressed in bulk, purified, and buffer exchanged to ready it for extracellular application.

The SpyTag-mCherry-labeled C1C2 channelrhodopsin was expressed in human embryonic kidney (HEK) cells, incubated with 25 μM Catcher-GFP protein for 45 min, washed and imaged. Maximum-intensity projections and single plane confocal images show that the SpyCatcher-GFP binds to the membrane-localized fraction of the Tag-C1C2-mCherry expressed in live cells, with minimal background (**Supplementary Figure 1A**). Intracellular Tag-C1C2-mCherry protein was not labeled by Catcher-GFP (**Supplementary Figure 1A**). Full field, single plane confocal images show that only cells expressing Tag-C1C2-mCherry are labeled with Catcher-GFP (**Supplementary Figure 1A).** Intracellular puncta or aggregates of Tag-C1C2-mCherry (**Supplementary Figure 1A**) could be due to oligomerization of mCherry (96). We chose mCherry because it is the

most commonly used red marker for opsins used in optogenetics (62). Because the SpyTag/SpyCatcher system is modular, any FPs can be substituted for mCherry and GFP, as long as they are spectrally distinguishable.

Labeling in live cells requires SpyTag display on the cellular surface and covalent binding to SpyCatcher. The placement of the SpyTag dictates its accessibility for labeling with SpyCatcher. In addition to the constructs discussed above that mediated stable and robust labeling with Catcher-GFP, a number of alternative constructs were built to test the requirements of the SpyTag/SpyCatcher system in live and fixed cells. As expected, Catcher-GFP applied to cultured cells expressing a C-terminal fusion of SpyTag to ChR2-mCherry does not label the inaccessible, intracellular SpyTag (**Supplementary Figure 3.2B**). However, when cells were permeabilized with paraformaldehyde (PFA), SpyCatcher-GFP could label the C-terminal SpyTag (**Supplementary Figure 3.2B**). Mutation of the reactive aspartic acid (D) residue in SpyTag to a non-reactive alanine (A) (Tag(DA)-C1C2-mCherry) leads to no observable labeling with Catcher-GFP when the SpyTag is expressed in HEK cells (**Figure 3.2A**), indicating that the covalent bond is required for stable labeling of the membrane-localized Tag-C1C2-mCherry. Placement of the SpyTag N-terminal to the signal peptide cleavage site (Tag^0-C1C2-mCherry) also leads to no observable labeling with Catcher-GFP when the SpyTagged construct is expressed in HEK cells (**Figure 3.2A**).

Labeling of cell surface displayed Tag with Catcher-GFP in complex media and at temperatures suitable for live cell applications. Catcher-GFP (2-50 μM) added directly to the medium of live cells expressing Tag-C1C2-mCherry shows significant labeling of the membrane-localized opsin (**Figure 1A,B** and **Supplementary Figure 1B-D**). SpyTag/SpyCatcher covalent binding on the surface of live cells is robust to different temperatures in the range 16-37^{o}C (**Supplementary Figure 1D**), consistent with reported binding results using purified SpyTag/SpyCatcher protein (94). Robust binding in live cells at different temperatures is particularly useful for temperature-dependent protocol such as heat-shock experiments in flies, zebra fish, and nematodes, i.e., (97-99).

In **Supplementary Figure 1B-D** the efficiency of the Catcher-GFP binding to the Tag-C1C2-mCherry is reported as the ratio of GFP fluorescence to mCherry fluorescence using measurements of individually selected cells. This binding efficiency metric is internally normalized for the total protein expression level. The results in **Supplementary Figure 1B** show Catcher-GFP binding is saturated at 25 μM, and therefore 25 μM Catcher was used for all subsequent experiments in cultured cells. A time course for Catcher-GFP labeling of Tag-C1C2-mCherry expressing cells in culture medium indicates that binding improves with increased incubation time up to one hour (**Supplementary Figure 1C**).

Addition of the N-terminal Tag and covalent labeling with the Catcher-GFP does not affect channelrhodopsin expression or *in vitro* function in neurons. Since the SpyTag/SpyCatcher system gave efficient labeling under optimal live cell conditions, we tested its impact on neuronal function in primary neuronal cultures commonly used for microbial opsin characterization and refinement (62). Application of the Catcher-GFP directly to neuronal medium at 37^{o}C for 1 hour followed by washing with MEM shows efficient membrane labeling and sustained cell health (**Figure 1B**). This labeling method provided efficient Catcher-GFP binding to membrane-localized Tag-C1C2-mCherry expression in neurons (**Figure 1B**). These data show distinct membrane labeling at the cell body as well as throughout the axon, dendrites and axon terminals (**Figure 1B**). Whole-cell patch-clamp recordings of neurons expressing C1C2-mCherry, Tag-C1C2-mCherry, and the labeled GFP-Catcher-Tag-C1C2-mCherry complex show no significant differences in photocurrent magnitude or wavelength sensitivity (**Figure 1D,E)** to that of cells expressing similar unlabeled opsin levels (**Figure 1C**), indicating that the N-terminal SpyTag has no significant effect on opsin properties. Thus Spy-tagged opsin constructs can be used for optogenetic applications and then labeled for follow-up analysis.

To verify that SpyTag can be applied to other channelrhodopsins we inserted SpyTag C-terminal to residue 24 of ReaChR and observed efficient expression and labeling with Catcher-GFP in primary cultured neurons (**Figure 3.2B**). Patch-clamp electrophysiological recordings indicate that tagging ReaChR-mCherry does not effect photocurrent magnitude or spectral properties (**Figure 3.2C**), similar to the measurements for the tagged C1C2-

mCherry in **Figure 1E**. To test the applicability of the system beyond microbial opsins, we added the SpyTag to the N-terminus of the tropomyosin-related kinase B (TrkB) receptor (100). We observed efficient labeling of the membrane-localized protein with Catcher-GFP in HEK cells and in primary cultured neurons (**Supplementary Figure 3.3**).

SpyTag/SpyCatcher can be used to screen libraries for membrane-localized ChRs. Because opsin membrane localization is a prerequisite for activity in most optogenetic applications, we have used the SpyTag/SpyCatcher system in 96-well plate format for pre-screening libraries of opsin variants for membrane localization. As shown in **Figure 3.2B**, the N-terminal Tag-ReaChR-mCherry construct shows good expression and efficient membrane localization. We used Tag-ReaChR-mCherry as a parent for preparing a library of opsin variants and tested the ability of the SpyTag/SpyCatcher membrane localization assay to eliminate mutants with lesser membrane localization. Two residue positions, E130 and N289, identified as being part of the putative channel gate (23), were targeted for saturation mutagenesis.

Site-saturation mutagenesis libraries were generated at the E130 and N289 positions. Plasmid DNA from 30 clones was purified for each library (74% coverage) and used to transfect cultured HEK cells in a 96-well format (**Figure 3.3A**). Forty-eight hours post transfection, Catcher-GFP was added to the media of expressing HEK cells to label the membrane-localized opsin (**Figure 3.3A**). Soluble Catcher-GFP was removed, the cells were washed with maintenance medium, and full field, low magnification (10x) images containing hundreds of transfected cells were analyzed for mCherry and GFP fluorescence (**Figure 3.3A,D**; **Supplementary Figure 3.4A**). The ratio of GFP/mCherry fluorescence (reflecting the fraction of protein that is membrane localized) for each screened variant was plotted vs. the mCherry fluorescence (total opsin expression) for the two libraries (**Figure 3.3B**). Variants from the N298 library generally showed much lower membrane localization compared to the parent (Tag-ReaChR-mCherry) and compared with the E130 library (**Figure 3.3B,C**).

Four variants showing membrane localization and expression equal to or above the parent Tag-ReaChR-mCherry ('hits') and two variants showing membrane localization significantly worse than the parent ('poor localizers') were selected from the E130 library (**Figure 3.3B**) and further characterized. Three 'poor localizer' variants from the N289 library were also selected. No variants from the N289 library gave membrane localization and expression equal to or above the parent so none were selected as 'hits' (**Figure 3.3B**). Selected variants were sequenced, re-streaked to obtain high purity DNA for each variant, and used to transfect HEK cells. Catcher-GFP labeling was carried out 48 hours post-transfection. Single-plane, confocal images of expressing, labeled cells of each variant show that each of the 'hits' have predominantly membrane-localized opsin (Tag-ReaChR (E130T, E130G, E130Q and E130L)) while all of the 'poor localizers' show the opsin protein split between intracellular and membrane localization (Tag-ReaChR E130Y and E130D) (**Figure 3.3D**; **Supplementary Figure 3.4B**). Quantification of GFP/mCherry fluorescence measurements of individual cells within a population confirms that the variants identified as 'hits' have membrane localization similar to the parent while variants identified as 'poor localizers' have significantly lower GFP/mCherry compared to the parent (**Figure 3.3F; Supplementary Figure 3.4E**). The mCherry fluorescence quantification shows that only one variant Tag-ReaChR (E130D) had significantly lower overall expression compared to Tag-ReaChR (**Figure 3.3F; Supplementary Figure 3.4E**).

Electrophysiology was used to compare photocurrents of the 'hits' and the 'poor localizers' of the E130 library (**Figure 3.3G**). 'Poor localizers' E130Y and E130D show weak currents, both peak and steady state, compared to the Tag-ReaChR parent under green light (590 nm) activation. This decrease in current is not due to a shift in spectral sensitivity. The maximum excitation wavelength for all variants is closest to 590 nm within the wavelengths tested ranging from 390-650 nm (**Figure 3.3H**). Further the decrease in current is not due to an altered reversal potential since the currents at all holding potentials are much lower for the 'poor localizers' when compared with the Tag-ReaChR. The 'hits', on the other hand, show both high and low currents (**Figure 3.3G**). This variability is to be expected since total photocurrents are a result of both membrane localization and channel conductance. These data suggest that variants Tag-ReaChR E130T and E130L may have

decreased single channel conductance resulting in low currents while variants Tag-ReaChR E130G and E130Q appear to have single channel conductance similar to the parent (Tag-ReaChR). Of particular interest is the variant Tag-ReaChR E130G which has no side chain at residue 130 while the parent has a large, negatively charged side chain, but both variant and parent appear to have similar ion conductance, while introduction of a polar, uncharged side chain (E130T) or a hydrophobic side chain (E130L) both results in what appears to be a strong decrease in the conductance of the channel.

These results indicate that the SpyTag/SpyCatcher system is a useful tool for screening libraries of opsin mutants for membrane localization. Opsin membrane localization is sensitive to mutations in the protein, and mutations at some residue positions have more drastic effects on expression and localization than others. This assay can facilitate pre-screening of ChRs libraries to eliminate variants with poor localization and enrich for functional ChRs for further analysis using low-throughput but precise methods such as patch-clamp electrophysiology. If hits are identified as having high expression and good membrane localization then using electrophysiology to characterize the hits enables identification of single amino acid substitutions that have significant affect on the channel's electrical properties (i.e. conductance) without the confounding variable of expression and membrane localization.

Stability of SpyTag/Catcher labeling enables monitoring of protein dynamics in living cells. We hypothesized that the Spy system would be sufficiently stable in live cells to enable observation of protein dynamics. Catcher-GFP was added directly to the medium of Tag-C1C2-mCherry-expressing cells for 1 hour, at which point the cells were washed and imaged for both mCherry fluorescence and GFP fluorescence (Day 1). Labeled cells were then incubated at 37^{o}C for an additional 24 hours and reimaged (Day 2) (**Supplementary Figure 3.5**). The SpyTag/Catcher labeling was strongest on Day 1, but significant labeling was visible after 24 hours (Day 2) (**Supplementary Figure 3.5**), and Catcher-GFP labeling was visible up to 3 days after the initial treatment (**Supplementary Figure 3.5**). These observations indicate that even in a rapidly dividing mammalian cell line the

SpyTag/SpyCatcher interaction is maintained at the cell surface over several days though there is a decrease in the observed level of Catcher-GFP.

Comparison of SpyTag/Catcher and SNAP-tag labeling methods. To test our hypothesis that N-terminal insertion of larger tags, i.e. SNAP-tag, can disturb the natural compartmentalization and localization of a membrane protein we compared the expression, membrane localization and photocurrents of the Tag-C1C2-mCherry construct with a SNAP-tag-C1C2-mCherry construct in HEK cells. The SNAP-tag-C1C2-mCherry was constructed with the SNAP-tag sequence inserted after the signal peptide sequence (residues 1-23) in the same N-terminal position as the SpyTag and the Tag-C1C-mCherry construct. The Tag-C1C2-mCherry construct is able to express and traffic to the plasma membrane more efficiently than the N-terminal SNAP-tag opsin fusion construct (SNAP-tag-C1C2-mCherry) in mammalian cell culture when imaged under the same imaging conditions (**Figure 3.4A and Figure 3.4B**). Due to the decrease in localization the SNAP-tag-opsin has decreased currents upon activation with 480 nm light (**Figure 3.4C**) in cells with similar levels of overall mCherry expression (**Figure 3.4D**). Though the SNAP-tag system has enabled post-translational labeling of a number of protein targets (101, 102) these results indicate that for tagging channel proteins such as opsin the SpyTag/SpyCatcher system has less effect on native protein trafficking though it should be noted that the performance of one labeling strategy over another is protein specific.

Use of SpyTag/SpyCatcher to label membrane proteins *in vivo*. Since all the components of the SpyTag/SpyCatcher labeling method are genetically encoded, it can be applied to living organisms. As proof-of-concept, we specifically expressed Tag-C1C2-mCherry in the gonad of the nematode *C. elegans* and demonstrated that Catcher-GFP labels cells within the organ (**Figure 3.5A**). The *C. elegans* gonad arms are shaped through the migration of distal tip cells (DTCs), two cells that cap each end of the tube-like structure (103). We generated transgenic nematodes that specifically expressed Tag-C1C2-mCherry in the DTCs using an endogenous *hlh-12* promoter and observed mCherry fluorescence both at the plasma membrane and in internal compartments (**Figure 3.5A**). Because the outer cuticle of the animal is not permeable to Catcher-GFP, the gonad was

dissected out, fixed, and exposed to a solution of purified Catcher-GFP. Tag-C1C2-mCherry expressing DTCs were the only cells in the gonad that were labeled by Catcher-GFP, and its localization was specific to the plasma membrane ($N = 5$, **Figure 3.5A**). In the control experiment, DTCs that did not express Tag-C1C2-mCherry were not labeled by Catcher-GFP ($N = 7$).

Since both SpyTag and SpyCatcher can be produced endogenously within the organism where the labeling reaction occurs, we then produced transgenic nematodes expressing Tag-C1C2-mCherry in the DTCs under the *hlh-12* promoter and Catcher-GFP under a heat-shock (HS) promoter. The *HS::SpyCatcher-GFP* construct was designed to be expressed in many tissues upon HS treatment and, due to its signal sequence, secreted extracellularly into the body cavity. At room temperature the DTCs expressed only Tag-C1C2-mCherry and no Catcher-GFP ($N = 15$, **Figure 3.5B**), three hours after a 33°C HS treatment, we observed specific Catcher-GFP labeling at the DTC plasma membrane ($N = 6$, **Figure 3.5B**). Initially we observed background cytoplasmic fluorescence from Catcher-GFP expression in the cells responsive to HS, however twenty-four hours after HS treatment, the DTC plasma membrane continues to be stably labeled by SpyCatcher-GFP ($N = 13$), and the background Catcher-GFP fluorescence was absent (**Figure 3.5B**). To demonstrate specificity of labeling, we HS-treated control animals with *HS::Catcher-GFP* but not *hlh-12::Tag-C1C2-mCherry* and observed no Catcher-GFP labeling of DTCs three hours ($N = 6$) or 24 hours ($N = 11$) after HS (**Figure 3.5B**).

Given that the SpyTagged opsin constructs described here are most useful for neuronal applications we investigated SpyTag/SpyCatcher labeling and function of Tag-ReaChR constructs in *C. elegans* neurons. *C. elegans* has 26 GABA-producing neurons, including 19 D-type neurons that reside in the ventral nerve cord and innervate dorsal and ventral body muscle (**Figure 3.6A,C**). Activation of these GABA neurons inhibits body muscle contractions and paralyzes the worm (104) (**Supplementary Figure 3.6A**). We made transgenic animals expressing Catcher-GFP under heat-shock control and also specifically expressing either Tag-ReaChR-mCherry or the mutant Tag-ReaChR(E130D)-mCherry in GABA neurons. The Tag-ReaChR(E130D)-mCherry mutant was identified in the

expression/membrane-localization screen to have poor expression and membrane localization. We used this low expressing mutant both to test the sensitivity of the SpyTag/SpyCatcher screen *in vivo* and to further validate the screening method's potential to identify high and low expressers. Although the same concentration of transgenes was delivered for both Tag-ReaChR constructs, we found that Tag-ReaChR-mCherry expression is brighter than Tag-ReaChR(E130D)-mCherry (**Figure 3.6A**) . The mCherry expression in neuronal cell bodies and processes was visible at 200x magnification in 47% (N=36) of animals carrying the wild-type Tag-ReaChR-mCherry construct, but only in 4% (N=47) of animals carrying the Tag-ReaChR(E130D)-mCherry construct (**Figure 3.6A,B**). Expression of Tag-ReaChR(E130D)-mCherry was visible at 1000x magnification in 28% (N=47) of animals implying that the worms are transgenic but expressing the opsin mutant at very low levels. In *C. elegans* Tag-ReaChR(E130D)-mCherry appears to be expressed at lower levels than the parent molecule with the bulk of the protein localizing to the cell body rather than the cell processes (**Figure 3.6A,D**). These data are consistent with the mammalian cell culture results. To test labeling of the Tag-ReaChR constructs we heat-shock treated both transgenic animals, and examined labeling of Tag-ReaChR-mCherry and Tag-ReaChR(E130D)-mCherry by Catcher-GFP 24 hours after heat-shock. We observed specific Catcher-GFP labeling of the Tag-ReaChR expressing GABA neurons and processes for both constructs, but consistent with their expression levels, the Catcher-GFP labeling was brighter in Tag-ReaChR-mCherry over Tag-ReaChR(E130D)-mCherry (**Figure 3.6D**)**.** These results indicate that the SpyTag/SpyCatcher assay can be used *in vivo* to measure varying levels of expression and to differentiate between high and low membrane localization.

We tested whether the tagged opsin construct described in this study, could be used *in vivo* to induce light activated behaviors. We measured the impact of the Tag-ReaChR-mCherry and Tag-ReaChR(E130D)-mCherry expression on the locomotion behavior of the animal upon light activation. We selected animals expressing high levels of Tag-ReaChR-mCherry based on mCherry visibility at 200x magnification, and of mutant Tag-ReaChR(E130D)-mCherry based on visibility at 1000x magnification. By individually assaying the animal's locomotion behavior in response to green light, we found that 100% of animals expressing

wild-type (N=11) or mutant (N=10) Tag-ReaChR-mCherry immediately became paralyzed upon green light activation and recovered movement when the light was turned off (**Supplementary Figure 3.6**). Low expressing animals tested showed no effective paralysis upon light activation. Animals expressing high levels of wild-type Tag-ReaChR-mCherry but grown without all trans-retinal (ATR) did not become paralyzed in response to green light (N=3). Catcher-GFP labeling of Tag-ReaChR-mCherry did not affect the ReaChR function as shown by the results that 100% of animals (N=6) exhibited paralysis in response to green light exposure 4 hours after heat-shock treatment.

3.4 Discussion

This work demonstrates the SpyTag/SpyCatcher as a versatile system for the characterization of membrane localization of channels and receptors in live cells and organisms. The irreversible covalent interaction between the surface-displayed SpyTag, fused to a membrane protein, and the extracellular, SpyCatcher-GFP is not affected by competing proteins in complex culture media or in cells *in vivo* and permits efficient long-term labeling without disturbing cell viability. N-terminal insertion of the SpyTag into the ReaChR (27) and C1C2 (23) ChRs had no significant effect on their expression levels, membrane localization, or photocurrents which is not the case for the SNAP-tag cell-surface labeling method tested.

An application of the SpyTag/SpyCatcher system validated here is screening membrane localization of opsins in mammalian cells in high throughput to support directed evolution experiments for the discovery of improved opsins (35, 36, 41, 59). Membrane localization of ChRs is crucial to their ability to mediate efficient neuronal modulation (105). We demonstrate that the SpyTag/SpyCatcher system can be used in a 96-well format to enrich mutant libraries for membrane localizing variants that are therefore worthy of detailed, but time-involved, electrophysiological characterization. This method enables screening libraries to identify a reduced number of candidates for detailed characterization. This is important because the number and complexity of characteristics of a useful opsin (speed,

wavelength sensitivity, photocurrent strength, ion selectivity, and reversal potential) require extensive variant-by-variant analysis (62).

We shows that the SpyTag/SpyCatcher system can be used in live cells to label membrane-localized receptors (TrkB). The long-term stability of labeling and the neutral impact on cellular viability make the SpyTag/SpyCatcher useful for monitoring endocytosis of receptors. This is especially relevant in receptor systems in which insertion and endocytosis are critical to altering neuronal excitability, e.g. AMPA or NMDA receptors (106). We have successfully applied this method for *in vivo* labeling of proteins in live *C. elegans*, while retaining protein function for subsequent behavioral assays. Even *in vivo* the SpyCatcher is able to label low levels of expression of the SpyTagged molecule. Given this work the SpyTag/SpyCatcher could be used between cells on the extracellular matrix, to track transient interactions during development, or in response to physiological changes in live animals (i.e. *C. elegans*). Our work described here is dedicated to labeling tagged heterologous membrane proteins, however, with recent advances in genome editing via, e.g. CRISPR/Cas9 (107) the SpyTag/SpyCatcher system could also be expanded to label endogenous proteins.

The SpyTag/SpyCatcher genetically encoded post-translational fusion system can be used as an affordable, highly specific, binding assay for live and fixed cells in culture and *in vivo*. The SpyTag/SpyCatcher system is between 20-50x less expensive than using SNAP-tag labeling probes (New England BioLabs, S9124S) and between 14-35x less expensive than using FLAG-tag/secondary antibodies (Sigma-Aldrich, F3165/Life Technologies, A27022). This cost advantage enables high-throughput screening and large tissue volume labeling for which the cost of the labeling molecule can be prohibitive. The SpyTag and SpyCatcher have a covalent, irreversible interaction which is advantages for experiments that require long experimental times, *in vivo* labeling, and to reduce the level of labeling variability from well-to-well for high-throughput screening. The labeling protein can be fused to any fluorescent protein or enzyme for detection and can be bulk-produced, making it a preferred option when large amounts of antibodies are required, for example staining of whole cleared organs or thick tissue slices (108, 109). The SpyTag and SpyCatcher are

both genetically encoded which allows for *in vivo* post-translational labeling something that is not possible with antibodies, SNAP-tag/CLIP-tag/Halo-tag or other labeling methods that rely on synthetic probes. Finally, we present the generation and validation of two SpyTagged, spectrally separate, channelrhodopsin molecules (SpyTag-C1C2 and SpyTag-ReaChR) which can be used for optogenetic experiments.

3.5 Experimental procedures

Ethics statement. All experiments using animals in this study were approved by Institutional Animal Care and Use Committee (IACUC) at the California Institute of Technology.

Generating constructs and site-saturation library. SpyTag/Catcher & SNAP-tag fusion constructs were generated through standard molecular biology cloning techniques. All constructs were verified by sequencing and reported in **Table S2**. Site-saturation libraries of the *pLenti-CMV/CaMKIIa::SpyTag-ReaChR-mCherry* parent were built using the 22c-trick method reported in (110) at position E130 and N298. Ten clones from each library we sequenced to test for library quality. DNA from individual clones was isolated and used to transfect HEK cells for further testing. For detailed methods see **Supplemental Methods**.

SpyCatcher production and labeling of HEK cells and primary neuronal cultures. Recombinant SpyCatcher for exogenous application was expressed and purified in bulk from *E. coli* strain *BL21(DE3)* harboring the *pQE80l-T5::6xhis-SpyCatcher-Elp-GFP* plasmid. Cells were grown at 37 ^{o}C in TB, expression was induced with 1 mM IPTG at 30 ^{o}C, and after 4 hours, cells were harvested. Protein purifiation was done on HiTrap columns (GE Healthcare, Inc.) following column manufactur's recommendations.

HEK cells and primary neuronal cultures were maintained and transfected using standard methods. For detailed methods see **Supplemental Methods**. Both HEK cells and neurons went through SpyCatcher labeling 48 hours post-transfection. Unless otherwise noted the SpyCatcher-GFP was added to the media of HEK cells at a final concentration of 25 μM and the cells were then incubated for 45 minutes – 1 hour at 25 ^{o}C. After labeling HEK cells were washed with D10 three to four times. Cells were then returned to incubate at 37 ^{o}C for 10 minutes to 1 hour before imaging. For more details on SpyCatcher labeling protocol for 96-well plate see **Supplemental Methods.** SpyCatcher labeling of neurons was carried out in 500 μl of the neuronal maintenance media in a 24-well plate. SpyCatcher-GFP was then added to each well of neurons for a final concentration of 25 μM. The neurons were then incubated with the SpyCatcher for 45 minutes – 1 hour at 37

^{o}C for labeling. After labeling cells were washed in Minimal Essential Media (MEM) three to four times. After washing the neurons were placed back into the stored neuronal maintenance media without SpyCatcher and incubated at 37 ^{o}C for 10 minutes to 1 hour before imaging.

C. elegans **experiments**. Transgenic *C. elegans* expressing each Tag-opsin construct were generated by DNA injection into *unc-119* mutant animals. A transgenic *C. elegans* line expressing heat-shock activated Catcher-GFP and cell-type specific expression of the tagged opsin was generated by co-injecting plasmid DNA of both constructs into *unc-119* mutant animals. To induce expression of Catcher-GFP *C. elegans* were heat-shock treated at 33°C for 15 minutes in a water bath. Following heat-shock, animals were allowed to recover at room temperature. At specific time points they were placed on an agar pad in 3 mM levamisole and imaged. For behavioral experiments transgenic animals expressing Tag-opsin constructs were grown on NGM plates with OP50 bacteria and all-trans retinal. L4-stage transgenic animals were placed on plates and grown in the dark for approximately 16 hours. To assay paralysis, animals were transferred individually onto plain NGM plates and their movement was monitored on a dissecting microscope (Leica) at 2.5x magnification for 10 s without green light, 5 s with green light illumination, and 10 s without green light. More details on generation and maintenance of SpyTag-C1C2-mCherry, SpyTag-ReaChR-mCherry, SpyTag-ReaChR(E130D)-mCherry, and SpyCatcher-GFP transgenic *C. elegans* strains, SpyCatcher-GFP staining of dissected *C. elegans* gonad, heat-shock treatment to induce SpyCatcher-GFP expression and locomotion assay evoked by green light can all be found in **Supplemental Methods.**

Electrophysiology. Conventional whole-cell patch-clamp recordings were done in cultured HEK cells and cultured rat hippocampal neurons at 2 days post transfection. For detailed methods see **Supplemental Methods**.

Fluorescence imaging and data analysis. Fluorescence analysis of single cells was done by manually selecting regions around each cell in ImageJ and fluorescence measurements were recorded for each region of interest (ROI). The same ROI was used for both the

mCherry and GFP fluorescence measurements in co-labeled cells. Fluorescence analysis and comparison between populations of cells expressing different opsin variants was done using a custom MATLAB script. For detailed methods see **Supplemental Methods**. Statistical methods- One-way ANOVA, unpaired student's *t*-tests and Dunnett's multiple comparison tests were performed using GraphPad Prism (version 6.04 for Windows, GraphPad Software, San Diego California USA, www.graphpad.com).

3.6 Figures

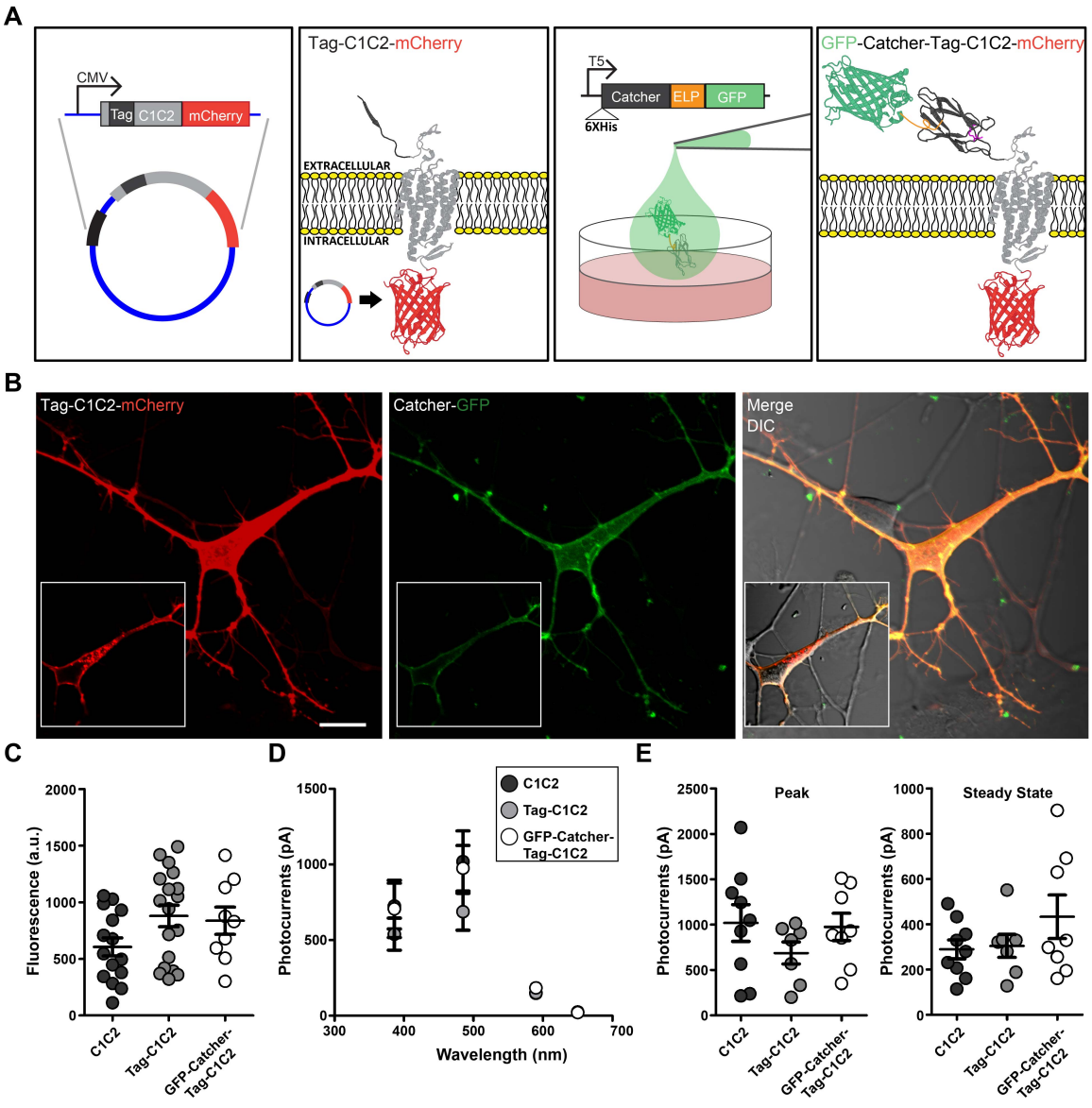


Figure 3.1. SpyTag fused to the N-terminus of C1C2 enables covalent binding of Catcher-GFP for membrane-localized Tag-C1C2 detection in live neurons without affecting light-induced currents.

(**A**) Construct design and labeling assay workflow. (left) Schematic of SpyTag fused to the N-terminus of C1C2-mCherry (Tag-C1C2-mCherry) under a CMV promoter for expression in mammalian cells. (middle left) Correctly folded Tag-C1C2-mCherry displays

the SpyTag extracellularly. (middle right) His-tagged SpyCatcher fused to a small elastin-like protein (ELP) and GFP (Catcher-GFP) with a T5 promoter for expression in *E. coli*. (right) Extracellular application of Catcher-GFP converts the membrane localized Tag-C1C2-mCherry to GFP-Catcher-Tag-C1C2-mCherry through formation of a covalent bond between the reactive lysine residue in SpyCatcher and the reactive aspartic acid residue in the surface-displayed SpyTag. (**B**) Maximum intensity projection of Tag-C1C2-mCherry expressing neurons (red), Catcher-GFP membrane-localized protein binding (green), and merge of red and green channels with DIC image of neuronal cells (inset: single plane confocal images of each) showing specific labeling of membrane-localized Tag-C1C2-mCherry. Only the cells expressing the Tag-C1C2-mCherry show binding of the Catcher-GFP. (**C**) Fluorescence measurements of mCherry in cultured neurons for C1C2-mCherry (N = 15), Tag-C1C2-mCherry (N = 18) and GFP-Catcher-Tag-C1C2-mCherry (N = 9) showing no significant difference. One-way ANOVA, P = 0.095. (**D**) Whole-cell recordings of peak photocurrents induced by different wavelengths in cultured neurons under voltage clamp. Neurons expressing C1C2-mCherry (N = 9), Tag-C1C2-mCherry (N = 7), and GFP-Catcher-Tag-C1C2-mCherry (N = 8) show similar spectral properties. (**E**) Peak and steady-state photocurrents induced by 480 nm light in cultured neurons under voltage clamp. Cells expressing C1C2-mCherry (N = 9), Tag-C1C2-mCherry (N = 7), and GFP-Catcher-Tag-C1C2-mCherry (N = 8) show no significant difference in peak or steady state currents. One-way ANOVA, peak currents: P = 0.4 and steady state currents: P = 0.3. All population data are plotted as mean ± SEM. Not significant (ns), $P > 0.05$. Scale bar, 10 μm.

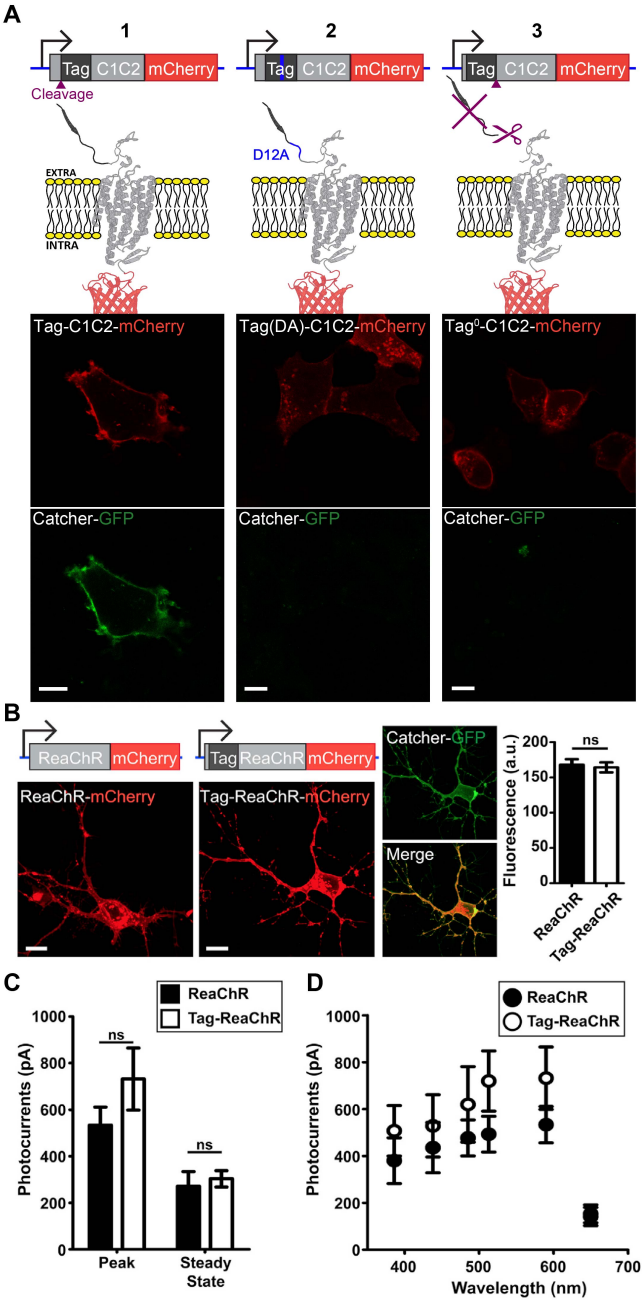


Figure 3.2. Opsin SpyTag fusion construct requirements for successful binding of SpyCatcher and application of the SpyTag/SpyCatcher to ReaChR.

(**A**) (top) Schematic of 3 different C1C2/SpyTag designs with corresponding labeling patterns (bottom). (1) SpyTag fused to the N-terminus of C1C2-mCherry after the signal peptide cleavage site results in expression of Tag-C1C2-mCherry with the SpyTag

displayed on the extracellular surface of the cell which successfully binds extracellularly applied Catcher-GFP. (2) SpyTag with the reactive aspartic acid (D12) residue mutated to alanine (A12) fused to the N-terminus of C1C2-mCherry after the signal peptide cleavage site results in expression of Tag(DA)-C1C2-mCherry. The mutated SpyTag does not bind to extracellular Catcher-GFP. (3) SpyTag fused to the N-terminus of C1C2-mCherry before the signal peptide cleavage site results in expression of C1C2-mCherry but no binding to extracellular Catcher-GFP. Single plane confocal images shown. (**B**) Maximum intensity projection of ReaChR-mCherry and Tag-ReaChR-mCherry expression in primary neuronal cultures under a CMV promoter. Application of Catcher-GFP to Tag-ReaChR-mCherry expressing neuron shows labeling. Fluorescence comparison of neurons expressing ReaChR-mCherry ($N = 6$) compared with neurons expressing Tag-ReaChR-mCherry ($N = 5$) shows no significant difference between the two opsin constructs (unpaired t-test, $P = 0.7$). (**C**) Whole-cell recordings of peak and steady-state photocurrents induced by 590 nm light under voltage clamp in neurons expressing ReaChR-mCherry ($N = 3$) and Tag-ReaChR-mCherry ($N = 5$) shows no significant difference (unpaired students *t*-test, peak: $P = 0.3$ and steady state: $P = 0.6$). (**D**) Peak photocurrents induced by different wavelengths of light under voltage clamp in neurons expressing ReaChR-mCherry ($N = 3$) and Tag-ReaChR-mCherry ($N = 5$). ReaChR-mCherry and Tag-ReaChR-mCherry show similar spectral properties. All population data are plotted as mean ± SEM. Not significant (ns), $P > 0.05$. Scale bar, 10 μm.

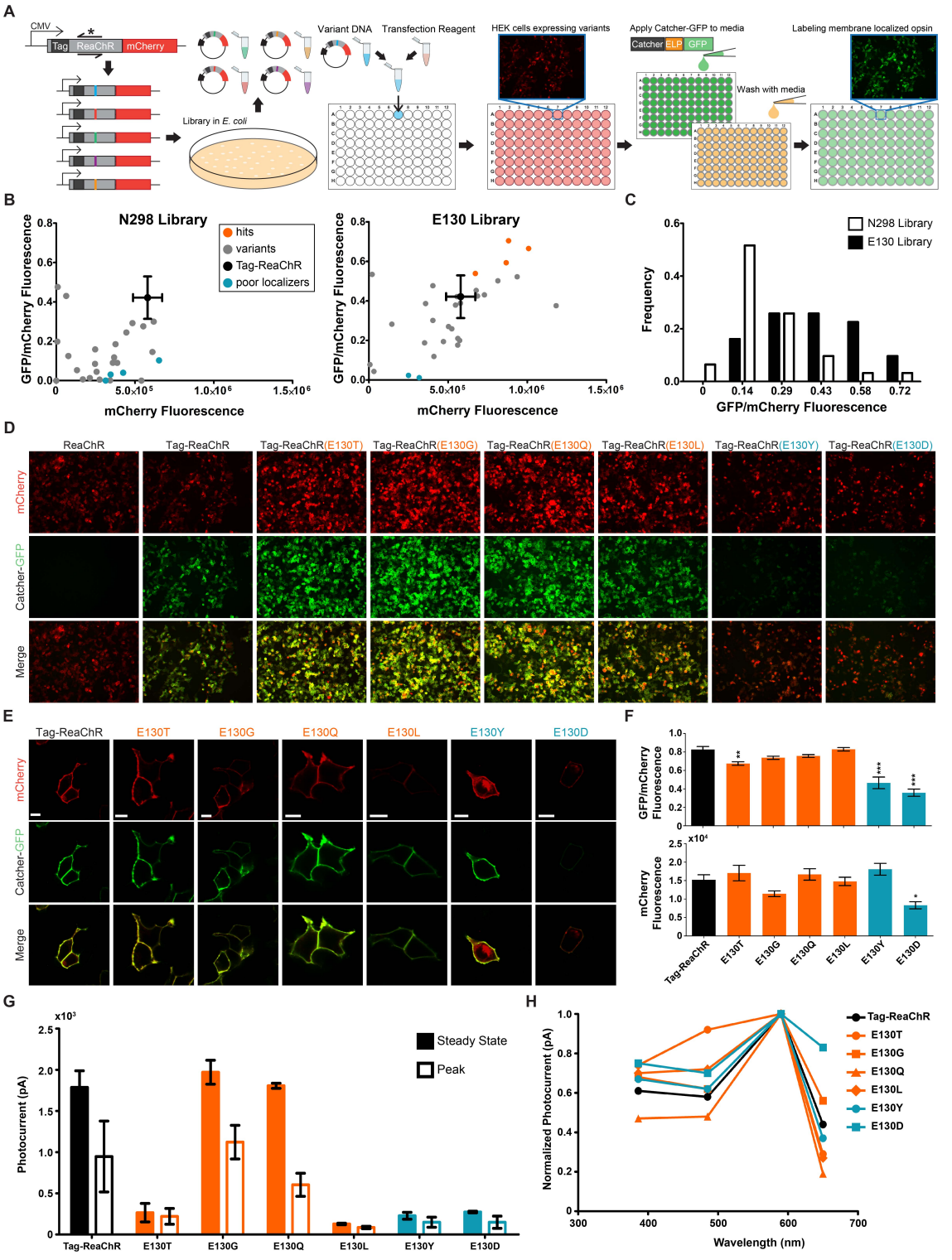


Figure 3.3. A screen for membrane localization based on SpyTag/SpyCatcher for optogenetics.

(**A**) Screening assay workflow. From left to right: Schematic of the SpyTag/SpyCatcher opsin membrane localization assay for screening in a 96-well format. Site-saturation mutagenesis of the *CMV::Tag-ReaChR-mCherry* backbone targeting specific amino acid locations. Transformation of the library into *E. coli*. Selection and isolation of plasmid DNA of individual clones. Transfection of HEK cells plated in a 96-well plate with each clone in a different well. Catcher-GFP is then added to each well, incubated for 1 hour and washed. Cells in each well are imaged for both mCherry fluorescence and GFP fluorescence. (**B**) GFP/mCherry fluorescence vs mCherry fluorescence for the two site-saturation libraries at amino acids N298 and E130 in ReaChR. Library 'variants' are shown in gray, 'hits' in orange, and 'poor localizers' in blue. The mean fluorescence with SEM of the Tag-ReaChR parent is shown in black (N = 4). (**C**) Distribution of GFP/mCherry fluorescence ratio for each of the two site-saturation libraries. (**D**) Example images from the screening process for non-tagged control (ReaChR), parent (Tag-ReaChR), Tag-ReaChR mutant 'hits' and Tag-ReaChR mutant 'poor localizers' from the E130 library. Full field, population images were taken for each tested variant and used to measure the GFP and mCherry fluorescence. Amino acid mutations at residue 130 are highlighted in orange for the 'hits' and in blue for the 'poor localizers' in the variants label. (**E**) Single plane confocal images of parent (Tag-ReaChR-mCherry) compared with the 'hits' and 'poor localizers' of mCherry (red), Catcher-GFP (green) and merge. (**F**) (top) GFP/mCherry fluorescence ratio or (bottom) mCherry fluorescence of Tag-ReaChR (N = 24) compared with ReaChR variants (E130T: N = 27, E130T: N = 72, E130Q: N = 43, E130L: N = 64, E130Y: N = 14, and E130D: N = 33) from single plane confocal images of HEK cells expressing the tagged opsins with intensity measurements made by selection of a region of interest around each cell and measurement of mean GFP and mCherry fluorescence across the region. Comparisons between Tag-ReaChR with each variants was done by Dunnett's Multiple Comparison Test. (**G**) Recordings of peak and steady-state photocurrents induced by 590 nm light under voltage clamp in HEK cells expressing Tag-ReaChR-mCherry (N = 6), each of the 'hits' (each variant, N = 3) and the 'poor localizers' (each variant, N = 3) from the E130 library. (**H**) Peak photocurrents induced by different wavelengths of light under voltage clamp in HEK expressing Tag-ReaChR-mCherry, each

of the ‘hits’ and the ‘poor localizers’ from the E130 library. Photocurrents are normalized to show spectral sensitivity. All population data are plotted as mean ± SEM. *P < 0.05, **P < 0.01, *** P < 0.001. Scale bar, 10 μm.

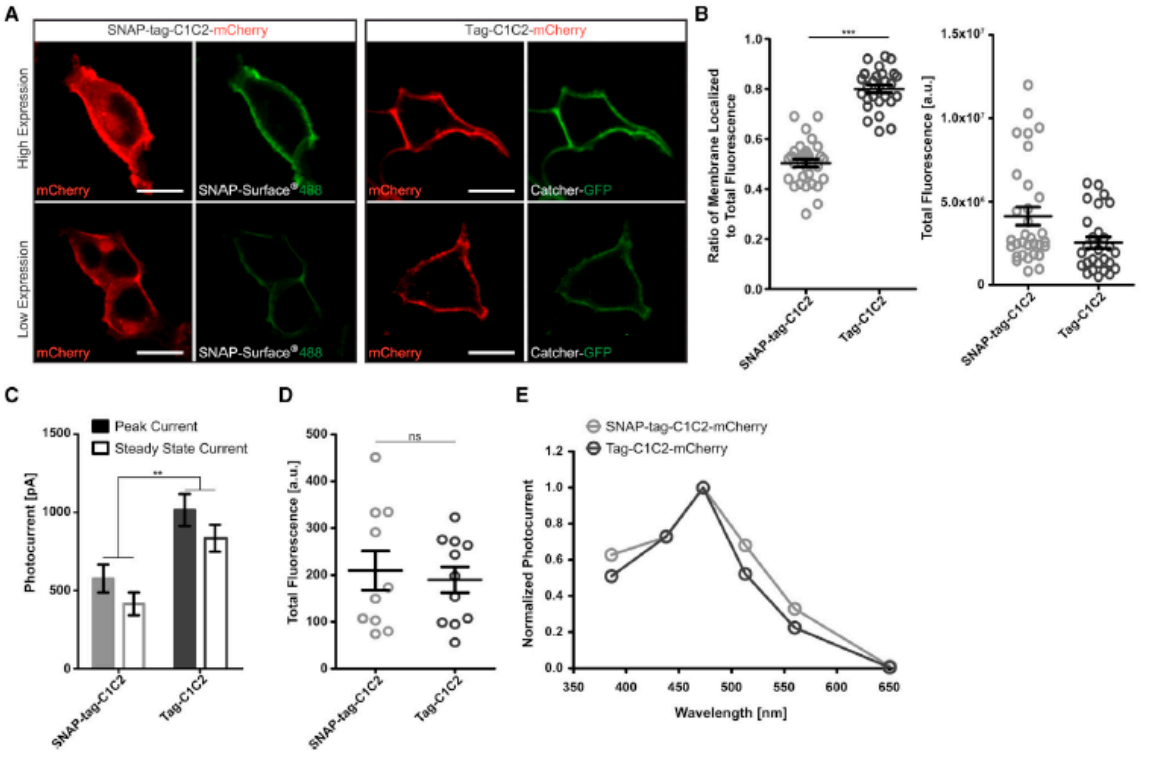


Figure 3.4. The N-terminal SpyTag opsin fusion construct (Tag-C1C2-mCherry) is able to express and traffic to the plasma membrane more efficiently than the N-terminal SNAP-tag opsin fusion construct (SNAP-tag-C1C2-mCherry) in mammalian cell culture.

(**A**) Fluorescence images of total opsin-mCherry expression (red) and successful labeling of membrane localized expression (green). Example cell with high expression (top) and low expression (bottom) comparing two different construct/labeling sets: SNAP-tag-C1C2-mCherry/SNAP-Surface®488 (left) and Tag-C1C2-mCherry/Catcher-GFP (right). (**B**) (left) Plot of the ratio of membrane localized fluorescence to total fluorescence of the SNAP-tag-C1C2-mCherry (N = 32 cells) vs Tag-C1C2-mCherry (N = 27 cells) expressing cells. The Tag-C1C2-mCherry construct shows a larger fraction of total expression localized to the plasma membrane while the SNAP-tag-C1C2-mCherry construct has a larger fraction of its total expression internally localized. There is a significant difference in the ratio of membrane-localized opsin between the two constructs. Unpaired t-test, $P < 0.0001$. (right) Plot of the total level of fluorescence of the SNAP-tag-C1C2-mCherry (N = 32 cells) vs

Tag-C1C2-mCherry (N = 27 cells) expressing cells. (**C**) Peak (filled bar) and steady-state (empty bar) photocurrents induced by 480 nm light in HEK cells under voltage clamp. Cells expressing SNAP-tag-C1C2-mCherry (N = 9), and Tag-C1C2-mCherry (N = 10) show a significant difference in peak and steady-state currents. Unpaired t-test, peak currents: P = 0.0053 and steady-state currents: P = 0.0019. (**D**) Total fluorescence measurements of mCherry in cultured HEK cells expressing either SNAP-tag-C1C2-mCherry (N = 10) or Tag-C1C2-mCherry (N = 11) used for whole-cell recordings show no significant difference. Unpaired t-test, P = 0.688. (**E**) Whole-cell recordings of peak photocurrents induced by different wavelengths in HEK cells under voltage clamp. HEK cells expressing SNAP-tag-C1C2-mCherry and Tag-C1C2-mCherry show similar spectral properties. All population data are plotted as mean ± SEM. Not significant (ns), $P > 0.05$. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$. Scale bar, 10 μm.

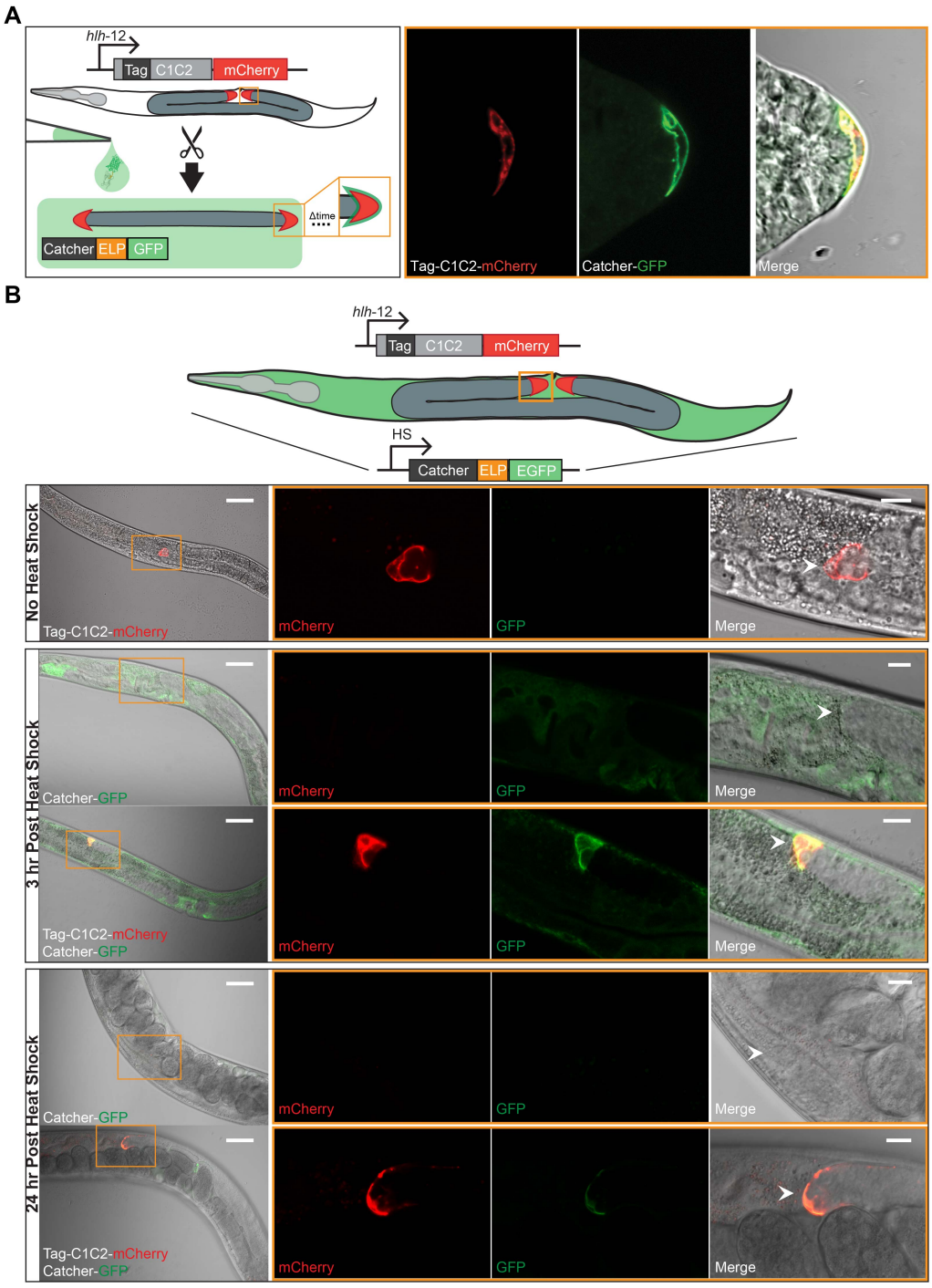


Figure 3.5. SpyTag fusion constructs shows efficient single-cell labeling with SpyCatcher in fixed and live *C. elegans*.

(**A**) (left) Schematic of Tag-C1C2-mCherry expression in the distal tip cells (DTCs) under the *hlh-12* promoter, dissection of the expressing *C. elegans* gonad and labeling of the dissected, fixed tissue with the Catcher-GFP. (right) Single plane confocal images of Tag-C1C2-mCherry expression in one DTC (red) with efficient labeling of Catcher-GFP (green) specific to the Tag-C1C2-mCherry expressing DTC. (**B**) (top) Schematic of transgenic *C. elegans* expressing Tag-C1C2-mCherry in the DTCs under the *hlh-12* promoter and Catcher-GFP under a heat-shock (HS) promoter. The *HS::Catcher-GFP* construct expresses Catcher-GFP in many tissue types upon HS treatment. Catcher-GFP is then secreted from cells into the body cavity. Single plane confocal images of a *C. elegans* expressing Tag-C1C2-mCherry in the DTC: without HS treatment show mCherry expression in the DTC without any Catcher-GFP expression and labeling; 3 hours post HS treatment shows mCherry expression in the DTC and significant Catcher-GFP expression throughout the body cavity with specific labeling of the Tag-C1C2-mCherry. While single plane confocal images of a *C. elegans* without Tag-C1C2-mCherry expression in the DTC 3 hours post HS treatment shows significant Catcher-GFP expression throughout the body cavity without specific labeling of the DTC, imaging 24 hours after HS shows decreased levels of GFP throughout the *C. elegans* while specific labeling of the DTC is achieved with Tag-C1C2-mCherry expression in the DTC. Scale bar, 20 μm.

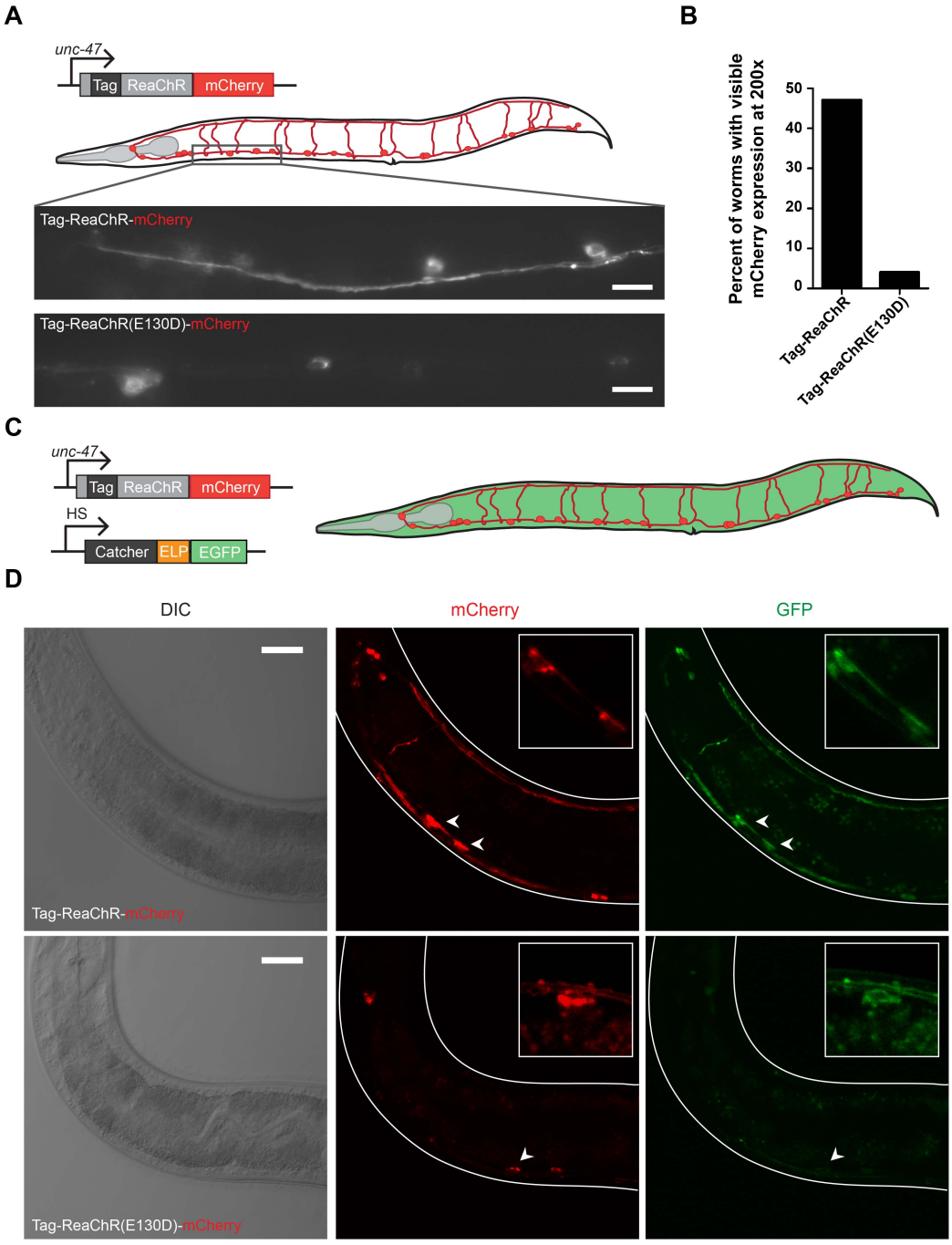


Figure 3.6. SpyTag opsin constructs expressed in GABA-producing neurons show efficient labeling with SpyCatcher in live *C. elegans* for both high expressing and low expressing SpyTag opsin constructs.

(**A**) (top) Schematic showing Tag-ReaChR-mCherry constructs expressed in the *C. elegans* 19 D-type GABA-producing neurons that reside in the ventral nerve cord and innervate dorsal and ventral body muscle. (bottom) Expression of both Tag-ReaChR-mCherry and Tag-ReaChR(E130D)-mCherry in cell bodies and fine processes of GABA-producing neurons in the ventral nerve cord. Scale bar, 10 μm. (**B**) Comparison of the expression levels of the Tag-ReaChR-mCherry and Tag-ReaChR(E130D)-mCherry constructs in *C. elegans* GABA-producing neurons characterized by mCherry visibility at 200x magnification. (**C**) Schematic showing both Tag-ReaChR-mCherry constructs expressed in the *C. elegans* 19 D-type GABA-producing neurons and Catcher-GFP expression and secretion from many tissue types post heat-shock. (**D**) Confocal images of (left) DIC, (middle) mCherry and (right) GFP for both Tag-ReaChR-mCherry and Tag-ReaChR(E130D)-mCherry constructs in *C. elegans* GABA-producing neurons 24 hr post heat-shock. Large images are maximum intensity projections of images that are power/gain matched for both constructs. Inset images show a single plane confocal image of individual cell(s) (indicated with arrow in large image). For the inset alone we increased the gain in low expresser for visibility. Scale bar, 20 μm.

3.7 Supplemental experimental procedures

3.7.1 SpyTag/SpyCatcher & SNAP-tag fusion constructs

The mammalian codon optimized *SpyTag* was first introduced into the N-terminus of *pLenti-CaMKIIa-C1C2-TS-EYFP* (**Supplementary Table 3.2**) after the signal peptide cleavage site (between amino acid position 23 and 24 in the C1C2 sequence) by overlap extension PCR using external primers plenti-CaMKII_F and plenti-CaMKII_R, and internal primers SpyTag_C1C2_F and SpyTag_C1C2_R (**Supplementary Table 3.3**). To generate the *pLenti-CMV/CaMKIIa::SpyTag-C1C2-TS-EYFP* (**Supplementary Table 3.1**) the assembly product was then inserted into the *BamHI*/*EcoRI* cut *pLenti-CaMKIIa-C1C2-TS-EYFP* vector (**Supplementary Table 3.2**). The *pLenti-CMV/CaMKIIa::SpyTag-C1C2-mCherry* construct (**Supplementary Table 3.1**) was built by first amplifying *SpyTag-C1C2-TS* from the *pLenti-CMV/CamMKIIa-SpyTag-C1C2-TS-EYFP* (**Supplementary Table 3.1**) construct using the plenti-CaMKII_F and TS_Rev primers (**Supplementary Table 3.3**) and amplifying the *TS-mCherry* from *pAAV-CaMKII-C1V1-TS-mCherry* (**Supplementary Table 3.2**) using TS_For and plenti-CaMKII_R (**Supplementary Table 3.3**). The fragments were then assembled using overlap extension PCR with plenti-CaMKII_F and plenti-CaMKII_R primers (**Supplementary Table 3.3**). The assembly product was then inserted into *BamHI*/*EcoRI* cut *pLenti-CaMKIIa-C1C2-TS-EYFP* (**Supplementary Table 3.2**) vector using Gibson assembly. A similar process was used to generate the *pLenti-CaMKIIa::C1C2-mCherry* construct only the initial amplification was done using the *pLenti-CamMKIIa::C1C2-TS-EYFP* backbone. Note that all vectors denoted as having a *CaMKII*, *CaMKIIa,* or *hSyn1* promoter also have an upstream *CMV* promoter. For the construct built for this work we have labeled the promoter as *CMV/CaMKIIa* since both promoters are present. The *CMV* promoter drives expression in transfections while the *CamKIIa* promoter would drive expression upon viral infection. These constructs can be used for both transfection of viral production and infection.

The mammalian codon optimized *SNAP-tag* sequence was first introduced into the N-terminus of *pLenti-CaMKIIa-C1C2-TS-mCherry* (**Supplementary Table 3.2**) after the

signal peptide cleavage site (between amino acid position 23 and 24 in the C1C2 sequence). The SNAP-tag sequence was amplified from $pSNAP_f$ vector (NEB, cat N9183S) with primers: C1C2-SNAP-NS-start and C1C2-SNAP-NS-end, and fused to C1C2 and mCherry with internal primers C1C2-NS-R and SNAP-CIC2-NS-mid, and external primers plenti-C1V1-3 and WPRE-R by overlap extension PCR method (**Supplementary Table 3.3**). To generate the *pLenti-CMV/CaMKIIa::SNAP-tag-C1C2-TS-mCherry* (**Supplementary Table 3.1**) the assembly product was then inserted into the *BamHI*/*EcoRI* cut *pLenti-CaMKIIa-C1C2-TS-mCherry* vector using Gibson assembly method (**Supplementary Table 3.2**).

Substitution of the aspartic acid, the reactive residue in the SpyTag, to the non-reactive alanine was done through mutation of the codon from *GAC* to *GCC*. This mutation was introduced through overlap extension PCR. The *SpyTag-C1C2-mCherry* was amplified into two separate fragments with the mutation introduced at the beginning of one fragment and the end of the other fragment using the C1C2_Spy_TagDA_F/plenti-CaMKII_R and plenti-CaMKII_F/C1C2_Spy_TagDA_R primer pairs (**Supplementary Table 3.3**). These fragments were assembled though PCR, digested with *BamHI*/*EcoRI*, and then ligated into the *BamHI*/*EcoRI* cut *pLenti-CaMKIIa-C1C2-TS-EYFP* vector (**Supplementary Table 3.3**) to produce the *pLenti-CMV/CaMKIIa::SpyTag(DA)-C1C1-TS-mCherry* construct.

To generate SpyTag-ChR2-mCherry variants *ChR2-mCherry* was amplified from *pLenti-CaMKIIa-hChR2(H134R)-mCherry-WPRE* (**Supplementary Table 3.2**) using pGP_Gib_ChR2_F and pGP_Gib_ChR2_R primers (**Supplementary Table 3.3**). Gibson assembly method was then used to insert the *ChR2-mCherry* amplification product into the *pGP-CMV-GCaMP6f* vector (**Supplementary Table 3.2**) cut with *BglII/XbaI*. This produced the *pGP-CMV::ChR2-mCherry* construct that was then used for all ChR2, SpyTag fusions. The C-terminal fusion of *SpyTag* to *ChR2-mCherry* (*pGP-CMV::ChR2-mCherry-SpyTag*) was generated by overlap-extension PCR by first amplifying the *ChR2-mCherry-SpyTag* in two parts with ChR2_SpyTag_F/ pGP-Gib_ R and pGP-Gib_ F/ ChR2_SpyTag_R primer pairs. The two amplified fragments were then assembled using ChR2_SpyTag_F/ ChR2_SpyTag_R primers. The assembly product was inserted into the

pGP-CMV-GCaMP6f vector (**Supplementary Table 3.2**) cut with *BglII/XbaI*. Stepping of the SpyTag at the N-terminal end of ChR2 was done using the same method using different SpyTag insertion primers labeled as SpyTag_ChR2_# based on the position described in **Supplementary Figure 3.4**.

ReaChR rhodopsin was fused to the mCherry reporter after a three-alanine residue linker, and a trafficking signal (TS) KSRITSEGEYIPLDQIDINV (43). The *ReaChR* gene was amplified from *AAV-EFIa-ReaChR-mCitrine-FLEX* vector (**Supplementary Table 3.2**) using ReaChR_fwd and ReaChR_rev primers (**Supplementary Table 3.3**). The *3xA-linker-TS-mCherry* was amplified from *pLenti-CAMKIIa-C1C2-TS-mCherry* (**Supplementary Table 3.1**) plasmid using pA_TS_mcherry_fwd and WPRE_rev primers (**Supplementary Table 3.3**). *ReaChR-TS-mCherry* was assembled using overlapping assembly PCR and digested with *EcoRI* and *BamHI*. Digested insert was ligated into an *EcoRI*/*BamHI* digested Lentiviral vector containing the *CMV*/*CaMKIIa* promoters and WPRE to obtain the *pLenti-CMV/CaMKIIa::ReaChR-TS-mCherry-WPRE* clone (**Supplementary Table 3.1**). *SpyTag* was inserted at the N-terminus of *ReaChR* after the signal peptide cleavage site (24 amino acids from the N-terminus) using overlap extension PCR with primers Spy_ReaChR_fwd, Spy_ReaChR_rev, WPRE_rev and CAMKIIa_fwd (**Supplementary Table 3.3**). Digestion and ligation of the assembled product into the template lentiviral vector produced the *pLenti-CMV/CaMKIIa::SpyTag-ReaChR-TS-mCherry-WPRE* clone (**Supplementary Table 3.1**).

The *pLenti-CMV/hSyn1::TrkB-mCherry-WPRE* vector (**Supplementary Table 3.1**) was built by Gibson assembly method. A lentiviral vector containing *human synapsin I* (*hSyn1*) promoter and *WPRE*, *pLenti-hSyn-eNpHR 3.0-EYFP* (**Supplementary Table 3.2**), was digested with *AgeI* and *EcoRI* enzymes and used as backbone for all TrkB constructs. *TrkB* was synthesized from GenScript USA Inc, fused with *EYFP* reporter and inserted into this lentiviral vector to build *pLenti-CMV/hSyn1::TrkB-EYFP-WPRE* vector. To replace the *EYFP* marker with *mCherry* the *mCherry* gene from *pLenti-CaMKIIa-ReaChR-TS-mCherry-WPRE* vector was amplified with 3xGS_mcherry_fwd and mcherry_rev primers (**Supplementary Table 3.3**). Assembly product of the *TrkB*, *mCherry* fusion was

generated using TrkB_fwd and 3xGS_TrkB_rev primers, and then inserted into the digested lentiviral vector. *pLenti-CMV/hSyn1*::SpyTag-TrkB-mCherry-WPRE* (**Supplementary Table 3.1**) was built by inserting the *SpyTag-GGSG* linker at the N-terminus of *TrkB* after the signal peptide cleavage site (31 amino acids from the N-terminus) using the overlapping primers spy_trkB_rev and spy_trkB_fwd, and assembled with end primers hsyn_fwd and mcherry_rev primers. This was then inserted into the template lentiviral vector containing *hSyn1* promoter at sites *AgeI* and *EcoRI* using Gibson assembly method.

To generate *C. elegans* expression plasmid *hlh-12::SpyTag-C1C2-mCherry*, 1.2 kb of the *hlh-12* 5' region was PCR amplified from genomic DNA using primers Worm1 and Worm2, and cloned into pPD49.26 (Fire vector) using *PstI* and *BamHI* restriction sites. Then, the *SpyTag-C1C2-mCherry* sequence was PCR amplified from plasmid *pLenti-CaMKIIa::SpyTag-C1C2-mCherry* (**Supplementary Table 3.1**) using primers Worm3 and Worm4, and was cloned into ppD49.26 *hlh-12* vector using the *KpnI* restriction site. Plasmid pSM*::unc-47::SpyTag-ReaChR-TS-mCherry* was generated by first PCR amplifying *SpyTag-ReaChR-TS-mCherry* from *pLenti-CaMKIIa::SpyTag-ReaChR-TS-mCherry* using primers Worm13 and Worm14 and inserting the PCR product into vector pSM::GFP (gift from Cori Bargmann) using *KpnI* and *EcoRI* restriction sites. pSM*::unc-47::SpyTag-ReaChR(E130D)-TS-mCherry* was constructed in the same way only by PCR amplifying *SpyTag-ReaChR(E130D)-TS-mCherry* from *pLenti-CaMKIIa::SpyTag-ReaChR(E130D)-TS-mCherry*. 1.5 kb of the *unc-47* 5' region was PCR amplified from genomic DNA using primers Worm15 and Worm16, and cloned into pSM::*SpyTag-ReaChR-TS-mCherry* using *FseI* and *AscI* restriction sites (**Supplementary Table 3.1).**

PCR fusion product *HS::lin-3 signal sequence::SpyCatcher-GFP* was constructed by PCR fusion of PCR products from a *HS::lin-3* plasmid and *SpyCatcher-GFP* plasmid. To generate *HS::lin-3* plasmid, a partial *lin-3* (*C. elegans* EGF) sequence was PCR amplified from genomic DNA using primers Worm5 and Worm6, and inserted into plasmid *pPD49.83* (Fire vector containing HS promoter) using the *KpnI* restriction site. *SpyCatcher-GFP* plasmid was generated by amplifying *SpyCatcher-GFP* sequence from

plasmid *pQE80l-T5::6xhis-SpyCatcher-Elp-GFP* using primers Worm7 and Worm8 and inserting into plasmid pPD49.83 using the *KpnI* restriction site. *HS::lin-3 signal sequence* was PCR amplified from the *HS::lin-3* plasmid using primers Worm9 and Worm10, and *SpyCatcher-GFP::unc-54 3' UTR* sequence was amplified from *SpyCatcher-GFP* plasmid using primers Worm11 (containing overlap to the *lin-3* signal sequence) and Worm12. *HS::lin-3 signal sequence* and *SpyCatcher-GFP::unc-54 3' UTR* PCR products were fused through a second PCR reaction using both products as templates and primers Worm9 and Worm12.

Plasmid *pSM-lin-3 signal sequence::SpyCatcher-GFP::unc-54 3'UTR* was generated by adding on the *lin-3 signal sequence* to *SpyCatcher-GFP::unc-54 3'UTR* through two PCR reactions, first using primers Worm17 and Worm19 and *SpyCatcher-GFP* plasmid as template. The PCR product was then amplified using primers Worm18 and Worm19. The product was cloned into *pSM-GFP* vector using the *KpnI* and *EcoRI* restriction sites. *HS* sequence was PCR amplified from *pPD49.83* using primers Worm20 and Worm21 and the product was cloned into *pSM-lin-3 signal sequence::SpyCatcher-GFP::unc-54 3'UTR* using *FseI* and *AscI* restriction sites to generate the final plasmid, *pSM-HS::lin-3 signal sequence::SpyCatcher-GFP::unc-54 3'UTR.*

To build the *pQE80l-T5::6xhis-SpyCatcher-Elp-GFP* construct new restriction sites were introduced for our system (for details see Sun *et al*.). The original restriction sites following His6 tag in pQE-80l were removed. This SpyCatcher-GFP construct was derived from SpyCatcher-Elp-SpyCatcher (pQE-BB) described in Sun *et al.* (111). The GFP gene with a TAA stop codon was inserted between *SacI* and *SpeI* sites to generate the final construct.

3.7.2 Generating site-saturation library from the Tag-ReaChR-mCherry parent

Primers designs are listed in **Supplementary Table 3.3** with degenerate residues highlighted in yellow. Mutations were introduced by overlap extension PCR of the *pLenti-CaMKIIa::SpyTag-ReaChR-TS-mCherry* parent backbone with external primers plenti-CaMKII_F/ plenti-CaMKII_R used for amplification and assembly. Assembly product was then digested with *EcoRI/BamHI* and ligated into *EcoRI/BamHI* cut *CaMKIIa::SpyTag-*

ReaChR-TS-mCherry vector. Each library was then transformed into *E. coli*, single colonies were picked and 2-5 ml cultures were grown for each variant. DNA for each variant was purified and the concentration of DNA for each variant was normalized to 100 ng/ul for transfection into HEK cells.

3.7.3 HEK cell maintenance and transfection

HEK 293F cell were cultured at 37 ^{o}C and 5% CO2 in D10 (Dulbecco's modified Eagle medium (DMEM) supplemented with 10% (vol/vol) FBS, 1% sodium bicarbonate, 1% sodium pyruvate, and penicillin-streptomycin). For low throughput confocal imaging constructs were transfected with Fugene6 into HEK cells according to the manufacturer's protocol plated at a density of 5,000 cells per cm^2 onto 12 mm- PolyDLysine coated coverslips at 18 hours post-transfection. The HEK cells were then left to adhere to coverslips and continue to express for another 30 hours (so total expression for 48 hours post transfection) before labeling with SpyCatcher and imaging. For the 96-well format screening HEK cells were seeded at low density in tissue culture treated 96-well plates (BD Falcon Microtest^{TM} 96). Cells were left to divide until they reached ~20-30% confluency. Library variants were transfected with Fugene6 into HEK cells according to the manufacturer's recommendations with one variant per well (with pre-normalized DNA concentration of each variant). Cells were then labeled with SpyCatcher 48 hours post transfection and imaged.

3.7.4 Preparation and transfection of primary neuronal cultures

Rat hippocampal cells were dissected from Wistar embryos (prenatal days E18, Charles-River Labs), and cultured at 37oC, 5% CO2 in Neurobasal media supplemented with B27, glutamine, and 2.5% FBS. 3 days after plating, glial growth was inhibited by addition of FUDR. Cells were transfected 4-5 days after plating with SpyTag-opsin variants using calcium chloride. Neurons were labeled with SpyCatcher and imaged 2-5 days after transfection.

3.7.5 SpyCatcher labeling of HEK cells in 96-well format

SpyCatcher labeling protocol for 96-well plate. To avoid any variability in labeling in the 96-well format screen a saturating concentration of the SpyCatcher (30 μM) was used for labeling experiments. A 75 μM SpyCatcher stock was made and 20 μl of the stock was added to 30 μl of D10 in each well for a final concentration of 30 μM SpyCatcher per well. The cells were then incubated with the labeling protein at room temperature for 45 minutes. After the labeling the cells were washed. To avoid complete removal of media from the cells 200 μl of fresh D10 was added to each well to dilute the SpyCatcher concentration and then 200 μl was removed from each well. This washing/dilution was repeated four times. After washing the 96-well plates of cells were returned to a 37 ^{o}C incubator and left for 30 minutes before imaging. For imaging of cells in each well the media was replaced with extracellular buffer (in mM: 140 NaCl, 5 KCl, 10 HEPES, 2 $MgCl_2$, 2 $CaCl_2$, 10 glucose; pH 7.35) to avoid the high autofluorecence of the D10.

3.7.6 SNAP labeling of HEK cells

SNAP-Surface® 488 was purchased from NEB (cat S9124S). Labeling of live HEK cells transfected with *pLenti-CaMKII::SNAP-tag-C1C2-TS-mCherry* was done following manufacturer's instructions for cellular labeling. In brief, the SNAP-Surface® 488 reconstituted in DMSO to make a 1 mM stock solution. The stock solution was then diluted 1:200 in D10 media to yield a labeling medium of 5 uM dye substrate. The SNAP-tag-C1C2-mCherry expressing HEK cells were then incubated in the labeling medium for 30 min at 37^{o}C. After labeling the cells were washed 3-4x with D10 media before confocal imaging.

3.7.7 Generating and maintaining SpyTag and SpyCatcher transgenic C. elegans strains

C. elegans strains were cultured at room temperature using standard protocols unless indicated otherwise (112). Strains used in this study were *him-5*(*e1490*) (113) and *unc-119(ed4)* (Maduro and Pilgrim, 1995). Transgenic *C. elegans* expressing Tag-C1C2-mCherry was generated by co-injecting plasmid *hlh-12::SpyTag-C1C2-mCherry* (14 ng), *unc-119* rescue plasmid (60 ng), and 1kb ladder carrier DNA (50ng) into *unc-119* mutant animals. A transgenic *C. elegans* line expressing heat-shock activated Catcher-GFP and

specific expression of Tag-C1C2-mCherry in DTCs was generated by co-injecting plasmid *hlh-12::SpyTag-C1C2-mCherry* (14 ng), PCR fusion product *HS::lin-3 signal sequence::SpyCatcher-GFP* (40 ng), 1kb ladder carrier DNA (50 ng), and *unc-119* rescue plasmid (60 ng), into *unc-119* mutant animals. Transgenic animals expressing heat-shock activated Catcher-GFP and either wild-type or mutant *SpyTag-ReaChR-TS-mCherry* in GABA neurons was generated by co-injecting plasmid *unc-47::SpyTag-ReaChR-TS-mCherry* (wild-type or mutant 90 ng), plasmid *HS::lin-3 signal sequence::SpyCatcher-GFP* (50 ng), 1kb ladder carrier DNA (50 ng), and *unc-119* rescue plasmid (60 ng), into *unc-119* mutant animals.

3.7.8 SpyCatcher staining of dissected C. elegans gonad

To extrude gonads from animals, hermaphrodites were placed in 6 mL of PBS (phosphate buffered saline) on a Superfrost plus microscope slide (Fisher Scientific) and cut below the pharynx with a razor blade as described previously (Chan and Meyer, 2006). 6 mL of 4% p-formaldehyde solution was added, sandwiched with a coverslip, and fixed for 10 minutes. The entire slide was then submersed in liquid nitrogen for a few minutes, and immediately upon removal, the coverslip was removed and the slide was washed with PBS three times. 30 mL of purified Catcher-GFP in PBS solution (20 μM) was applied to the fixed gonads on the slide and incubated for 30 minutes at room temperature. The slide was washed 3x5 minutes with PBS and imaged after mounting with Vectashield mounting media (Vector Laboratories).

3.7.9 Heat-shock treatment to induce SpyCatcher expression

C. elegans strain carrying transgenes *hlh-12::SpyTag-C1C2-mCherry* and *HS::lin-3 signal sequence::SpyCatcher-GFP* was heat-shock treated at 33°C for 15 minutes in a water bath. *C. elegans* strain carrying transgenes *HS::lin-3 signal sequence::SpyCatcher-GFP and either wild-type or mutant unc-47::SpyTag-ReaChR-TS-mCherry* were heat-shock treated at 33°C for 30 minutes. Following heat-shock, animals were allowed to recover at room temperature. At specific time points they were placed on an agar pad in 3 mM levamisole and imaged.

3.7.10 Locomotion assay evoked by green light

Animals expressing either wild-type or mutant *unc-47::SpyTag-ReaChR-TS-mCherry* were grown on NGM plates with OP50 bacteria and all-trans retinal. 150mL of OP50 culture alone or with 100mM all trans-retinal (0.15mL of 100mM stock in ethanol; Sigma-Aldrich) was added to NGM plates and dried for several hours in the dark. L4-stage transgenic animals were placed on plates and grown in the dark for approximately 16 hours. To assay paralysis, animals were transferred individually onto plain NGM plates and their movement was monitored on a dissecting microscope (Leica) at 2.5x magnification for 10 s without green light, 5 s with green light illumination, and 10 s without green light. Green light (650±13 nm) was generated using LED illumination using a Lumencor SPECTRAX light engine at a power of 1 mW. White light illumination, which was constant throughout the experiment, was filtered to remove blue/green light. Paralysis upon illumination was scored as a positive.

3.7.11 Fluorescence Imaging

For non-confocal imaging of cultured neurons expressing different opsin variants a Zeiss Axio Examiner.D1 microscope with a 20x 1.0 NA water immersion objective (Zeiss W Plan Apochromat 20x/1.0 DIC D=0.17 M27 75mm) was used. Images of neurons were taken before electrophysiological recordings and the images we analyzed for fluorescence level comparison between variants. Imaging of the mCherry fusion fluorescence was excited with 650±13 nm, and imaging of the GFP label fluorescence was excited with 485±20 nm. Both wavelengths of light were generated with LED illumination using a Lumencor SPECTRAX light engine with quad band 387/485/559/649 nm excitation filter, quad band 410/504/582/669 nm dichroic mirror, and quad band 440/521/607/700 nm emission filter (all SEMROCK).

Confocal imaging was preformed on a Zeiss LSM 780 Confocal Microscope. Imaging of live cultured HEK cells and neurons was preformed with a Zeiss W Plan-APOCHROMAT 20x/1.0 DIC(UV) Vis-IR objective. Imaging of live *C. elegans* was preformed using a Zeiss LD LCI Plan-APOCHROMAT 25x/0.8 Imm Korr DIC M27 objective. GFP

fluorescence was excited with a 488 nm laser and mCherry fluorescence was excited with a 561 nm laser. Fluorescence emission was imaged using the LSM 780's GaAsP detectors with a detection range of 499-606 nm for GFP and 578-695 nm mCherry. Imaging was done with excitation and emission measurements of GFP and mCherry done on separate tracks to avoid crossover. Imaging settings were matched across experiments to enable comparison.

Full population images of cells in 96-well plates were taken with a Leica DM IRB microscope and the Leica microsystems objective HC PL FL 10x/0.30 PH1. Cells were illuminated with LEJ ebq 50 ac mercury lamp. GFP fluorescence was imaged with SEMROCK Blue light filter set: SEMROCK BrightLine® single-band filter set with BrightLine® single-band bandpass excitation filter (482/18 nm), emission filter (520/28) and 495 nm edge BrightLine® single-edge dichroic beamsplitter. mCherry fluorescence was imaged with Leica's N2.1 filter cube with bandpass excitation filter (515-560 nm), longpass suppression filter (590 nm) and dichromatic mirror (580).

3.7.12 Electrophysiology

Conventional whole-cell patch-clamp recordings were done in cultured HEK cells and cultured rat hippocampal neurons at least 2 days post transfection. Cells were continuously perfused with extracellular solution at room temperature (in mM: 140 NaCl, 5 KCl, 10 HEPES, 2 $MgCl_2$, 2 $CaCl_2$, 10 glucose; pH 7.35) while mounted on the microscope stage. Patch pipettes were fabricated from borosilicate capillary glass tubing (1B150-4; World Precision Instruments, Inc., Sarasota. FL) using a model P-2000 laser puller (Sutter Instruments) to resistances of 2-5 MΩ. Pipettes were filled with intracellular solution (in mM): 134 K gluconate, 5 EGTA, 10 HEPES, 2 $MgCl_2$, 0.5 $CaCl_2$, 3 ATP, 0.2 GTP. Whole-cell patch-clamp recordings were made using a Multiclamp 700B amplifier (Molecular Devices, Sunnyvale, CA), a Digidata 1440 digitizer (Molecular Devices), and a PC running pClamp (version 10.4) software (Molecular Devices) to generate current injection waveforms and to record voltage and current traces.

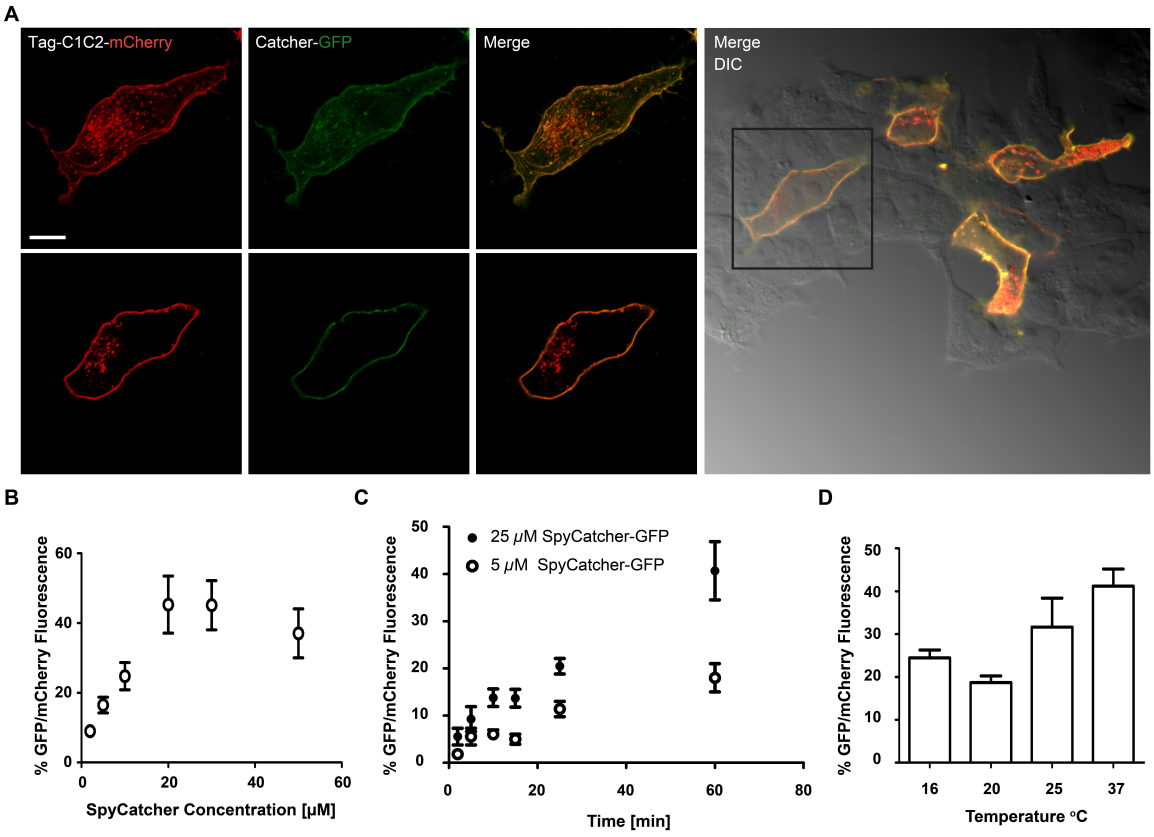
Patch-clamp recordings were done with short light pulses to measure photocurrents. Photocurrents induced by green light were measured using 590±25 nm LED at 1 mW mm^{-2} for ReaChR-mCherry and Tag-ReaChR-mCherry variants. While C1C2-mCherry and Tag-C1C2-mCherry variant's photocurrents were induced by cyan light were measured using 485±20 nm LED at 0.3 mW mm^{-2}. Photocurrents were recorded from cells in voltage clamp held at -50 mV with 3-10 light pulse trains (0.5 s each pulse; 2 s apart). Both wavelengths were produced using LED illumination using a Lumencor SPECTRAX light engine with quad band 387/485/559/649 nm excitation filter, quad band 410/504/582/669 nm dichroic mirror and quad band 440/521/607/700 nm emission filter (all SEMROCK).

Action spectra measurements were performed for the following wavelengths: 386±23 nm, 485±20 nm, 590±25 nm, and 650±13 nm with light intensity matched across all experiments at 0.1 mW mm^{-2}. Each light pulse was delivered for 0.6 s with 10 s breaks between light pulses. All wavelengths were produced using LED illumination from a SPECTRAX light engine (Lumencor). Cell health was monitored through holding current and input resistance.

3.7.13 Data analysis

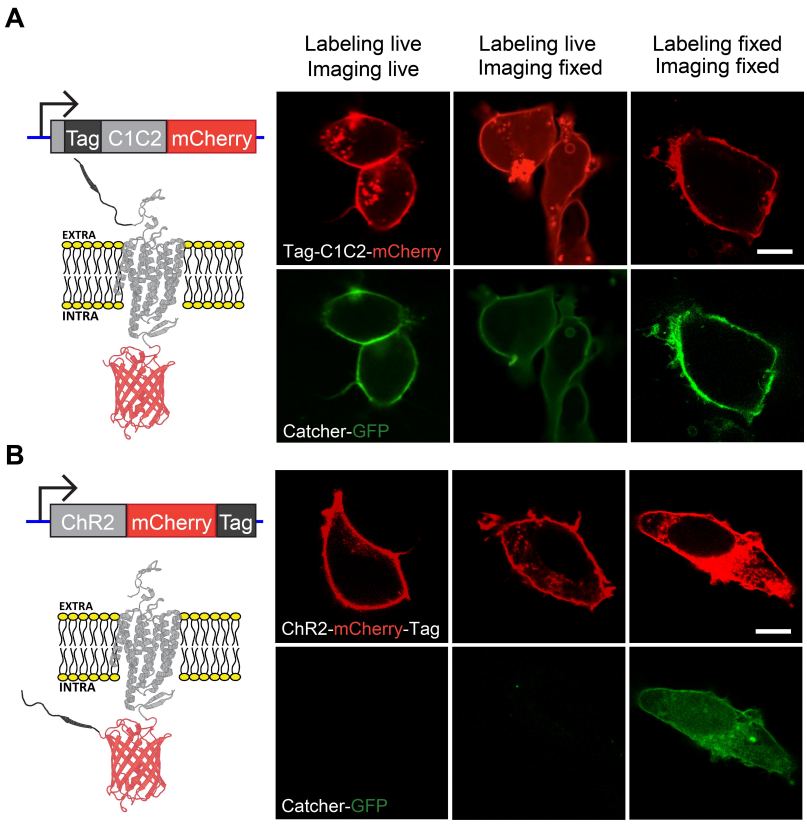
A MATLAB script was written to compare area above a threshold level of fluorescence in a population of cells. This was done for both mCherry fluorescence and GFP fluorescence. The mCherry-above-threshold-area was then used to normalize the GFP-above-threshold-area so that the density of cells within the image was not a confounding factor. The ratio of GFP-above-threshold-area to mCherry-above-threshold-area was the metric used to compare across the libraries reported in **Figure 3.3**.

3.8 Supplementary figures and tables

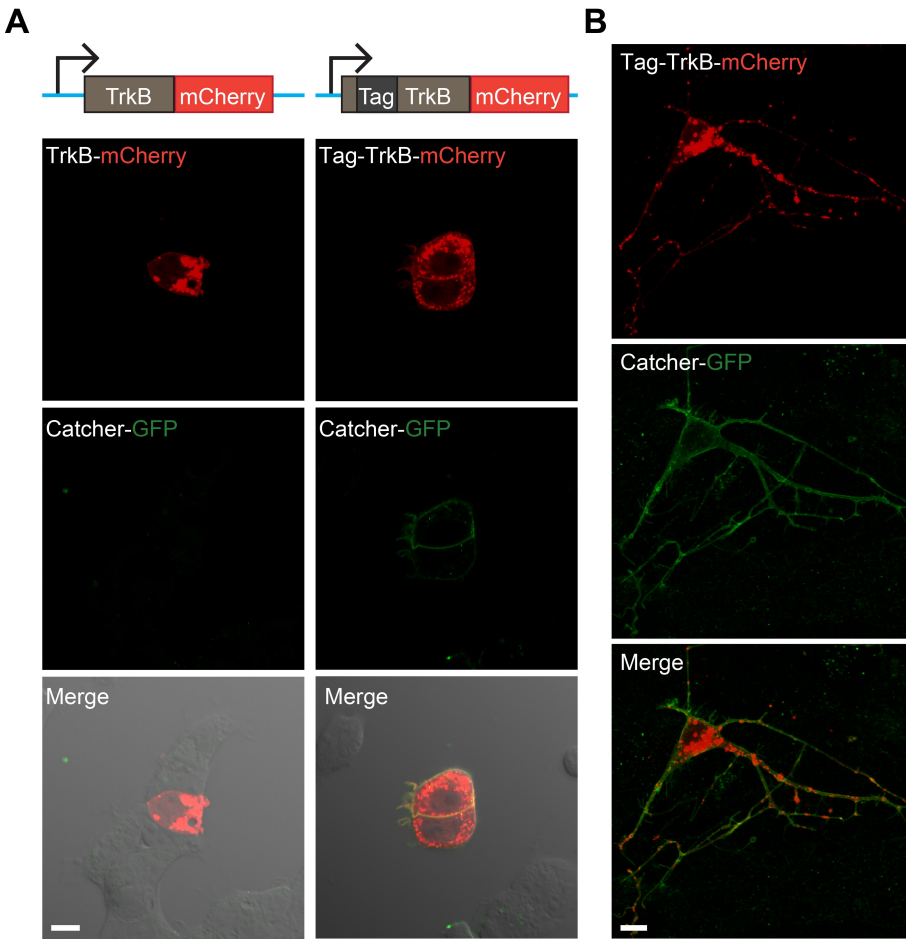


Supplementary Figure 3.1. Catcher-GFP labeling of membrane-localized Tag-C1C2-mCherry in live HEK cells and optimization of SpyTag/SpyCatcher binding efficiency in complex media used for mammalian cell cultures. (**A**) Top row: (left) Maximum intensity projection of Tag-C1C2-mCherry expression in HEK cells (red), (middle left) Catcher-GFP membrane localized protein binding (green) and (middle right) a merge. Bottom row shows single plane confocal images of cell in each channel. (right) Single plane confocal image of a population of HEK cells with only a fraction of cells expressing Tag-C1C2-mCherry. Black box indicates cell shown to the left. Only the cells expressing the Tag-C1C2-mCherry show binding of the Catcher-GFP. (**B**) Effect of different concentrations of extracellular Catcher-GFP. Plot shows quantification of GFP fluorescence relative to mCherry fluorescence of individual labeled Tag-C1C2-mCherry expressing cells. Fluorescence measurements were obtained from single plane confocal images of Catcher-GFP bound to membrane-localized Tag-C1C2-mCherry after treatment

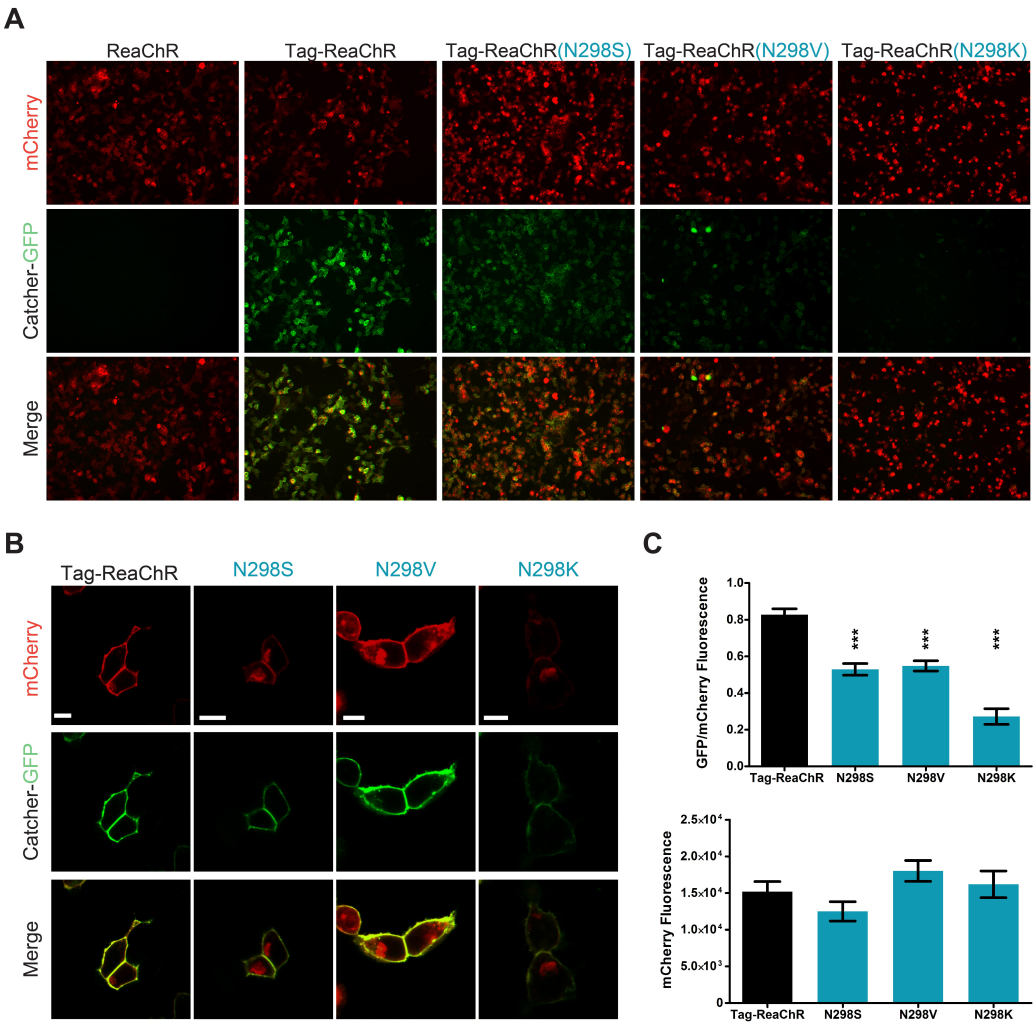
of HEK cells with Catcher-GFP for 1 hour in D10 medium (N = 12-14 cells for each concentration). (**C**) Testing different incubation times from 2-60 min at 25^{o}C. Plot shows the percent of GFP fluorescence relative to mCherry fluorescence of individual Tag-C1C2-mCherry expressing cells covalently bound to Catcher-GFP after incubation of Tag-C1C2-mCherry expressing cells with either 5 μM (empty circles) or 25 μM (filled circles) Catcher-GFP (N = 11-14 for each time point). (**D**) Effect of temperature from 16 to 37 ^{o}C. Plot of the percent of GFP fluorescence relative to mCherry fluorescence of individual Tag-C1C2-mCherry expressing cells bound to Catcher-GFP after incubation of Tag-C1C2-mCherry expressing cells with 25 uM Catcher-GFP for 1 hour (N = 9-14 for each temp). All population data are plotted as mean ± SEM.



Supplementary Figure 3.2. SpyTag/SpyCatcher system works with both live and fixed cultured cells and can be used to identify the signal peptide of ChR2 and its positioning can affect ChR2 membrane localization. (**A**) (left) Schematic of N-terminal SpyTagged opsin construct (Tag-C1C2-mCherry) in the cell membrane with the SpyTag displayed on the extracellular surface. (**B**) (left) Schematic of C-terminal SpyTagged opsin construct (ChR2-mCherry-Tag) in the cell membrane with the SpyTag displayed on the intracellular side of the cell. (**A**) & (**B**) (right) Single plane confocal images of the two opsin constructs with varying labeling and fixation methods. Column 1: both constructs show expression of the tagged ChR-mCherry. With extracellular application of Catcher-GFP only the N-terminal SpyTag shows Catcher-GFP binding since the Catcher-GFP cannot penetrate the membrane to label the C-terminal SpyTag. Column 2: fixation in paraformaldehyde (PFA) has no effect of the membrane-localized tagging after covalent binding of the Catcher-GFP. Column 3: fixation with PFA permeabilizes the cells allowing Catcher-GFP to get through the membrane and then covalently bind total ChR-mCherry for both the N-terminal and C-terminal SpyTagged constructs. Scale bar, 10 μm.

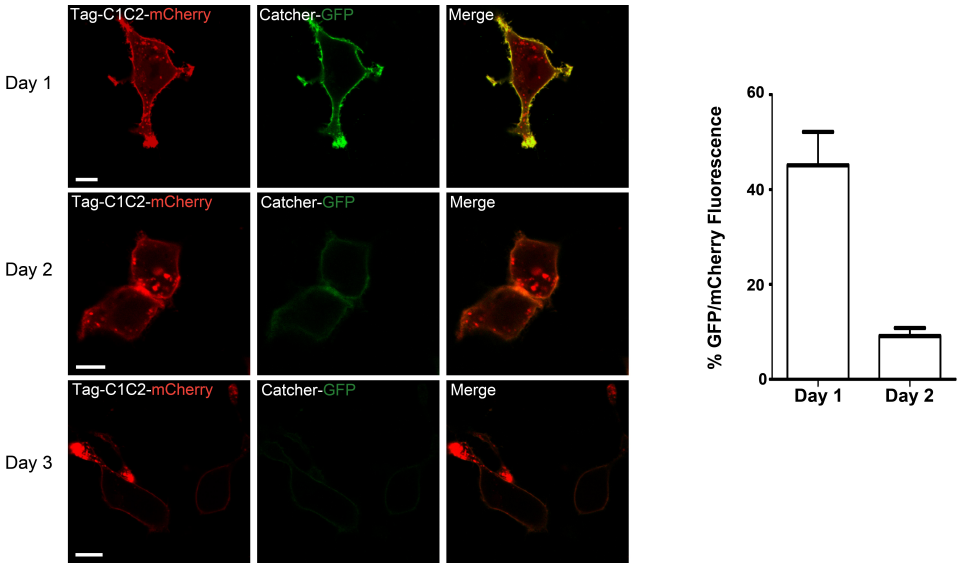


Supplementary Figure 3.3. SpyTag/SpyCatcher labeling of TrkB receptor transfected in HEK cells and neurons. (**A**) The SpyTag was placed at the N-terminus after the signal peptide cleavage site of the TrkB-mCherry fusion. Single plane confocal images of HEK cells expressing TrkB-mCherry and Tag-TrkB-mCherry (red) after 1-hour incubation with Catcher-GFP (green). Only the Tag-TrkB-mCherry expressing cells show labeling with Catcher-GFP. (**B**) Maximum intensity projection of the Tag-TrkB-mCherry expressed in primary neuronal cultures (red) labeled with Catcher-GFP (green). Scale bar, 10 μm.



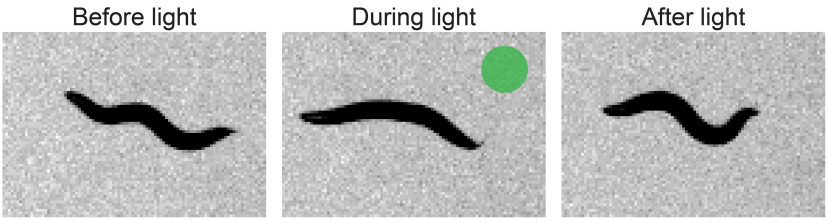
Supplementary Figure 3.4. Characterization of a subset of variants with poor membrane localization identified in the SpyTag/SpyCatcher screen of the ReaChR N298 library. (**A**) Example images from the screening process for non-tagged control (ReaChR), parent (Tag-ReaChR), and Tag-ReaChR mutant 'poor localizers' from the N298 library. Full field, population images were taken for each tested variant and used to measure the GFP and mCherry fluorescence. Amino acid mutations at the 298 residue position are highlighted in blue for the 'poor localizers' in the variants labeled as in Figure 3.3*D*. (**B**) Single plane confocal images of parent (Tag-ReaChR-mCherry) compared with the 'poor localizers' of mCherry (red), Catcher-GFP (green) and merge. All 'poor localizers' show high levels of internal mCherry localization. (**C**) (top) GFP/mCherry fluorescence ratio or (bottom) mCherry fluorescence of Tag-ReaChR (N = 24) compared

with ReaChR variants (N298S: $N = 44$, N298V: $N = 68$, and N298K: $N = 26$) from single plane confocal images of HEK cells expressing the tagged opsins with intensity measurements made by selection of a region of interest around each cell and measurement of mean GFP and mCherry fluorescence across the region. Comparisons between Tag-ReaChR parent with each variant was done by Dunnett's Multiple Comparison Test. All population data are plotted as mean ± SEM. $*P < 0.05$, $**P < 0.01$, $*** P < 0.001$. Scale bar, 10 μm.



Supplementary Figure 3.5. Long-term stability of SpyTag/SpyCatcher labeling. Single plane confocal images of Tag-C1C2-mCherry expression (red) and Catcher-GFP membrane-localized protein binding (green) and merge. Day 1 is imaged shortly after a 1-hour incubation of Catcher-GFP with HEK cells expressing Tag-C1C2-mCherry in D10 and washing with MEM. Cells were then left in D10 at 37°C for 24 hours and imaged again for Tag-C1C2-mCherry expression (Day 2). Cells were then left in D10 at 37°C for another 24 hours and imaged again for Tag-C1C2-mCherry expression (Day 3). (right) Plot of the percent of GFP/mCherry fluorescence of individual Tag-C1C2-mCherry expressing cells covalently bound to Catcher-GFP on Day 1 vs Day 2 ($N = 12$ for each day). All population data are plotted as mean ± SEM. Scale bar, 10 μm.

A



B

	Catcher-GFP labeling?	ATR?	Fraction worms showing light induced paralysis
Tag-ReaChR-mCherry	NO	YES	11/11
Tag-ReaChR-mCherry	YES	YES	6/6
Tag-ReaChR-mCherry	NO	NO	0/3
Tag-ReaChR(E130D)-mCherry	NO	YES	10/10

Supplementary Figure 3.6. Functional characterization of Tag-opsin constructs in locomotion behavioral assay in *C. elegans*. (**A**) Three frames of video of a *C. elegans* expressing Tag-ReaChR-mCherry specifically in GABA-producing neurons (19 D-type neurons) before (left), during (middle) and after (right) green light stimulation. Activation of these GABA neurons paralyses the worm. Activation of Tag-ReaChR-mCherry with green light shows clear induction of paralysis as shown by the posture change evident during light stimulation. (**B**) Table showing the fraction of worms with high opsin expression that have light induced paralysis under different conditions.

Supplementary Table 3.1. Comparison between size of SpyTag with other covalent labeling methods.

Tag Name	Size [amino acids]	Reference
SpyTag	13	(94)
SpyTag optimized	10	(53)
SpyCatcher	139	(94)
SpyCatcher optimized	84	(53)
SNAP-Tag	181	(90, 93)
CLIP-Tag	181	(88)
Halo Tag	295	(87)
GFP	238	(80)

Supplementary Table 3.2. Summary of constructs built with protein product name used in the text.

Construct	**Protein**
pLenti-CMV/CaMKIIa::SpyTag-C1C2-TS-EYFP	Tag-C1C2-EYFP
pLenti-CMV/CaMKIIa::SpyTag-C1C2-TS-mCherry	Tag-C1C2-mCherry
plenti-CMV/CaMKIIa::SNAP-tag-C1C2-TS-mCherry	SNAP-tag-C1C2-mCherry
pLenti-CMV/CaMKIIa::C1C2-TS-mCherry	C1C2-mCherry
pQE80l-T5::6xhis-SpyCatcher-Elp-GFP	Catcher-GFP
pLenti-CMV/CaMKIIa::SpyTag(DA)-C1C1-TS-mCherry	Tag(DA)-C1C1-mCherry
pLenti-CMV/CaMKIIa::SpyTag(0)-C1C1-TS-mCherry	Tag^{0}-C1C1-mCherry
pGP-CMV::ChR2-mCherry	ChR2-mCherry
pGP-CMV::ChR2-mCherry-SpyTag	ChR2-mCherry-Tag
pLenti-CMV/CAMKIIa::ReaChR-TS-mCherry-WPRE	ReaChR-mCherry
pLenti-CMV/CAMKIIa::SpyTag-ReaChR-TS-mCherry-WPRE	Tag-ReaChR-mCherry
pLenti-CMV/hSyn1::TrkB-3xGS linker- mCherry-WPRE	TrkB-mCherry
pLenti-CMV/hSyn1::SpyTag-TrkB-mCherry-WPRE	Tag-TrkB-mCherry
hlh-12::SpyTag-C1C2-mCherry	Tag-C1C2-mCherry
HS::lin-3 signal sequence::SpyCatcher-GFP	Catcher-GFP
pSM*::unc-47::SpyTag-ReaChR-TS-mCherry*	Tag-ReaChR-mCherry
pSM*::unc-47::SpyTag-ReaChR(E130D)-TS-mCherry*	Tag-ReaChR(E130D)-mCherry

Supplementary Table 3.3. Addgene plasmids with accession codes used for construct designs used in this paper.

Construct	**Addgene #**
pAAV-CaMKIIa-C1V1 (t/t)-TS-mCherry	35500
pLenti-CaMKIIa-C1C2-TS-EYFP	35520
pLenti-CaMKIIa-hChR2(H134R)-mCherry-WPRE	20943
pGP-CMV-GCaMP6f	40755
pLenti-hSyn-eNpHR 3.0-EYFP[*]	26775

Chapter 4

STRUCTURE-GUIDED SCHEMA RECOMBINATION GENERATES DIVERSE CHIMERIC CHANNELRHODOPSINS

A version of this chapter has been published as (114)

4.1 Abstract

Integral membrane proteins (MPs) are key engineering targets due to their critical roles in regulating cell function. In engineering MPs it can be extremely challenging to retain membrane localization capability while changing other desired properties. We have used structure-guided SCHEMA recombination to create a large set of functionally diverse chimeras from three sequence-diverse channelrhodopsins (ChRs). We chose 218 ChR chimeras from two SCHEMA libraries and assayed them for expression and plasma membrane localization in human embryonic kidney cells. The majority of the chimeras express, with 89% of the tested chimeras outperforming the lowest-expressing parent; 12% of the tested chimeras express at even higher levels than any of the parents. A significant fraction (23%) also localize to the membrane better than the lowest-performing parent ChR. Most (93%) of these well-localizing chimeras are also functional light-gated channels. Many chimeras have stronger light-activated inward currents than the three parents, and some have unique off-kinetics and spectral properties relative to the parents. An effective method for generating protein sequence and functional diversity, SCHEMA recombination can be used to gain new insights into sequence-function relationships in MPs.

4.2 Introduction

Integral membrane proteins (MPs) serve diverse and critical roles in controlling cell function. Their receptor, channel, and transporter functions make MPs common targets for pharmaceutical discovery and important tools for studying complex biological processes (2, 115-117). Biochemical studies of MPs and their engineering for biotechnological

applications are often limited by poor expression and membrane localization in heterologous systems (118, 119). Unlike soluble proteins, MPs must go through the additional steps of membrane targeting and insertion as well as rigorous post-translational quality control (120, 121). Functional diversity depends on sequence diversity, but it is challenging to design highly diverse variants that retain membrane localization while at the same time revealing other useful functionality (122). To address this challenge, we demonstrate that structure-guided SCHEMA recombination (123) can create functional MP chimeras from related yet sequence-diverse channelrhodopsins (ChRs). The resulting chimeric ChRs retain their ability to localize to the plasma membrane of mammalian cells but exhibit diverse, potentially useful functional properties.

ChRs are light-gated ion channels with seven transmembrane alpha-helices. They were first identified in photosynthetic algae, where they serve as light sensors in phototaxic and photophobic responses (124, 125). ChR's light-sensitivity is imparted by a covalently linked retinal chromophore (7). With light activation, ChRs open and allow a flux of ions across the membrane and down the electrochemical gradient (126). When ChRs are expressed in neurons, their light-dependent activity can stimulate action potentials, allowing cell-specific control over neuronal activity (10, 127). This has led to extensive application of these proteins as tools in neuroscience (117). The functional limitations of available ChRs have led to efforts to engineer and/or discover novel ChRs e.g. ChRs activated by far red light, ChRs with altered ion specificity, or ChRs with increased photocurrents with low light intensity (126). The utility of any ChR, however, depends on its ability to express in eukaryotic cells of interest and localize to the plasma membrane. Our goal is to generate sequence-diverse ChRs whose functional features are useful for neuroscience applications and have not been found in natural environments.

MP engineering is still in its infancy when compared to soluble protein engineering. Significant progress in increasing microbial expression and stability of MP's has been made using high-throughput screening methods to identify variants with improved expression from large mutant libraries (119, 128-130). The main motivation was to generate MP mutants that are stable and produced in sufficient quantities for

crystallographic and biochemical characterization. This pioneering work demonstrated that MP expression in *E. coli* and yeast can be enhanced by directed evolution. Because there is not a good method for high throughput screening of ChR function, however, we chose to focus on introduction of sequence diversity using structure-guided SCHEMA recombination.

SCHEMA recombination offers a systematic method for modular, rational diversity generation that conserves the protein's native structure and function but allows for large changes in sequence (131-133). SCHEMA divides structurally-similar parent proteins into blocks that, when recombined, minimize the library-average disruption of the tertiary protein structure (123). Two different structure-guided recombination methods have been developed: one restricts blocks to be contiguous in the polypeptide sequence (123, 134), while the other allows for design of structural blocks that are non-contiguous in the polypeptide sequence but are contiguous in 3D space (135). SCHEMA has enabled successful recombination of parental sequences with as low as 34% identity (136), which is not possible using random DNA recombination methods such as DNA shuffling (137). SCHEMA recombination has been used to create a variety of functionally diverse soluble proteins (136, 138-141), but it has not yet been applied to MP engineering. Our goals in this study were to 1) test whether structure-guided recombination produces chimeric MPs that express and localize; 2) measure the fraction of chimeric sequences in a SCHEMA library that express and localize; and 3) assess the functional diversity of the MPs that successfully localize to the membrane.

We used SCHEMA to design two libraries of chimeric ChRs, using three parental ChRs having 45-55% amino acid sequence identity. The parent ChRs show different levels of expression and localization in mammalian cells, differences in channel current strength, and differences in the optimal wavelength for channel activation. The SCHEMA recombination libraries, one contiguous and the other non-contiguous, were designed with 10 blocks, yielding an overall library size of 2×3^{10}, or more than 118,000 possible sequences. On average, chimeras are 73 mutations from the closest parent. We chose and synthesized a set of 218 chimeric genes from these libraries and assayed the proteins for

expression and membrane localization in mammalian cells. Our results offer new insight into the sequence dependence of ChR expression and localization, and reveal new functional variation in diverse, well-localizing ChR chimeras. We show that SCHEMA recombination can rapidly and efficiently generate functionally-diverse MPs.

4.3 Results

Parents for ChR chimera library. Since the initial discovery and characterization of channelrhodopsins ChR1 (9) and ChR2 (8) from the alga *Chlamydomonas reinhardtii*, a number of ChRs have been isolated and characterized, e.g., VChR1 (28), VChR2 (142, 143), MvChR1 (144), CaChR1 (32), DChR (2), and PsChR (145). *De novo* transcriptome sequencing of 127 species of algae led to the discovery of fourteen new ChRs that express and function in mammalian neurons (41). To create new ChRs by SCHEMA recombination, we chose CsChrimsonR (41), C1C2 (23), and CheRiff (14) as parents. These three ChRs are representative of the available sequence diversity and share 45-55% amino acid identity (**Figure 4.1*A***). CsChrimsonR (CsChrimR) is a fusion between the N-terminus of CsChR from *Chloromonas subdivisa* and the C-terminus of CnChR1 from *Chlamydomonas noctigama* and contains a single mutation (K176R) that improves the off-kinetics (the time it takes the channel to close after it is exposed to light) (41). C1C2 is a fusion between ChR1 (N-terminal) and ChR2 (C-terminal), both from *Chlamydomonas reinhardtii* (23). C1C2 is the only ChR with a solved crystal structure, making it a useful parent for structure-guided recombination. CheRiff is SdChR, from *Scherffelia dubia* with a single mutation (E154A) that speeds up the off-kinetics and provides a blue-shifted peak in the action spectrum (the current strength achieved by different wavelengths of light) (14). These three parental sequences are fully functional in mammalian cells and have distinct spectral properties. The peak activation wavelengths for CsChrimR, C1C2, and CheRiff are 590 nm, 480 nm, and 460 nm, respectively.

Quantifying ChR expression and localization. Fluorescent protein fusions have been used extensively as markers for ChR expression (43). To quantify ChR expression, we fused the red fluorescent protein, mKate2.5 (mKate) (96), to the C-termini of the ChRs. To

quantify membrane insertion and plasma membrane localization, we used the SpyTag/SpyCatcher labeling method (78). Briefly, SpyTag is a 13-amino acid tag that forms a covalent bond with its interaction partner, SpyCatcher (94). For each ChR, SpyTag was cloned after the native N-terminal signal sequence. This tag is displayed on the extracellular surface of the cell if the ChR is correctly localized to the plasma membrane. Surface-exposed SpyTag can be quantified using exogenously added SpyCatcher protein fused to GFP, which specifically and covalently binds to the SpyTag of correctly localized SpyTag-ChR. Using these methods, we assayed ChR expression (mKate fluorescence: **Figure 4.1*B***) and localization (GFP fluorescence: **Figure 4.1*C***) in human embryonic kidney (HEK) cells and measured the localization efficiency, or fraction of total protein localized, using the ratio of GFP fluorescence signal to mKate fluorescence signal (**Figure 4.1*D***).

HEK cells were transfected in a 96-well plate format, labeled with SpyCatcher-GFP, and imaged for mKate and GFP fluorescence as described in **Methods**. For the three parental ChRs, images have been processed by cell segmentation to show the distribution of protein expression and localization levels across the population of expressing cells. Alternative image processing, measuring the whole population intensity, was used to quantify the expression (mean mKate intensity), plasma membrane localization (mean GFP intensity), and localization efficiency (mean mKate intensity / mean GFP intensity) of each ChR construct (see **Methods**). The whole-population intensity measurements provide a single intensity measurement for each property for a given population of expressing cells. There is significant cell-to-cell variability in transient transfections. To account for this, we measured the properties of each ChR in quadruplicate and calculated the deviation of single intensity measurements between these replicates.

Expression, localization, and localization efficiency of the three parent ChRs. Figure 4.1*B-D* shows the expression, localization, and localization efficiency of each parent protein in HEK cells. Each parent ChR has an easily distinguishable signature expression and localization profile that can be seen in example images and in the distributions of expression, localization, and localization efficiency for the three parents (**Figure 4.1*B-D***).

Both CsChrimR and C1C2 have very high expression levels with large cell-to-cell variation, whereas CheRiff expresses at a significantly lower yet consistent level (**Figure 4.1*B***). CsChrimR has the highest level of localization, whereas CheRiff and C1C2 have lower localization levels (**Figure 4.1*C***). Localization efficiency shows a different ranking among the parent proteins: CheRiff has the highest localization efficiency and C1C2 has the lowest (**Figure 4.1*D***). The wide range in parent ChR mean expression, localization, and localization efficiency should facilitate generation of chimeras with different levels of these properties.

SCHEMA recombination library design. Using the three ChR parents, the known structure of C1C2, and the SCHEMA algorithm (123, 134), we designed two 10-block recombination libraries. SCHEMA is a scoring function that predicts block divisions that minimize the disruption of protein structure when swapping homologous sequence elements among parental proteins. SCHEMA works by defining pairs of residues that are in 'contact' and identifying a block design (size and location of sequence blocks) that minimizes the average number of broken amino acid contacts in the resulting library. Two residues are defined to be in contact if they contain non-hydrogen atoms that are within 4.5 Angstroms of each other. If a chimera inherits a contacting pair that is not present in a parent sequence, that contact is said to be broken. Contacts can only be identified in regions of the ChR protein with reliable structural information. The C1C2 structure provides such information for part of the N-terminal extracellular domain (residues 49-84), the 7-helix integral membrane domain (residues 85-312), and the intracellular C-terminal beta-turn (residues 313-342) (23). A parental alignment was made for the structurally modeled residues of C1C2 (49-342) and homologous regions of CheRiff (23-313) and CsChrimR (48-340) (**Supplementary Figure 4.1**). The full contact map calculated from the C1C2 structure is shown in **Figure 4.2*A***. Only contacts between non-conserved residues are relevant for the library design (**Figure 4.2*B***), because only these can be broken upon recombination. Though contacts are distributed throughout the ChR structure, the non-conserved contacts are far denser at the termini and on the outer surface of the protein; these are the areas of the protein with the most sequence diversity (**Figure 4.2**).

Two SCHEMA libraries were designed: contiguous (123, 134) and non-contiguous (135). Contiguous libraries are designed so that blocks are contiguous in the amino acid sequence, while non-contiguous libraries swap blocks in the three-dimensional structure that are not necessarily contiguous in the primary structure. Using the parental alignment and the contact map, SCHEMA generates a list of possible library designs with a minimized library-average disruption score, the E-value, i.e. the average number of broken parental contacts per chimera in the library. A 10-block contiguous library was selected (**Figure 4.2*C***) with roughly even-length blocks (14-43 residues), a relatively low average E-value (25), and whose sequences have an average of 73 mutations from the nearest parent. The selected 10-block non-contiguous library has a low average E-value (23), block sizes comparable to the contiguous library, and an average of 71 mutations from the nearest parent (**Figure 4.2*D***). The non-contiguous library design also maintains the presumptive dimer interface (see **Methods)**. For these libraries, the 'mutations' introduced into any one parent are limited to the non-conserved residues of the other two parents. Each of the 10-block, 3-parent libraries gives 59,049 possible chimeras (3^{10}), for a total of 118,098 possible chimeras.

The two library designs both place block boundaries in positions that may not be obvious in the protein structure. For example, that several boundaries appear in the middle of alpha-helices indicates that naïve chimeragenesis by simply swapping elements of secondary structure would be more disruptive than design based on conservation of native contacting residue pairs. To test this, we calculated the average E-value for libraries with block boundaries within the loops between transmembrane alpha-helices such that the N-terminal domain, the C-terminal domain, and each helix form separate blocks for a total of 9 blocks. Within the loops, there are multiple possible locations for block boundaries. We built 128 different designs with block boundaries within loops and calculated library average E-values that range from 36 to 43. These values are significantly higher than those for the SCHEMA designs and indicate that naïve helix swapping is more disruptive than SCHEMA recombination.

Production of chimeras for characterization. We chose a set of 223 sequences from the recombination libraries for gene synthesis and characterization of expression and localization properties of the ChRs in mammalian cells. This set included all 120 proteins with single-block swaps from both libraries. These chimeras consist of 9 blocks of one parent and a single block from one of the other two parents. An additional 103 sequences were designed to maximize mutual information (146) between chosen chimeras and the remainder of the chimeric library, using the rationale described in Romero *et al.* (140). Seventeen of these sequences were designed with a constraint on the number of mutations from the nearest parent (<40 mutations). This set, referenced as the "maximally informative with mutation cap", provided chimeras composed of, on average, six blocks of one dominant parent and four blocks of a mix of the other two parents. The remaining 86 of the "maximally informative" sequences are highly diverse, consisting of blocks from all three parents and containing, on average, 84 mutations when compared to the most sequence-related parent. This set of 223 genes was synthesized and cloned in a mammalian expression vector at Twist Bioscience, Inc. Two hundred and fifteen of the designed sequences were synthesized successfully and cloned into the expression vector; with the three parent sequences, this gave a total of 218 sequences for the library characterization studies.

Localization and expression of ChR chimeras. HEK cell expression and localization were measured for each chimera using at least 150 and up to 100,000 transfected cells from at least four replicate HEK cell transfections. Chimeras were benchmarked to the lowest performing parent. CheRiff is the lowest performing parent for expression and localization, and C1C2 is the lowest performing parent for localization efficiency. The majority (89%) of the chimeras have higher expression levels than the lowest parent (**Figure 4.3*A***) while a lower number, amounting to 23%, have higher localization levels than the lowest parent (**Figure 4.3*B***). 44% of the chimeras have better localization efficiency than the lowest parent (**Figure 4.3*C***). The difference between the number of chimeras that *express* well and the number of chimeras that *localize* well suggests that the sequence demands for localization are more stringent.

Measurements show no clear correlation between chimera expression and localization (**Supplementary Figure 4.2*A***), and chimeras localize more frequently if they are only a single-block swap away from the nearest parent (<40 mutations) (**Supplementary Figure 4.2*B***). On the other hand, most chimeras express, even with as many as 108 mutations from the nearest parent (**Supplementary Figure 4.2*C***). Only 9% of the sequences in the 'maximally informative' set localize as well as the lowest localizing parent, while 24% of the 'maximally informative mutation cap' set localize as well as the lowest localizing parent, and 33% of the sequences with a single block swap localize as well as the lowest parent (**Figure 4.4*A***). Thus sequences from the 'maximally informative' set are less likely to localize than the sequences with single-block swaps or sequences with a mutation cap. These results highlight the difficulty of finding highly mutated ChR sequences (>40 mutations from the nearest parent) that localize well. Nonetheless we found 51 new ChRs in this test set of 218 that localize to the plasma membrane at least as well as the worst parent, and 8 of those are more than 40 mutations away from the closest parent. Although less diverse than the 'maximally informative' chimeras, the single-block-swap chimeras still contain on average 15 mutations when compared to the closest parent. This is a significant amount of diversity to introduce while still maintaining localization, given that even a single mutation can destroy a protein's ability to fold or function (133).

Performance ranking of chimera sequences for each property of interest (expression, localization, and localization efficiency) shows that sequences dominated by CheRiff generally rank low in expression but have the highest rankings for localization efficiency (**Figure 4.3*E*,*G***), while sequences dominated by CsChrimR have the highest ranking for localization (**Figure 4.3*F***). These trends are seen for both the contiguous and non-contiguous libraries (**Supplementary Figure 4.3**). No clear patterns or specific blocks of sequence emerge from the data that determine chimera performance, suggesting that each sequence/structural block behaves differently in different contexts. However, the single-block-swapped chimeras offer insight into the sequence dependence of properties in the context of the parental ChRs.

We also wanted to compare the two library design strategies. Both the contiguous and non-contiguous SCHEMA recombination libraries have the same number of blocks, similar average disruption scores (E-values) (25 and 23, respectively), similar average number of mutations (73 and 71, respectively), but different design strategies. We found that chimeras show similar ranges in measured properties whether they were designed to be contiguous in the primary or tertiary structure (**Supplementary Figure 4.4**). These results suggest that, for ChRs, library design is less important than the average disruption score and average number of mutations per chimera. For soluble proteins, the average disruption score and average number of mutations of SCHEMA libraries have been shown to correlate with the fraction of the recombination library that does not fold and function (136).

Comparison of chimeras with good localization. Chimeras with single-block swaps indicate which individual blocks increase localization (**Figure 4.4*B***), expression (**Supplementary Figure 4.5*B***), and localization efficiency (**Supplementary Figure 4.5*D***). For both the CheRiff and C1C2 parents, there is a single-block swap from CsChrimR that results in a chimera with large improvements in localization (**Figure 4.4*B***). Interestingly, the block from CsChrimR that boosts CheRiff's localization is different from the CsChrimR block that improves C1C2's localization: the former contains the CsChrimR N-terminus and an associated extra-cellular loop and the latter contains the first and (structurally adjacent) seventh CsChrimR helices. In fact, the CsChrimR block that causes a nearly two-fold increase in C1C2's localization causes a two-fold *decrease* in CheRiff localization when chimeras are compared to their respective dominant parent. This result stresses again the importance of context when assessing the sequence dependence of a property as complex as localization.

There are also single blocks from both the CheRiff and C1C2 parents that significantly increase localization of CsChrimR (**Figure 4.4*B***). This is interesting because both the CheRiff and C1C2 parents have lower localization levels than the CsChrimR parent. This result illustrates recombination's ability to produce progeny that outperform all of the parental sequences. The three single-block swaps that produce chimeras that outperform CsChrimR are at the N-terminus, first helix, and second helix (**Figure 4.4*C***). It is expected

that swapping the N-terminus of the protein could influence localization (147), but it is not clear why the first and second helix swaps are important for localization. Finally, there are two "maximally informative mutation cap" sequences that also outperform the top parent, CsChrimR (**Figure 4.4*A***). These chimeras have blocks from all three parents spread across the protein sequence (**Figure 4.4*C***).

Functional characteristics of chimeras that localize. Seventy-five chimeras with localization levels above or within one standard deviation of the CheRiff parent or localization efficiency above or within one standard deviation of the C1C2 parent were analyzed for other functional characteristics. Each chimera was expressed in HEK cells and its light-inducible currents were measured using patch-clamp electrophysiology in voltage clamp mode upon sequential exposure to three different wavelengths of light (473, 560, and 650 nm). ChRs have a characteristic light-activated current trace with an initial peak in inward current occurring immediately after light exposure followed by a decay of inward current to a constant, or steady state, current (**Figure 4.5; inset**). The majority of tested chimeras were functional, with only five of the 75 tested chimeras having light-activated steady-state inward currents less than 20 pA (**Figure 4.5**). Different chimeras are optimally activated by different wavelengths. All 70 of the active chimeras are activated by 473 nm light, whereas only 18 chimeras show robust activation with 650 nm light (**Figure 4.5**). When activated with 473 nm light, ten chimeras have stronger peak and steady-state photocurrents than the parental protein with the strongest photocurrents (CsChrimR) (**Figure 4.5*C***), demonstrating again that recombination can generate MPs that outperform any of the parents.

Though localization is a prerequisite for channel function, a chimera that localizes well does not necessarily provide stronger currents than a chimera that localizes less well. In addition to the amount of protein in the membrane, the channel's conductance properties also affect current strength. The mutations in these ChR sequences could cause a change in channel conductance. To test if changes in current strength are due to differences in localization or conductance, we compared the measured localization and peak current strength for each chimera (**Supplementary Figure 4.6**). That we did not find a strong

positive correlation between these two measurements suggests that differences in chimera currents are dominated by changes in their conductance. That is, as long as an adequate fraction of a ChR is able to localize to the plasma membrane, the major factor determining current strength is the chimera's specific conductance properties, which is sequence-dependent and can be tuned by mutation.

ChR chimeras with altered photocurrent properties. Analysis of the photocurrent properties of single-block-swap chimeras activated with 473 nm light show that there are many single-block changes to both the CheRiff and C1C2 parent that cause large increases in current strength (**Figure 4.6*A***). The CheRiff parent shows large increases in current strength with single blocks from either C1C2 or CsChrimR, while C1C2 performs best with single blocks from CheRiff, even though CheRiff has the weakest currents of the three parents. Comparison of the sequences of these highly functional chimeras shows that single blocks swapped at many different positions in the ChR sequence can have a positive effect on current strength and that no single block position alone accounts for the improved currents (**Figure 4.6*B***).

Significant effort has been taken to find ChR sequences with red-shifted properties (activation by ~650 nm light), because red light has enhanced tissue penetration and decreased phototoxicity when compared to higher energy blue light (28, 41). Three natural ChRs have been shown to be activated with red light: CsChR/Chrimson (41), VChR1 (28), and MChR1 (144). Here we show that recombination generates many chimeras that are activated with 650 nm light and that have significant sequence diversity when compared to their red-light activated parent (a mean of 15 and as many as 70 mutations) (**Figure 4.5*A*, Figure 4.6*A***). All the single-block-swap chimeras capable of producing photocurrents with 650 nm light have CsChrimR as the dominant parent (**Figure 4.6*A***). The CsChrimR parent can tolerate single-block swaps from either C1C2 or CheRiff at many positions in the ChR sequence and still retain strong currents activated by 650 nm light (>50 pA peak current) (**Figure 4.6*B***), showing that none of its single block positions is necessary for CsChrimR's red light-activated current.

Some chimeras have novel spectral properties, exhibited by none of the three parent ChRs. One multi-block-swap chimera from the maximally informative set, for example, shows strong activation with 560 nm light but atypical properties once the light is turned off (**Figure 4.6*C***). This chimera shows a gradual increase in inward current once the green light is turned off, followed by a very slow decrease in current. This inward current can be turned off with 473 nm light, causing a brief depolarization, then a decrease in inward current while the 473 nm light is on. Once the 473 nm light is turned off, there is a brief depolarization followed by a decrease in current to baseline levels. When activated by 473 nm light without pre-exposure to 560 nm light, this chimera produces inward currents with unusual light-off behavior (**Supplementary Figure 7*A***). Sequential 1-second exposures to 560 nm light causes continued depolarization (**Supplementary Figure 7*C***). This type of bi-stable excitation, step function opsin (SFO) has been reported previously, in ChRs generated with site-directed mutagenesis at a single position (C128) in ChR2 (38). However this SFO is activated by blue (470 nm) light and terminated by green (542 nm) light (38). The unusual light-off behavior, with inward currents that continue to increase ~0.5 s after the light has been turned off, suggests an altered photocycle (38).

4.4 Discussion

SCHEMA uses structural information to guide the choice of block boundaries for creating libraries of chimeric proteins from homologous parents. Both conservative and innovative, recombination generates large changes in sequence without destroying the features required for proper folding, localization, and function. Recombination is conservative because the sequence diversity source has passed the bar set by natural selection for fold and function. Recombination thus introduces limited diversity and at positions which are tolerant to mutation e.g. at the protein termini or the surface interacting with the lipid bilayer. In contrast, conserved functional residues and those in the structural core experience little or no change upon recombination. The sequence changes that are made can nonetheless lead to new functional properties that may not be selected for in nature.

In the largest screen of ChR sequences and properties to date, we found that a high proportion of chimeras made by recombining three parent integral membrane ChRs retain the ability to localize to the plasma membrane and exhibit high photocurrents despite having an average of 43 mutations with respect to the closest parent. In HEK cells, 89% of the 218 tested chimeras expressed at least as well as the lowest performing parent, and 23% localized better than the lowest performing parent. Moreover, 70 out of 75 well-localizing chimeras show light-activated inward currents. The innovative nature of SCHEMA recombination was observed in ChR expression, localization, and photocurrents under activation by 473 nm light, for which 5-15% of the tested chimeras outperformed the best-performing parent. In particular, six single-block-swap chimeras showed between a 1.5 to 2-fold increase in photocurrent relative to the parent with the strongest photocurrents (CsChrimR) when activated by 473 nm light. From one of the heavily mutated chimeras, we also discovered that the photophysical properties of a ChR can be modified dramatically and unexpectedly.

Recombination can create sequences with properties that may not be selected in nature. For example, red wavelengths do not penetrate to the water depths typically occupied by algae, and thus red-light activated ChR's are rare in nature, with only three natural such ChRs discovered to date (28, 41, 144). We purposefully biased our recombination libraries by choosing a red-light activated parent, CsChrimR and found a number of sequence-diverse progeny that were also red-light activated. Although the retinal binding pockets of the two blue-shifted parents are nearly identical, almost half of the residues in the retinal-binding pocket of CsChrimR are different. Including CsChrimR as a parent thus allowed us to explore sequence diversity in this vital region of the protein and enrich for properties desirable for neuroscience applications but not necessarily favored in nature. This type of enrichment in recombination libraries depends on the choice and availability of parent proteins.

Two of the parent proteins for this study came from the 61 ChR homologs that were discovered from *de novo* transcriptome sequencing of 127 species of algae (41). Of the 50 of these ChR homologs assayed for expression and photocurrents in HEK cells, 25

produced photocurrents while the other 25 did not. Fourteen of these sequences were then characterized and shown to retain function in mammalian neurons (41). Although interesting and useful genes can to be found in nature, it is not always clear where to look for them. SCHEMA recombination, on the other hand, offers a systematic, straightforward method for generating artificial diversity from a set of natural sequences. Furthermore, the type of systematic diversity in a recombination library is useful for analyzing how sequence features determine protein properties. Such analysis is greatly simplified by the greatly reduced sequence space (i.e., 10 blocks with only 3 possible sequences at each block).

This ChR chimera dataset offers insights into the robustness of ChR expression, localization, and function to changes in sequence. Although almost all the chimeric sequences express, localization is more rare, indicating that the sequence and structural constraints on localization are greater than those on expression. Among sequences that successfully localize, most are functional light-activated channels, but there is significant sequence-based variability in activation wavelength and conductance. This suggests that membrane localization is a principal hurdle to engineering ChR sequences with novel functions. Simply extrapolating the fraction of well-localized chimeras in our 218-chimera sample set to the overall library, we could expect 10,000-27,000 of the 118,000 chimeras to localize to the membrane.

The ability to predict which sequences are likely to localize will remove a key roadblock to identifying novel, functional sequences. Changes throughout the ChR protein can enhance localization and photocurrents, and no single sequence block determines the observed improvements. This suggests that each sequence/structural block behaves differently in different contexts. For certain soluble protein properties (e.g. thermostability), it has been shown that block contributions are additive, i.e. context independent, and that chimera stability can be predicted using linear regression (139, 140, 148, 149). Our data suggest that ChR localization and photocurrent properties, however, require a more complex model to account for the nonlinear dependence of function on block sequence. Our future work will explore the use of statistical models to provide sequence/structure insights into the features that determine localization and photocurrent properties, to predict the properties of all

118,000 sequences in the recombination libraries, and to engineer novel ChR sequences with desirable properties.

4.5 Materials and methods

Design and construction of parental ChRs and recombination library. The three ChR parent genes were built using a consistent vector backbone (pFCK) (37) with the same promoter (CMV), trafficking signal (TS) sequence (38), and fluorescent protein (mKate2.5) (39). For the SpyTag/SpyCatcher membrane localization assay, it was necessary to add the SpyTag sequence close to the N-terminus of each of the parental proteins but C-terminal to the signal peptide sequence cleavage site. Assembly-based methods and traditional cloning were used for vector construction and parental gene insertion. Annotated GenBank files are included as supplemental materials for the three SpyTagged parental constructs used in this study.

SCHEMA was used to design 10-block contiguous and non-contiguous recombination libraries of the three parent ChRs that minimize the library-average disruption of the ChR structure (123, 134, 135). Both recombination library designs were made using software packages for calculating SCHEMA energies openly available at http://cheme.che.caltech.edu/groups/fha/Software.htm. The SCHEMA software outputs the amino acid sequences of all chimeras in a library. The amino acid sequence for each chimera chosen for experimental testing was converted into a nucleotide sequence such that all chimeras had consistent codon usage. Gene sequences for the 223-chimera set were synthesized by Twist Bioscience, Inc., cloned in the pFCK vector by a homology based cloning strategy, and transformed into Stbl3 cells (Invitrogen) or Endura cells (Lucigen). Individual clones were picked and sequence verified by NGS. Purified plasmid DNA of each chimera was prepared for HEK cell transfection.

Measuring ChR expression, localization, and photocurrents. HEK 293T cells were transfected with purified, ChR variant DNA using Fugene6 reagent according to the manufacturer's recommendations. Cells were given 48 hours to express before being assayed for expression, localization, or photocurrents. To assay localization level,

transfected cells were subjected to the SpyCatcher-GFP labeling assay, as described in Bedbrook *et al*.. Transfected HEK cells were then imaged for mKate and GFP fluorescence using a Leica DMI 6000 microscope. We used conventional whole-cell patch-clamp recordings in transfected HEK cells to measured light-activated inward currents using methods and equipment described in (15).

4.6 Figures

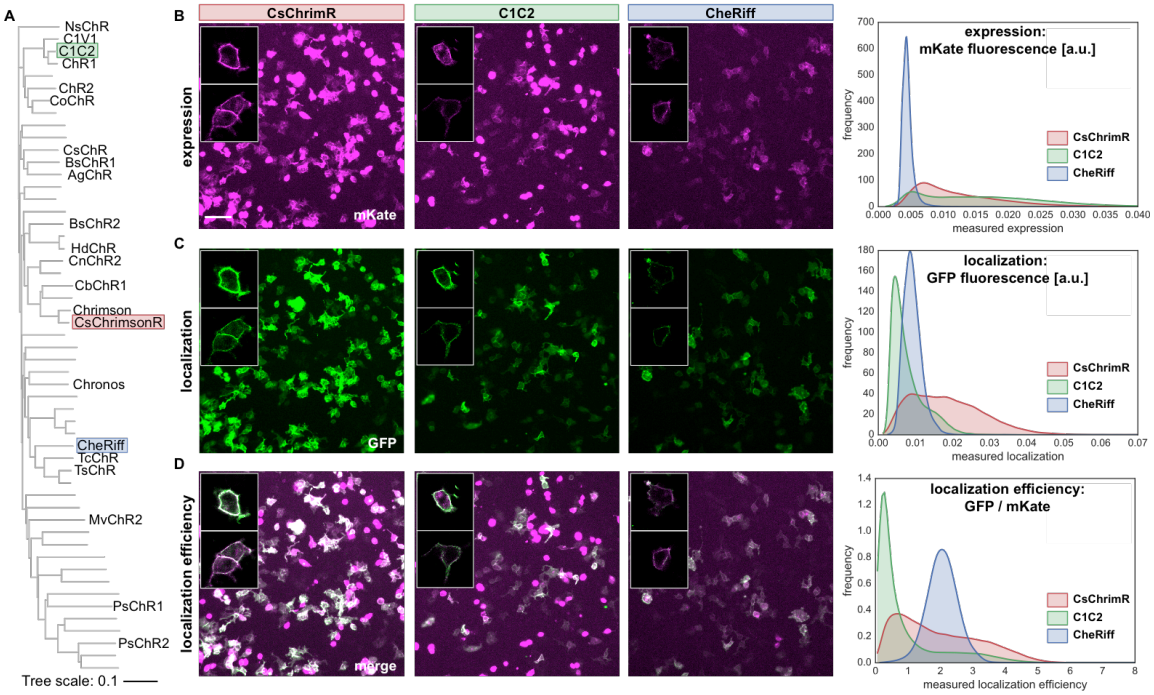


Figure 4.1. Parental ChRs and their properties. (**A**) Phylogenetic tree of published ChR sequences. Sequences with an alias (e.g. NsChR) have been characterized for expression and functionality in HEK cells and/or mammalian neurons. The three parental sequences (C1C2, CsChrimsonR and CheRiff) are highlighted. (**B-D**) HEK cells were transfected with a parental ChR. Membrane-localized ChR was labeled using SpyCatcher-GFP assay, and ChR expression was measured using mKate. HEK cell populations were imaged and processed to measure expression (mean mKate fluorescence [a.u.]), plasma membrane localization (mean GFP fluorescence [a.u.]), and localization efficiency (mean GFP fluorescence / mean mKate fluorescence). Example images show population expression (**B**), localization (**C**), and localization efficiency (**D**) for each parental construct. Scale bar: 100 μm. Insets show confocal images for a few representative cells expressing each parental construct. HEK cell population images were segmented and the ChR expression, localization, and localization efficiency were measured for each cell. The distribution of these properties for the population of transfected cells is plotted for each parent using kernel density estimation for smoothing.

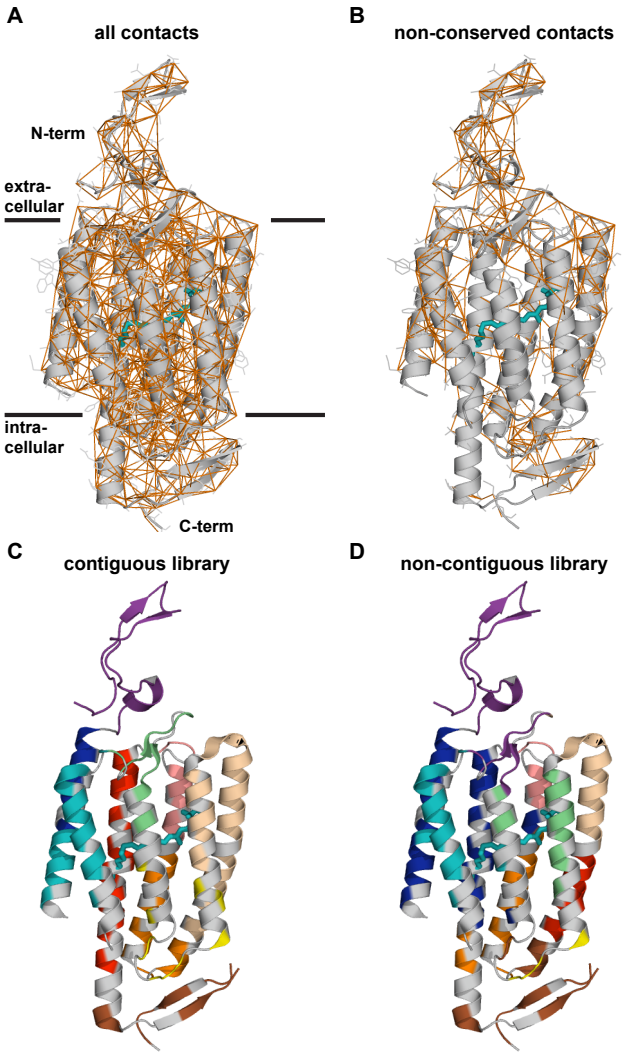


Figure 4.2. Structure-guided recombination library design. (**A**) Contact map highlighting all amino acids within 4.5 Å of each other (orange lines) in the ChR structure. (**B**) For library design we only considered those contacts that can be broken when a different parent block is inserted. Contiguous and non-contiguous libraries were built using the three parental ChRs. The structural cartoon representation of the two libraries is shown for both the contiguous library (**C**) and non-contiguous library (**D**). Residues conserved among the parents are shown in gray, and the different sequence blocks are color-coded. All-trans-retinal (ATR) is shown covalently linked to the protein by the conserved lysine residue using a teal-colored stick representation.

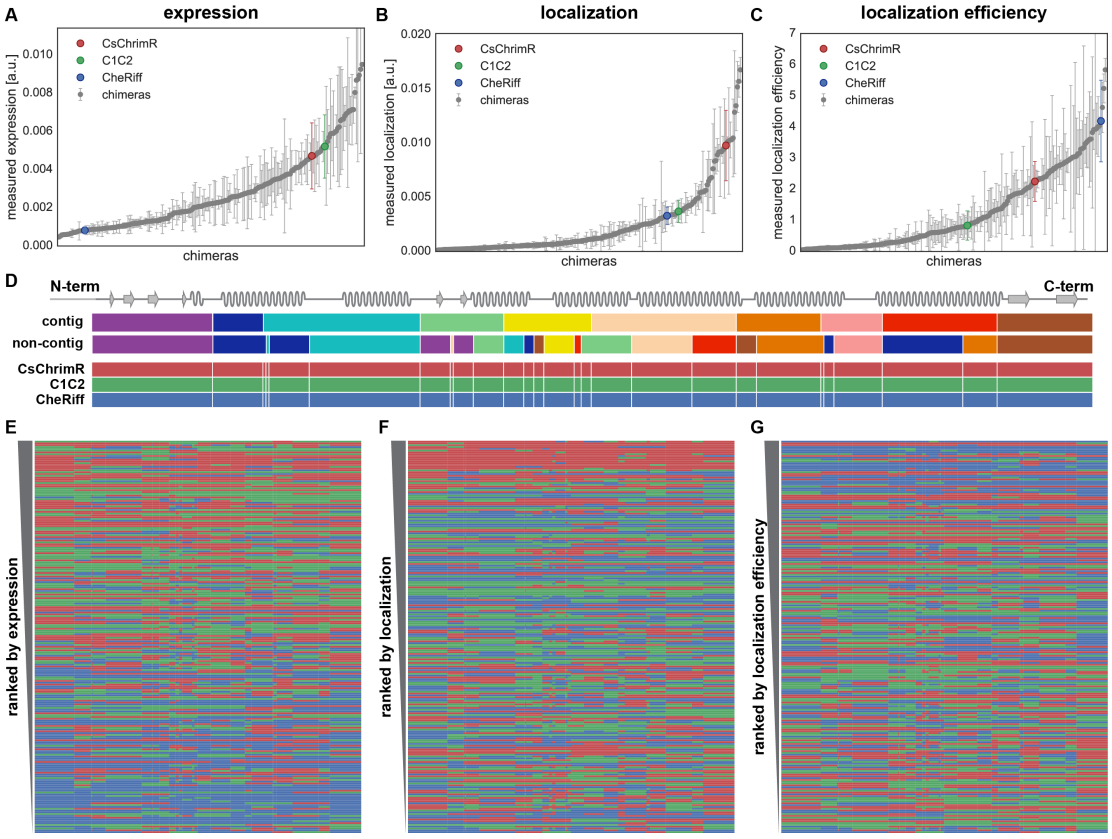


Figure 4.3. Chimera expression, localization, and localization efficiency. (**A**-**C**) show the measured expression (mean mKate fluorescence [a.u.]) (**A**), localization (mean GFP fluorescence [a.u.]) (**B**), and localization efficiency (mean mKate/GFP fluorescence) (**C**), respectively, of all 218 chimeras with the properties of the three parental constructs highlighted in color. Error bars represent the SD of measurements from, at least, quadruplicate replicates with each replicate representing >150 transfected cells. Each chimera is ranked according to its performance for each property (expression, localization, and localization efficiency) in ascending order. (**D**) shows the contiguous (contig) and non-contiguous (non-contig) ten-block library designs with each block in a different color aligned with a schematic of the ChR secondary structure. The block coloring of the contig and non-contig block designs match **Figure 4.2** and **Supplementary Figure 4.1,** although, for clarity, the conserved locations are not shown in gray. Block boundaries (white lines) for the combined contiguous and non-contiguous library designs are shown on the three

parents below the individual library designs. (**E-G**) show the block identity of the chimeras ranked according to their performance for each given property with the best ranking chimera at the top of the list. Each row represents a chimera. The colors represent the parental origin of the block (red – CsChrimR, green – C1C2, and blue – CheRiff).

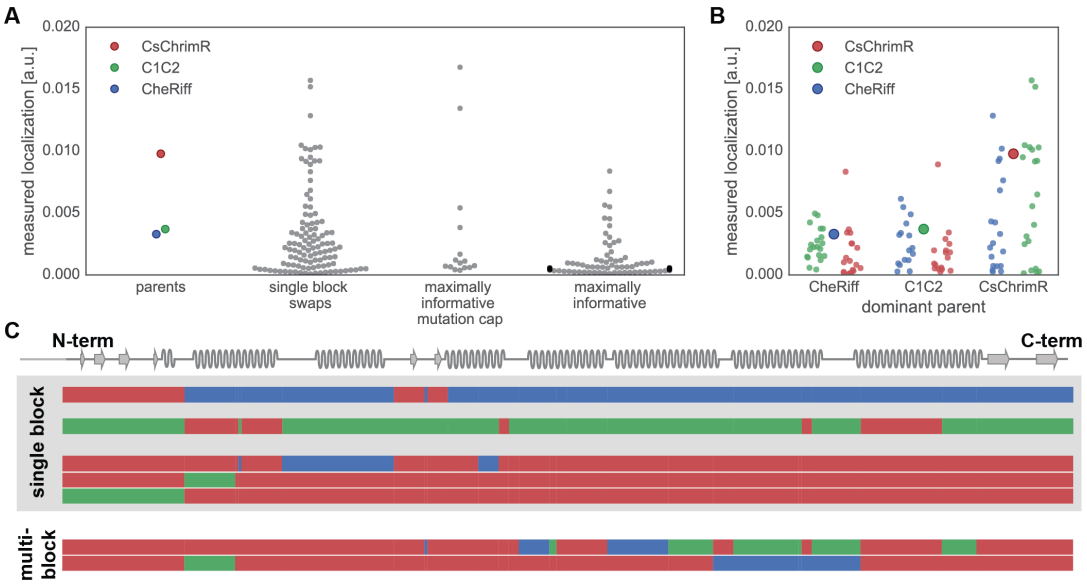


Figure 4.4. Comparison of membrane localization for different chimeras. (**A**) Swarm plots of measured localization for the parent constructs and each chimera set: single-block swaps, maximally informative with mutation cap, and maximally informative. Chimera data are plotted as gray points; parental data are highlighted in color. (**B**) Comparison of measured localization of single-block-swap chimeras relative to their dominant parent. Each single-block-swap chimera is grouped based on the dominant parent with data points colored according to the identity of the single block being swapped into the dominant parent (red – CsChrimR block, green – C1C2 block, and blue – CheRiff block). The large point in each group shows the performance of the dominant parent. (**C**) Shows the block identity of selected single-block-swap and multi-block-swap chimeras aligned with the ChR secondary structure. The top two single-block-swap chimeras are the top performing chimeras for the CheRiff and C1C2 dominant parents. The bottom three single-block-swap chimeras are the top performing single-block swaps in the CsChrimR dominant parent. The two multi-block-swap chimeras are the top two performing chimeras in the 'maximally informative with mutation cap chimera set'. Each row represents a chimera. The three different colors represent blocks from the three different parents (red – CsChrimR, green – C1C2, and blue – CheRiff).

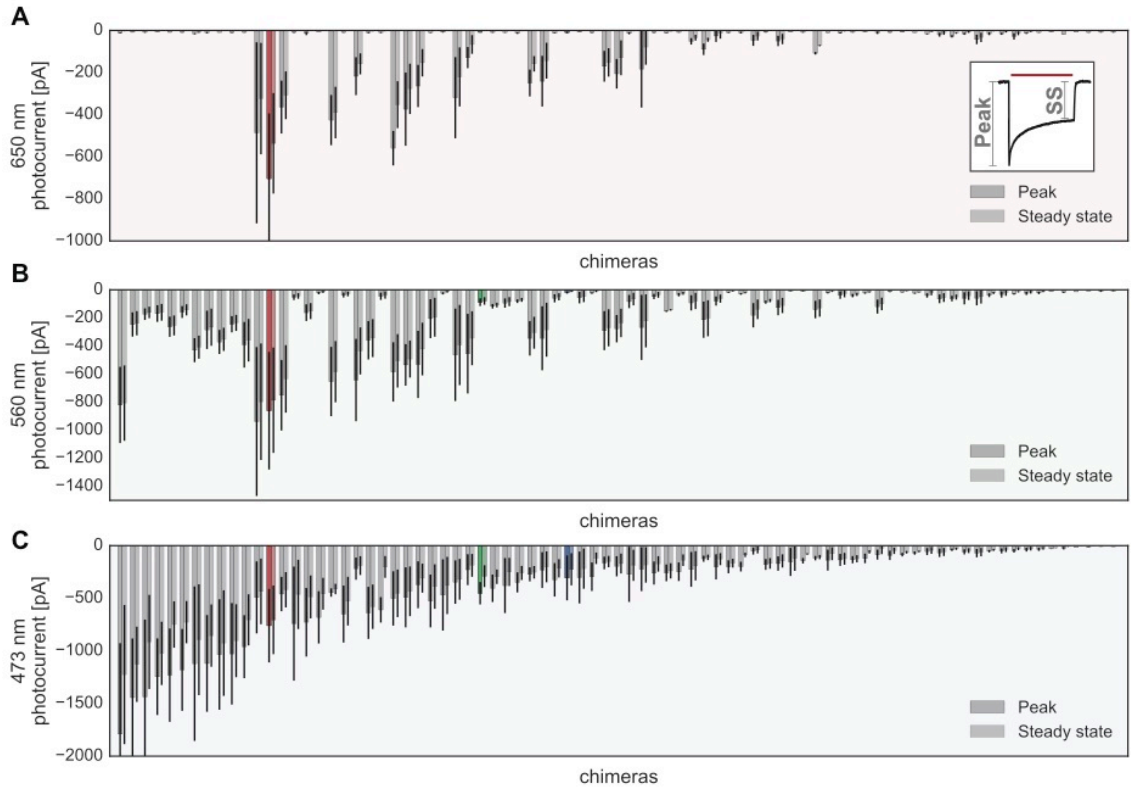


Figure 4.5. Chimera photocurrents with 650 nm, 560 nm, and 473 nm light. Peak and steady-state photocurrents induced by a 1 s exposure to 650 nm (**A**: red shading), 560 nm (**B**: green shading), and 473 nm (**C**: blue shading) wavelength light for each chimera measured. Inset shows the canonical ChR peak vs steady- state (SS) inward current observed when the channel is exposed to light. All chimera data are plotted as gray bars and parental data are highlighted in color (red – CsChrimR, green – C1C2, and blue – CheRiff). Peak and steady-state current are measured for N = 4-10 cells for each chimera. Bars show the mean and error bars represent SD of measured cells for both peak and steady-state current.

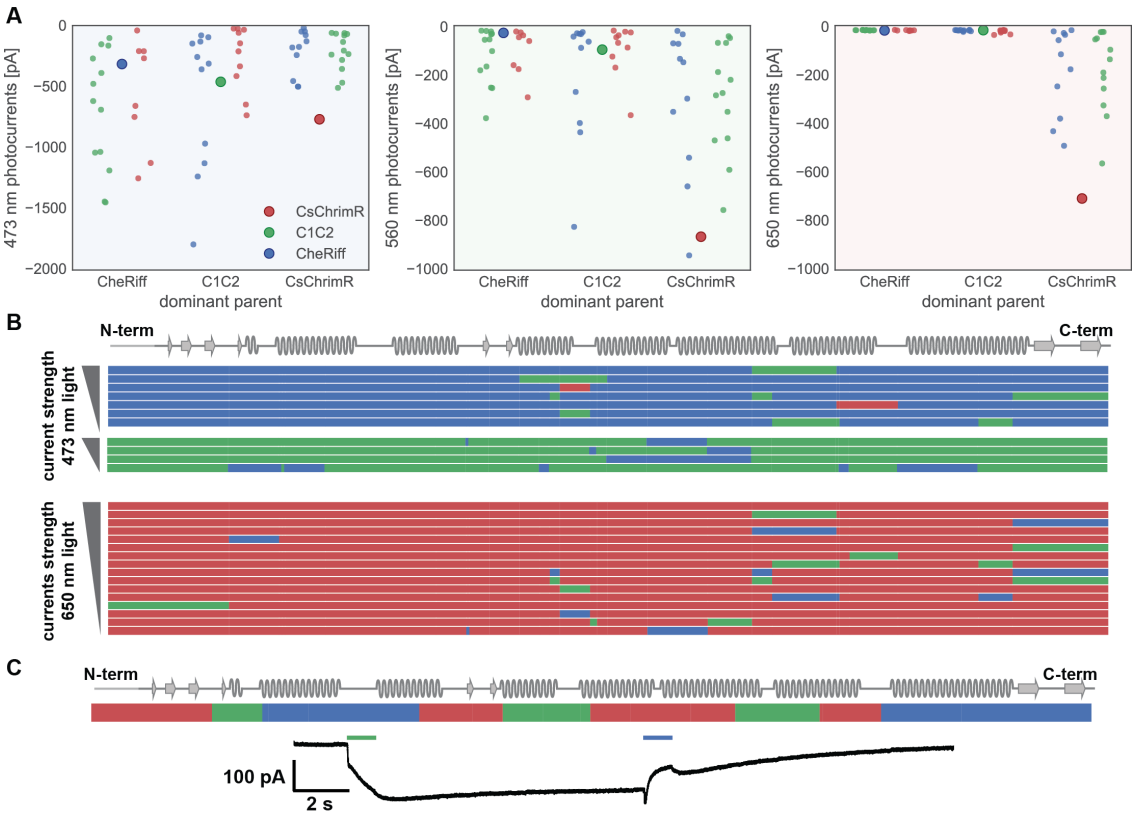


Figure 4.6. Comparison of chimeras with significantly altered photocurrent properties. (**A**) Peak photocurrent for each single-block-swap chimera grouped based on the dominant parent with data points colored based on the identity of the single block being swapped in (red – CsChrimR block, green – C1C2 block, and blue – CheRiff block). The large point in each group shows the performance of the dominant parent. (**B**) Shows the block identity of top performing single-block-swap chimeras aligned with the ChR secondary structure. Single-block-swap chimeras that outperform CsChrimR with 473 nm light are shown (top six performing single-block-swap chimeras with the CheRiff dominant parent and the top four performing single-block-swap chimeras with the C1C2 dominant parent). All chimeras that produce photocurrents >50 pA upon 650 nm light exposure are also shown. These single-block-swap chimeras all have the CsChrimR dominant parent. Chimeras are grouped based on the identity of the dominant parent and ranked based on photocurrent with either 473 nm light or 650 nm light. For the non-contiguous design, a single (structural) block may be disconnected along the primary sequence. Thus single-

block-swap chimeras from the non-contiguous library may have new sequence elements in more than one location along the primary sequence. Each row represents a chimera. The colors represent the parental origin of the block (red – CsChrimR, green – C1C2, and blue – CheRiff). (**C**) One multi-block-swap chimera has novel light-activation properties relative to the parents. This ChR chimera is activated by 560 nm light and closes with 473 nm light. The chimera block identity is shown.

4.7 Supporting information

4.7.1 Parental ChR constructs

Each of the three ChR library parent genes was built using a consistent vector backbone (pFCK) with the same promoter (CMV), trafficking signal (TS) sequence, and fluorescent protein (mKate). We used the pFCK vector from the construct FCK-CheRiff-eGFP [addgene plasmid #51693 (14)]. A TS sequence (43) was inserted between the opsin and the fluorescent protein. The TS sequence has been shown to enhance opsin membrane trafficking (43). The GFP was replaced with mKate2.5 (96). Use of a red fluorescent protein as the marker for the opsin expression enabled use of SpyCatcher-GFP labeling for membrane-localized proteins. mKate2.5 is a monomeric far-red fluorescent protein that shows no aggregation. The mKate2.5 sequence was synthesized by IDT with overhangs for cloning into the desired vector system.

For the SpyTag/SpyCatcher membrane localization assay it was necessary to add the SpyTag sequence close to the N-terminus of each of the parental proteins and C-terminal to the signal peptide sequence cleavage site. For C1C2 an optimal position of the SpyTag had already been published. The SpyTag-C1C2 gene was amplified from the construct pLenti-CaMKIIa-SpyTag-C1C2-TS-mCherry (78) and inserted into the pFCK backbone. For CheRiff and CsChrimR, it was necessary to test various N-terminal SpyTag locations. The CheRiff gene was first amplified from FCK-CheRiff-eGFP [addgene plasmid #51693 (14)] and the SpyTag sequence was added at different N-terminal positions by assembly PCR methods. The CsChrimR gene was built by assembly of the Cs N-terminal sequence (synthesized by IDT) with the C-terminal end of ChrimsonR amplified from the FCK-ChrimsonR-GFP construct [addgene plasmid #59049 (41)]. The sequence of CsChrimR was designed to be identical to the previously published sequence (41). The SpyTag sequence was then inserted at different positions in the N-terminal region of the protein using assembly PCR methods. We tested 3 different pFCK-SpyTag-CheRiff-TS-mKate designs and three different pFCK-SpyTag-CsChrimR-TS-mKate designs and selected the

design that showed expression and localization levels most similar to the non-tagged parent.

Assembly-based methods and traditional cloning were used for vector construction and parental gene insertion. Annotated GenBank files are included as supplemental materials for the three SpyTagged parental constructs used in this study.

4.7.2 Library design

SCHEMA was used to design recombination libraries of the three parental ChRs to minimize the library-average disruption of the ChR structure (123, 136, 139). For this library, the SCHEMA predicted block definitions were not modified. This 10-block library had roughly even-length blocks [14-43 residues], a relatively low average E-value, 25, and whose sequences have an average of 73 mutations from the nearest parent. For the non-contiguous library, the SCHEMA predicted block definitions were modified to group the N- or C-terminal domains into single blocks, maintain the presumptive dimer interface, and minimize the number of small blocks (less than 5 mutations). Specifically, a 13-block non-contiguous recombination library was generated for which two N-terminal blocks were combined, two C-terminal blocks were combined, two of four blocks in TM 5 were combined, and two residues of TM 3 were switched to the same block as TM 4 (where TM 3 and 4 make up the dimer interface observed for C1C2). The two loops that were not modeled in the C1C2 structure, between TM 1 and TM 2 and in the beta-turn of the C-terminal motif, were added to the block containing TM2 and the C-terminal block, respectively. The un-modeled residues of the N- and C-termini were added to the N- and C-terminal blocks. The resulting non-contiguous library had 10 blocks, an average E-value of 23, an average of 71 mutations, and block size similar to the contiguous library (**Figure 4.2*C*,*D***).

Among the three ChR parents, 5 unique N-linked glycosylation sites have been predicted by the NetNGlyc 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/) and GlycoEP servers (150). C1C2 harbors four of these sites with by far the highest confidence at each site. With one exception, the putative N-linked glycosylation sites do not overlap with recombination

block borders. The exception site (SpyTag-C1C2 N95) is located in between the N-terminal domain and the first TM helix.

Contiguous recombination design was done using a software package for calculating SCHEMA energies and running the RASPP algorithm (134) openly available at http://cheme.che.caltech.edu/groups/fha/Software.htm (151). Non-contiguous recombination design was done using a software package for performing non-contiguous protein recombination (135) openly available at http://cheme.che.caltech.edu/groups/fha/Software.htm (152). Both software packages are written in the Python programming language.

4.7.3 Construction of chimeras

The SCHEMA software outputs the amino acid sequences of all chimeras in a library. The amino acid sequence for each chimera chosen for experimental testing was converted into a nucleotide sequence using the following method to define codon usage:

1. Align the amino acid sequence to the C1C2 parent.

2. Assign conserved amino acids in the alignment to the C1C2 parental codon.

3. Assign non-conserved amino acids to the parental codon from which the amino acid is derived.

This method was used for all chimeras to ensure that codon usage was consistent. Once amino acid sequences were converted into nucleotide sequences, additional 3' and 5' sequences containing a BamHI and a NotI restriction enzyme cut site, respectively, were appended to the gene sequence. These sequences were necessary for cloning in the pFCK vector using either restriction ligation or homology-based cloning strategies. Gene sequences for the 223-chimera set were synthesized by Twist Bioscience, Inc. using its proprietary silicon-based DNA writing technology. After assembly, each fragment was cloned in the pFCK vector by homology based cloning strategy and transformed into Stbl3 cells (Invitrogen) or Endura cells (Lucigen). Individual clones were picked and sequenced

by NGS. Perfect clones were stored as individual glycerol stocks. Eight of the single-block swap sequences failed either the synthesis or cloning steps; these were not included in the chimera set.

Purified plasmid DNA of each chimera was prepared for HEK cell transfection. Each construct was streaked onto LB-amp plates from a glycerol stock, an individual colony from each construct was picked and used to inoculate a 5 ml LB-ampicillin liquid media. Cultures were then grown overnight to reach saturation. Plasmid DNA for each construct was then purified using the QIAprep Spin Miniprep Kit. DNA concentrations for all constructs were measured and normalized prior to HEK cell transfection.

4.7.4 HEK cell maintenance and transfection

HEK 293T cells were cultured at 37^{o}C and 5% CO_2 in D10 (Dulbecco's modified Eagle medium (DMEM) supplemented with 10% (vol/vol) FBS, 1% sodium bicarbonate, and 1% sodium pyruvate). For 96-well transfections, HEK cells were plated on PolyDLysine-coated glass-bottom 96-well plates at 20-30% confluency. Cells were left to divide until they reached 70-80% confluency. HEK cells were then transfected with one library variant per well at a pre-normalized DNA concentration using Fugene6 reagent according to the manufacturer's recommendations. Cells were given 48 hours to express and then subjected to the SpyCatcher-GFP labeling assay and imaged.

4.7.5 Recombinant SpyCatcher-GFP expression and purification

The SpyCatcher-GFP was produced from a previously published construct – *pQE80l-T5::6xhis-SpyCatcher-Elp-GFP* – for details see Bedbrook *et al.*. *E. coli* expression strain *BL21(DE3)* harboring the *pQE80l-T5::6xhis-SpyCatcher-Elp-GFP* plasmid was grown at 37^{o}C in TB medium to an optical density of 0.6-0.8 at 600 nm, and protein expression was induced using 1 mM Isopropyl β-D-1-thiogalactopyranoside at 30^{o}C. After 4 hours of induction, cells were harvested and frozen at -80^{o}C prior to protein purification. Protein purification was carried out using HiTrap columns (GE Healthcare, Inc.) following the

column manufacture's recommendations. Protein was buffer exchanged into sterile PBS at 4^{o}C. Protein was stable through multiple freeze/thaws and over many months.

4.7.6 SpyCatcher labeling of HEK cells

HEK cells were subjected to SpyCatcher labeling 48 hours post-transfection. Labeling was done in a 96-well format using multichannel pipettes. SpyCatcher-GFP was added directly into the D10 media of wells containing HEK cells at a final concentration of 30 μM and the cells were then incubated for 45 min at 25^{o}C. To avoid variability in labeling in the 96-well format screen, we used a saturating concentration of the SpyCatcher (30 μM) for labeling experiments. After labeling, HEK cells were washed with D10 three times, and then cells were incubated at 37^{o}C for 1 hour to allow any remaining SpyCatcher to diffuse off of the well surface. For cell imaging, D10 media was replaced with extracellular buffer (in mM: 140 NaCl, 5 KCl, 10 HEPES, 2 $MgCl_2$, 2 $CaCl_2$, 10 glucose; pH 7.35) to avoid the high autofluorecence of the D10. Cells were washed two times with extracellular buffer to fully remove any residual D10 before imaging.

4.7.7 Imaging and image processing of ChR expression and localization

Imaging of ChR expression and localization was done using a Leica DMI 6000 microscope. Four positions in each well were imaged in all 96-well plates using a fully-automated system with motorized stage and automated z-focus. Three channels were imaged at each position (mKate, GFP, and bright-field). Cell segmentation was done using CellProfiler (153), an open source image processing software, and whole population intensity measurements was done using custom image processing scripts written using open-source packages in the SciPy ecosystem (154-156). Both processing methods require a series of filtering steps and background subtraction. Whole population intensity measurements required a thresholding step when defining a pixel mask for image processing. We used wells containing non-transfected HEK cell that went through the labeling experiment as a background for establishing a threshold. A threshold was set to two standard deviations above the mean intensity values calculated in these background wells for each channel (mKate and GFP). For each image, a mask was defined for each

channel (mKate and GFP) as the pixels above a set threshold. The masks for the two channels were then combined so that the mask included any pixel that was above threshold in the GFP channel or the mKate channel. This combined pixel mask was used to calculate the mean mKate fluorescence intensity (expression) and mean GFP fluorescence intensity (localization) across the pixels in the mask. The ratio mean mKate intensity / mean GFP intensity is the localization efficiency.

4.7.8 Electrophysiology for ChR photocurrents

Conventional whole-cell patch-clamp recordings were done in cultured HEK cells at two days post transfection. Cells were continuously perfused with extracellular solution at room temperature (in mM: 140 NaCl, 5 KCl, 10 HEPES, 2 $MgCl_2$, 2 $CaCl_2$, 10 glucose; pH 7.35) while mounted on the microscope stage. Patch pipettes were fabricated from borosilicate capillary glass tubing (1B150-4; World Precision Instruments, Inc., Sarasota. FL) using a model P-2000 laser puller (Sutter Instruments) to resistances of 2-5 MΩ. Pipettes were filled with intracellular solution (in mM): 134 K gluconate, 5 EGTA, 10 HEPES, 2 $MgCl_2$, 0.5 $CaCl_2$, 3 ATP, 0.2 GTP. Whole-cell patch-clamp recordings were made using a Multiclamp 700B amplifier (Molecular Devices, Sunnyvale, CA), a Digidata 1440 digitizer (Molecular Devices), and a PC running pClamp (version 10.4) software (Molecular Devices) to generate current injection waveforms and to record voltage and current traces.

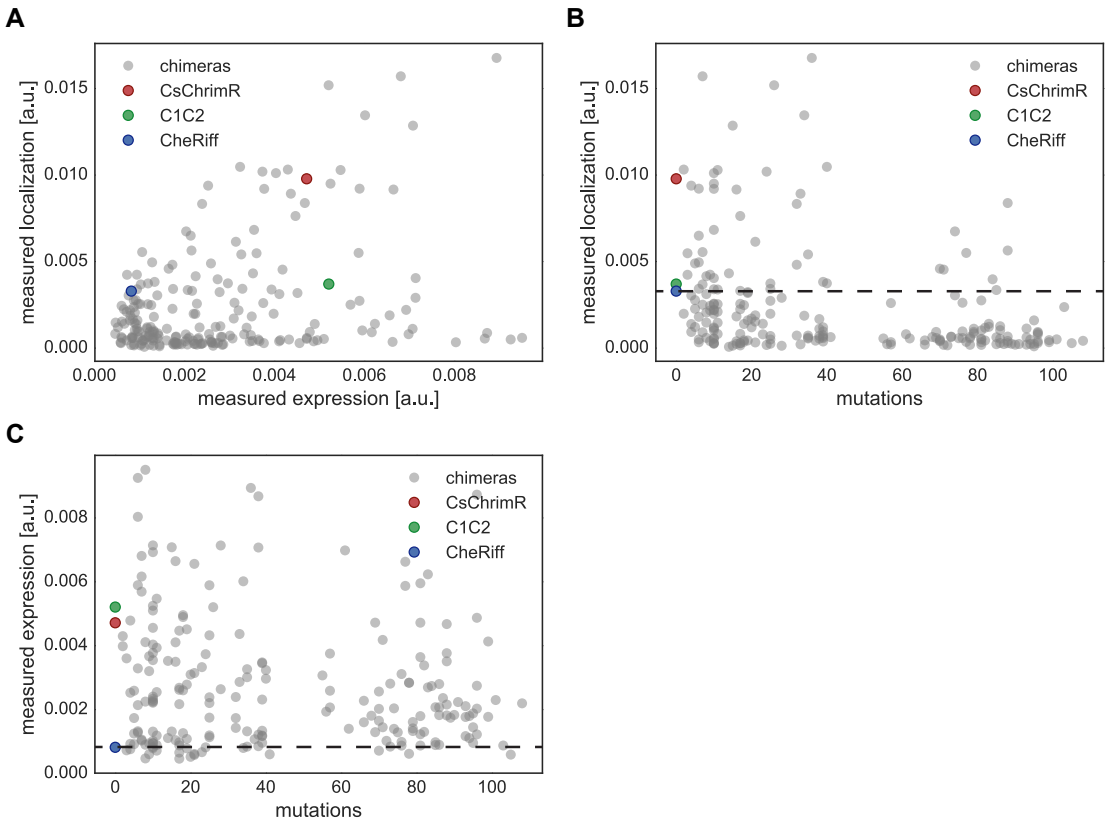
Patch-clamp recordings were done with short light pulses to measure photocurrents. Photocurrents for each chimera were induced by three different wavelengths of light (473±10 nm, 560±25 nm, and 650±13 nm) at 2 mW (~0.1 mW mm^{-2}). Photocurrents were recorded from cells in voltage clamp held at -50 mV with one light pulse for 1 s with each wavelength of light tested sequentially with 2 min between light exposures. Because ChRs show some level of desensitization to light after continued light exposure, we ran all colors in one direction (red → green → blue) and then again in the other direction (blue → green → red). The means peak and steady state currents were calculated for each color between the two trials for a given cell. Light wavelengths were produced using LED illumination using a Lumencor SPECTRAX light engine with quad band 387/485/559/649 nm

excitation filter, quad band 410/504/582/669 nm dichroic mirror and quad band 440/521/607/700 nm emission filter (all SEMROCK).

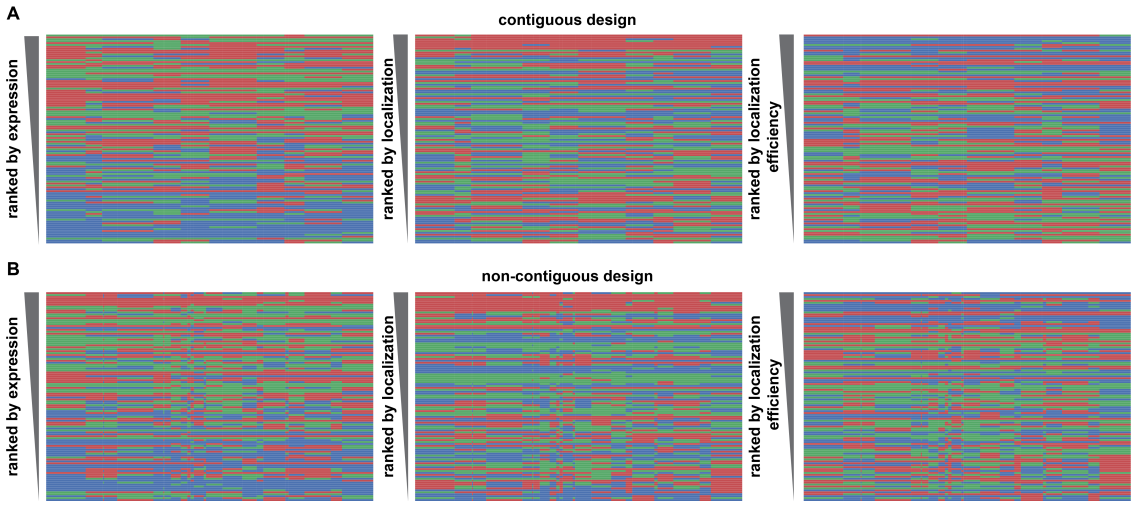
Electrophysiology data was analyzed using custom data processing scripts written using open-source packages in the Python programming language to do baseline adjustments, find the peak inward currents, and find the steady state currents.

4.8 Supplemental figures

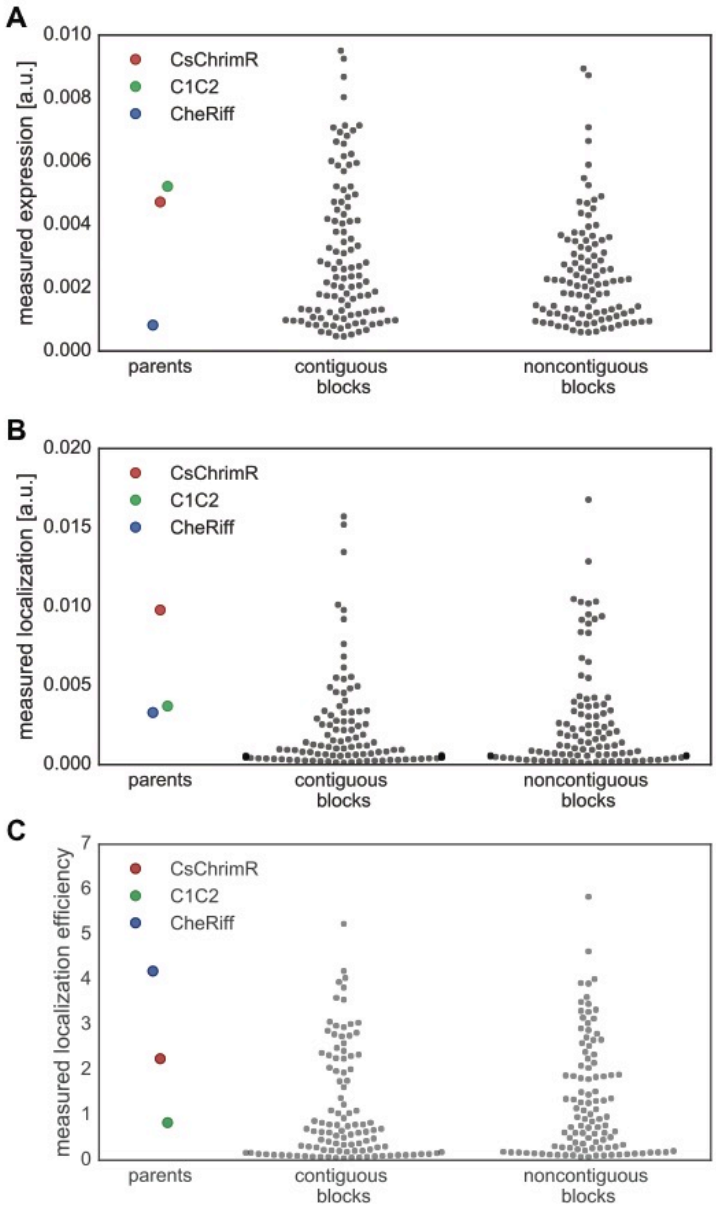
Supplementary Figure 4.1. Amino acid alignment of parental sequences and recombination block designs. Alignment showing the contiguous and non-contiguous block designs. Each color represents a different block, and white shows the conserved residues. Amino acids thought to be important for ChR spectral properties are bolded and underlined. The conserved lysine residue that participates in a Schiff base linkage with retinal is highlighted in red text. The secondary structure is shown below the alignment.



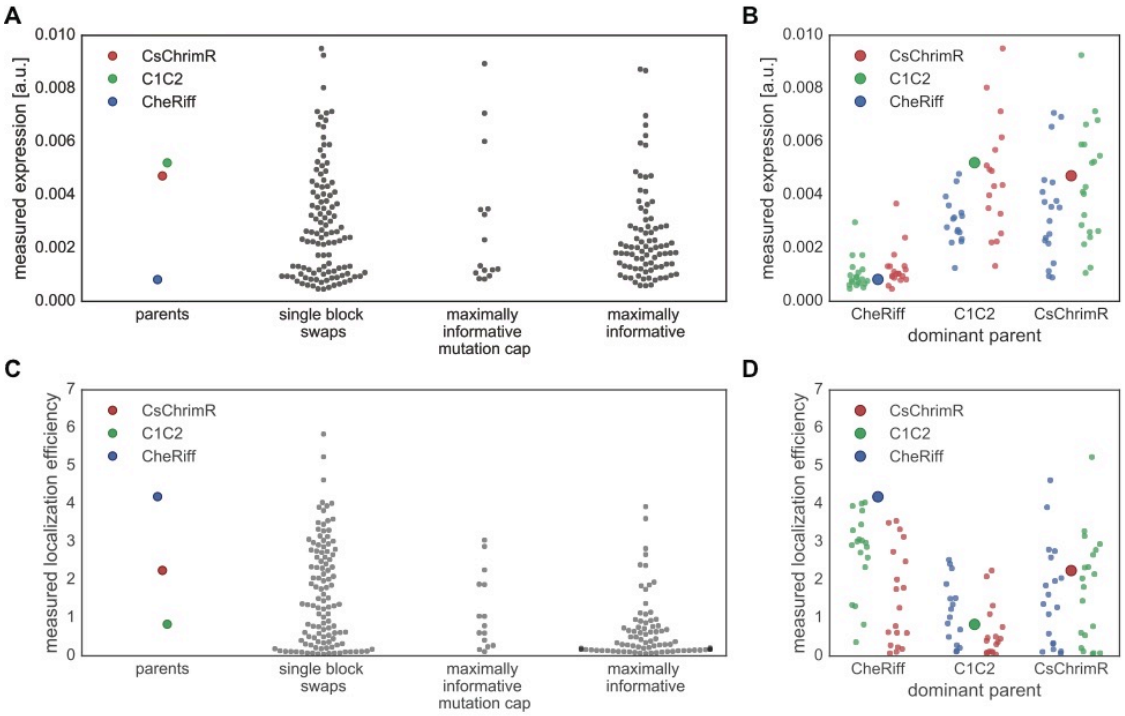
Supplementary Figure 4.2. Interdependencies of chimera properties. Chimera data are plotted as gray points and parental data points are highlighted in color (red – CsChrimR, green – C1C2, and blue – CheRiff). (**A**) Plot of measured localization (mean GFP fluorescence [a.u.]) vs measured expression (mean mKate fluorescence [a.u.]) shows no clear correlation. (**B**) Plot of measured localization vs number of mutations from closest parent. (**C**) Plot of measured expression vs number of mutations from closest parent. Dashed lines in (**B**) and (**C**) show the measured properties of the lowest-performing parent (CheRiff).



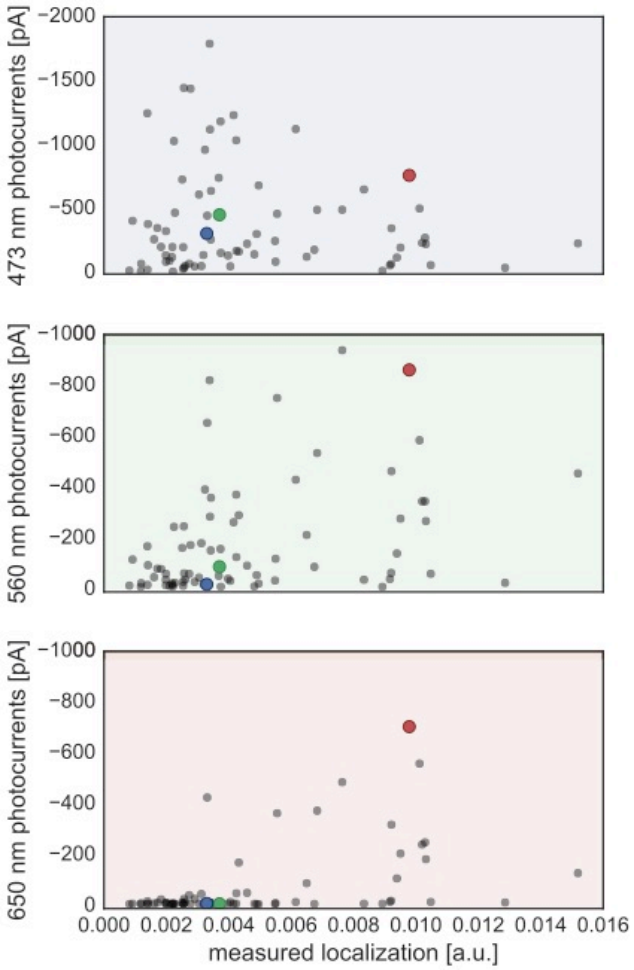
Supplementary Figure 4.3. Chimeras from the contiguous and non-contiguous libraries, ranked by expression, localization, and localization efficiency. Block identity of the chimeras ranked according to performance for each given property with the best ranking chimera at the top of the list for the contiguous (**A**) and non-contiguous (**B**) library chimeras. Each row represents a chimera. The colors represent the parental origin of the block (red – CsChrimR, green – C1C2, and blue – CheRiff). The properties shown are measured expression (mean mKate fluorescence [a.u.]), localization (mean GFP fluorescence [a.u.]), and localization efficiency (mean mKate/GFP fluorescence).



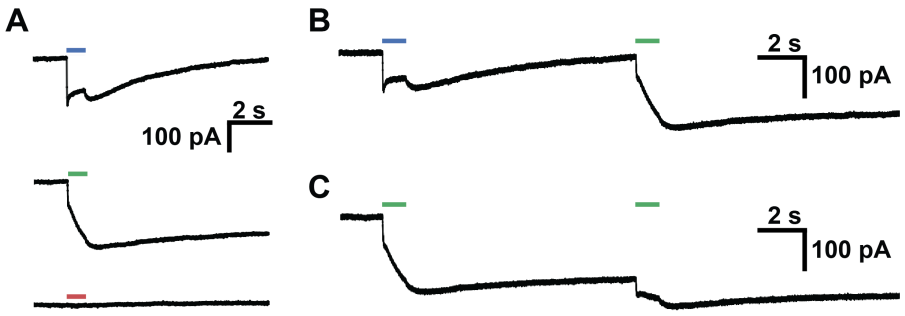
Supplementary Figure 4.4. Comparison of chimeras from the contiguous and non-contiguous recombination libraries. Swarm plot showing each chimera's expression (mean mKate fluorescence [a.u.]) (**A**), localization (mean GFP fluorescence [a.u.]) (**B**), and localization efficiency (mean mKate/GFP fluorescence) (**C**) for the contiguous and non-contiguous recombination libraries. Chimera data are plotted as gray points and parental data points are highlighted in color (red – CsChrimR, green – C1C2, and blue – CheRiff).



Supplementary Figure 4.5. Comparison of measured expression and membrane localization efficiency for each chimera set. Swarm plots of expression (mean mKate fluorescence [a.u.]) (**A**) and localization efficiency (mean mKate/GFP fluorescence) (**C**) showing measurements for each data set compared with parents: single-block swaps, maximally informative with mutation cap, and maximally informative. Chimera data are plotted as gray points and parental data points are highlighted in color (red – CsChrimR, green – C1C2, and blue – CheRiff). Comparison of single-block-swap chimeras measured expression (**B**) and localization efficiency (**D**) relative to the dominant parent. Each single-block-swap chimera is grouped based on the dominant parent with data points colored based on the identity of the single block being swapped in (red – CsChrimR block, green – C1C2 block, and blue – CheRiff block). The large point in each group shows the performance of the dominant parent.



Supplementary Figure 4.6. Photocurrents versus measured localization for all tested chimeras. Chimera data are plotted as gray points and parental data points are highlighted in color (red – CsChrimR, green – C1C2, and blue – CheRiff). Plot of measured photocurrents vs measured localization (mean GFP fluorescence [a.u.]) for three different wavelengths: 473 nm (top – blue shading), 560 nm (middle – green shading), and 650 nm (bottom – red shading).



Supplementary Figure 4.7. One multi-block-swap chimera with unique properties. (**A**) Chimera photocurrents upon 1 s exposure to 473 nm (top), 560 nm (middle), and 650 nm (bottom) light. (**B**) Sequential activation of chimera with 473 nm and then 560 nm light. (**C**) Sequential activation of chimera with 560 nm and then 560 nm light.

Chapter 5

MACHINE LEARNING TO DESIGN INTEGRAL MEMBRANE CHANNELRHODOPSINS FOR EFFICIENT EUKARYOTIC EXPRESSION AND PLASMA MEMBRANE LOCALIZATION

A version of this chapter has been published as (45)

5.1 Abstract

There is growing interest in studying and engineering integral membrane proteins (MPs) that play key roles in sensing and regulating cellular response to diverse external signals. A MP must be expressed, correctly inserted and folded in a lipid bilayer, and trafficked to the proper cellular location in order to function. The sequence and structural determinants of these processes are complex and highly constrained. Here we describe a predictive, machine-learning approach that captures this complexity to facilitate successful MP engineering and design. Machine learning on carefully-chosen training sequences made by structure-guided SCHEMA recombination has enabled us to accurately predict the rare sequences in a diverse library of channelrhodopsins (ChRs) that express and localize to the plasma membrane of mammalian cells. These light-gated channel proteins of microbial origin are of interest for neuroscience applications, where expression and localization to the plasma membrane is a prerequisite for function. We trained Gaussian process (GP) classification and regression models with expression and localization data from 218 ChR chimeras chosen from a 118,098-variant library designed by SCHEMA recombination of three parent ChRs. We use these GP models to identify ChRs that express and localize well and show that our models can elucidate sequence and structure elements important for these processes. We also used the predictive models to convert a naturally occurring ChR incapable of mammalian localization into one that localizes well.

5.2 Introduction

As crucial components of regulatory and transport pathways, integral membrane proteins (MPs) are important pharmaceutical and engineering targets (115). To be functional, MPs must be expressed and localized through a series of elaborate sub-cellular processes that include co-translational insertion, rigorous quality control, and multi-step trafficking to arrive at the correct topology in the correct sub-cellular location (120, 121, 157). With such a complex mechanism for production, it is not surprising that MP engineering has been hampered by poor expression, stability, and localization in heterologous systems (44, 118, 158). To overcome these limitations, protein engineers need a tool to predict how changes in sequence affect MP expression and localization. An accurate predictor would enable us to design and produce MP variants that express and localize correctly, a necessary first step in engineering MP function. A useful predictor would be sensitive to subtle changes in sequence that can lead to drastic changes in expression and localization. Our goal here was to develop data-driven models that predict the likelihood of a MP's expression and plasma membrane localization using the amino acid sequence as the primary input.

For this study, we focus on channelrhodopsins (ChRs), light-gated ion channels that assume a seven transmembrane helix topology with a light-sensitive retinal chromophore bound in an internal pocket. This scaffold is conserved in both microbial rhodopsins (light-driven ion pumps, channels, and light sensors – type I rhodopsins) and animal rhodopsins (light-sensing G-protein coupled receptors – type II rhodopsins) (7). Found in photosynthetic algae, ChRs function as light sensors in phototaxic and photophobic responses (124, 125). On photon absorption, ChRs undergo a multi-step photo-cycle that allows a flux of ions across the membrane and down the electrochemical gradient (126). When ChRs are expressed transgenically in neurons, their light-dependent activity can stimulate action potentials, allowing cell-specific control over neuronal activity (10, 127) and extensive applications in neuroscience (117). The functional limitations of available ChRs have spurred efforts to engineer or discover novel ChRs (126). The utility of a ChR, however, depends on its ability to express and localize to the plasma membrane in eukaryotic cells of interest, and changes to the amino acid sequence frequently abrogate

localization (44). A predictor for ChRs that express and localize would be of great value as a pre-screen for function.

The sequence and structural determinants for membrane localization have been a subject of much scientific investigation (159-161) and have provided some understanding of the MP sequence elements important for localization, such as signal peptide sequence, positive charge at the membrane–cytoplasm interface (the "positive-inside" rule (162)), and increased hydrophobicity in the transmembrane domains. However, these rules are of limited use to a protein engineer: there are too many amino acid sequences that follow these rules but still fail to localize to the plasma membrane (see **5.3 Results**). MP sequence changes that influence expression and localization are highly context-dependent: what eliminates localization in one sequence context has no effect in another, and subtle amino acid changes can have dramatic effects (44, 160, 163). In short, sequence determinants of expression and localization are not captured by simple rules.

Accurate atomistic physics-based models relating a sequence to its level of expression and plasma membrane localization currently do not exist, in large measure due to the complexity of the process. Statistical models offer a powerful alternative. Statistical models are useful for predicting the outcomes of complex processes because they do not require prior knowledge of the specific biological mechanisms involved. That being said, statistical models can also be constructed to exploit prior knowledge, such as MP structural information. Statistical models can be trained using empirical data (in this case expression or localization values) collected from known sequences. During training, the model infers relationships between input (sequence) and output (expression or localization) that are then used to predict the properties of unmeasured sequence variants. The process of using empirical data to train and select statistical models is referred to as machine learning.

Machine learning has been applied to predicting various protein properties, including solubility (164, 165), trafficking to the periplasm (166), crystallization propensity (167), and function (168). Generally, these models are trained using large data sets composed of literature data from varied sources with little to no standardization of the experimental

conditions, and trained using many protein classes (i.e. proteins with various folds and functions), because their aim is to identify sequence elements across all proteins that contribute to the property of interest. This generalist approach, however, is not useful for identifying subtle sequence features (i.e. amino acids or amino acid interactions) that condition expression and localization for a specific class of related sequences, the ChRs in this case. We focused our model building on ChRs, with training data collected from a range of ChR sequences under standardized conditions. We applied Gaussian process (GP) classification and regression (169) to build models that predict ChR expression and localization directly from these data.

In our previous work, GP models successfully predicted thermal stability, substrate binding affinity, and kinetics for several soluble enzymes (170). Here, we asked whether GP modeling could accurately predict mammalian expression and localization for heterologous integral membrane ChRs and how much experimental data would be required. For a statistical model to make accurate predictions on a wide range of ChR sequences, it must be trained with a diverse set of ChR sequences (169). We chose to generate a training set using chimeras produced by SCHEMA recombination, which was previously demonstrated to be useful for producing large sets (libraries) of diverse, functional chimeric sequences from homologous parent proteins (132). We synthesized and measured expression and localization for only a small subset (0.18%) of sequences from the ChR recombination library. Here we use these data to train GP classification and regression models to predict the expression and localization properties of diverse, untested ChR sequences. We first made predictions on sequences within a large library of chimeric ChRs; we then expanded the predictions to sequences outside that set.

5.3 Results

The ChR training set. The design and characterization of the chimeric ChR sequences used to train our models have been published (44); we will only briefly describe these results. Two separate, ten-block libraries were designed by recombining three parental ChRs (CsChrimsonR (CsChrimR) (41), C1C2 (23), and CheRiff (14)) with 45-55% amino

acid sequence identity and a range of expression, localization, and functional properties (**Supplementary Figure 5.1**) (44). Each chimeric ChR variant in these libraries is composed of blocks of sequence from the parental ChRs. These libraries were prepared by the SCHEMA algorithm to define sequence blocks for recombination that minimize the library-average disruption of tertiary protein structure (123, 139). One library swaps contiguous elements of primary structure (contiguous library), and the second swaps elements that are contiguous in the tertiary structure but not necessarily in the sequence (non-contiguous library (135)). The two libraries have similar, but not identical, element boundaries (**Supplementary Figure 5.1A**) and were constructed in order to test whether one design approach was superior to the other (they gave similar results). These designs generate 118,098 possible chimeras (2×3^{10}), which we will refer to as the recombination library throughout this paper. Each of these chimeras has a full N-terminal signal peptide from one of the three ChR parents.

Two hundred and eighteen chimeras from the recombination library were chosen as a training set, including all the chimeras with single-block swaps (chimeras consisting of 9 blocks of one parent and a single block from one of the other two parents) and multi-block-swap chimera sequences designed to maximize mutual information between the training set and the remainder of the chimeric library. Here, the 'information' a chimera has to offer is how its sequence, relative to all previously tested sequences, changes ChR expression and localization. By maximizing mutual information, we select chimera sequences that provide the most information about the whole library by reducing the uncertainty (Shannon entropy) of prediction for the remainder of the library, as described in (140, 146). The 112 single-block-swap chimeras in the training set have an average of 16 mutations from the most closely related parent, while the 103 multi-block-swap chimeras in the training set have an average of 73 mutations from the most closely related parent (**Table 5.1**). While the multi-block-swap chimeras provide the most sequence diversity to learn from, they are the least likely to express and localize given their high mutation levels. The single-block-swap chimeras offer less information to learn from due to their sequence redundancies with other chimeras in the training set, but are more likely to express and localize.

Genes for these sequences were synthesized and expressed in human embryonic kidney (HEK) cells, and their expression and membrane localization properties were measured (**Supplementary Figure 5.1B**) (44). The expression levels were monitored through a fluorescent protein (mKate) fused to the C-termini of the ChRs. Plasma-membrane localization was measured using the SpyTag/SpyCatcher labeling method, which exclusively labels ChR protein that has its N terminus exposed on the extracellular surface of the cell (78). The training set sequences displayed a wide range of expression and localization properties. While the majority of the training set sequences express, only 33% of the single-block-swap chimeras localize well, and an even smaller fraction (12%) of the multi-block-swap chimeras localize well, emphasizing the importance of having a predictive model for membrane localization.

First we explored whether ChR chimera properties could be predicted based on basic biological properties, specifically, signal peptide sequence and hydrophobicity in the transmembrane (TM) domains. Each chimera in the library has one of the three parental signal peptides. Although the signal peptide sequence does affect expression and localization (**Supplementary Figure 5.2A**), chimeras with any parental signal peptide can have high or low expression and localization. Thus, the identity of the signal peptide alone is insufficient for accurate predictions of the ChR chimera properties. We then calculated the level of hydrophobicity within the 7-TM domains of each chimera. With very weak correlation between increasing hydrophobicity and measured expression and localization (**Supplementary Figure 5.2B**), hydrophobicity alone is also insufficient for accurate prediction of ChR chimera properties. These models do not accurately account for the observed levels of expression or localization (**Supplementary Figure 5.1**). Therefore, we need more expressive models to predict expression and localization from the amino acid sequences of these MPs.

Using GP models to learn about ChRs. Our overall strategy for developing predictive machine-learning models is illustrated in **Figure 5.1**. The goal is to use a set of ChR sequences and their expression and localization measurements to train GP regression and classification models that describe how ChR properties depend on sequence and predict the

behavior of untested ChRs. GP models infer predictive values from training examples by assuming that similar inputs (ChR sequence variants) will have similar outputs (expression or localization). We quantify the relatedness of inputs (ChR sequence variants) by comparing both sequence and structure. ChR variants with few differences are considered more similar than ChR variants with many differences. We define the sequence similarity between two chimeras by aligning them and counting the number of positions at which they are identical. For structural comparisons, a residue-residue 'contact map' was built for each ChR variant, where two residues are in contact if they have any non-hydrogen atoms within 4.5 Å. The maps were generated using a ChR parental sequence alignment and the C1C2 crystal structure, which is the only available ChR structure (23), with the assumption that ChR chimeras share the overall contact architecture observed in the C1C2 crystal structure. The structural similarity for any two ChRs was quantified by aligning the contact maps and counting the number of identical contacts (170). Using these metrics, we calculated the sequence and structural similarity between all ChRs in the training set relative to one another (218 x 218 ChR comparisons).

These similarity functions are called kernel functions and specify how the functional properties of pairs of sequences are expected to covary (they are also known as covariance functions). In other words, the kernel is a measure of similarity between sequences, and we can draw conclusions about unobserved chimeras on the basis of their similarity to sampled points (169). The model has high confidence in predicting the properties of sequences that are similar to previously sampled sequences, and the model is less confident in predicting the properties of sequences that are distant from previously sampled sequences.

To build a GP model, we must also specify how the relatedness between sequences will affect the property of interest, in other words how sensitive the ChR properties are to changes in relatedness as defined by the sequence/structure differences between ChRs. This is defined by the form of the kernel used. We tested three different forms of sequence and structure kernels: linear kernels, squared exponential kernels, and Matérn kernels (see **5.5 Methods**). These different forms represent the kinds of functions we expect to observe for the protein's fitness landscape (i.e. the mapping of protein sequence to protein function).

The linear kernel corresponds to a simple landscape where the effects of changes in sequence/structure are additive and there is no epistasis. The two non-linear kernels represent more rugged, complex landscapes where effects may be non-additive. Learning involves optimizing the form of the kernel and its hyperparameters (parameters that influence the form of kernel) to enable accurate predictions. The hyperparameters and the form of the kernel were optimized using the Bayesian method of maximizing the marginal likelihood of the resulting model. The marginal likelihood (i.e. how likely it is to observe the data given the model) rewards models that fit the training data well while penalizing model complexity to prevent overfitting.

Once trained with empirical data, the output of the GP regression model is a predicted mean and variance, or standard deviation, for any given ChR sequence variant. The standard deviation is an indication of how confident the model is in the prediction based on the relatedness of the new input relative to the tested sequences.

We used GP models to infer links between ChR properties and ChR sequence and structure from the training data. We first built GP binary classification models. In binary classification, the outputs are class labels i.e. ‘high’ or ‘low’ localization, and the goal is to use the training set data to predict the probability of a sequence falling into one of the two classes (**Figure 5.1**). We also built a GP regression model that makes real-valued predictions, i.e. amount of localized protein, based on the training data (**Figure 5.1**). After training these models, we verify that their predictions generalize to sequences outside of the training set. Once validated, these two models can be used in different ways. A classification model trained from localization data can be used to predict the probability of highly diverse sequences falling into the ‘high’ localization category (**Figure 5.1**). The classification model can only predict if a sequence has ‘high’ vs ‘low’ localization, and it cannot be used to optimize localization. The regression model, on the other hand, can be used to predict sequences with ‘optimal’ properties; for example, a regression model trained from localization data can predict untested sequences that will have very high levels of localization (**Figure 5.1**).

Building GP classification models of ChR properties. The training set data (**Supplementary Figure 5.1**) were used to build a GP classification model that predicted which of the 118,098 chimeras in the recombination library would have 'high' vs 'low' expression, localization, and localization efficiency. The training set includes multi-block swaps chosen to be distant from other sequences in the training set in order to provide information on sequences throughout the recombination library. A sequence was considered 'high' if it performed at least as well as the lowest performing parent, and it was considered 'low' if it performed worse than the lowest performing parent. Because the lowest performing parent for expression and localization, CheRiff, is produced and localized in sufficient quantities for downstream functional studies, we believe this to be an appropriate threshold for 'high' vs 'low' performance. For all of the classification models (**Figure 5.2** and **Supplementary Figure 5.3**), we used kernels based on structural relatedness. For the expression classification model, we found that a linear kernel performed best, i.e. achieved the highest marginal likelihood. This suggests that expression is best approximated by an additive model weighting each of the structural contacts. Localization and localization efficiency required a non-linear kernel for the model to be predictive. This more expressive kernel allows for non-linear relationships and epistasis and also penalizes differing structural contacts more than the linear kernel. This reflects our intuitive understanding that localization is a more demanding property to tune than expression, with stricter requirements and a non-linear underlying fitness landscape.

Most of the multi-block-swap sequences from the training set did not localize to the membrane (44). We nonetheless want to be able to design highly mutated ChRs that localize well because these are most likely to have interesting functional properties. We therefore used the localization classification model to identify multi-block-swap chimeras from the library that had a high predicted probability (>0.4) of falling into the 'high' localizer category (**Figure 5.2D**). From the many multi-block-swap chimeras predicted to have 'high' localization, we selected a set of 16 highly diverse chimeras with an average of 69 amino acid mutations from the closest parent and called this the 'exploration' set (**Supplementary Figure 5.4**). We synthesized and tested these chimeras and found that the model had accurately predicted chimeras with good localization (**Figure 5.2** and **Figure**

5.3): 50% of the exploration set show 'high' localization compared to only 12% of the multi-block-swap sequences from the original training set, even though they have similar levels of mutation (**Table 5.1**) (chimeras in the exploration set have on average 69 ± 12 amino acid mutations from the closest parent, versus 73 ± 21 for the multi-block-swap chimeras in the training set). The classification model provides a four-fold enrichment in the number of chimeras that localize well when compared to randomly-selected chimeras with equivalent levels of mutation. This accuracy is impressive given that the exploration set was designed to be distant from any sequence the model had seen during training. The model's performance on this exploration set indicates its ability to predict the properties of sequences distant from the training set.

The data from the exploration set were then used to better inform our models about highly diverse sequences that localize. To characterize the classification model's performance, we calculated the area under the receiver operating characteristic (ROC) curve (AUC). A poorly performing model would not do better than random chance, resulting in an AUC of 0.5, while a model that perfectly separates the two classes will have an AUC of 1.0. The revised models achieved AUC up to 0.87 for "leave-one-out" (LOO) cross-validation, indicating that there is a high probability that the classifiers will accurately separate 'high' and 'low' performing sequences for the properties measured. The AUC is 0.83 for localization, 0.77 for localization efficiency and 0.87 for expression for LOO cross-validation predictions (**Supplementary Figure 5.5**).

To further test the models, we then built a verification set of eleven chimeras, designed using the localization model. This verification set was composed of four chimeras predicted to be highly likely to localize, six chimeras predicted to be very unlikely to localize, and one chimera with a moderate predicted probability of localizing (**Supplementary Figure 5.4**). The measured localization (**Figure 5.2E**) and localization efficiency (**Supplementary Figure 5.3B**) of the chimeras in the verification set show clear differences, 'high' vs 'low', consistent with the model predictions (**Table 5.1**). The verification sets consist exclusively of chimeras with 'high' measured expression, which is consistent with the model's predictions (**Figure 5.2B**). The model perfectly classifies the eleven chimeras as either

'high' or 'low' for each property (expression, localization, or localization efficiency) as shown in plots of predicted vs measured properties (**Figure 5.2B** and **2E** and **Supplementary Figure 5.3B**) and by perfect separation in ROC curves i.e. AUC = 1.0 (**Supplementary Figure 5.5**). These models are powerful tools that can confidently predict whether a chimera will have 'high' or 'low' expression (**Figure 5.2C**), localization (**Figure 5.2F**), and localization efficiency (**Supplementary Figure 5.3C**). Of the 118,098 chimeras in the recombination library, 6,631 (5.6%) are predicted to have a probability > 0.5 of 'high' localization, whereas the vast majority of chimeras (99%) are predicted to have a probability > 0.5 of 'high' expression.

Building a regression model for ChR localization. The classification model predicts the probability that a sequence falls into the 'high' localizer category, but does not give a quantitative prediction as to how well it localizes. Our next goal was to design chimera sequences with optimal localization. Localization is considered optimal if it is at or above the level of CsChrimR, the best localizing parent, which is more than adequate for *in vivo* applications using ChR functionality to control neuronal activity (41). A regression model for ChR plasma membrane localization is required to predict sequences that have optimal levels of localization. We used the localization data from the training and exploration sets to train a GP regression model (**Figure 5.4A**). The diversity of sequences in the training data allows the model to generalize well to the remainder of the recombination library. For this regression model, we do not use all of the features from the combined sequence and structure information; instead, we used L1 linear regression to select a subset of these features. The L1 linear regression identifies the sequence and structural features that most strongly influence ChR localization. Using this subset of features instead of all of the features improved the quality of the predictions (as determined by cross-validation). This indicates that not all of the residues and residue-residue contacts have a large influence on localization of ChR. We then used a kernel based on these chosen features (specific contacts and residues) for GP regression. The regression model for localization showed strong predictive ability as indicated by the strong correlation between predicted and measured localization for LOO cross-validation (correlation coefficient, R > 0.76) (**Figure 5.4A**). This was further verified by the strong correlation between predicted and measured

values for the previously-discussed verification set (R > 0.9) (**Figure 5.4A**). These cross-validation results suggest that the regression model can be used to predict chimeras with optimal localization.

We used the localization regression model to predict ChR chimeras with optimal localization using the Lower Confidence Bound (LCB) algorithm, in which the predicted mean minus the predicted standard deviation (LB1) is maximized (171). The LCB algorithm maximally exploits the information learned from the training set by finding sequences the model is most certain will be good localizers. The regression model was used to predict the localization level and standard deviation for all chimeras in the library, and from this the LB1 was calculated for all chimeras (**Figure 5.4B**). We selected four chimeras whose LB1 predictions for localization were ranked in the top 0.1% of the library (**Supplementary Figure 5.4**). These were constructed and tested (**Figure 5.3** and **Supplementary Figure 5.6**). Measurements showed that they all localize as well as or better than CsChrimR (**Figure 5.3** and **Figure 5.4A** and **Table 5.1**). Cell population distributions of the optimal set show properties similar to the CsChrimR parent, with one chimera showing a clear shift in the peak of the distribution towards higher levels of localization (**Supplementary Figure 5.7**). These four sequences differ from CsChrimR at 30 to 50 amino acids (**Supplementary Figure 5.4**).

We were interested in how predictive the GP localization models could be with fewer training examples. To assess the predictive ability of the GP models as a function of training set size, we sampled random sets of training sequences from the dataset, trained models on these random sets, then evaluated the model's performance on a selected test set (**Supplementary Figure 5.8**). As few as 100 training examples are sufficient for accurate predictions for both the localization regression and classification models. This analysis shows that the models would have been predictive with even fewer training examples than we chose to use.

Sequence and structure features that facilitate prediction of ChR expression and localization. In developing the GP regression model for localization, we used L1-

regularized linear regression to identify a limited set of sequence and structural features that strongly influence ChR localization (**Figure 5.4**). These features include both inter-residue contacts and individual residues and offer insight into the structural determinants of ChR localization. To better gauge the relative importance of these features, L2-regularized linear regression was used to calculate the positive and negative feature weights, which are proportional to each feature's inferred contribution to localization. While not as predictive as the GP regression model because it cannot account for higher-order interactions between features, this linear model has the advantage of being interpretable.

When mapped onto the C1C2 structure, these features highlight parts of the ChR sequence and structural contacts that are important for ChR localization to the plasma membrane (**Figure 5.5**). Both beneficial and deleterious features are distributed throughout the protein, with no single feature dictating localization properties (**Figure 5.5**). Clusters of heavily weighted positive contacts suggest that having structurally proximal CsChrimR-residue pairs are important in the N-terminal domain (NTD), between the NTD and TM4, between TM1 and TM7, and between TM3 and TM7. CsChrimR residues at the extracellular side of TM5 also appear to aid localization, although they are weighted less than CheRiff residues in the same area. Beneficial CheRiff contacts and residues are found in the C-terminal domain (CTD), the interface between the CTD and TM5-6, and in TM1. C1C2 residues at the extracellular side of TM6 are also positively weighted for localization, as are C1C2 contacts between the CTD and TM3-4 loop. From the negatively weighted contacts, it is clear that total localization is harmed when CheRiff contributes to the NTD or the intracellular half of TM4 and when CsChrimR contributes to the CTD. Interestingly, positive contacts were formed between TM6 from C1C2 and TM7 from CheRiff, but when the contributions were reversed (TM6 from CheRiff TM7 from C1C2) or if CsChrimR contributed TM6, strong negative weights were observed. Not surprisingly, the sequence and structure of optimal localizers predicted by GP regression (**Figure 5.4**) largely agree with the L2 weights (**Supplementary Figure 5.9**).

Using this strategy for model interpretation (L1 regression for feature selection followed by L2 regression), we can also weight the contributions of residues and contacts for ChR

expression (**Supplementary Figure 5.10** and **Supplementary Figure 5.11**). There is some overlap between the heavily weighted features for ChR expression and the features for localization, which is expected because more protein expressed means more protein available for localization. For example, both expression and localization models seem to prefer the NTD from CsChrimR and the extracellular half of TM6 from C1C2, and both disfavor the NTD and the intra-cellular half of TM4 from CheRiff. While the heavily-weighted expression features are limited to these isolated sequence regions, localization features are distributed throughout the protein. Moreover, the majority of heavily-weighted features identified for expression are residues rather than contacts. This is in contrast to those weighted features identified for localization, which include heavily-weighted residues and structural contacts. This suggests that sequence is more important in determining expression properties, which is consistent with the largely sequence-dependent mechanisms associated with successful translation and insertion into the ER membrane. In contrast, both sequence and specific structural contacts contribute significantly to whether a ChR will localize to the plasma membrane. Our results demonstrate that the model can 'learn' the features that contribute to localization from the data and make accurate predictions on that property.

Using the GP regression model to engineer novel sequences that localize. We next tested the ChR localization regression model for its ability to predict plasma-membrane localization for ChR sequences outside the recombination library. For this, we chose a natural ChR variant, CbChR1, that expresses in HEK cells and neurons but does not localize to the plasma membrane and thus is non-functional (41). CbChR1 is distant from the three parental sequences, with 60% identity to CsChrimR and 40% identity to CheRiff and C1C2. We optimized CbChR1 by introducing minor amino acid changes predicted by the localization regression model to be beneficial for membrane localization. To enable measurement of CbChR1 localization with the SpyTag-based labeling method, we substituted the N-terminus of CbChR1 with the CsChrimR N-terminus containing the SpyTag sequence downstream of the signal peptide to make the chimera CsCbChR1 (78). This block swap did not change the membrane localization properties of CbChR1 (**Figure 5.6C**). Using the regression model, we predicted localization levels for all the possible

single-block swaps from the three library parents (CsChrimR, C1C2 and CheRiff) into CsCbChR1 and selected the four chimeras with the highest Upper Confidence Bound (UCB). These chimeras have between 4 and 21 mutations when compared with CsCbChR1. Unlike the LCB algorithm, which seeks to find the safest optimal choices, the UCB algorithm balances exploration and exploitation by maximizing the sum of the predicted mean and standard deviation.

The selected chimeras were assayed for expression, localization, and localization efficiency. One of the four sequences did not express; the other three chimeras expressed and had higher localization levels than CsCbChR1 (**Figure 5.6B**). Two of the three had localization properties similar to the CheRiff parent (**Figure 5.6B**). Images of the two best localizing chimeras illustrate the enhancement in localization when compared with CbChR1 and CsCbChR1 (**Figure 5.6C** and **Supplementary Figure 5.12**). This improvement in localization was achieved through single-block swaps from CsChrimR (17 and 21 amino acid mutations) (**Figure 5.6A**). These results suggest that this regression model can accurately predict minor sequence changes that will improve the membrane localization of natural ChRs.

5.4 Discussion

The ability to differentiate the functional properties of closely related sequences is extremely powerful for protein design and engineering. This is of particular interest for protein types that have proven to be more recalcitrant to traditional protein design methods, e.g. MPs. We show here that integral membrane protein expression and plasma membrane localization can be predicted for novel, homologous sequences using moderate-throughput data collection and advanced statistical modeling. We have used the models in four ways: 1) to accurately predict which diverse, chimeric ChRs are likely to express and localize at least as well as a moderately-performing native ChR; 2) to design ChR chimeras with optimized membrane localization that matched or exceeded the performance of a very well-localizing ChR (CsChrimR); 3) to identify the structural interactions (contacts) and

sequence elements most important for predicting ChR localization; and 4) to identify limited sequence changes that transform a native ChR from a non-localizer to a localizer.

Whereas 99% of the chimeras in the recombination library are predicted to express in HEK cells, only 5.6% are predicted to localize to the membrane at levels equal to or above the lowest parent (CheRiff). This result shows that expression is robust to recombination-based sequence alterations, whereas correct plasma-membrane localization is much more sensitive. The model enables accurate selection of the rare, localization-capable, proteins from the nearly 120,000 possible chimeric library variants. In future work we will show that this diverse set of several thousand variants predicted to localize serves as a highly enriched source of functional ChRs with novel properties.

Although statistical models generalize poorly as one attempts to make predictions on sequences distant from the sequences used in model training, we show that it is possible to train a model that accurately distinguishes between closely related proteins. The tradeoff between making accurate predictions on subtle sequence changes vs generalized predictions for significantly different sequences is one we made intentionally in order to achieve accurate predictions for an important and interesting class of proteins. Accurate statistical models, like the ones described in this paper, could aid in building more expressive physics-based models.

This work details the steps in building machine-learning models and highlights their power in predicting desirable protein properties that arise from the intersection of multiple cellular processes. Combining recombination-based library design with statistical modeling methods, we have scanned a highly functional portion of protein sequence space by training on only 218 sequences. Model development through iterative training, exploration, and verification has yielded a tool that not only predicts optimally performing chimeric proteins, but can also be applied to improve related ChR proteins outside the library. As large-scale gene synthesis and DNA sequencing become more affordable, machine-learning methods such as those described here will become ever more powerful tools for protein engineering offering an alternative to high-throughput assay systems.

5.5 Materials and methods

The design, construction, and characterization of recombination library chimeras is described in Bedbrook *et al.* (44). Briefly, HEK 293T cells were transfected with purified ChR variant DNA using Fugene6 reagent according to the manufacturer's recommendations. Cells were given 48 hours to express before expression and localization were measured. To assay localization level, transfected cells were subjected to the SpyCatcher-GFP labeling assay, as described in Bedbrook *et al*. (78). Transfected HEK cells were then imaged for mKate and GFP fluorescence using a Leica DMI 6000 microscope (for cell populations) or a Zeiss LSM 780 confocal microscope (for single cells: **Supplementary Figure 5.12**). Images were processed using custom image processing scripts for expression (mean mKate fluorescence intensity) and localization (mean GFP fluorescence intensity). All chimeras were assayed under identical conditions.

For each chimera, net hydrophobicity was calculated by summing the hydrophobicity of all residues in the TM domains. The C1C2 crystal structure was used to identify residues within TM domains (**S2B Figure**), and the Kyte & Doolittle amino acid hydropathicity scale (172) was used to score residue hydrophobicity.

GP modeling

Both the GP regression and classification modeling methods applied in this paper are based on work detailed in (170). Romero *et al.* applied GP models to predict protein functions and also defined protein distance using a contact map. We have expanded on this previous work. Regression and classification were performed using open-source packages in the SciPy ecosystem (173-175). Below are specifics of the GP regression and classification methods used in this paper. The hyperparameters and the form of the kernel were optimized using the Bayesian method of maximizing the marginal likelihood of the resulting model.

<u>GP regression</u>

In regression, the problem is to infer the value of an unknown function $f(x)$ at a novel point x_* given observations y at inputs X. Assuming that the observations are subject to independent identically distributed Gaussian noise with variance σ_n^2, the posterior distribution of $f_* = f(x_*)$ for Gaussian process regression is Gaussian with mean

$$\bar{f}_* = k_*^T (K + \sigma_n^2 I)^{-1} y \quad (1)$$

and variance

$$v_* = k(x_*, x_*) - k_*^T (K + \sigma_n^2 I)^{-1} k_* \quad (2)$$

where

1. K is the symmetric, square covariance matrix for the training set, where $K_{ij} = k(x_i, x_j)$ for x_i and x_j in the training set.

2. k_* is the vector of covariances between the novel input and each input in the training set, where $k_{*i} = k(x_*, x_i)$.

We found that results could be improved by first performing feature selection with L1-regularized linear regression and then only training the GP model on features with non-zero weights in the L1 regression. The hyperparameters in the kernel functions, the noise hyperparameter σ_p and the regularization hyperparameter were determined by maximizing the log marginal likelihood:

$$\log p(y|X) = -\frac{1}{2} y^T (K + \sigma_n^2 I)^{-1} y - \frac{1}{2} \log|K + \sigma_n^2 I| - \frac{n}{2} \log 2\pi, \quad (3)$$

where n is the dimensionality of the inputs.

GP classification

In binary classification, instead of continuous outputs y, the outputs are class labels $y_i \in \{+1, -1\}$, and the goal is to use the training data to make probabilistic predictions

$\pi(x_*) = p(y_* = +1|x_*)$. Unfortunately, the posterior distribution for classification is analytically intractable. We use Laplace's method to approximate the posterior distribution. There is no noise hyperparameter in the classification case. Hyperparameters in the kernels are also found by maximizing the marginal likelihood.

GP kernels for modeling proteins

Gaussian process regression and classification models require kernel functions that measure the similarity between protein sequences. A protein sequence s of length l is defined by the amino acid present at each location. This information can be encoded as a binary feature vector x_{se} that indicates the presence or absence of each amino acid at each position. The protein's structure can be represented as a residue-residue contact map. The contact-map can be encoded as a binary feature vector x_{st} that indicates the presence or absence of each possible contacting pair. The sequence and structure feature vectors can also be concatenated to form a sequence-structure feature vector.

We considered three types of kernel functions $k(s_i, s_j)$: linear kernels, squared exponential kernels, and Matérn kernels. The linear kernel is defined as

$$k(s, s') = \sigma_p^2 x^T x', \tag{4}$$

where σ_p is a hyperparameter that determines the prior variance of the fitness landscape. The squared exponential kernel is defined as

$$k(s, s') = \sigma_p^2 \exp\left(-\frac{\|x-x'\|_2^2}{l}\right), \tag{5}$$

where l and σ_p are also hyperparameters and $|\cdot|_2$ is the L2 norm. Finally, the Matérn kernel with $v = \frac{5}{2}$ is defined as

$$k(s, s') = \left(1 + \frac{\sqrt{5\|x-x'\|_2^2}}{l} + \frac{5\|x-x'\|_2^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5\|x-x'\|_2^2}}{l}\right), \tag{6}$$

where l is once again a hyperparameter.

L1 regression feature identification and weighting

To identify those contacts in the ChR structure most important in determining chimera function (here, localization) we used L1 regression. Given the nature of our library design and the limited set of chimeras tested, there are certain residues and contacts that covary within our training set. The effects of these covarying residues and contacts cannot be isolated from one another using this data set and therefore must be weighted together for their overall contribution to ChR function. By using the concatenated sequence and structure binary feature vector for the training set we were able to identify residues and contacts that covary. Each individual set of covarying residues and contacts was combined into a single feature. L1 linear regression was then used to weight features as either zero or non-zero in their contribution to ChR function. The level of regularization was chosen by LOO cross-validation. We then performed Bayesian ridge linear regression on features with non-zero L1 regression weights using the default settings in scikit-learn (176). The Bayesian ridge linear regression weights were plotted onto the C1C2 structure to highlight positive and negative contributions to ChR localization (**Figure 5.5**) and ChR expression (**Supplementary Figure 5.11**).

5.6 Figures and tables

Table 5.1. Comparison of size, diversity, and localization properties of the training set and subsequent sets of chimeras chosen by models in the iterative steps of model development.

Set	Count	Mutations mean ± stdev	Percent with good localization*	Localization mean ± stdev ($x10^{-3}$)
training – parents	3	0	100%	5.6 ± 3.0
training – single-block swap	112	15 ± 9	33%	3.2 ± 3.4
training – multi-block swap	103	73 ± 21	12%	1.5 ± 2.5
exploration	16	69 ± 12	50%	4.8 ± 4.7
verification – high performing	4	29 ± 17	100%	8.0 ± 1.6
verification – low performing	7	67 ± 12	0%	0.89 ± 0.73
optimization	4	43 ± 6	100%	14 ± 3.5

* 'good localization' is localization at or above that of the lowest-performing parent, CheRiff

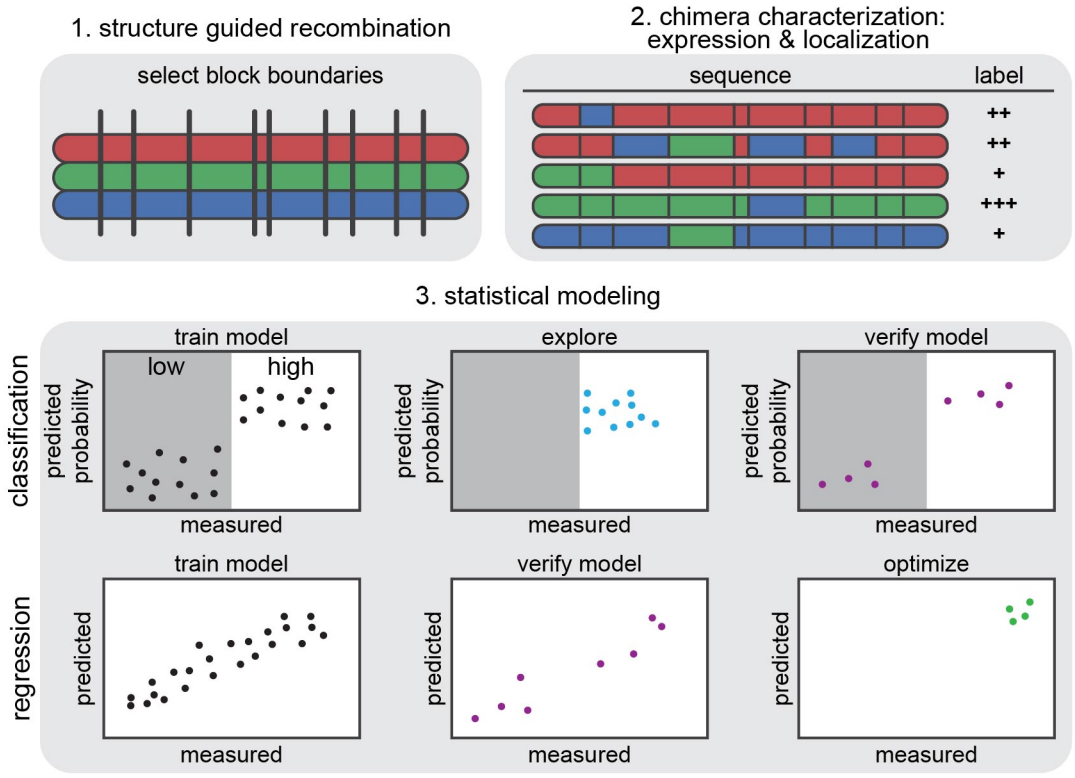


Figure 5.1. General approach to machine learning of protein (ChR) structure-function relationships: diversity generation, measurements on a training set, and modeling. (**1**) Structure-guided SCHEMA recombination is used to select block boundaries for shuffling protein sequences to generate a sequence-diverse ChR library starting from three parent ChRs (shown in red, green, and blue). (**2**) A subset of the library serves as the training set. Genes for these chimeras are synthesized and cloned into a mammalian expression vector, and the transfected cells are assayed for ChR expression and localization. (**3**) Two different models, classification and regression, are trained using the training data and then verified. The classification model is used to explore diverse sequences predicted to have 'high' localization. The regression model is used to design ChRs with optimal localization to the plasma membrane.

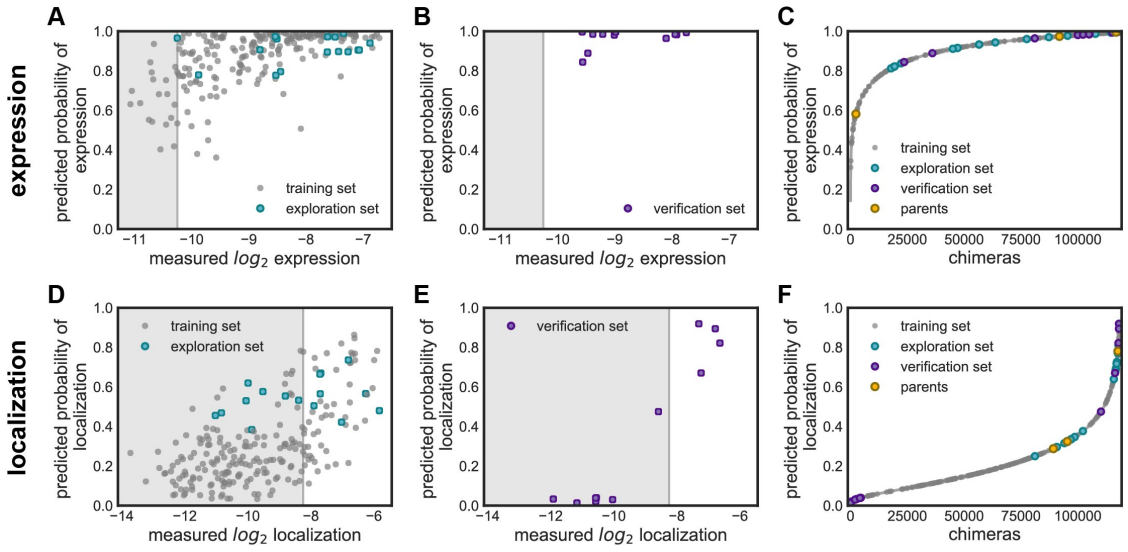


Figure 5.2. GP binary classification models for expression and localization. Plots of predicted probability vs measured properties are divided into 'high' performers (white background) and 'low' performers (gray background) for each property (expression and localization). (**A**) & (**D**) Predicted probability vs measured properties for the training set (gray points) and the exploration set (cyan points). Predictions for the training and exploration sets were made using LOO cross-validation. (**B**) & (**E**) Predicted probabilities vs measured properties for the verification set. Predictions for the verification set were made by a model trained on the training and exploration sets. (**C**) & (**F**) Predicted probability of 'high' expression, and localization for all chimeras in the recombination library (118,098 chimeras) made by models trained on the data from the training and exploration sets. The gray line shows all chimeras in the library, the gray points indicate the training set, the cyan points indicate the exploration set, the purple points indicate the verification set, and the yellow points indicate the parents. (**A-C**) Show expression and (**D-F**) show localization. For all plots, the measured property is plotted on a $\log_2$ scale.

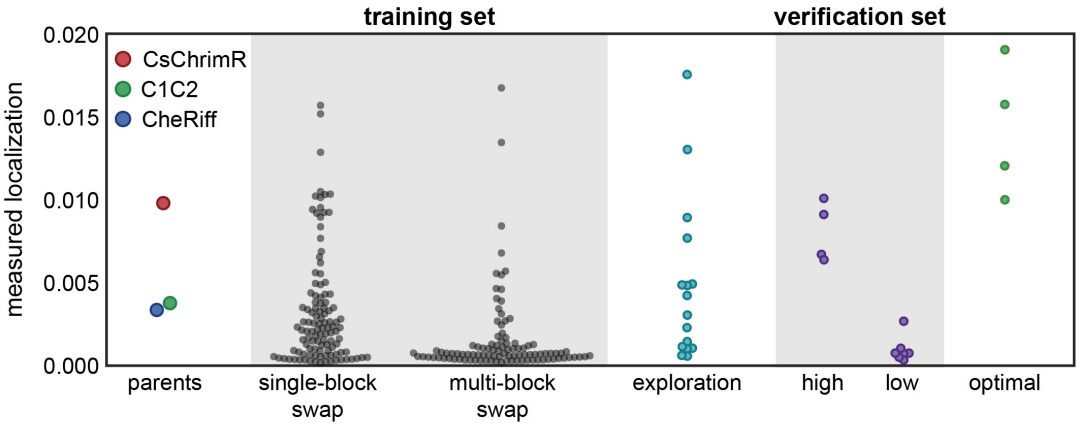


Figure 5.3. Comparison of measured membrane localization for each data set. Swarm plots of localization measurements for each data set compared with parents: training set, exploration set, verification set, and optimization set.

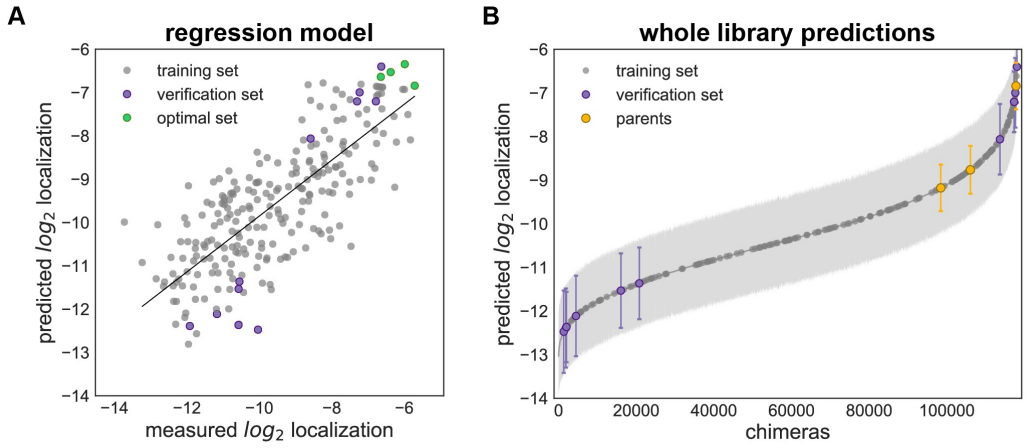


Figure 5.4. GP regression model for localization. (**A**) Predicted vs measured localization for the combined training and exploration sets (gray points), verification set (purple points), and the optimal set (green points). Predictions for the training and exploration sets were made using LOO cross-validation; predictions for the verification and optimal sets were made by a model trained on data from the training and exploration sets. There is a clear correlation between predicted and measured localization. The combined training and exploration sets showed good correlation ($R > 0.73$) as did the verification set ($R > 0.9$). (**B**) Predicted localization values of all chimeras in the recombination library (118,098 chimeras) based on the GP regression model trained on the training and exploration sets. The gray line shows all chimeras in the library, the gray points indicate the training set and exploration sets, the purple points indicate the verification set, and the yellow points indicate the parents. Error bars (light gray shading) show the standard deviation of the predictions. For all plots, the predicted and measured localization are plotted on a $\log_2$ scale.

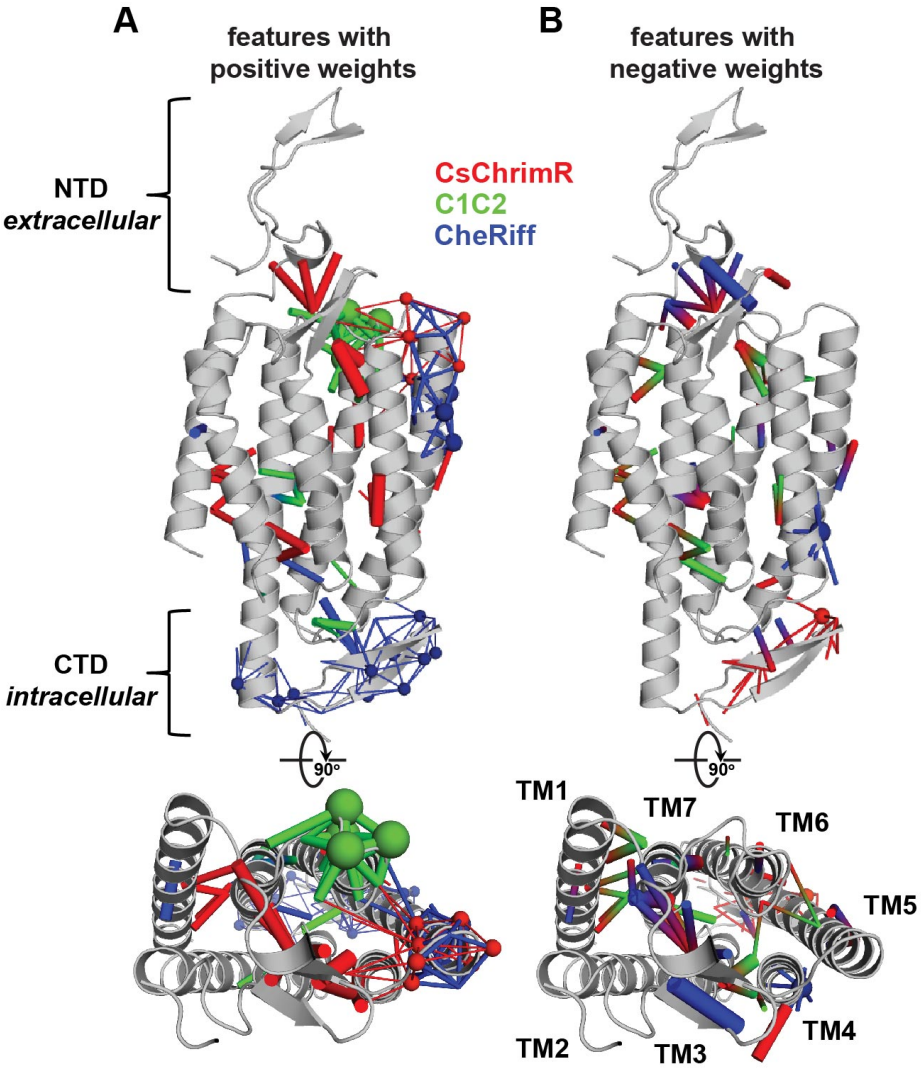


Figure 5.5. Sequence and structural contact features important for prediction of ChR localization. Features with positive (**A**) and negative (**B**) weights are displayed on the C1C2 crystal structure (grey). Features can be residues (spheres) or contacts (sticks) from one or more parent ChRs. Features from CsChrimR are shown in red, features from C1C2 are shown in green, and features from CheRiff are shown in blue. In cases where a feature is present in two parents, the following color priorities were used for consistency: red above green above blue. Sticks connect the beta carbons of contacting residues (or alpha

carbon in the case of glycine). The size of the spheres and the thickness of the sticks are proportional to the parameter weights. Two residues in contact can be from the same or different parents. Single-color contacts occur when both contributing residues are from the same parent. Multi-color contacts occur when residues from different parents are in contact. The N-terminal domain (NTD), C-terminal domain (CTD), and the seven transmembrane helices (TM1-7) are labeled.

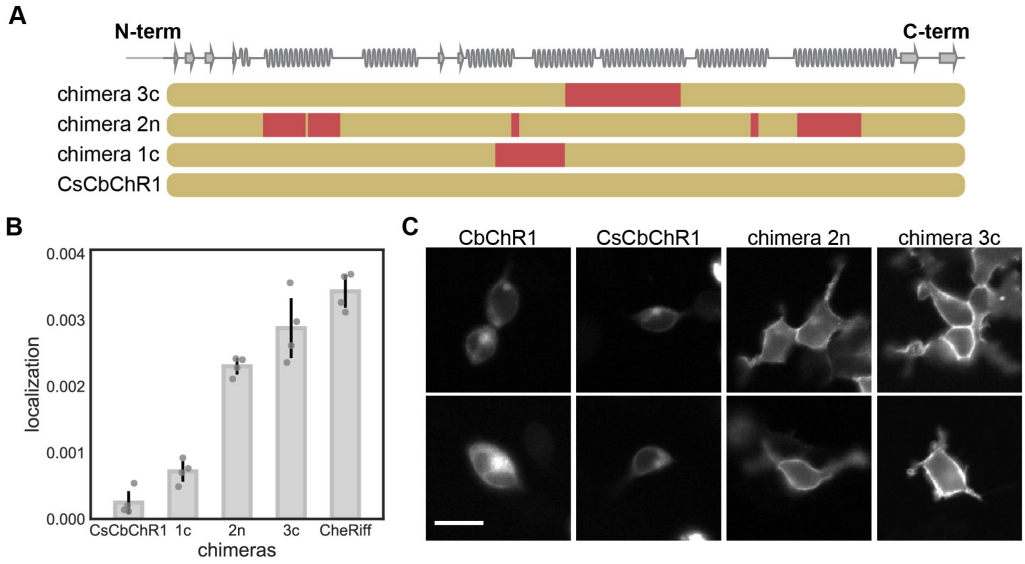
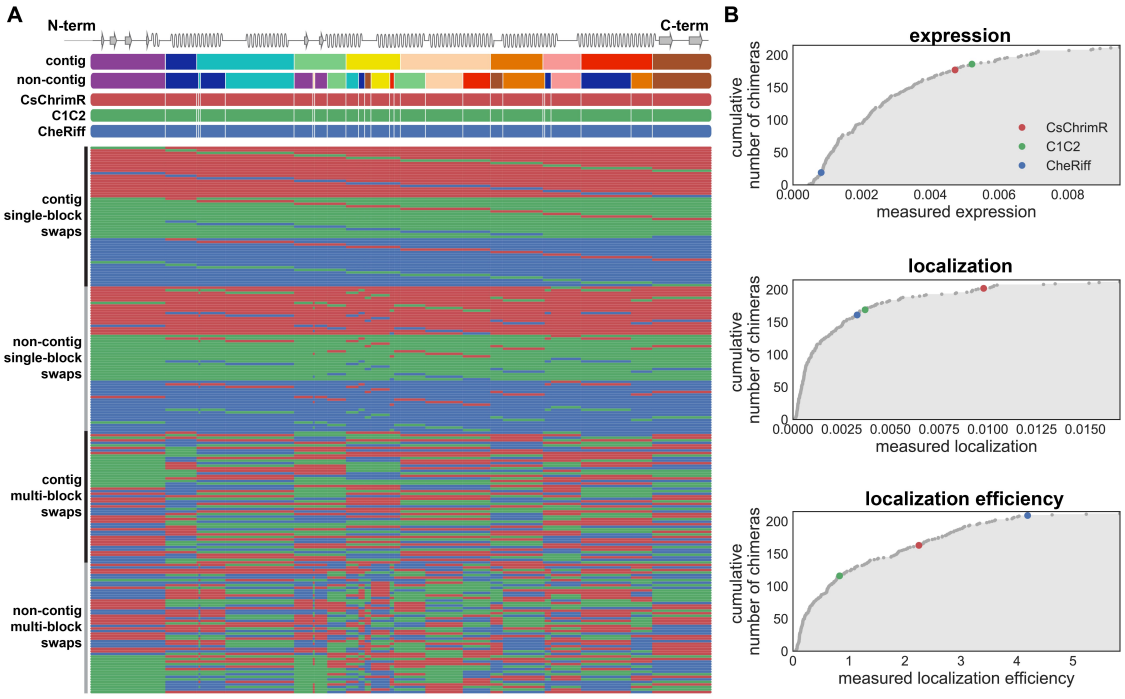
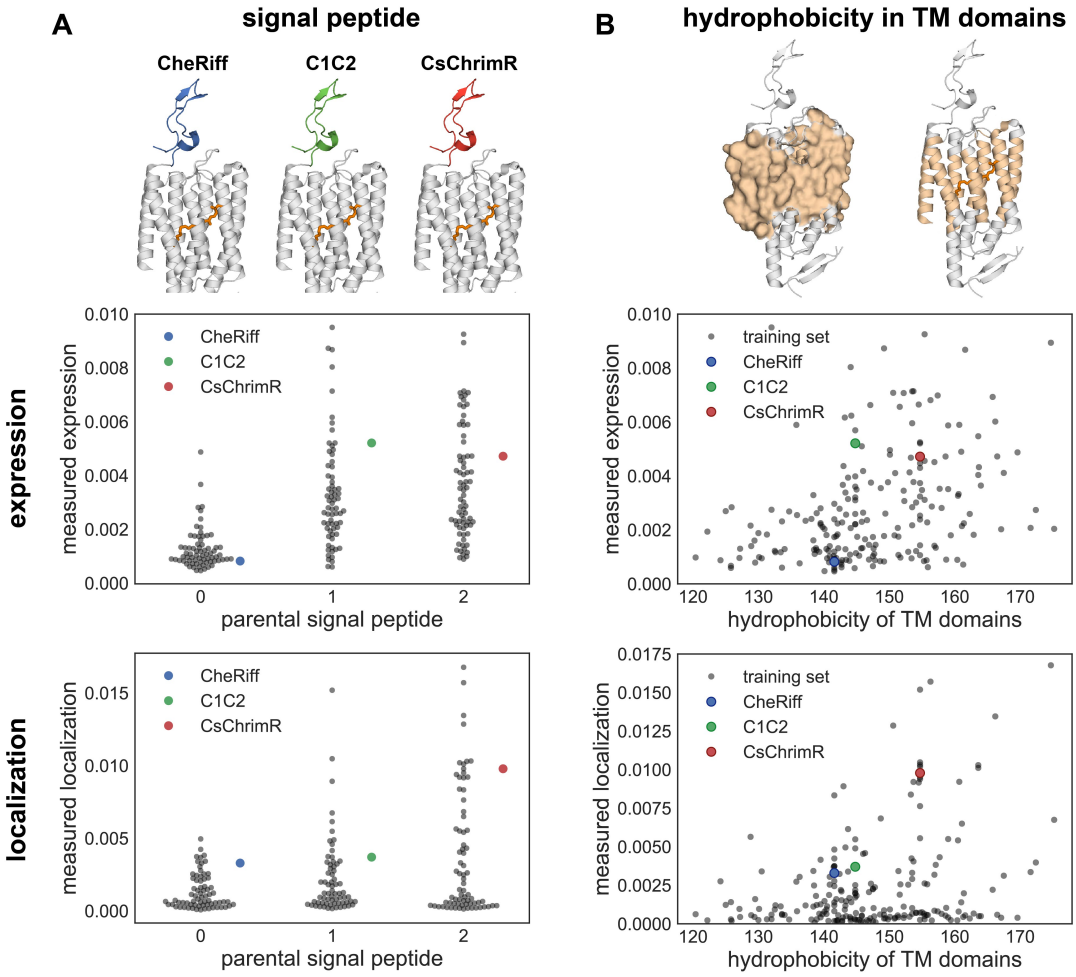


Figure 5.6. GP regression model enables engineering of localization in CbChR1. (**A**) Block identities of the CsCbChR1 chimeras. Each row represents a chimera. Yellow represents the CbChR1 parent and red represents the CsChrimR parent. Chimeras 1c, 2n, and 3c have 4, 21, and 17 mutations with respect to CsCbChR1, respectively. (**B**) Plot of measured localization of CsCbChR1 compared to three CsCbChR1 single-block-swap chimeras and the CheRiff parent. (**C**) Two representative cell images of mKate expression of CbChR1 and CsCbChR1 compared with top-performing CsCbChR1 single-block-swap chimeras show differences in ChR localization properties – chimera 2n and chimera 3c clearly localize to the plasma membrane. Scale bar: 20 μm.

5.7 Supplementary Figures

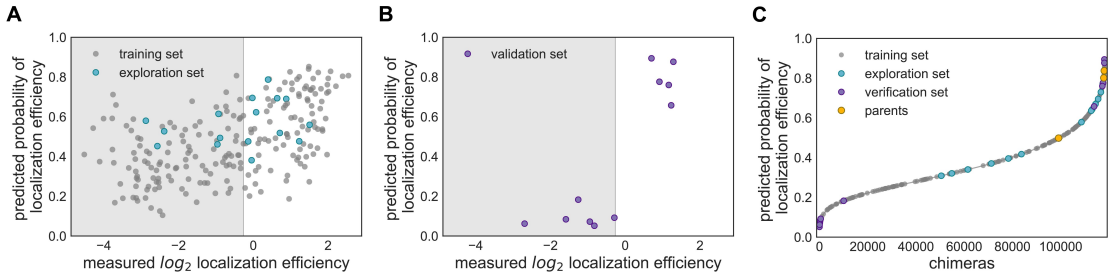


Supplementary Figure 5.1. Chimera sequences in training set and their expression, localization, and localization efficiencies. (**A**) (top) shows blocks (different colors) for the contiguous (contig) and non-contiguous (non-contig) library designs and also shows block boundaries (white lines) for the combined contiguous and non-contiguous library designs on the three parental ChRs aligned with a schematic of the ChR secondary structure. (bottom) Sequences of training set chimeras showing block identities. The colors represent the parental origin of the block (red – CsChrimR, green – C1C2, and blue – CheRiff). (**B**) Cumulative distributions of the measured expression, localization, and localization efficiency of all 218 chimeras with the three parental constructs highlighted in color (5).

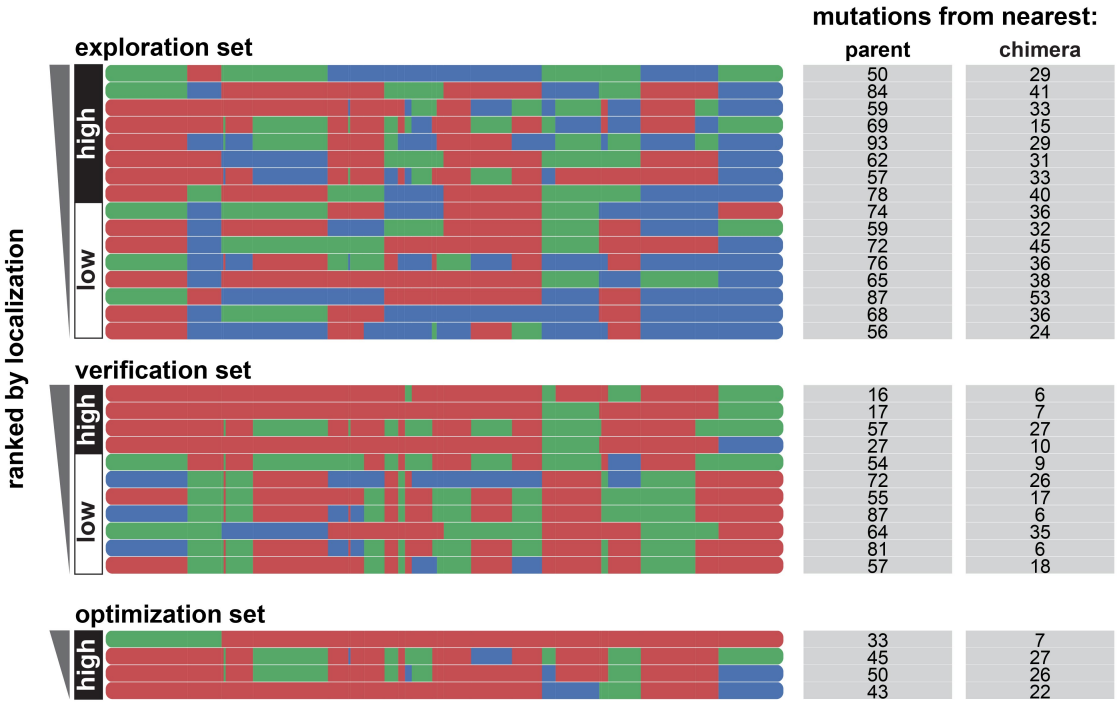


Supplementary Figure 5.2. Chimera expression and localization cannot be predicted from simple rules. Expression and localization measurements are plotted with chimeras grouped based on (**A**) signal peptide sequence identity and (**B**) hydrophobicity in the transmembrane (TM) domains. (**A**) Each chimera in the training set is grouped based on its signal peptide identity, which could be the CheRiff (0), C1C2 (1), or CsChrimR (2) signal peptide. The measured expression and localization are shown for each chimera in each of the three groups. (**B**) The measured expression and localization with respect to the calculated level of hydrophobicity within the 7-TM domains of each chimera.

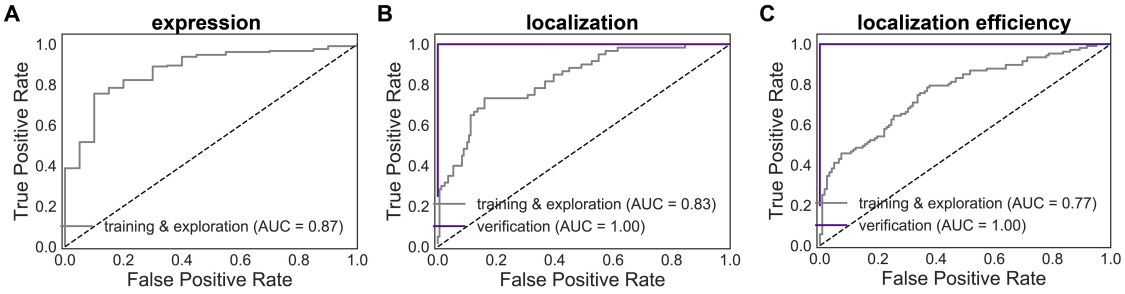
Hydrophobicity was calculated in the region of the protein highlighted in the surface rendering on the ChR structure.



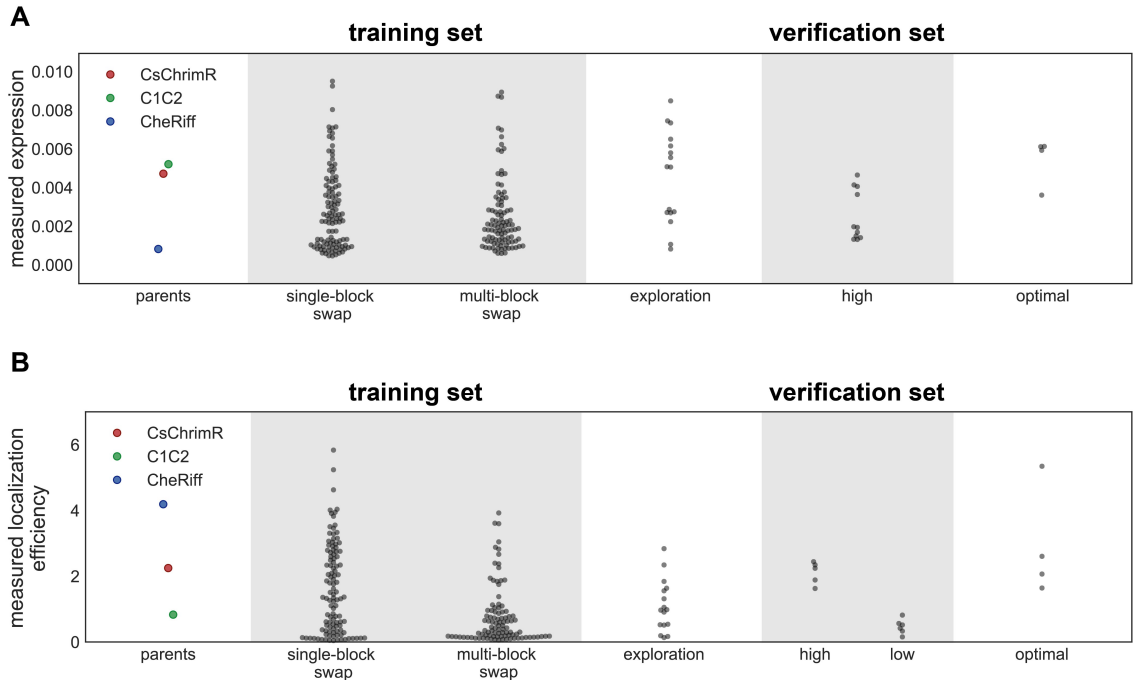
Supplementary Figure 5.3. GP binary classification model for localization efficiency. Plots of predicted probability vs measured localization efficiency are divided into 'high' performers (white background) and 'low' performers (gray background) for localization efficiency. (**A**) Predicted probability vs measured localization efficiency for the training set (gray points) and the exploration set (cyan points). Predictions for the training and exploration sets were made using LOO cross-validation. (**B**) Predicted probabilities vs measured localization efficiency for the verification set. Predictions for the verification set were made by a model trained on the training and exploration sets. (**C**) Probability of 'high' localization efficiency for all chimeras in the recombination library (118,098 chimeras) made by a model trained on the data from the training and exploration sets. The gray line shows all chimeras in the library, the gray points indicate the training set, the cyan points indicate the exploration set, the purple points indicate the verification set, and the yellow points indicate the parents. For all plots, the measured localization efficiency is plotted on a $\log_2$ scale.



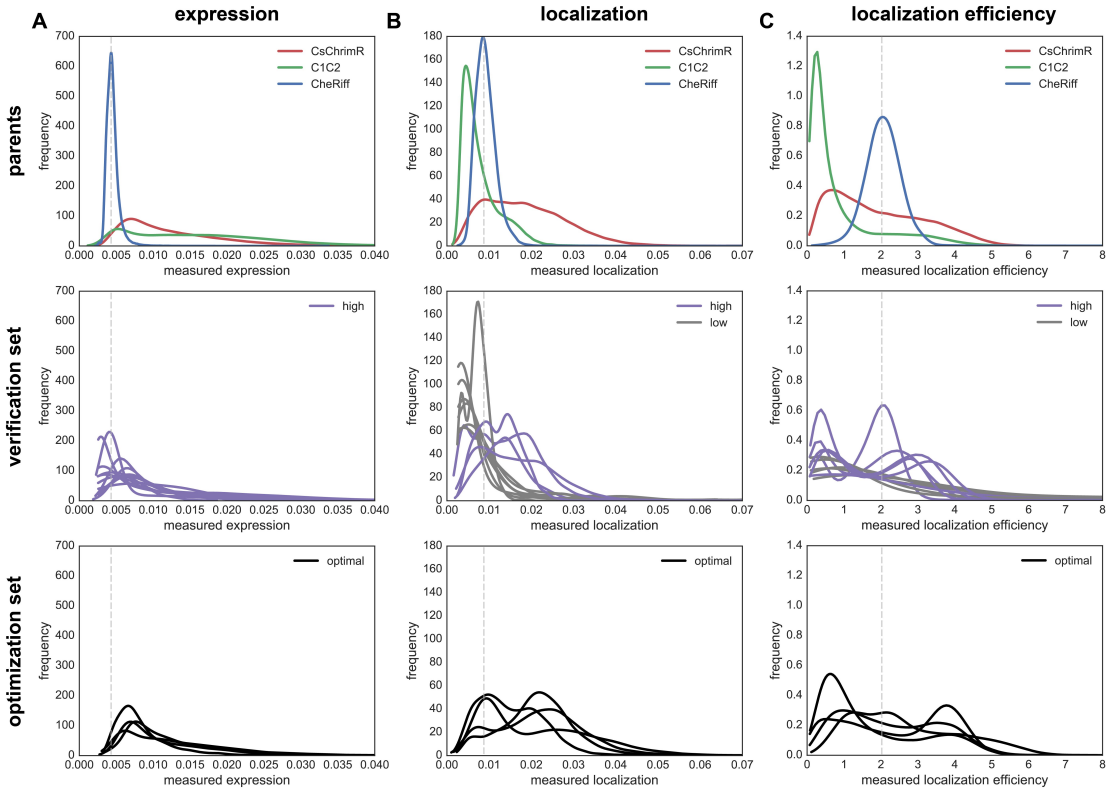
Supplementary Figure 5.4. Chimera block identities for exploration, verification, and optimization sets. Block identity of chimeras from each set ranked according to their performance for localization with the best ranking chimera listed at the top of the list. 'High' and 'low' indicates those chimeras had a high predicted probability of localization vs a low predicted probability of localization. Each row represents a chimera. The three different colors represent blocks from the three different parents (red – CsChrimR, green – C1C2, and blue – CheRiff). The number of mutations from the nearest parent and the number of mutations from the nearest previously tested chimera from the library are shown for each chimera.



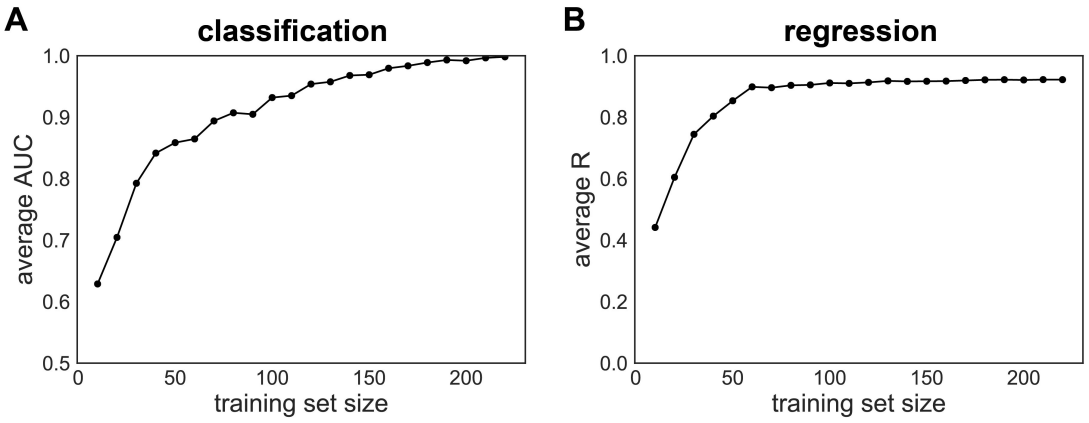
Supplementary Figure 5.5. ROC curves for GP classification expression, localization, and localization efficiency models. ROC curves show true positive rate vs false positive rate for predictions from the expression (**A**), localization (**B**), and localization efficiency (**C**) classification models. The gray line shows the ROC for the combined training and exploration sets. The purple line shows the ROC for the verification set. The verification sets consist exclusively of chimeras with 'high' expression so no verification ROC curve for expression is shown. Predictions for the training and exploration sets were made using LOO cross-validation, while predictions for the verification set were made by a model trained on the training and exploration sets. Calculated AUC values are shown in the figure key.



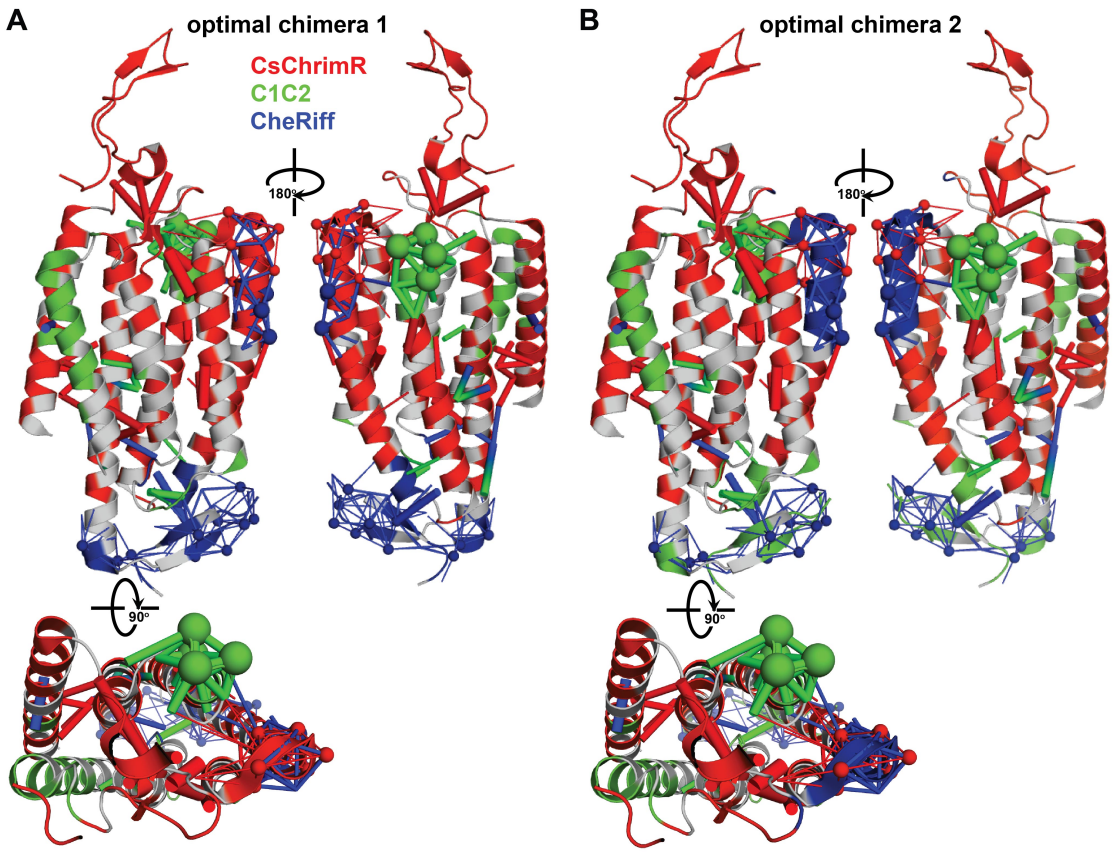
Supplementary Figure 5.6. Comparison of measured expression and localization efficiency for each data set. Swarm plots of expression (**A**) and localization efficiency (**B**) measurements for each data set compared with parents: training set, exploration set, verification set, and optimization set.



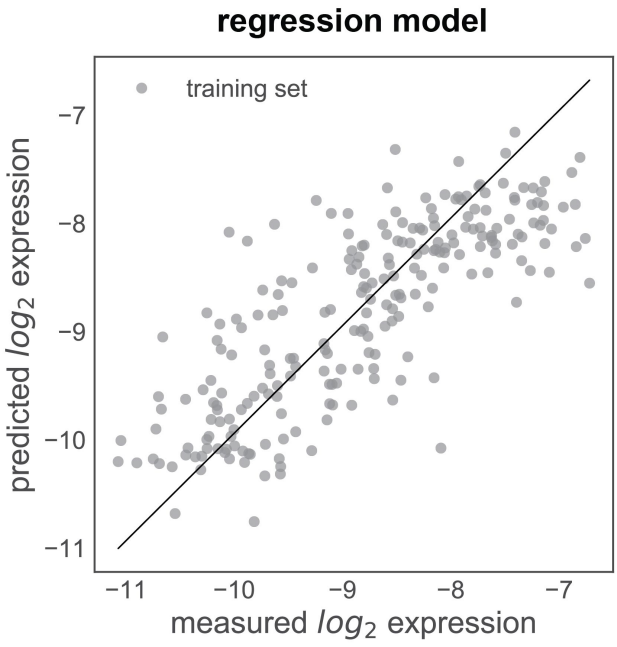
Supplementary Figure 5.7. Cell population distributions of expression, localization, and localization efficiency properties for each chimera in the verification and optimization sets compared with parents. The distribution of expression (**A**), localization (**B**), and localization efficiency (**C**) for the population of transfected cells is plotted for each parent (top row), each chimera in the verification set (middle row), and each chimera in the optimization set (bottom row) using kernel density estimation for smoothing. Parents are plotted in red (CsChrimR), green (C1C2), and blue (CheRiff). Chimeras in the verification set are plotted in gray if they were predicted to be 'low' or purple if they were predicted to be 'high' in each property. The vertical, gray, dashed line indicates the mean behavior of the CheRiff parent for each property.



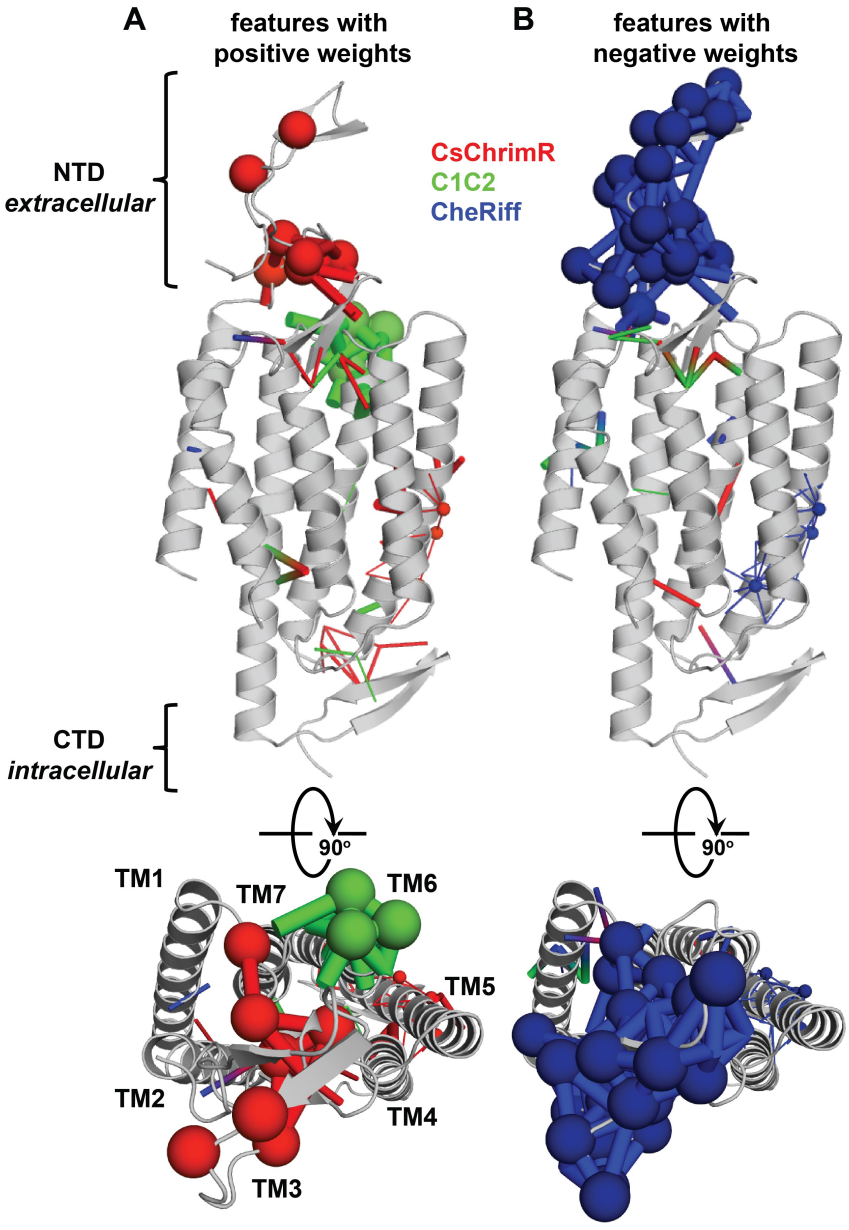
Supplementary Figure 5.8. Predictive ability of GP localization models as a function of training set size. We trained GP models on random training sets of various sizes sampled from our data and evaluated their predictive performance on a fixed test set of sequences for the classification (**A**) and regression (**B**) localization models. The predictive performance of the classification model is described by AUC for the test set (**A**), while the predictive performance of the regression model (**B**) is described by the correlation coefficient (R-value) for the test set. For each training set size, the results are averaged over 100 random samples.



Supplementary Figure 5.9. Important features for prediction of ChR localization aligned with chimeras with optimal localization. Features with positive weights from the localization model (**Fig 5**) are displayed on the C1C2 crystal structure which is colored based on the block design of two different chimeras, (**A**) n1_7 and (**B**) n4_7, from the optimization set. Features can be residues (spheres) or contacts (sticks) from one or more parent ChRs. Features/blocks from CsChrimR are shown in red, features/blocks from C1C2 are shown in green, and features/blocks from CheRiff are shown in blue. Gray positions are conserved residues. Sticks connect the beta carbons of contacting residues (or alpha carbon in the case of glycine). The size of the spheres and the thickness of the sticks are proportional to the parameter weights.

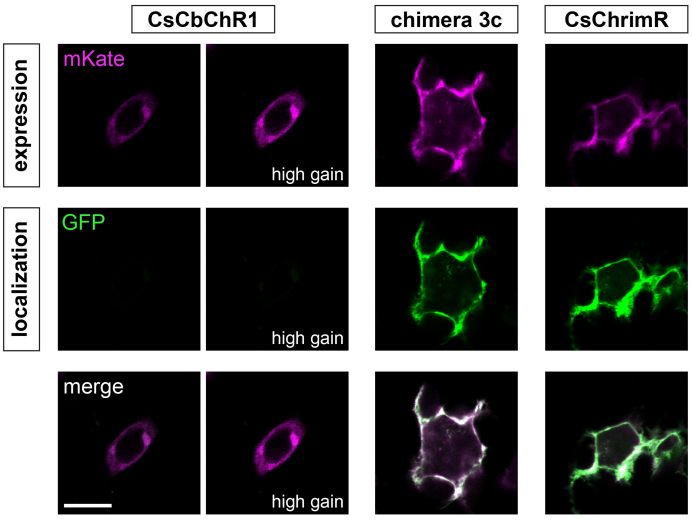


Supplementary Figure 5.10. GP regression model for ChR expression. Shows the GP regression model predicted vs measured expression for the combined training and exploration sets (gray points). Predictions for the training and exploration sets were made using LOO cross-validation. The predicted and measured expression are plotted on a $\log_2$ scale. The combined training and exploration sets showed good correlation ($R > 0.70$).



Supplementary Figure 5.11. Sequence and structure features important for prediction of ChR expression. Features with positive (**A**) and negative (**B**) weights are displayed on the C1C2 crystal structure (grey). Features can be residues (spheres) or contacts (sticks) from one or more parent ChRs. Features from CsChrimR are shown in red, features from C1C2 are shown in green, and features from CheRiff are shown in blue. In cases where a

feature is present in two parents, the following color priorities were used for consistency: red above green above blue. Sticks connect the beta carbons of contacting residues (or alpha carbon in the case of glycine). The size of the spheres and the thickness of the sticks are proportional to the parameter weights. Two residues in contact can be from the same or different parents. Single-color contacts occur when both contributing residues are from the same parent. Multi-color contacts occur when residues from different parents are in contact. The N-terminal domain (NTD), C-terminal domain (CTD), and the seven transmembrane helices (TM1-7) are labeled.



Supplementary Figure 5.12. Localization of engineered CbChR1 variant chimera 3c. Representative cell confocal images of mKate expression and GFP labeled localization of CsCbChR1 compared with top-performing CsCbChR1 single-block-swap chimera (chimera 3c), and top-performing parent (CsChrimR). CsCbChR1 shows weak expression and no localization, while chimera 3c expresses well and clearly localizes to the plasma membrane as does CsChrimR. Gain was adjusted in CsCbChR1 images to show any low signal. Scale bar: 10 μm.

Chapter 6

MACHINE LEARNING TO ENGINEER 'DESIGNER' CHANNELRHODOPSINS FOR MINIMALLY INVASIVE OPTOGENETICS

6.1 Introduction

Channelrhodopsins (ChRs) are light-gated ion channels found in photosynthetic algae, which upon transgenic expression in neurons enable light-dependent activation of neuronal activity (17). These channel proteins have been widely applied as tools in neuroscience research in the field of optogenetics (117); however, functional limitations of available ChRs prohibits or limits a number of optogenetic applications. In their algal hosts, ChRs serve as sunlight sensors in phototaxic and photophobic responses (17). Because these channels have evolved to use sunlight for functional activation, they have broad activation spectra in the visible range (400-650 nm) and require high-intensity light for activation [~1 mW mm^{-2}, which is the average intensity of sunlight on the earth's surface]. ChRs are naturally low-conductance channels requiring on the order of 10^5-10^6 functional ChRs expressed in the membrane of a neuron to produce sufficient light-dependent depolarization to induce neuronal activation (20). When applied to the mouse brain, ChRs require ~1-15 mW light delivered <100 μm from the target cell population to reliably activate action potentials (19). This confines light-dependent activation to a small volume of brain tissue [approximately a cubic millimeter (40)]. 'Optogenetic access' to large brain volumes or the entire brain without the need to implant invasive fibers for light delivery (i.e. non-invasive optogenetic excitation) would be highly desirable.

Our goal has been to engineer enhanced ChRs to overcome the above-mentioned limitations and extend what is currently possible with optogenetic excitation experiments. Engineering useful and interesting ChRs requires overcoming three major challenges. First, rhodopsins are trans-membrane proteins that are inherently difficult to engineer because the sequence and structural determinants of membrane protein expression and plasma-

membrane localization are highly constrained and poorly understood (44, 45). Second, protein properties of interest for neuroscience applications are assayed using very low-throughput patch-clamp electrophysiology, preventing the use of high-throughput assay approaches required for directed evolution experiments. And third, *in vivo* application of these improved tools require either retention or optimization of multiple protein properties in a single protein tool; for example, we must optimize expression and localization of these algal membrane proteins in mammalian cells while at the same time optimizing kinetics, photocurrents, and spectral properties (11). This challenging protein-engineering problem demands a method for designing ChRs with specific combinations of desirable properties without having to screen hundreds to thousands of ChR variants for their functional properties.

Since the first discovery and application of ChR2 for neuronal activation, there has been a diversity of ChR variants published, including variants discovered from nature (41), variants engineered through recombination (27, 44) and point mutagenesis (29, 38), as well as variants resulting from more rational design approaches (177). Studies of these different variants coupled with structural information (23) and molecular dynamic simulations (36) has established some understanding of the mechanics and sequence features important for specific ChR properties (17, 177). Despite this useful work, it is still not possible to predict the functional properties of new ChR sequences and therefore not trivial to design new ChR variants with a desired combination of functional properties.

Our approach has been to leverage the significant literature of ChR variants (both natural and engineered) to train statistical models that enable the design of new, highly-functional ChR variants. These models take as their input sequence and structural information for a given ChR variant and then output a prediction of the ChR's functional properties based on sequence. To train the models, we collect a dataset of functional properties from ChR sequence variants. The models use the training data to learn how sequence and structural elements map to functional properties. The resulting models approximate the ChR 'fitness landscape' for each given property (132, 170). Once known, the mapping of ChR sequence to functional properties can be used to predict the functional behavior of untested ChR

sequence variants. The models can be used to then select sequence variants predicted to have optimal combinations of desired properties (e.g. good membrane localization, strong photocurrents, and red-shifted light activation).

We train models in this manner and find that they very accurately predict the functional properties of untested ChR sequences. We used these models to successfully engineer 30 'designer' ChR variants with specific combinations of desired properties. A number of ChR variants identified from this work have unprecedented conductance and light sensitivity. These superconducting variants may change the way optogenetics experiments can be done by enabling less-invasive activation of populations of cells throughout the nervous system. We have characterized these low-light sensitive, super-conducting ChRs for applications in the mammalian brain. This work is a convincing demonstration of the power of machine-learning guided protein engineering for a very difficult to engineer class of proteins.

6.2 Results

6.2.1 Dataset of ChR sequence variants and corresponding functional properties

In previous work, we explored the use of structure-guided recombination (123, 135) of three highly-functional ChR parents: CsChrimsonR (CsChrimR) (41), C1C2 (23), and CheRiff (14) by building two 10-block recombination libraries with a theoretical size of ~120,000 ChR variants (i.e. $2x3^{10}$) (44). Measuring expression, localization, and photocurrent properties of a subset of these chimeric ChR variants showed that these recombination libraries provide a rich source of functional sequence diversity (44). This work produced 75 ChR variants with measured photocurrent properties, the largest single source of published ChR functional data. In subsequent work, we generated an additional 22 ChR variants from the same recombination libraries (45), which we have now characterized via patch clamp electrophysiology for functional properties. Together, we have 97 ChR sequence variants with measured functional properties from the two recombination libraries, providing the primary dataset used for model training in this work. We supplemented this dataset with data from other published sources including 23 ChR

variants from nature, 16 point-mutant ChR variants, and 36 recombination variants from various recombination libraries. The previously published data produced by other labs were not collected under the same experimental conditions as data collected in our hands, so it is not valid for comparison for absolute ChR properties (i.e. photocurrent strength); however, these data do provide useful binary information: is the sequence variant functional or not. Thus, we used published data from other sources when training binary classification models for ChR variant function.

Our primary interest was the optimization of three ChR photocurrent properties: photocurrent strength, wavelength sensitivity, and off kinetics (**Figure 6.1A**). Enhancing the photocurrent strength of ChRs would enable strong currents and thus reliable neuronal activation even under low light conditions. As metrics of photocurrent strength, we use peak and steady-state photocurrent (**Figure 6.1A**). Altering (narrowing or broadening) or shifting ChR's activation wavelength sensitivity could enable multiplexed application of ChRs (41). As a metric for each ChR's activation spectrum, we use the normalized current strength with green light (550 nm) (**Figure 6.1A**). Different off-kinetic properties can be useful for different applications; fast off kinetics is useful for high-frequency stimulation (178), slow off kinetics is correlated with increased light sensitivity (20, 29, 38), and very slow off kinetics can be used for constant depolarization [Step-function opsins (38)]. We use two parameters to characterize the off kinetics: the time to reach 50% of the light activated current, and the decay rate, τ_{off} (**Figure 6.1A**). In addition to functional properties, it is also necessary to optimize or maintain plasma-membrane localization because membrane localization is a prerequisite for ChR function (45).

As inputs for the machine-learning models, we consider both ChR sequence and structure. The ChR sequence information is simply encoded in its amino acid sequence, but for structural comparisons, we need to convert the 3D structural information into a form that is convenient for modeling. To do this, we encode structural information as a residue-residue 'contact map'. Two residues are in contact if they have any non-hydrogen atoms within 4.5 Å in the C1C2 crystal structure (23). These 'contacts' are considered as potential interactions that may be important for structural and functional integrity. This structural

encoding assumes that ChR chimeras share the overall contact architecture observed in the C1C2 crystal structure. For a given ChR, the contact map is simply a list of contacting amino acids with their relative positions, so a single contact can be described by: [('A134'), ('M1')].

6.2.2 Training Gaussian process (GP) classification and regression models

Using the ChR sequence/structure and functional data as inputs, we trained Gaussian process (GP) classification and regression models (**Figure 6.1**). GP models have successfully predicted thermal stability, substrate binding affinity, and kinetics for several soluble enzymes (170), and, more recently, ChR membrane localization (45). For a detailed description of the GP model architecture and properties used for protein engineering see (45, 170). Briefly, these models infer predictive values from training examples by assuming that similar inputs (ChR sequence variants) will have similar outputs (photocurrent properties). To quantify the relatedness of inputs (ChR sequence variants), we compare both sequence and structure. We define the sequence and structural similarity between two chimeras by aligning them and counting the number of positions at which they are identical (170).

We first trained a binary classification model to predict if a ChR sequence will be functional using all 97 training sequences from our recombination library as well as data from 75 sequence variants published from other groups. A ChR sequence was considered to be functional if its photocurrents were >0.1 nA upon light exposure. This was a threshold we set as an approximate lower bound for conductance necessary to activate neuronal activity. We then used this trained classification model to predict whether uncharacterized ChR sequence variants were functional (**Figure 6.1A**). To verify that the classification model is capable of accurate predictions, we performed 20-fold cross validation on the training data set and measured an area under the receiver operator curve (AUC) of 0.78, indicating good predictive power (**Table 6.1**).

Next, we trained three regression models, one for each of the ChR photocurrent properties of interest: photocurrent strength, wavelength sensitivity of photocurrents, and off kinetics

(**Figure 6.1A**). For these models, we exclusively used data collected from our ChR recombination libraries. Once trained, these models were used to predict photocurrent strength, wavelength sensitivity of photocurrents, and off kinetics of new, untested ChRs sequence variants. Again, to test whether these models make accurate predictions, we performed 20-fold cross validation on the training dataset and observed high correlation between predicted and measured properties as indicated by R values between 0.65-0.9 for all models (**Table 6.1**).

6.2.3 Selection of designer ChRs using trained models

A 'designer' ChR is defined as a ChR predicted by our models to have extraordinary properties. We used a tiered approach (**Figure 6.1B**) to select designer ChRs. Our first step was to eliminate all ChR sequences predicted to not localize to the plasma membrane or predicted to be non-functional. To do this, we used the ChR function classification model (described above) along with our previously published ChR localization classification model (45) to predict the probability of localization and function for each ChR sequence in the 120,000 variant recombination library. Not surprisingly, most ChR sequence variants were predicted to not localize and not function. Given the limitation of our ChR functionality assay (patch-clamp electrophysiology), we are only interested in assaying ChR sequence variants that are very likely to localize and function. We set a threshold for the product of the predicted probabilities of localization and function; any ChR sequence above that threshold would be considered for the next tier of the process (**Figure 6.1A**). We selected a conservative threshold of either 0.4 or 0.5 (**Figure 6.1A**). This first step eliminates the vast majority of the 120,000 variant library with only 136 sequence variants passing the 0.5 threshold and 1,161 sequence variants passing the 0.4 threshold (**Figure 6.1**).

The model training data made clear that the higher the mutation rate, the less likely it was that a sequence would be functional. We wanted to select the more diverse sequences predicted to function by the classification models. We selected 22 ChR variants that passed the 0.4 threshold that were highly diverse multi-block-swap sequences (i.e. containing on

average 70 mutations from the closest parent). These 22 sequences were synthesized, cloned into the expression vector, expressed in HEK cells, and their photocurrent properties were measured with patch-clamp electrophysiology. 59% of the tested sequences were functional (**Figure 6.2A**), compared to 38% functional sequences in multi-block swap sequences with the same mutation rate, but not predicted by the model. This validates the use of the classification model for making accurate predictions on novel functional sequences, even for those sequences that are more diverse than those previously tested.

For the second tier of the selection process, we used the three regression models to predict the photocurrent strength, wavelength sensitivity of photocurrents, and off kinetics for each of the remaining 1,161 ChR sequence variants. From these predictions, we selected ChR sequence variants predicted to have the highest photocurrent strength, most red-shifted or blue-shifted activation wavelengths, and variants with a range of off kinetics from very fast to very slow. We selected 28 designer ChRs with different combinations of desirable properties that were all predicted to be highly functional (photocurrents > 0.2 nA) and capable of good membrane localization.

The 28 designer ChR variants were selected, synthesized, and cloned into expression vectors, expressed in HEK cells, and characterized for their photocurrent properties with patch-clamp electrophysiology. For each of the designer ChR variants, the three measured photocurrent properties correlated very well with the model predictions (R>0.9 for all models) (**Figure 6.2B**, **Table 6.1**). This outstanding performance on a novel set of sequences demonstrates the power of this data-driven predictive method for engineering designer ChRs with specific sets of properties. As a negative control, we selected two ChR variant sequences from the recombination library that the model predicted would be non-functional (ChR_29_10 and ChR_30_10). As predicted, these sequences were indeed non-functional (**Figure 6.3A**). Interestingly, these non-functional sequences are a single-block swap from two of the most highly functional ChR recombination variants tested and demonstrates how easily ChR functionality can be destroyed by incorporating minor diversity and the value of predictive models as a guide for navigating ChR sequences space.

6.2.4 The machine-guided search identified ChR variants with a large range of useful functional properties

We assessed photocurrent amplitude, wavelength sensitivity, and off kinetics of the designer ChRs and the three parental ChRs [CsChrimR (41), CheRiff (14), and C1C2 (23)] as controls (**Figure 6.3**). For this analysis, we included the top performing ChRs from the classification localization model (ChR_9_4) and classification function model (ChR_25_9) with the 28 regression-model predicted ChRs for a total of 30 highly functional model predicted ChRs as well as the two negative control ChRs (ChR_29_10, ChR_30_10). Of the 30 model-predicted ChRs, we found 13 variants with significantly higher blue-light activated photocurrents than the top-performing parent (CheRiff) (**Figure 6.3A**). Six variants exhibit significantly higher green-light activated photocurrents than CsChrimR (**Figure 6.3A**). Eight variants have larger red-light activated photocurrents when compared with the blue light activated parents (CheRiff and C1C2), though none significantly out-perform CsChrimR (**Figure 6.3A**). Both ChR variants predicted to be non-functional by the models produce <0.03 nA currents.

Characterization of the 30 designer ChRs revealed that their off-kinetics properties fall into a range spanning 4 orders of magnitude (τ_{off} = 10 ms – 1 min) (**Figure 6.3B**). This range is quite remarkable given that all the designer ChRs are built from sequence blocks of three parents that have very similar off-kinetic properties (ranging from τ_{off} = 30-50 ms). We found that 5 designer ChRs have significantly faster off kinetics than the fastest parent while 16 ChRs show significantly slower off kinetics (**Figure 6.3B**). Four ChRs have particularly slow off-kinetics with τ_{off} > 1 s.

6.2.5 Detailed characterization of top designer ChRs

Although all of the designed ChRs are functional, some stand out as having altered off-kinetics, novel spectral properties, or significantly enhanced photocurrents. Short, 1 ms, exposures to blue light elicits distinct profiles from each selected ChR: ChR_21_10 turns off rapidly, ChR_25_9 and ChR_11_10 turn off more slowly, and ChR_15_10 exhibits little decrease in photocurrent 0.5 s after the light was turned off (**Figure 6.4D**).

Three designer ChRs exhibit interesting spectral properties. ChR_28_10 has a red-shifted spectra matching CsChrimR's, demonstrating that even though we are incorporating sequence elements from blue-shifted ChRs into CsChrimR, we can still generate red-shifted activation spectra (**Figure 6.4C**). Two of the designer ChRs show novel spectral properties. ChR_11_10 exhibits a significant broadening of its activation spectra with very strong currents from 400 nm – 570 nm light and even strong currents when activated with 580 nm light (**Figure 6.4C**). ChR_25_9, on the other hand, shows narrowing of its activation spectra relative to the parental spectra, with a peak at 485 nm light (**Figure 6.4C**).

We assessed the light-sensitivity of the designer ChRs with enhanced photocurrents by measuring photocurrent strength at various irradiances (**Figure 6.4B**). We will refer to these high-photocurrent ChRs are 'superconducting' ChRs. All superconducting ChRs show significantly larger currents at all intensities of light tested. The superconducting ChRs also show little decrease in photocurrent over the range of intensities tested (10^{-1} – 10^{1} mW/mm^2) suggesting that the opsin's photocurrents were saturated and that much lower intensities are required to see a photocurrent drop off (**Figure 6.4B**).

We also compared superconducting designer ChRs with ChR2(H134R) (11, 42), an enhanced photocurrent point mutant of ChR2 commonly used for *in vivo* optogenetics, and CoChR *(*from *Chloromonas oogama*) (41) which was reported to be one of the highest conducting ChRs with blue light. The three superconducting ChRs (ChR_25_9, c9_4, ChR_11_10) show significantly larger peak and steady-state currents compared with ChR2, and significantly larger steady-state currents when compared with CoChR with 2 mW/mm^2 485 nm light (**Figure 6.5A**). CoChR did have very strong peak currents, similar to the superconducting ChRs, but rapidly drops off to a much lower steady-state level (**Figure 6.5A**). At lower light intensities (6.5 x 10^{-2} mW/mm^2), the superconducting ChRs show significantly larger peak and steady-state photocurrents than both ChR2(H134R) and CoChR (**Figure 6.5B**). These results demonstrate the potential of these superconducting opsins for optogenetic activation with very low light levels. Lower light power is less invasive because high power, continuous light exposure is phototoxic.

6.2.6 Validation of designer ChRs for neuroscience applications

For validation in cultured neurons and acute brain slices, the superconducting ChRs and ChR2(H134R) were built into AAV viral vectors with the hSyn promoter, TS sequence (43), and eYFP marker and packaged in the engineered AAV-PHP.eB capsid (179). When expressed in cultured neurons, the superconducting ChRs display beautiful membrane localization and expression throughout the neuron (plasma membrane of the cell body and processes) (**Figure 6.6A**). We assessed the light-sensitivity of the superconducting designer ChRs in cultured neurons by measuring photocurrent strength across a large range of light irradiance (10^{-3} – 10 mW/mm^2) using ChR2(H134R) as a comparison (**Figure 6.6B**). Consistent with the pervious experiments, we observe a large increase in the light sensitivity for the superconducting ChRs compared with ChR2(H134R). Both superconducting opsins tested exhibit >200pA photocurrent at the lowest irradiance tested, 10^{-3} mW/mm^2, while at the equivalent irradiance, ChR2(H134R) exhibits undetectable photocurrents (**Figure 6.6B**). The superconducting ChRs reach >1 nA photocurrents with ~10^{-2} mW/mm^2 light, a four-fold improvement over ChR2(H134R)'s photocurrents measured at equivalent irradiance (**Figure 6.6B**). Our characterization of ChR2(H134R)'s light sensitivity and photocurrent strength is consistent with previously published results from other labs (11, 14).

We then assessed neuronal spike fidelity with varying irradiance using ChR2(H134R) for comparison. We observed a 10-10^2 fold decrease in the light intensity required for robust spiking with the superconducting opsins when compared with ChR2(H134R) (**Figure 6.6A,C**). Spike fidelity was validated using both 1 ms light pulses and 5 ms light pulses. These results demonstrate that for neuronal activation, the super conducting designer ChRs require 1-2 orders of magnitude lower light intensity than ChR2(H134R). The superconducting ChRs show robust light-induced firing from 2-20 Hz frequency activation, but have reduced spike fidelity at higher frequency stimulation.

We performed direct intracranial injections into the mouse prefrontal cortex (PFC) of either ChR_25_9, c9_4, or ChR2(H134R) packaged in AAV-PHP.eB (**Figure 6.6D**). After

allowing 3-5 weeks for expression, we measured light-sensitivity in ChR expressing neurons in acute brain slices. As expected, we observe enhanced light sensitivity and photocurrents in the superconducting ChRs when compared with ChR2(H134R) consistent with results observed in cultured neurons (**Figure 6.6D**).

6.2.7 Designer ChRs to enable minimally invasive neuronal excitation

We were particularly interested in determining if these light sensitive, superconducting ChRs could provide optogenetic activation of relatively large volumes of tissue with minimally invasive gene delivery. To deliver opsins to large regions of the brain we used systemic delivery of rAAV-PHP.eB packaging the selected ChRs (**Figure 6.6E**). AAV-PHP.eB is capable of efficiently passing the blood-brain barrier (179), so systemic injection results in expression of opsins throughout the brain (this could be targeted to specific populations with promoters or Cre-lines). Systemic delivery for brain-wide expression is a powerful technique as it avoids the invasive nature of intracranial surgery and targets large volumes of the brain (179). The limitation of systemic delivery is that the number of virus particles transducing neurons in the brain is much lower than with direct injections, and therefore the copy number of ChRs in the cell is lower. This can be an issue with low-conductance channels (e.g. ChR2), but we hypothesized that our superconducting ChRs would overcome this limitation and allow for large-volume optogenetic excitation. We first validated this approach by measuring light-sensitivity in opsin expressing neurons in acute brain slices after systemic delivery (**Figure 6.6E**). As expected, we did find stronger currents in super conducting opsin expressing cells relative to ChR2(H134R). We also observed higher spike fidelity with lower light levels when compared with ChR2(H134R).

Ongoing work in this project is to evaluate the optogenetic efficiency of the superconducting opsins after systemic delivery using behavioral and brain slice experiments. We are attempting optogenetic activation of the motor cortex and superficial layers of the suprachiasmatic nucleus though the scull. Given the light sensitivity of these ChRs, we hope they can be used for non-invasive optogenetic excitation with systemic delivery of AAV-PHP.eB packaging the ChRs coupled with fiber placement on the top of

the skull. Non-invasive optogenetic excitation is an exciting next step in the field of optogenetics. We are also taking advantage of the superconducting opsins to activate the nervous system outside the brain by testing both optogenetic activation of vagal nerve projections and optogenetic activation of the cardiac nervous system after systemic delivery of the superconducting ChRs with ChR2(H134R) as a control. Previous attempts with ChR2 for these experiments have been unsuccessful due to insufficient copy number of the low-conductance ChR2 channel for neuronal activation. We hope that the super-conducting ChRs will overcome this limitation.

6.3 Discussion

We have applied machine learning to overcome a challenging protein-engineering problem. Directed evolution methods have frequently proven to be incredibly powerful for optimizing protein properties; however, directed evolution methods are typically limited to protein properties for which high-throughput screens are available. For protein properties where high-throughput screening is not available or possible, there is no obvious and robust method for engineering. Our approach, data-driven learning of ChR properties, takes a relatively unbiased view of the protein and enables efficient discovery of highly functional and novel ChR variants with relatively little data. In this approach we approximate the fitness landscape of the protein and use it to efficiently search sequence space to select for the top performing variants for a given property. We first eliminate the non-functional sequences, allowing focus on the local peaks scattered throughout the fitness landscape and ignoring the valleys. Then using regression models, we predict sequences that lie on the fitness peaks. We are able to do this for multiple properties simultaneously to build useful ChR tools. This is a generally applicable platform to engineer difficult to screen proteins.

Designing useful ChRs for *in vivo* applications requires optimization of multiple properties; machine learning provides a platform for such ‘simultaneous optimization’ and we were able to build designer variants each with a combination of diverse properties that follow our engineering specifications. Using a relatively ‘limited’ sequence space, we were able to generate variants with large variations in functional properties from off-kinetics of 10 ms to

1 min and photocurrents that far exceed any of the parental constructs or other commonly used ChRs.

We have designed a number of superconducting ChRs with unprecedented light sensitivity and validated their application for *in vivo* optogenetics and demonstrating the activation of large tissue volumes with relatively low light. This could be particularly useful when activating large brain nuclei in mice, or in model systems with larger brains where brain nuclei span much larger tissue volumes relative to the mouse brain (e.g. rats or non-human primates). These opsins could also be particularly useful for optogenetic activation of the peripheral nervous system using systemic delivery of AAV-PHP.eB or AAV-PHP.S (179) packaging the designer ChRs. The superconducting properties of these ChRs could overcome the limitations of the low per cell copy number of ChRs after systemic delivery. Our main ongoing goal is to validate these superconducting tools for non-invasive optogenetics to enable optogenetic activation of brain areas by simply placing fiber optic cables on the skull of mice after systemic delivery of ChRs. This method would enable neuronal excitation with high temporal precision without invasive intracranial surgery for virus delivery or fiber optic implantation and could be particularly useful for relatively superficial brain areas where intracranial surgery is technically difficult due to anatomical constraints (e.g. the dorsal raphe nucleus). Validation of these applications is ongoing, and if successful, could change the way optogenetics experiments are done.

6.4 Materials and methods

6.4.1 Ethics statement

All experiments using animals in this study were approved by Institutional Animal Care and Use Committee (IACUC) at the California Institute of Technology.

6.4.2 Construct design and characterization

The design, construction, and characterization of recombination library chimeras is described in Bedbrook *et al.* (44). Briefly, HEK 293T cells were transfected with purified ChR variant DNA using Fugene6 reagent according to the manufacturer's recommendations. Cells were given 48 hours to express the ChRs before photocurrent measurements. Primary neuronal cultures of rat hippocampal cells (Wistar pups) were prepped at postnatal days 0-1 (Charles-River Labs), and cultured at 37^{o}C, 5% CO_2 in Neurobasal media supplemented with B27, glutamine, and 2.5% FBS. Cells were transduced 3-4 days after plating with AAV-PHP.eB packaging ChR2(H134R), ChR_11_10, ChR_25_9, or ChR_9_4 4-5. Neurons were patched 10-14 days after transduction.

6.4.3 Patch-clamp electrophysiology

Conventional whole-cell patch-clamp recordings were performed in transfected HEK cells, transduced neurons, and acute brain slices to measure light-activated inward currents. Photocurrent recordings were done from cells in voltage clamp held at -70 mV with short light pulses to measure photocurrents. Electrophysiology data was analyzed using custom data processing scripts written using open-source packages in the Python programming language to do baseline adjustments, find the peak and steady state inward currents, off kinetic properties, and spike fidelity.

6.4.4 Imaging

Imaging of ChR expression in neuronal cultures and in brain slices was performed using a Zeiss LSM 880 confocal microscope.

6.4.5 AAV production and purification

Production of recombinant AAV-PHP.eB packaging pAAV-hSyn-X-TS-eYFP-WPRE (X = ChR2(H134R), ChR_11_10, ChR_25_9, and ChR_9_4) was done following method described in (180). Briefly, triple transfection of HEK293T cells (ATCC) using polyethylenimine (PEI). Viral particles were harvested from the media and cells. Virus was then purified over iodixanol (Optiprep, Sigma; D1556) step gradients (15%, 25%, 40%, and 60%). Viruses were concentrated and formulated in phosphate buffered saline (PBS). Virus titers were determined by measuring the number of DNase I–resistant vg using qPCR with linearized genome plasmid as a standard.

6.4.6 Gaussian process modeling

Both the GP regression and classification modeling methods applied in this paper are based on work detailed in (45, 170). Regression and classification were performed using open-source packages in the SciPy ecosystem (173-175). Gaussian process regression and classification models require kernel functions that measure the similarity between protein sequences. We considered three types of kernel functions: squared exponential kernels, Matérn kernels, and polynomial kernels. The hyperparameters and the form of the kernel were optimized using the Bayesian method of maximizing the marginal likelihood of the resulting model.

6.5 Tables & figures

Model	Cross validation	Test set
Classification: function	AUC = 0.78	AUC = 1.0
Regression: peak photocurrent (current strength)	R = 0.65	R = 0.92
Regression: off kinetics	R = 0.75	R = 0.97
Regression: norm. green current (wavelength sensitivity)	R = 0.90	R = 0.96

Table 6.1. Evaluation of prediction accuracy for different ChR property models. Calculated AUC or R value after 20-fold cross validation on training set data for either classification or regression models. The test set for both the classification and regression models was the 28 ChR sequences predicted to have useful combinations of diverse properties. Accuracy of model predictions on the test set is evaluated by AUC (for classification model) or R value (for the regression models).

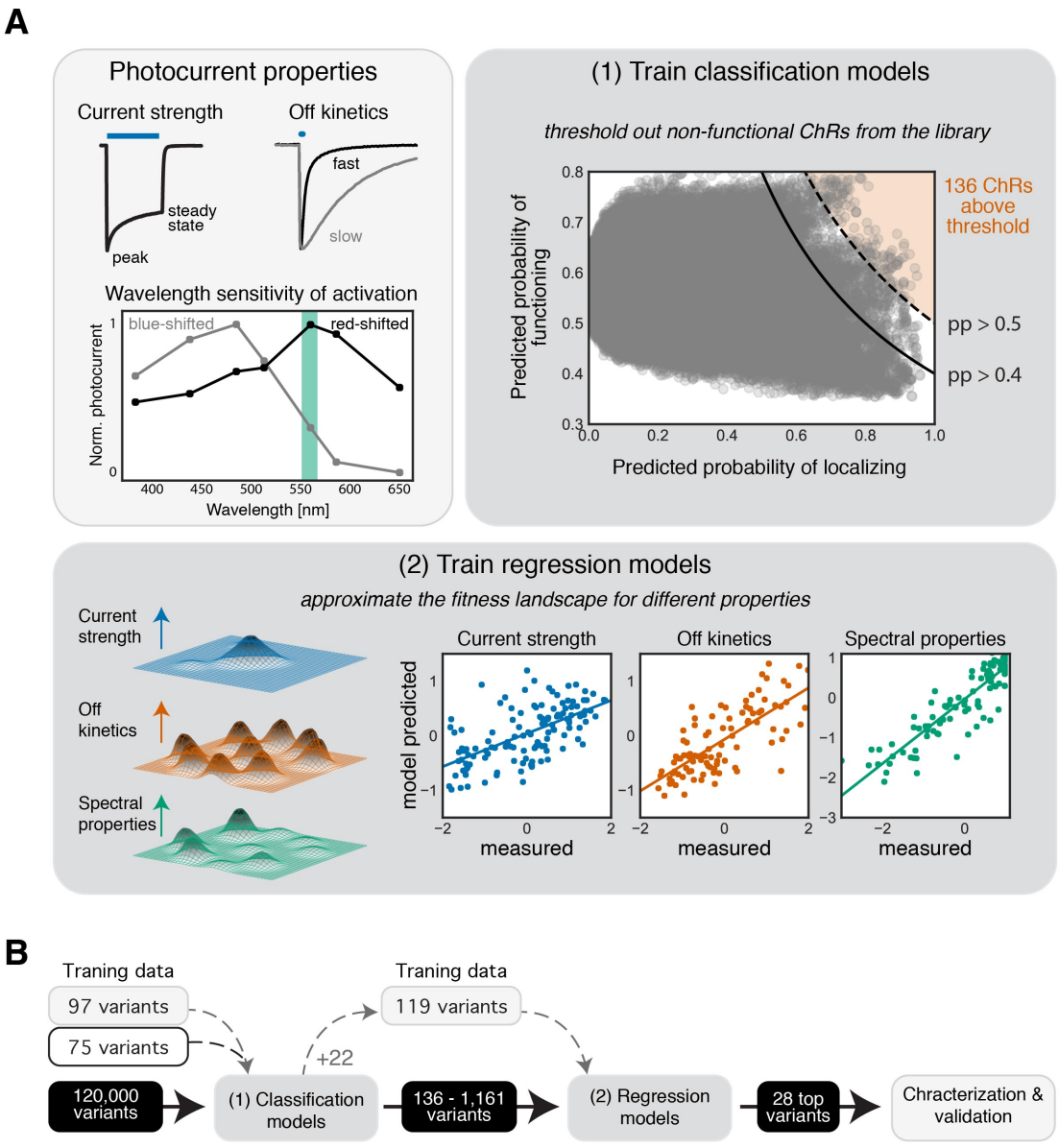


Figure 6.1. Machine-learning guided optimization of ChR photocurrent strength, off kinetics, and wavelength sensitivity of activation. (**A**) Upon light exposure, ChRs rapidly open and reach a *peak* inward current, with continuous light exposure, ChRs desensitize reaching a lower *steady-state* current. Both peak and steady state current are used as metrics for photocurrent strength To evaluate ChR off kinetics, the current decay after a 1 ms light exposure are fit to a monoexponential decay curve and use the decay rate (τ_{off}) as a metric for off kinetics. We also use the time to reach 50% of the light exposed current after light removal as a metric for off kinetics. ChRs are optimally activated by one wavelength

of light and less activated as one moves further from that optimal wavelength. Most ChRs are ‘blue shifted’ with their wavelength of peak activation at ~450-480 nm. Some ChRs are ‘red shifted’ with a wavelength of peak activation between 520-650 nm. We use the normalized photocurrent with green (560 nm) light as a metric for wavelength sensitivity of activation. Variant selection was carried out in tiers (1) Using trained classification models for predicting membrane localization and ChR function to eliminate all the non-localizing and non-functioning opsins in the library, (2) using regression models to approximate the fitness landscape for each property of interest for the recombination library. Models are trained with photocurrent properties for each ChR in the training set such that the model predicted properties correlate well with measured properties. (**B**) Schematic of the trajectory of the machine-learning guided engineering of designer ChRs. The classification function model was trained with 97 variants from our recombination libraries and 75 variants from previously published ChRs.

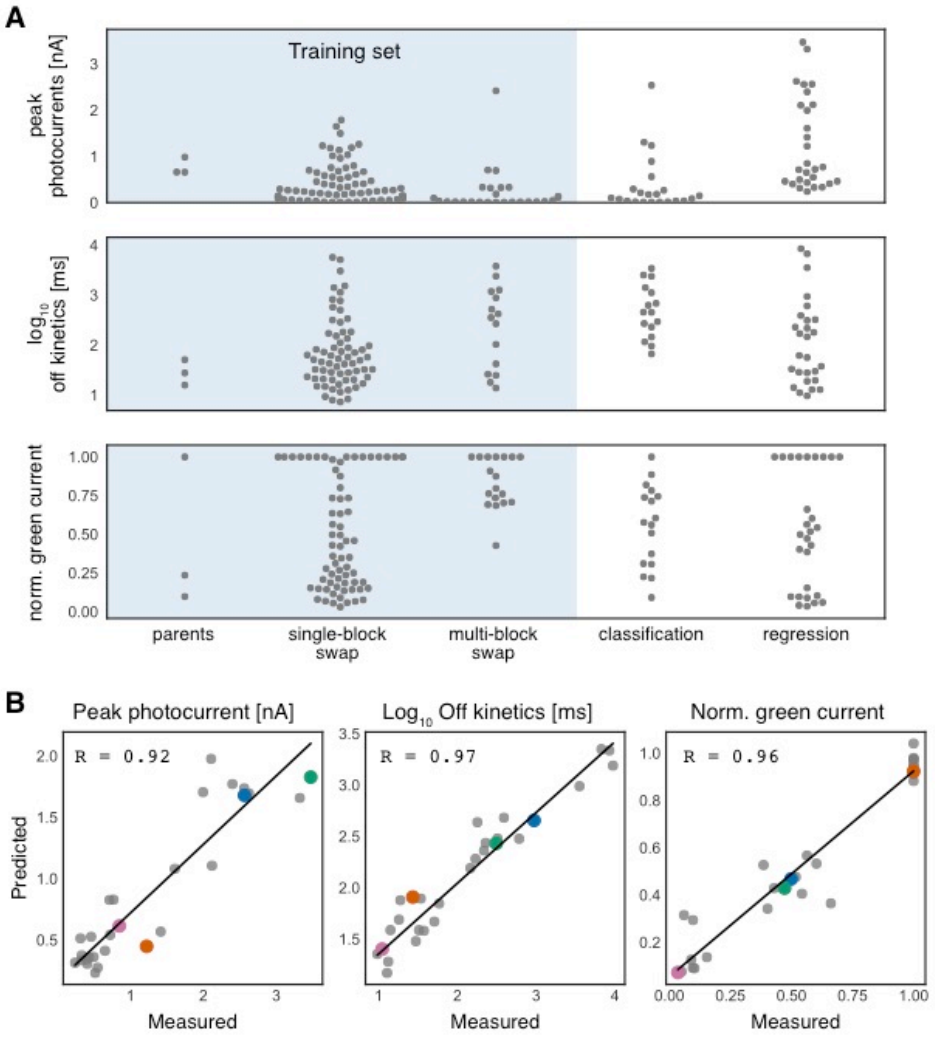


Figure 6.2. Training machine-learning models to predict ChR properties of interest based on sequence and structure enables design of ChR variants with specific collections of desirable properties. (**A**) Measurements of training set ChRs' and model-predicted ChRs', peak photocurrent, off kinetics, and normalized green current. Each gray point is a ChR variant. Training set data is shaded in blue. Mean number of mutations for each set is above the plots. (**B**) Model predictions vs measured property for peak photocurrent, off kinetics, and normalized green current of the 28 designer ChRs shows strong correlation evaluated by R values. Specific ChR variants are highlighted in color to show their predicted and measured properties for all three models. Blue, ChR_12_10, green, ChR_11_10, orange, ChR_28_10, pink, ChR_5_10.

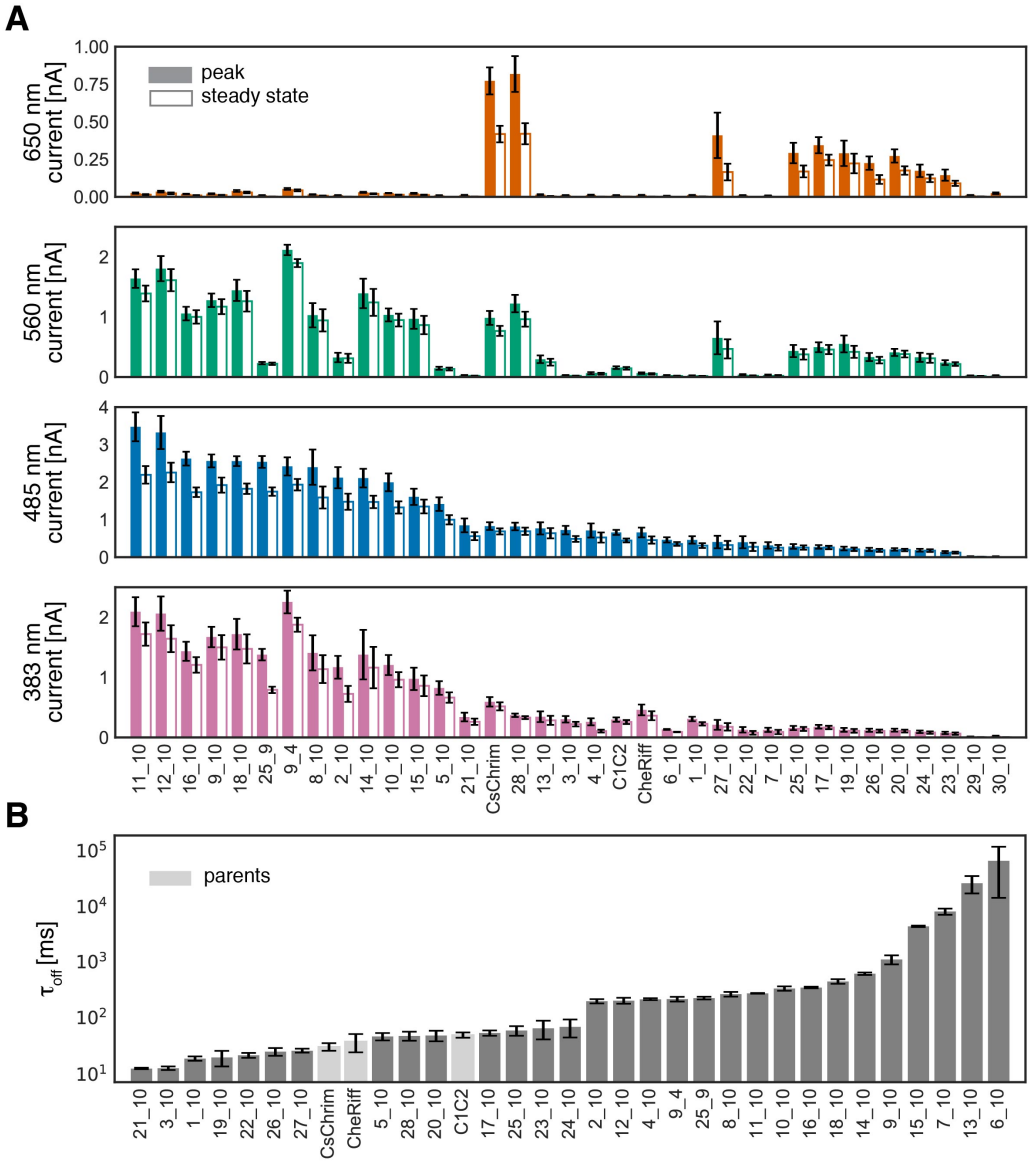


Figure 6.3. The model predicted ChRs exhibit a large range of properties often far exceeding the parents' functional diversity for the same properties. (**A**) Measured peak and steady state photocurrents of all designer ChRs with different wavelengths of light. 383 nm light at 1.5 mW mm^{-2}, 485 nm light at 2.3 mW mm^{-2}, 560 nm light at 2.8 mW mm^{-2}, and 650 nm light at 2.2 mW mm^{-2}. (**B**) Calculated decay (τ_{off}) rate after a 1 ms exposure to light (485 nm light at 2.3 mW mm^{-2}) as a metric for off kinetics for each of the designer ChRs. Parent ChRs are highlighted in light gray.

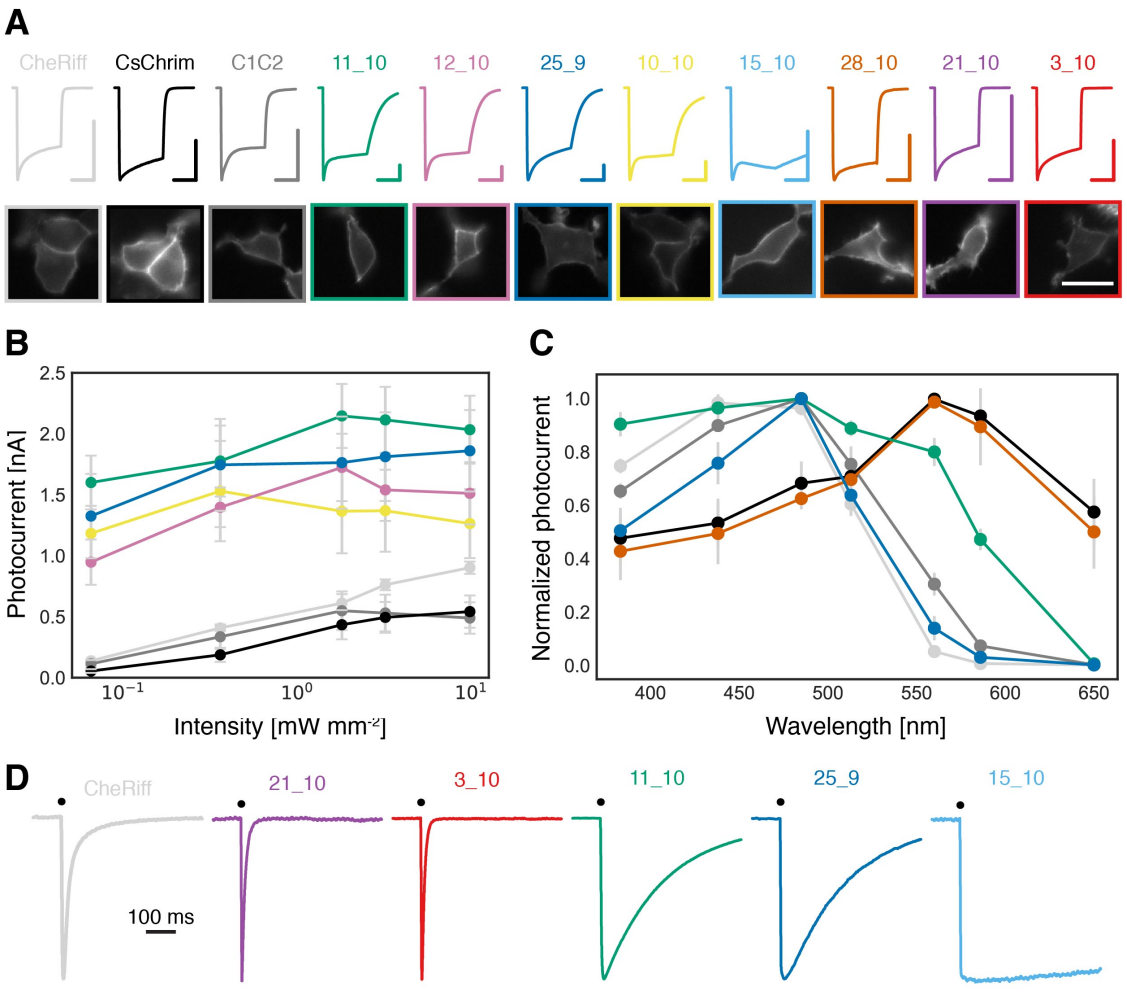


Figure 6.4. Characterization of select designer ChR variants for properties of interest for neuroscience applications demonstrates that our top variants outperform the parental ChRs. (**A**) Current trace after 0.5 s light exposure for each selected ChRs with corresponding HEK cell expression and localization. Vertical colored scale bar for each ChR current trace represents 0.5 nA, and horizontal scale bar represents 250 ms. Different color traces are labeled with each variant's name. The color for each variant as presented in (A) are kept for all other panels of the figure. (**B**) Peak photocurrent strength with varying light irradiances of select ChR variants compared with parental ChRs. (**C**) Normalized photocurrents to measure wavelength sensitivity of activation for select ChRs compared with parental ChRs. (**D**) Trace of current decay after 1 ms light exposure for select ChRs compared with CheRiff show the diversity of off-kinetic properties produced from designer ChR variants.

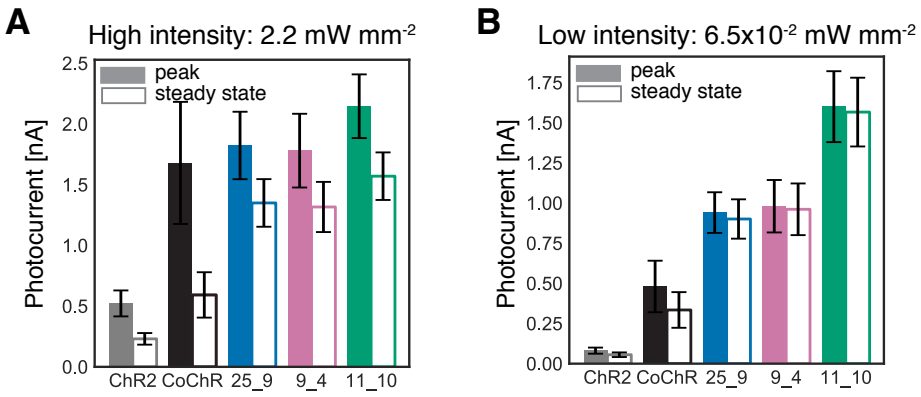


Figure 6.5. Comparison of superconducting ChRs with ChR2(H134R), a commonly used ChR for *in vivo* optogenetics, and CoChR *(*from *Chloromonas oogama*) reported to be one of the highest conducting ChRs with blue light. Peak and steady-state current measurements show that ChR_25_9, ChR_9_4, and ChR_11_10 all exhibit significantly stronger photocurrent peak and steady state current than ChR2(H134R) at both light intensities tested. While CoChR has comparable peak current to the superconducting opsins in high intensity light conditions, CoChR exhibits significantly lower steady state currents under the same conditions. CoChR is also less sensitive to low light intensity, as shown by its significantly lower peak and steady state currents in low light conditions compared with the superconducting opsins.

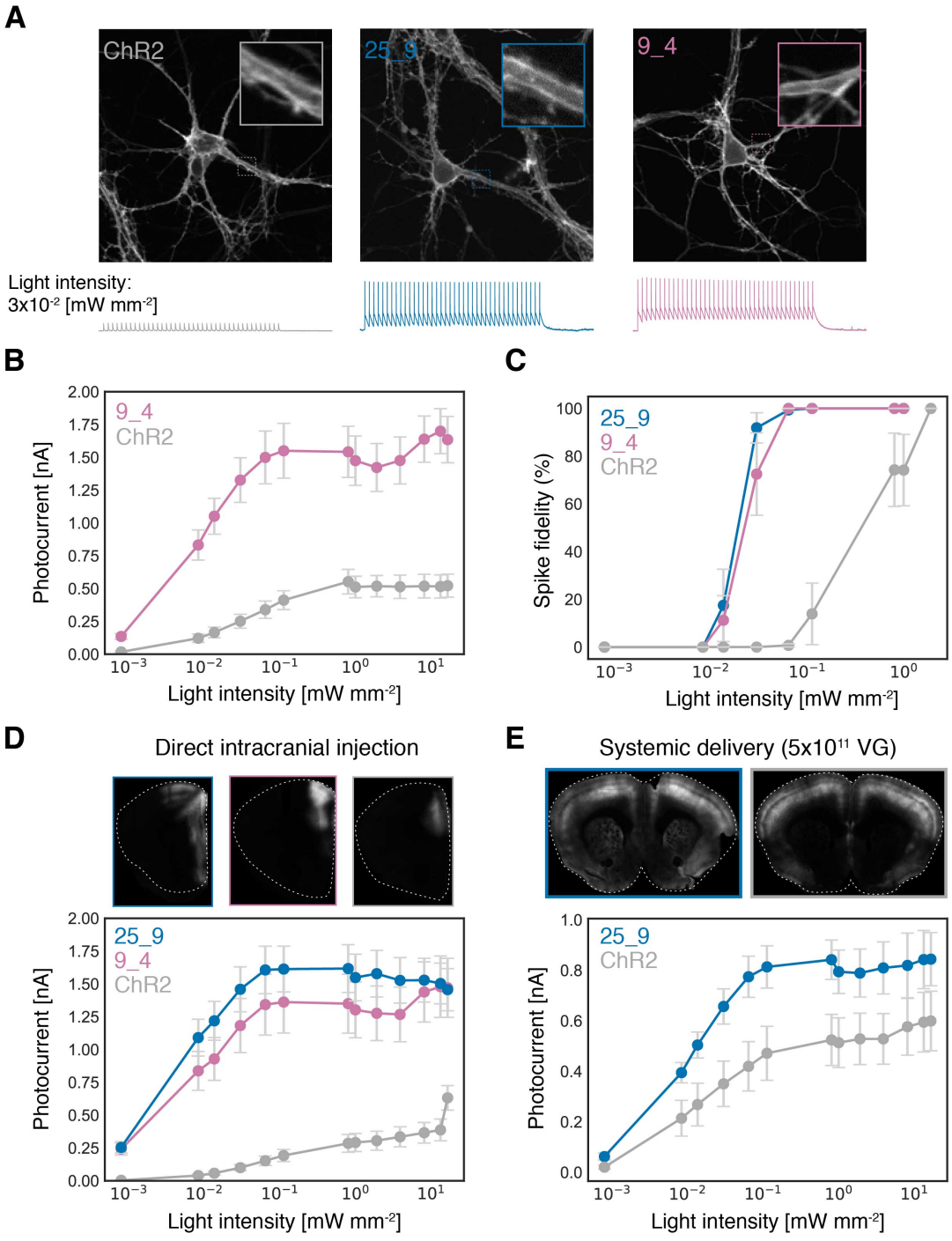


Figure 6.6. Application of superconducting ChR variants in cultured neurons (**A-C**) and in acute brain slices (**D-E**) demonstrates that these variants outperform commonly used ChR2(H134R). (**A**) In cultured neurons ChR_25_9 and ChR_9_4 exhibit good expression and membrane localization similar to ChR2(H134R). Both ChR_25_9 and ChR_9_4

produce robust light-induced neuronal firing at 2 Hz with 5 ms pulsed blue light stimulation with low light intensity ($3x10^{-3}$ mW mm^{-2}), while under matched conditions, ChR2(H134R) exhibits only sub-threshold light-induced depolarization and no neuronal firing. (**B**) ChR_9_4 exhibits improved current strength with all light intensities tested relative to ChR2(H134). (**C**) Both ChR_25_9 and ChR_9_4 exhibit 100% spike fidelity with 1-2 orders of magnitude lower intensities of light compared with ChR2(H134R). (**D**) Direct intracranial injection of AAV-PHP.eB packaging ChR_25_9, ChR_9_4, or ChR2(H134R) into the PFC resulted in local expression at the injection site. Consistent with the neuronal culture data, both ChR_9_4 and ChR_25_9 exhibit improved current strength with all light intensities tested relative to ChR2(H134). (**E**) After systemic delivery of either ChR_25_9 or ChR2(H134R) at packaged in AAV-PHP.eB at $5x10^{11}$ vg/animal, ChR_25_9 shows improved current strength relative to ChR2(H134R) in opsin expressing neurons. Both current and voltage recordings were done with whole-cell patch clamp measurements. vg, viral genomes.

Chapter 7

VIRAL STRATEGIES FOR TARGETING THE CENTRAL AND PERIPHERAL NERVOUS SYSTEMS

7.1 Abstract

Recombinant viruses allow for targeted transgene expression in specific cell populations throughout the nervous system. The adeno-associated virus (AAV) is among the most commonly used viruses for neuroscience research. Recombinant AAVs (rAAVs) are highly versatile and can package most cargo composed of desired genes within the capsid's ~5-kb carrying capacity. Numerous regulatory elements and intersectional strategies have been validated in rAAVs to enable cell type–specific expression. rAAVs can be delivered to specific neuronal populations or globally throughout the animal. The AAV capsids have natural cell type or tissue tropism and trafficking that can be modified for increased specificity. Here, we describe recently engineered AAV capsids and associated cargo that have extended the utility of AAVs in targeting molecularly defined neurons throughout the nervous system, which will further facilitate neuronal circuit interrogation and discovery.

7.2 Introduction

Owing to its massive complexity, the nervous system cannot be understood holistically without first understanding its parts. The approach of contemporary systems neuroscience is to use genetic dissection of individual neuronal circuit*s* to understand function and behavior within the context of the whole nervous system. Genetic elements regulate transcription or translation to build different cell types. Manipulating and harnessing these genetic elements allow for specific targeting of transgene expression to neuronal subpopulations within the highly heterogeneous cell populations in diverse tissues throughout the body. Today, systems neuroscience uses a rapidly expanding set of genetically encoded tools: cell markers to identify cells, actuators to control cell functions, and sensors to monitor cell state.

Neuroscientists want specificity but often also face the challenge of delivering genetic cargo to target neuronal subpopulations distributed across the central nervous system (CNS) or peripheral nervous system (PNS). Genetically encoded cell type specificity must be coupled to advanced methods for gene delivery to the desired cell population. Delivery can be achieved through transgenesis, physical methods (including electroporation and DNA particle bombardment), and viral vectors, the subject of this review. Viral vectors offer a versatile and fast platform for delivering transgenes and testing new genetically encoded tools. Here, we review the viral vectors and methods that enable genetic access to neuronal subpopulations throughout the nervous system.

The recombinant vectors used are based on viruses that evolved clever strategies for hijacking cellular machinery for transduction, transcription, gene expression, replication, and, in some neurotropic viruses, transneuronal spread. The nonpathogenic adeno-associated virus (AAV) is among the most commonly used viruses for neuroscience and therapeutic applications.

7.3 Why AAVs for neuroscience?

Recombinant adeno-associated viruses (AAVs) are widely used as vehicles for gene transfer to the nervous system for anatomical and functional circuit mapping and modulation (**Figure 7.1**). Since its discovery in 1965, the AAV's life cycle, components, structure, and mechanism of transduction have been studied in depth and reviewed extensively (especially for the AAV2 serotype) (181). These early studies enabled the development of recombinant AAV (rAAV) vectors for gene delivery applications (182) and ultimately led to their optimization as a safe gene delivery system that is easy to produce in a standard laboratory setting (183, 184). Most AAV serotypes transduce terminally differentiated or nondividing cells, such as neurons, with high efficiency; some serotypes also transduce astrocytes, oligodendrocytes, endothelial cells, and ependymal cells (185-189), making AAVs an excellent choice for stable, long-term expression of transgenes in cells throughout the nervous system (190).

The natural AAV genome is composed of two genes: *rep*, which encodes four nonstructural proteins for replication, and *cap*, which encodes the viral capsid proteins and the assembly-activating protein, which chaperones capsid assembly (191). The genome is single-stranded and flanked by inverted terminal repeats (ITRs, 145 bp) that form T-shape hairpins and contain the *cis*-elements required for genome replication, genome packaging, and the generation of stable episomes (**Figure 7.2**). Current rAAV vector systems deliver the desired transgenes into cells with minimal risk of replication and further infection because the *rep* and *cap* genes are deleted from the recombinant vectors' genome; these functions are provided in *trans* during viral vector production. The transgene of interest replaces the *rep* and *cap* genes in between the ITRs, which demarcate the genome and ensure that the genetic cargo is packaged into the capsid. Any sequence less than 4.7 kb can be inserted between the ITRs for packaging and delivery. The process of transduction is shown in **Figure 7.2**.

rAAVs are relatively low risk for the user and well tolerated by the subject. Their low integration rate minimizes possible disruption of host genomic components but makes AAVs poorly suited for lineage tracing studies and persistent gene expression in dividing cells. For these applications, recombinant oncogenic retroviral or lentiviral vectors are preferred.

AAVs facilitate cell type–specific expression of genetically targeted markers (fluorescent proteins), actuators (chemogenetic, optogenetic, and thermogenetic tools), neuronal activity sensors (calcium and voltage indicators), and gene editing tools such as CRISPR–Cas9 (192, 193). Over 600 recombinant AAV genomes harboring various markers, actuators, and sensors with different regulatory elements are available on Addgene (https://www.addgene.org/). With the development of helper-virus-free production methods, AAVs are now relatively easy to produce and can be concentrated to high titers [10^{11}–10^{14} viral genomes (vg)/ml] (194, 195). Common vector cores (e.g., from the University of Pennsylvania, University of North Carolina, Stanford University, Salk Institute for Biological Studies, and now Addgene) provide purified virus for natural and engineered AAV serotypes (currently >16 serotypes), packaging a diversity of genes under

different promoters. The use of AAVs for transgene delivery in large-scale, standardized Mouse Connectivity Database project at the Allen Institute for Brain Science is evidence of their prominent role in neuroscience (http://connectivity.brain-map.org/) (101). These recombinant viruses are well tolerated, produce persistent transgene expression, and facilitate relatively rapid testing in behavioral experiments.

7.4 Targeted expression in the central and peripheral nervous system with AAVs

The utility of AAV vectors for circuit studies depends on our ability to employ them for transgene expression within specific neuronal populations. Three factors contribute to AAV-mediated specificity: the delivery method, the packaged cargo, and the capsid protein (**Figure 7.3**). Different combinations of these factors provide numerous options for achieving the desired expression, from expression in a highly specific set of 10–1,000 neurons in a confined nucleus to global expression throughout both CNS and PNS. To aid in the design of customized vectors for cell type specificity, we have summarized several delivery methods (Delivery, **Supplemental Table 7.1**), genomic regulatory elements (Cargo, **Supplemental Table 7.2**), and application of commonly used serotypes (Capsid, **Supplemental Table 7.3**).

7.4.1 It's all in the delivery

To enable targeted delivery of AAVs to desired regions or neuronal populations within the nervous system, specific anatomical properties can be harnessed. Site-directed intracranial injections, the most common method, can be precisely localized with stereotaxic coordinates. Injection volume, rate, flow orientation/shaping, viral titer and formulation buffer, and AAV serotype influence the spread and transduction pattern of the virus. Time postinjection is also a factor in the expression level of transduced cells; generally, expression increases over time but plateaus by 4-6 weeks. However, specific expression dynamics depend on the promoter used and the transgene being expressed. Intraparenchyal brain infusions result in focal transgene delivery, with highest expression at the injection site and decreasing expression farther from the injection site. This spacially limited and nonhomogeneous expression pattern can be problematic when studying neuronal

populations extending over large brain volumes. Alternatively, AAV injection into the cerebrospinal fluid (e.g., intracerebroventricular, intracisterna magna, or intrathecal injections) has been explored as a method for achieving wide distribution within the brain, spinal cord, and dorsal root ganglia (DRG) (196-203).

Unlike direct injections, systemic injections through either the lateral tail vein or the retro-orbital venous sinus (180, 204) can provide widespread delivery of transgenes throughout the CNS. The difficulty in targeting the brain via systemic delivery lies in bypassing the highly selective blood–brain barrier (BBB). A few naturally occurring AAVs (e.g., AAV9, AAVrh8, and AAVrh10) can pass the BBB at low efficiency, resulting in sparse transgene expression, targeting primarily astrocytes, within the CNS when delivered in adult mice (204, 205). CNS transduction was most efficient when AAVs were delivered intravenously in neonatal mice (205, 206). Attempts to improve transduction in the adult CNS by transiently opening the BBB have included administration of AAVs following the use of chemical measures (207, 208), focused ultrasound for localized, temporary BBB disruption (209), and seizure induction (seizure-compromised BBB) (210).

Recent engineering efforts have yielded several AAV variants including, AAV-AS (211), AAV-PHP.B (180), and AAV-PHP.eB (179), that can efficiently transduce the CNS via systemic delivery in adult mice without additional means to open the BBB. Systemic delivery of AAV-PHP.eB can transduce ≥50% of several neuron populations in the brain with a vector dose of 5×10^{12} vg/kg (179), enabling efficient delivery of transgenes throughout the CNS via a single noninvasive injection in the adult mouse.

Site-directed injection in the PNS is possible for some targets, although the PNS lacks a standard coordinate system, making it challenging to accurately and reproducibly target specific ganglia or nerves. Nevertheless, AAV injections into peripheral nerves or ganglia have been used successfully. AAVs have been delivered to the nodose and jugular ganglia to monitor and modulate the vagus nerve and to map vagal projections to the lung and gut (212, 213). Delivery of AAVs by direct DRG injection, intrasciatic injection, or intrathecal administration has been used for studies of chronic pain in mice and rats (214-216). Direct

injection of AAVs into the lumbar dorsal horn enabled targeting and interrogation of a brainstem–spinal cord circuit important for pain modulation (217). Intramyocardial delivery of AAV9 vectors enabled optogenetic control of cardiac pacing (218, 219).

Intravenous delivery is a powerful alternative for certain peripheral neuron populations that are difficult to access surgically (e.g., DRG, nodose ganglia, sympathetic chain ganglia, and cardiac ganglia) or are widely distributed (e.g., the enteric nervous system). Systemic administration of AAV9 has been used to transduce PNS neurons and many peripheral organs (179, 219-221). Recent work from Chan et al. (2017) produced an engineered AAV variant, AAV-PHP.S, that, when delivered systemically, efficiently and broadly transduces PNS neuron populations (179).

The newly engineered AAVs described above enable systemic gene expression throughout the CNS and PNS. Neuroscientists new to the use of systemically delivered AAVs should be aware of several considerations associated with their use. First, the amount of vector required is 1–2 orders of magnitude higher than that required for localized injections. Second, systemically delivered AAVs have increased exposure to immune surveillance; so if either the transgene product or the capsid is immunogenic, transgene expression and transduced cells may be lost over time (222). Third, in addition to transduction of cells within the CNS and PNS, the promiscuity of many AAV vectors leads to transduction of the liver, heart, muscle, and numerous other organs (180, 223, 224). Therefore, specifically targeting gene expression to cells of the CNS or PNS after systemic delivery requires the use of gene regulatory elements or recombinase-based intersectional strategies.

7.4.2 Cargo

To facilitate constitutive or inducible transgene expression in specific cell types within the nervous system, various transcriptional or translational regulatory elements and recombinase recognition target elements can be included within the recombinant viral genome.

Specific promoters and enhancers. The restricted expression of many endogenous gene products is enabled by promoters and enhancers distributed over large genomic lengths. Identifying promoters, enhancers, and other regulatory elements that are short enough to be compatible with AAVs' limited packaging capacity, yet still capable of cell type–specific expression, remains a challenge (225-227). One approach uses regulatory elements from naturally concise genomes, such as that of fugu (*Takifugu rubripes*) (228). The large-scale Pleiades Promoter Project has identified Mini-Promoters (MiniPs) for various expression targets (227, 229). A subset of these elements have been validated for use in rAAVs and show conserved expression specificity in a range of cell types (**Supplemental Table 7.2**) (229, 230). Two MiniPs, Ple67 (containing *FEV* regulatory regions, specific to serotonergic cells) and Ple155 [containing Purkinje cell protein 2 (*PCP2*) regulatory regions, specific to Purkinje cells] (230), along with promoters for glial fibrillary acidic protein (GFAP) (231), myelin basic protein (MBP) (232), human Synapsin I (hSyn1) (233), and tyrosine hydroxylase (TH) (179) (**Supplemental Table 7.2**), have also been validated for cell type–specific expression after systemic delivery in either rAAV-PHP.B or eB (179) (**Figure 7.3*c***). Specific enhancer elements have been used for interneuron-specific expression via AAV transgene delivery via direct intracranial injections in multiple species [mDlx1/2 (234), mDlx5/6 (235)] or via systemic delivery in AAV-PHP.eB in the adult mouse [mDlx5/6 (179)]. Recent enhancer element screens (236) could be a resource for short enhancer elements that are compatible with AAVs and provide cell type–specific expression, following validation for use in AAVs.

Both promoter specificity and promoter strength are important to consider for neuroscience applications. Some commonly used transgenes (e.g., GCaMP and Cre recombinase) are not well tolerated when expressed at high levels and may benefit from promoters capable of low or tunable expression. Recombinases (e.g., Cre) used in intersectional expression strategies are highly efficient and may provide only cell type–restricted recombinase activity with tightly regulated promoters given that low-level leaky or transient expression is sufficient to permanently induce Cre-dependent transgene expression. Conversely, many neuronal actuators (e.g., ChR2, NpHR, and Arch) require high expression to produce sufficient ionic flux for neuronal activity modulation.

Inducible promoter systems [e.g., tetracycline (Tet)-inducible systems (237)] allow temporally controlled gene expression and can be packaged within the rAAV genome (179, 238, 239). Activity-dependent promoters allow for transgene expression in activated neurons via immediate early genes that are rapidly upregulated following increases in neuronal firing (240, 241). Further validation of existing promoters and enhancers and development and identification of improved promoters and enhancers compatible with AAVs are a necessary next step for high-resolution cell type–specific targeting.

miRNA target sequences. Transgene expression can also be regulated posttranscriptionally by inserting tandem copies of short microRNA (miRNA)-target sequences (miRNA-TSs) within the 3′ untranslated region of the rAAV genome. miRNAs are short, noncoding regulatory RNAs involved in RNA-mediated posttranscriptional gene silencing by binding complementary sequences in protein-coding mRNAs. Various miRNAs are expressed in a tissue-specific manner and have an important role in maintaining tissue-specific functions and differentiation (242, 243). Inclusion of an miRNA-TS within a transgene cassette can be used to reduce transgene expression in tissues or cells where the complementary miRNA is expressed. In this way, miRNA-TSs can be used to reduce off-target tissue expression after systemic rAAV delivery. For example, miRNA-TSs complementary to miR-122, enriched in liver hepatocytes, and miR-1, enriched in cardiac and skeletal muscle, when incorporated into the recombinant genome (244), successfully reduce transgene expression in both heart and liver while maintaining transgene expression in the CNS without perturbing the function of the endogenous miRNA (244). Numerous studies have used miRNA-TSs incorporated within AAV genomes to reduce expression in off-target tissues 2- to 100-fold with little to no effect on expression in target cell types (**Supplemental Table 7.2**). Multiple copies (3× is typically used) of a given miRNA-TS enhance silencing. Care should be taken to ensure that the endogenous miRNAs are not sequestered by saturating levels of the miRNA-TS delivered with the transgene (245). Given their small size, multiple copies of five to six different miRNA-TSs can easily fit into an AAV vector. More work is needed to identify CNS and PNS cell type–specific miRNA expression patterns, as already demonstrated for cell populations within the retina (246).

Intersectional expression strategies. To achieve refined cell type specificity or to bypass the AAV packaging limit, researchers have outlined a few strategies. In typical two-component intersectional transgene expression systems, the transgene of interest is encoded within a rAAV vector (first component) but is silent in the absence of a second component (i.e., inducer) that is, for example, driven by full-length, cell type–specific genetic regulatory elements incorporated within engineered animal cells. The rAAV transgene is expressed only within the subset of cells expressing the second component. Commonly used two-component systems compatible with AAVs are presented in **Supplemental Table 7.2**; chief among them are recombination-based systems, such as the bacteriophage P1-derived Cre-*lox* recombination system. However, the use of Cre is largely confined to mice. Numerous Cre mouse lines have been made available for specific neuronal cell types (e.g., the Jackson Laboratory Cre Repository and the GENSAT Project at the Rockefeller University), while other model systems (e.g. rats and non-human primates) have few or no engineered Cre line options.

Recombinase systems based on two or more regulatory elements or administration routes can also be combined to enable conditional transgene expression, which is useful because most cell types are not defined by a single genetic feature. The Cre-*lox* recombinase system does not cross-react with the flippase (Flp)-flippase recognition target (FRT) recombinase system (247). Cre and Flp can be expressed under different regulatory elements, and both recombinase target sites can be encoded in the AAV-packaged transgene. This approach, intronic recombinase sites enabling combinatorial targeting (INTRSECT), has been used to target specific subtypes of inhibitory interneurons in mammalian hippocampus, enabling transgene expression exclusively in the subpopulation of cells containing both parvalbumin and somatostatin (247). This technology has the potential to increase the cell- or region-specific resolution of genetically encoded transgenes for neuroscience applications (248).

Two-component systems can be coupled with systemic virus delivery to achieve sparse but strong transgene expression throughout the CNS or PNS. For example, the first component, carrying the transgene of interest, is delivered at a high titer and is under the control of an inducer [e.g., the tet-off transactivator (tTA) (238)], which is co-introduced at a

controllable low titer. Only the subset of cells containing the sparsely delivered tTA expresses the transgene of interest that, because of its delivery at high titer, is expressed at sufficient levels to produce stochastic sparse labeling akin to Golgi staining but with the advantage of being genetically encoded and targetable to specific cell types. This approach was used to deliver multiplexed florescent proteins for neuronal tracing, and was named vector-assisted spectral tracing (VAST) (179).

Alone and in combination, gene regulatory elements such as cell type–specific or temporally controlled promoters, two-component systems, and miRNA-TSs can be combined to enhance cell type specificity throughout the nervous system.

7.4.3 Capsid

The AAV capsid is made up of 60 protein monomers of viral capsid protein 1 (VP1), VP2, and VP3, in approximately a 1:1:10 ratio, assembled into an icosahedral structure roughly 25 nm in diameter (249). The amino acid sequence of the capsid influences its biodistribution and in vivo cell type- and tissue-specific tropism. AAVs typically use cell surface proteoglycans (e.g., sialic acid, galactose, or heparin sulfate) as primary receptors and cell surface proteins (e.g., fibroblast growth factor receptor, laminin, or integrins) as secondary receptors (250). The tropism and biodistribution of many AAV serotypes have been analyzed, revealing patterns of transduction that vary by brain region (251-253). Natural AAV capsids capable of cell type–specific transduction have not been found. However, AAV capsids can be engineered to produce variants with altered properties (see below). Though none of the current set of engineered AAV capsids are truly neuronal subtype specific, many of them favor defined cell types (180, 254, 255), and may provide an additional degree of specificity when used to package transgene cassettes harboring cell type–specific regulatory elements.

7.5 Engineering designer AAV capsids

Advancing the utility of AAVs for neuroscience has required optimization of the protein capsid for altered tropism, improved transduction efficiency, increased BBB crossing, and

enhanced axonal trafficking. The prominence of AAV as a delivery vehicle for gene therapy (256) has led to many efforts in capsid engineering, providing many useful approaches and an empirical view of the various AAV properties accessible through capsid engineering. To identify enhanced capsids, it is necessary to (*a*) find or create sequence diversity within the capsid proteins and (*b*) assess the resulting capsids for the desired properties. The number of sequence variants in such characterization experiments dictates the assay approach (high- versus low-throughput), and the desired property dictates whether optimal capsids can be identified by screening or selection.

7.5.1 Natural and engineered AAV capsid diversity

AAV capsid diversity was initially sourced from nature. Since the discovery of wild-type AAV2 as a contaminant in human-derived laboratory preparations of adenovirus, AAVs have been recovered from various animals, such as humans and nonhuman primates (257), pigs (258), and even snakes (259). A collection of more than 100 unique, full-length *cap* genes from human and nonhuman primate tissue shows most sequence divergence at the outer surface of the viral particle (257). Thirteen natural AAVs have been characterized for neuronal transduction and are available through vector cores for packaging rAAV cassettes (260, 261). rAAVs optimal for specific neuroscience applications are highlighted in **Supplemental Table 7.3**.

Capsid sequence diversity can also be created in the laboratory by either random or rational diversification approaches. Capsid sequences can be recombined to build chimeric capsids, mostly through random DNA-shuffling techniques (210, 262, 263). Random mutagenesis methods [e.g., error-prone polymerase chain reaction (PCR)] can be used to incorporate nucleotide changes throughout the capsid (264). Structural data can guide capsid diversification methods; to date, 25 AAV capsid structures, both wild type (249) and engineered (265), have been collected in the Protein Data Bank. Capsid proteins have a highly conserved eight-stranded β-barrel core and expansive loops that connect the β-strands that share less homology and create unique surface features (**Figure 7.4*a***). These structures guide targeted mutagenesis or insertions at specific residues likely important for

various properties; for example, surface-exposed tyrosine-to-phenylalanine substitutions enhance transduction and reduce proteasome-mediated degradation (266). Insertions or replacements of random peptides in the variable loop regions can dramatically alter AAV tropism and biodistribution and enhance axonal trafficking (179, 180, 255, 267, 268). To aid targeted and rational mutagenesis approaches, researchers have elucidated sequence elements required for specific functional properties through high-throughput mutagenesis and sequencing. Specifically, double alanine mutational scanning along parts of the AAV9 capsid protein and systematic hexapeptide swapping from various AAV serotypes into AAV2 at positions 441–484 and 571–604 mapped the capsid amino acids important for structural integrity, receptor binding, tropism, neutralization, and blood clearance (269). These sequence-function maps are useful for future targeted capsid diversification. They have already enabled the conversion of AAV2's natural heparan sulfate proteoglycan binding to galactose binding through substitution of 10 residues from AAV9, and have enabled the design of liver-detargeted variants of AAV9 (269). The optimal diversification method depends on the capsid property being sought (270), and multiple diversification strategies are often combined to achieve the desired goal (255, 268).

7.5.2 Selection and screening of optimal capsids

Once a diverse set of capsids is obtained, it is necessary to screen or select for specific capsids capable of the desired function. For small-scale sets of sequences, the transduction characteristics of each capsid can be assessed via systemic and direct injection (211, 224), and whole-body clearing methods can facilitate assessment of AAV capsid biodistribution (109, 180, 271, 272). High-throughput in vivo selection methods using PCR- or adenovirus-based amplification of capsid sequences from specific tissues can pull out capsid sequences with specific tissue biodistribution or blood clearance after systemic delivery or direct intracranial injections. Reiterated selection is used to enrich for the most effective capsids. However, this amplification and enrichment are not sufficient for selection of cell type–specific capsids.

Cre recombination–based AAV targeted evolution (CREATE) enables selective recovery of capsids that transduce specific cell populations expressing Cre (**Figure 7.4**) (180). In the presence of Cre, a sequence adjacent to the *cap* gene is inverted, and this inversion can be detected by PCR-mediated amplification. In the absence of Cre recombination, amplification does not occur. This method was first applied to select for capsid variants that cross the BBB to transduce cells within the CNS. Two rounds of selective amplification in Cre^{+} transgenic mice produced four designer viruses for intravenous delivery: AAV-PHP.A efficiently transduces CNS astrocytes but with reduced tropism for peripheral organs, and its use was confounded by low vector production. AAV-PHP.B and two additional variants (AAV-PHP.B2 and AAV-PHP.B3) efficiently transduce neurons and astrocytes (180). More recently, an enhanced AAV-PHP.B variant, AAV-PHP.eB, that transduces neurons and glia was found by implementing the same method in multiple Cre lines (Vglut2-IRES-Cre, Vgat-IRES-Cre, and GFAP-Cre mice) in parallel (179). AAV-PHP.eB achieves more efficient adult CNS neuron transduction after systemic delivery, reducing the amount of required virus. The CREATE method also yielded the variant AAV-PHP.S, which can efficiently transduce peripheral neurons (179). The CREATE method should be compatible with localized injections in any available Cre line or Cre delivery scheme.

Amplification-based methods provide positive selective pressure but alone do not enable the selection of capsids that are detargeted from specific tissues or cell types. Next-generation sequencing (NGS)-based methodologies may overcome this limitation. Deep sequencing of AAV capsid genes pre- and postselection permits researchers to analyze the level of enrichment of a capsid sequence in multiple tissues or cell types. Capsids exhibiting targeted tropisms should be enriched in one cell type or tissue and not in others. NGS-guided screening has identified a capsid that specifically transduces the endothelium of the pulmonary vasculature (273, 274). The NGS-based method AAV Barcode-Seq screens and characterizes the phenotypes of many AAVs via DNA barcode tags to label AAV variants (269). This method allows for simultaneous characterization of over a hundred different capsid variants by pooling sequences and then identifying them by their

barcode tag, and it is compatible with many in vitro and in vivo capsid phenotypic assays relevant to neuroscience.

7.6 Application of designer AAVs for widespread delivery to neuronal circuits

Systemically delivered rAAV-PHP.B and rAAV-PHP.eB decrease the need to cross transgenic animals for brain-wide expression experiments (**Figure 7.5**), and enable the delivery of multiple activity sensors with controlled labeled-cell density. For example, systemic delivery of the BBB-crossing variant rAAV-PHP.B permitted noninvasive, widespread expression of fluorescent neuronal activity sensors throughout molecularly defined cortical neurons to capture the dynamics of large-scale neuronal populations during behavior (275). Delivery of rAAV-PHP.B carrying GCaMP6 sensors (56) targeted to inhibitory cells (via Gad2-Cre line) and jRCaMP1b (276) targeted to excitatory cells (via CaMKIIα promoter) within the rAAV genome allowed simultaneous recording of inhibitory and excitatory dynamics during learned behavior (**Figure 7.5*b***). Even though jRCaMP1b transgenic lines are not yet available, widespread cell type–restricted jRCaMP1b expression was possible through the use of rAAV-PHP.B, highlighting the utility of virally mediated transgenesis for rapid and inexpensive validation and optimization of new tools in vivo. Hillier et el. (2017) used rAAV-PHP.B packaging of GCaMP6 for widespread labeling of neurons in the visual cortex after systemic delivery and reported that the vector yielded no nuclear expression of GCaMP6 for at least 10 weeks, indicative of the stable health of cells after rAAV-PHP.B–mediated GCaMP6 expression (277). Administration of the rAAV-PHP.B vector via a single noninvasive intravenous injection enables widespread and long-lasting neural expression, which is particularly useful for therapeutic genes. This was shown in the adult mouse nervous system, in which rAAV-PHP.B carrying GBA1 reversed α-synuclein pathology (278), outlining a powerful approach for studying neurodegenerative diseases with widespread brain pathology.

Tracing the morphology of individual cells is an important goal of neuroscience, and previous methods introduced AAV-based, multicolor dense labeling strategies (279).

However, it remains computationally challenging to perform cell segmentation and tracing on densely labeled samples. To aid in this challenge, the rAAV-PHP.eB and rAAV-PHP.S serotypes can be used for widespread, sparse stochastic delivery of combinations of different fluorescent proteins to specific neuron populations (179). The authors achieved high color diversity while reducing the fraction of labeled cells by using the two-component VAST system with the rAAV-PHP.S or rAAV-PHP.eB virus (**Figure 7.6*a*,*c***). Sparse but strong labeling provided by systemic delivery of two-component expression systems may also be advantageous for wide-field imaging of fluorescent activity sensors as a means of increasing the signal-to-noise ratio by reducing background fluorescence from adjacent sensor-expressing cells (280). Sparse labeling has also been achieved with direct intercranial injection; injection of highly diluted rAAV packaging Cre and simultaneous delivery of a high-titer of rAAV packaging a Cre-dependent fluorescent reporter produced high intensity labeling of only 10–50 neurons within the brain at the site of injection (281). Using this sparse and high-intensity labeling strategy, fine, long-range axonal collaterals could be detected with methods for submicron resolution, whole mouse brain imaging to enable reconstruction of individual neurons across the entire brain (281).

7.7 Viral strategies for targeting specific neuronal subpopulations via connectivity

Targeting specific subpopulations of neurons on the basis of both connectivity and cell type allows for high-resolution circuit interrogation. Cell type specificity alone cannot isolate functionally connected subpopulations because neurons of the same cell type may project to distinct areas of the brain executing unique functions and behavioral outputs (282, 283). Connectivity-based targeting can isolate transgene expression to relevant subpopulations of neurons but requires viruses capable of retrograde transport as well as retrograde and anterograde transsynaptic trafficking.

rAAVs are capable of retrograde and anterograde transport (284-287), though inherent levels are insufficient for robust connectivity-based transgene delivery. Notably, the engineered variant rAAV2-retro (268) enables efficient retrograde access to projection neurons (**Figure 7.6*b***), providing a more flexible alternative to the commonly used

replication-incompetent Canine adenovirus-2 (CAV-2) (288). rAAV2-retro does not jump synapses; instead, it is transported retrogradely (from axon terminals to the soma) within individual neurons, which is sufficient for circuit-based transgene applications such as targeting projection neurons (217, 289-292). Injection of rAAV2-retro results in expression both in cell bodies at the injection site and in neurons with axonal terminals at the injection site (i.e., the projection neurons) with varying efficiency depending on the specific circuit (268). Two-component systems that couple rAAV2-retro injections with a second component targeted to the cell bodies of the projection neurons can be used to achieve strong expression of the transgene of interest selectively in neurons with the targeted projection (289). In contrast to retrograde transduction, transsynaptic anterograde trafficking of AAV is often too inefficient to be detected through the delivery of reporter or actuator genes. However, with Cre gene delivery it is possible to permanently mark cells in at least a subset of downstream neurons (293). Further mechanistic understanding and optimization of AAV-mediated anterograde trafficking are necessary to broadly apply this technique.

Although AAVs are not naturally efficient at transsynaptic spread, many other neurotropic viruses, such as rabies virus (RV), vesicular stomatitis virus (VSV), and herpes simplex virus (HSV), are. Transsynaptic trafficking in these viruses is endowed by their replicative life cycle. RV-based systems has enabled input–output connectivity mapping of circuits throughout the nervous system (294-298) and single-cell-initiated monosynaptic tracing (299). One challenge with RV is its virulence; it causes cell death in as little as 1–2 weeks after infection (300), limiting the use of RV to short-term experiments such as input–output mapping (301). An RV strain with enhanced neurotropic properties and reduced toxicity was recently identified (302), and an alternative RV-based strategy has been developed in which the virus deactivates itself through proteolysis approximately 1 week after infection (303). Recombinant VSVs (304) and a Cre-dependent H129 strain of HSV-1 (305) have also been used for virally mediated anterograde trafficking, although these viruses also cause cell toxicity.

7.8 Outlook

7.8.1 AAVs for nontraditional model organisms and nontransgenic Animals

The application of genetic tools to nontransgenic animals is often hindered by the lack of vectors capable of safe, efficient, and specific delivery to the desired targets. Genetically tractable model organisms have been critical to progress in neuroscience, largely because of the ability to introduce novel genes and remove native genes from these organisms. No single model is optimal for understanding all components of the nervous system (306). Viruses may facilitate study of a diversity of model systems by enabling delivery of genes into nontraditional model organisms and nontransgenic animals.

Numerous studies have demonstrated the efficacy of common rAAV serotypes for transgene expression in a large array of warm-blooded vertebrates. Further discovery or engineering of AAVs with improved transduction efficiencies in nontraditional mammals (e.g., tree shrews and marmosets), birds (e.g., quails, chickens, and finches), fish (e.g., cichlids, zebrafish, and killi fish), and invertebrates (e.g., drosophilids, jellyfish, mollusks, leeches, and planarians) would enable the use of genetically encoded tools in these and other species. Alternatively, the use of viruses that are more naturally promiscuous (e.g., via VSV pseudotyping) may be more productive. VSVs, and lentiviruses pseudotyped with VSV glycoprotein, efficiently infect invertebrates such as the box jellyfish and fruit fly (*Drosophila melanogaster*) as well as vertebrates, such as the seahorse and many mammals (307). Regardless of the vector or route, virally mediated gene transfer to nontraditional model organisms would be a powerful addition for future studies of neuroscience.

7.8.2 Engineering designer AAVs for new neuroscience applications

Despite the successful application of the large array of rAAVs to neuroscience research, much remains to be achieved. Enhancing the already powerful CREATE-based screening system by incorporating NGS would aid the ongoing search for improved AAV capsids (**Figure 7.7**). The use of biochemistry assays probing for vector-receptor interactions and X-ray crystallography, cryogenic electron microscopy techniques (308), and macromolecular modeling [e.g. Rosetta] for delineating vector structures will help researchers understand how engineered AAVs (e.g., rAAV-PHP.B and rAAV-PHP.eB) can

efficiently cross the BBB, which in turn would benefit future capsid engineering. Capsid engineering may also overcome some of the current limitations of AAV capsids such as packaging capacity, specificity, transsynaptic transport, and access to embryonic tissue (**Figure 7.8**). Significant resources are also being devoted toward engineering capsids for gene transfer to humans (gene therapy).

The small packaging capacity of the capsid is a substantive barrier to some AAV applications. For example, *CRISPR/Cas9* driven by a cell type–specific promoter can exceed the ~4.7-kb limit on recombinant genome size (309). Engineered AAVs capable of packaging larger genomes would permit the use of larger cell type–specific promoters; the incorporation of multiple unique regulatory elements for combinatorial control, including, for example, a drug-controlled off/on switch for safer cargo production; and the packaging of multiple transgenes or large transgenes into a single capsid. Although modifying the packaging capacity of the capsid would be highly enabling, it is a challenging engineering proposition that likely requires drastic alteration of the capsid structure.

Cell type–specific capsids would mitigate dependence on Cre transgenics or on the limited range of cell type–specific promoters compatible with the size constraint of the AAV, and would allow gene delivery to specific cell types in wild-type animals. The engineering of AAVs capable of brain-region-specific delivery would allow for noninvasive gene transfer to specific anatomical brain regions without direct intracranial injection. Although genetic regulatory elements and Cre lines provide cell type–specific expression, they do not, in general, provide region-specific expression after systemic delivery. Neurons of a specific cell type are often distributed across separate anatomical regions of the nervous system with distinct functions. Combining region-specific AAVs with chemogenetics would allow investigators to modulate neuronal activity in a minimally invasive manner, a long-standing goal for neuroscience research.

Neurodevelopmental studies would benefit greatly from viral vectors that can transduce the embryonic brain in utero via the maternal vasculature. Embryonic gene delivery is limited by the placental barrier, which obstructs the transfer of many systemically delivered

molecules. Gene delivery to the embryonic brain requires invasive surgeries or in utero DNA electroporation, both of which pose risks to the mother and the embryos and provide limited tissue coverage and nonuniform expression. Given that capsid engineering has enabled efficient transport across the BBB, further engineering could result in an AAV capable of crossing the placental barrier.

Recombinant AAVs are currently being evaluated for human gene therapy (310-313). AAVs were the first vectors approved for use in humans to treat lipoprotein lipase deficiency [2012, European Regulatory Commission (256)] and recently the FDA approved an AAV-based therapy treatment for Leber congenital amaurosis; other AAV trials (e.g., for Parkinson's disease, spinal muscular atrophy, and hemophilia types A and B) are ongoing. The naturally occurring AAVs used to date have low specificity and largely overlapping tropisms, limited BBB permeability, some level of immunogenicity, and susceptibility to neutralization by preexisting antibodies, motivating past and ongoing capsid engineering (270). Use of new model systems such as human-derived brain organoids (314, 315) for selecting AAV properties could greatly accelerate engineering for human gene therapy applications.

7.9 Conclusions

Viruses have changed the way neuroscience research is done. Notably, selection methods such as CREATE are compatible with most described capsid diversification methods, can be used with either direct or systemic injection, and can be done in genetically less tractable organisms by introducing Cre to aid in the selection of capsid variants. A new generation of AAV capsid tools coupled with customizable regulatory elements and alternative viral delivery routes has the potential to significantly extend the utility of AAVs in targeting molecularly defined neurons throughout the nervous system across species.

7.10 Figures

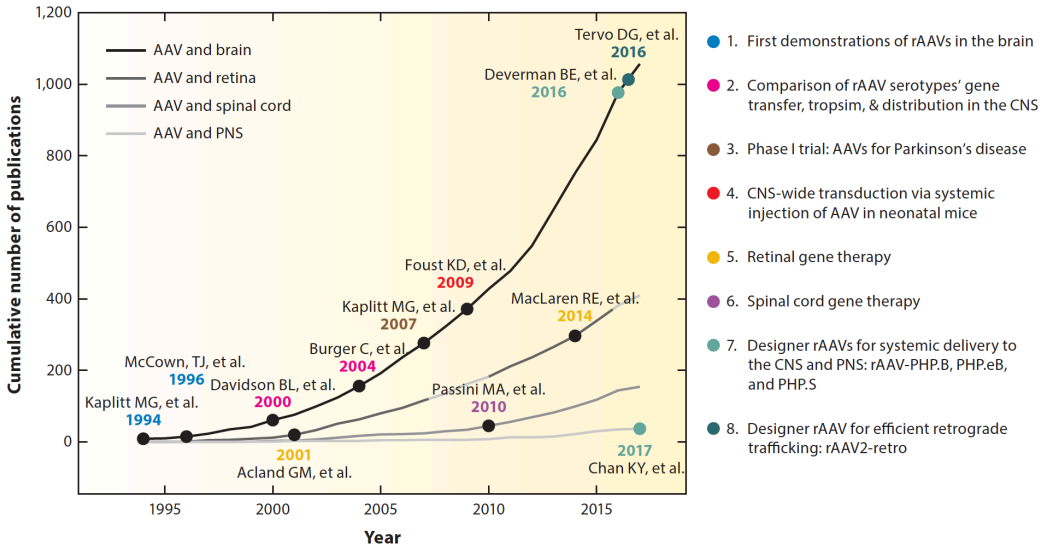


Figure 7.1 Overview of AAV use in the nervous system. Cumulative number of PubMed publications with the words *AAV* and *brain*, *AAV* and *retina*, *AAV* and *spinal cord*, or *AAV* and *PNS* highlights the use of AAVs in the brain in neuroscience research. Research involving the use of AAVs in the PNS, lags behind. The most-cited papers (*black circles*) highlight hallmark developments, starting with the first reported use of rAAVs for targeting neurons in the brain (316). Recent publications of key designer rAAVs for neuroscience applications are also highlighted. Abbreviations: AAV, adeno-associated virus; CNS, central nervous system; PNS, peripheral nervous system; rAAV, recombinant AAV.

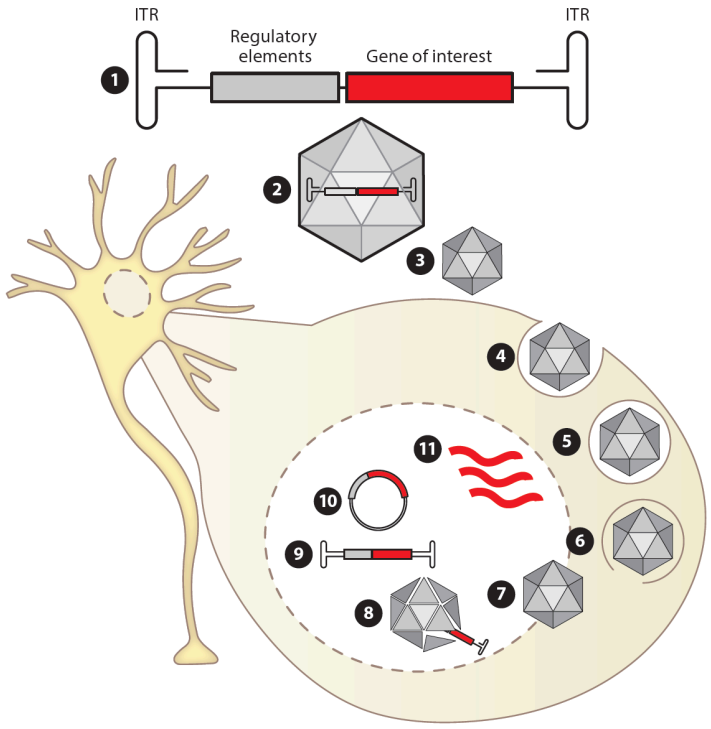


Figure 7.2 rAAV transduction of a neuron. (1) The recombinant genome. The regulatory elements (e.g., promoter, enhancers, and miRNA target sites) and gene of interest are inserted between the ITRs. (2) Packaging of the recombinant genome into capsids. (3) Binding of the capsid to receptors on the cell surface. (4) Receptor-mediated endocytosis of the viral particles. (5) Subcellular trafficking within endosomal compartments. (6) Endosomal escape. (7) Nuclear entry. (8) Genome release. (9) Second-strand synthesis. (10) Genome stabilization as episomal DNA, which often form concatemers in cells that are transduced by multiple virions. (11) Transgene expression, using cellular machinery. Abbreviations: ITR, inverted terminal repeat; miRNA, microRNA; rAAV, recombinant AAV.

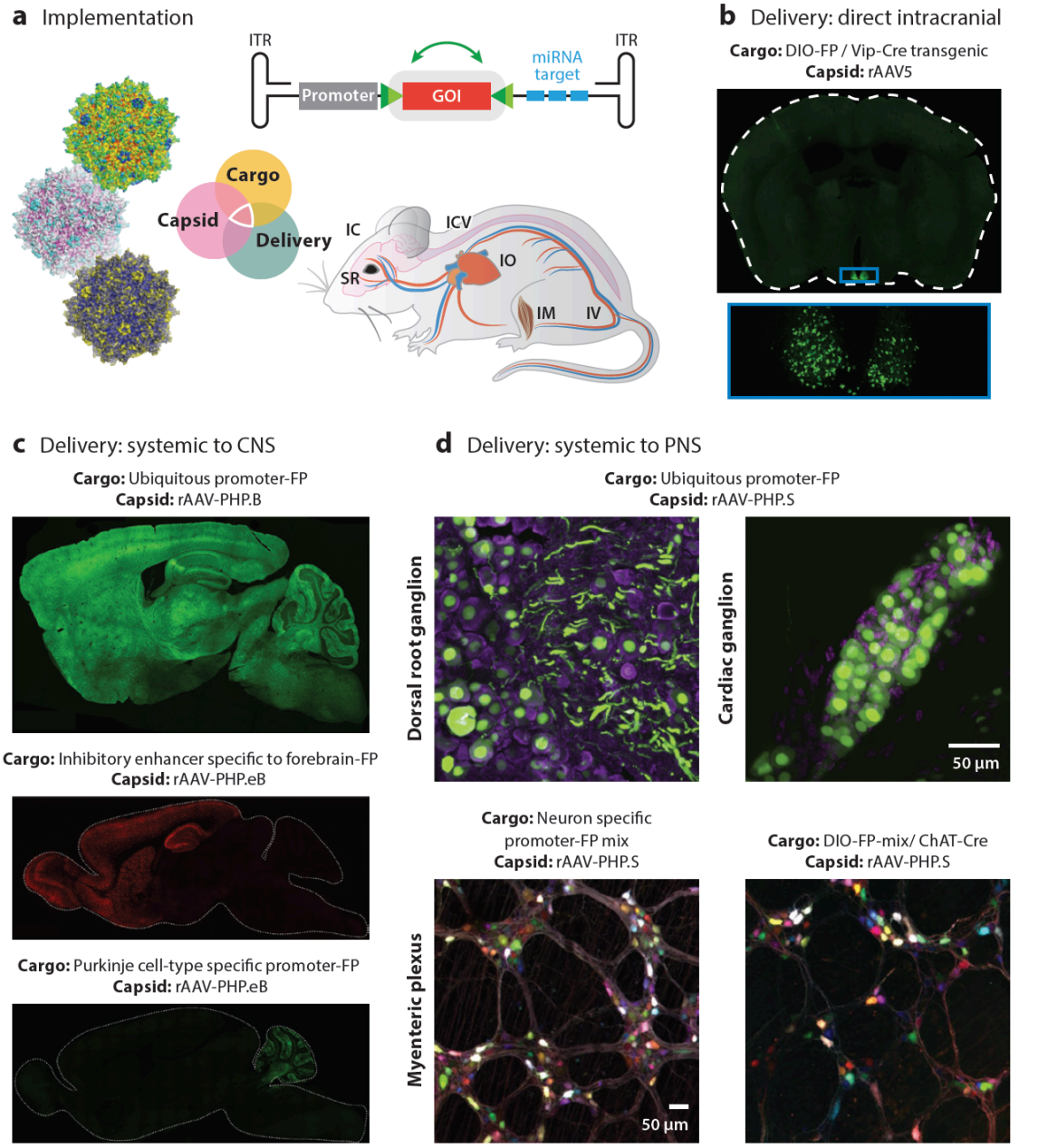


Figure 7.3 Methods for cell type–restricted expression in the CNS and PNS. (*a*) The AAV capsid, recombinant cargo, and anatomical delivery method dictate cell type–specific or region-specific expression. Delivery methods include intracranial (IC), intra-CSF (ICV), intravascular (IV), intramuscle (IM), intraorgan (IO), subretinal (SR), intravitreal, and intranasal delivery. (*b*) Targeted delivery to achieve local, cell type–specific transduction within GABAergic neurons in the suprachiasmatic nucleus through IC injection of a

rAAV-carrying a Cre-dependent transgene in a VIP-Cre transgenic animal (A. Kahan and V. Gradinaru unpublished data). (*c*) Systemic delivery for AAV-mediated transduction throughout the CNS for expression (*top*) throughout the whole brain (rAAV-PHP.B with the ubiquitous CAG promoter) (180), (*middle*) specific to the forebrain (rAAV-PHP.eB with a *Dlx5/6* enhancer specific to forebrain GABAergic interneurons), or (*bottom*) in Purkinje cells (rAAV-PHP.eB with the Ple155 promoter specific to Purkinje cells) (179). (*d*) Systemic delivery through a single IV injection to achieve expression throughout PNS neuron populations (*top*) by use of rAAV-PHP.S carrying a transgene (GFP) driven by a ubiquitous promoter (*magenta*, DAPI-stained nuclei) (179). (*Bottom*) Labeling diffuse neuronal populations within the PNS after single injection of rAAV-PHP.S–carrying a transgene under a neuron-specific promoter or cell type–specific labeling by rAAV-PHP.S–carrying a Cre-dependent transgene injected into a transgenic animal (Chat-Cre) (179). Abbreviations: AAV, adeno-associated virus; CNS, central nervous system; CSF, cerebrospinal fluid; GFP, green fluorescent protein; PNS, peripheral nervous system; rAAV, recombinant AAV; CAG, synthetic promoter containing the cytomegalovirus early enhancer element, first exon and first intron of chicken beta-actin gene, and the splice acceptor from the rabbit beta-globin gene; Dlx, distal-less homeobox promoter; VIP, vasoactive intestinal polypeptide.

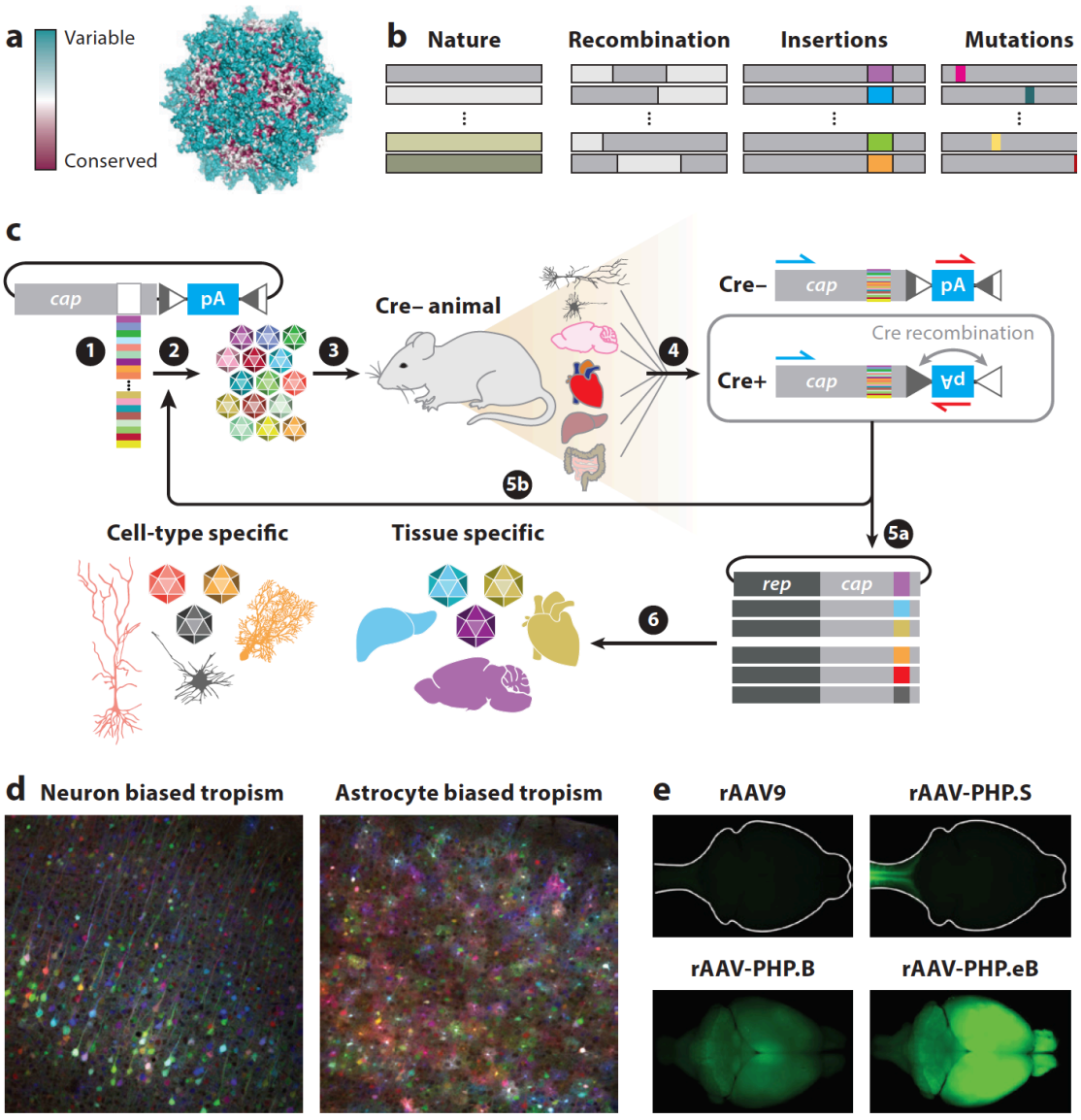


Figure 7.4 Engineering designer AAVs for neuroscience. (*a*) AAV capsid structural model highlighting the most variable (*cyan*) to most conserved (*magenta*) capsid regions. (*b*) Methods for diversifying the capsid sequence: isolate sequences from nature, recombination, insertions, and random mutagenesis. (*c*) The CREATE method: (1) build DNA library; (2) produce virus of capsid library; (3) inject library and, after the virus has had time to transduce cells, harvest the tissue of interest (e.g., brain, heart, intestine, liver) and homogenize or use cell-sorting methods; (4) Cre-dependent PCR amplification for genome recovery; (5a) transfer capsid sequences to a *rep-cap* helper for individual variant

testing, or (5b) if the remaining diversity is too high, repeat to enrich for most efficient capsid variants; (6) validate enriched capsid properties in vivo by the use of reporters and biodistribution assays (e.g., viral genome qPCR and imaging cleared tissues) (271). (*d*) CREATE-based evolution of AAV-PHP.B into variants with tropisms biased toward neurons (*left*) and astrocytes (*right*). Delivery of three fluorescent proteins shows distinct cell morphology clearly. (*e*) Engineered capsids capable of different biodistribution properties relative to those of the AAV9 parent after the CREATE method (179). Abbreviations: AAV, adeno-associated virus; CREATE, Cre recombination–based AAV targeted evolution; qPCR, quantitative polymerase chain reaction.

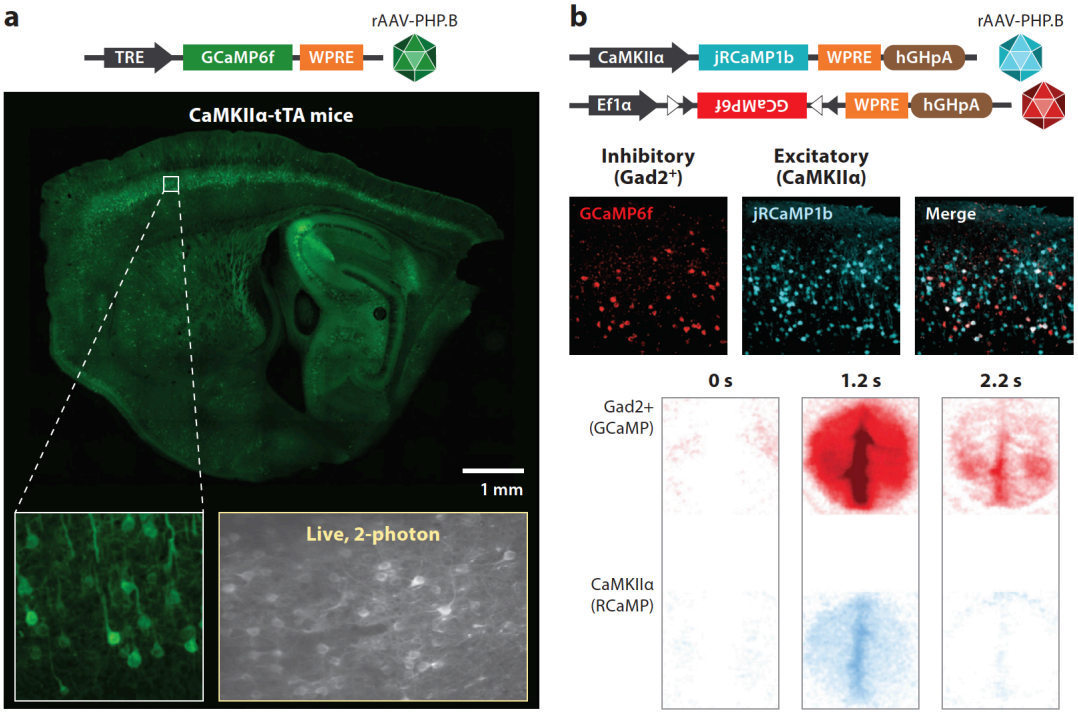


Figure 7.5 Widespread AAV-mediated delivery for recording neuronal activity dynamics during behavior. (*a*) Systemic delivery of GCaMP6 via rAAV-PHP.B results in strong expression in neurons throughout the brain while maintaining cell health for recording activity. GCaMP6f driven by tTA-dependent TRE promoter and packaged into rAAV-PHP.B results in strong GCaMP6f expression after systemic delivery to CaMKIIα-tTA transgenic mice. Sagittal image of a mouse brain with inset showing (*left*) individual cells and (*right*) in vivo, live two-photon image of cortical layers 2 and 3 from the same animal (D.Y. Tsao, F.J. Luongo, B.E. Deverman, V. Gradinaru, unpublished data). (*b*) Simultaneous imaging of inhibitory and excitatory dynamics during learned behavior. (*Top*) Diagram of viral genetic strategy for expression of GCaMP6f in all inhibitory neurons and jRCaMP1b in CaMKIIα-expressing (primarily excitatory) neurons, both of which are delivered with rAAV-PHP.B. Expression of jRCaMP1b and GCaMP6f in the cortex. (*Bottom*) Dual-color imaging of a 7-mm window view of the cortex for simultaneous imaging of two populations of neurons. Three frames showing time points of a video sequence of average fluorescence, simultaneously recorded in $Gad2^+$ and CaMKIIα populations, across trials in one mouse upon odor delivery (275). Abbreviations: AAV,

adeno-associated virus; Ef1α, elongation factor 1 alpha promoter; rAAV, recombinant AAV; tTA, tet-off transactivator; TRE, Tet response element; CaMKIIα, calmodulin-dependent protein kinase II promoter; GCaMP6f; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element.

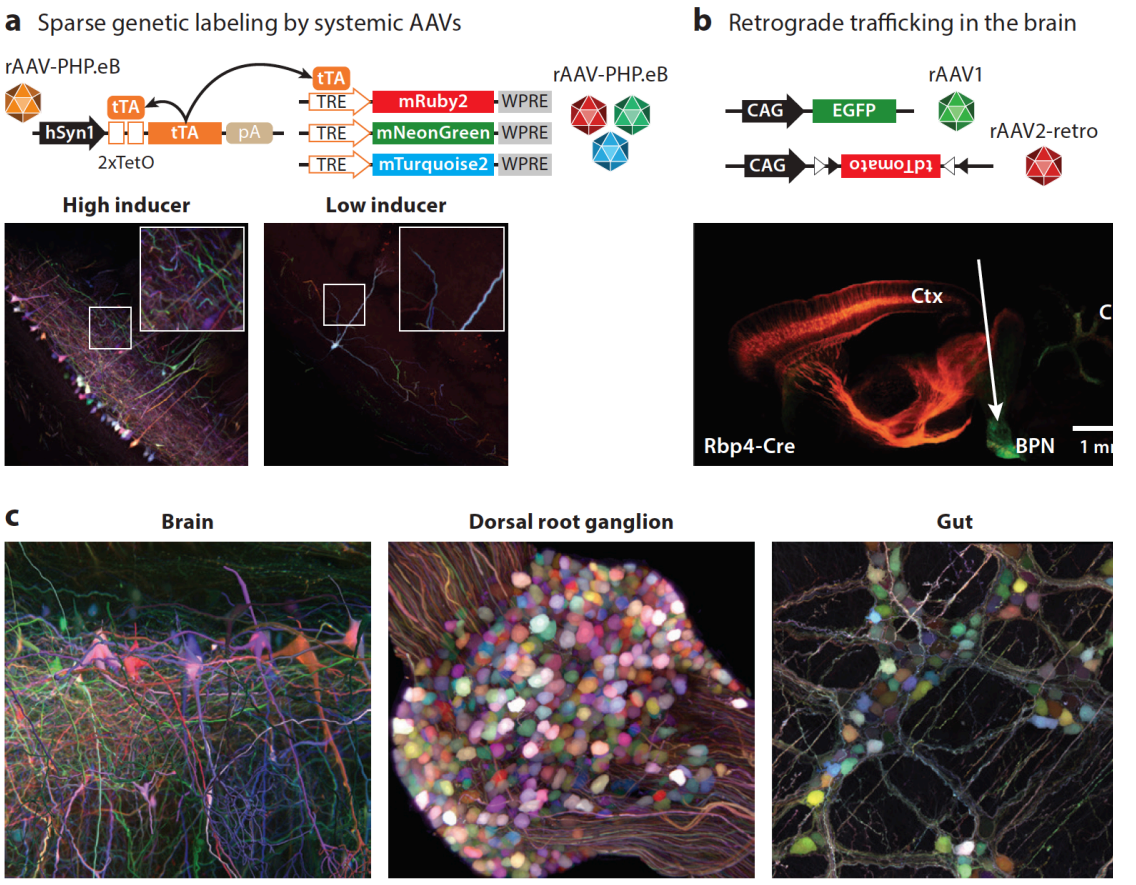


Figure 7.6 Designer AAVs for neuronal morphology and connectivity. (*a*) A method for Golgi staining–like (sparse, stochastic) genetic labeling by VAST. VAST uses systemic rAAV-PHP.eB to co-deliver an inducer (tTA) genome at either a high (*left*) or a low (*right*) titer and three different inducible fluorescent protein genomes at a high titer (179). (*b*) Engineered rAAV2-retro enables efficient labeling of the corticopontine tract throughout the rostro-caudal axis after coinjection of rAAV1-CAG-EGFP (*green*) and rAAV2-retro-DIO-CAG-tdTomato (*red*) in the basal pontine nuclei, in a layer 5–specific Cre mouse line (Rbp4_KL100 Cre) 3 weeks after injection (268). Panel adapted from Neuron, Vol 92/Issue 2, D. Gowanlock R. Tervo, Bum-Yeol Hwang, Sarada Viswanathan, Thomas Gaj, Maria Lavzin, Kimberly D. Ritola, Sarah Lindo, Susan Michael, Elena Kuleshova, David Ojala, Cheng-Chiu Huang, Charles R. Gerfen, Jackie Schiller, Joshua T. Dudman, Adam W. Hantman, Loren L. Looger, David V. Schaffer, and Alla Y. Karpova, A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons, p372-382, Copyright

(2016), with permission from Elsevier. (*c*) Multiplexed gene expression throughout the nervous system via engineered rAAV-PHP.eB and rAAV-PHP.S viruses (179). Abbreviations: AAV, adeno-associated virus; BPN, basal pontine nuclei; CAG, synthetic promoter containing the cytomegalovirus early enhancer element, first exon and first intron of chicken beta-actin gene, and the splice acceptor from the rabbit beta-globin gene; Cb, cerebellum; Ctx, cortex; DIO, double-floxed inverted; EGFP, enhanced green fluorescent protein; hSyn1, human Synapsin I promoter; pA, polyadenylation site; rAAV, recombinant AAV; TRE, Tet response element; tTA, tet-off transactivator; VAST, vector-assisted spectral tracing; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element.

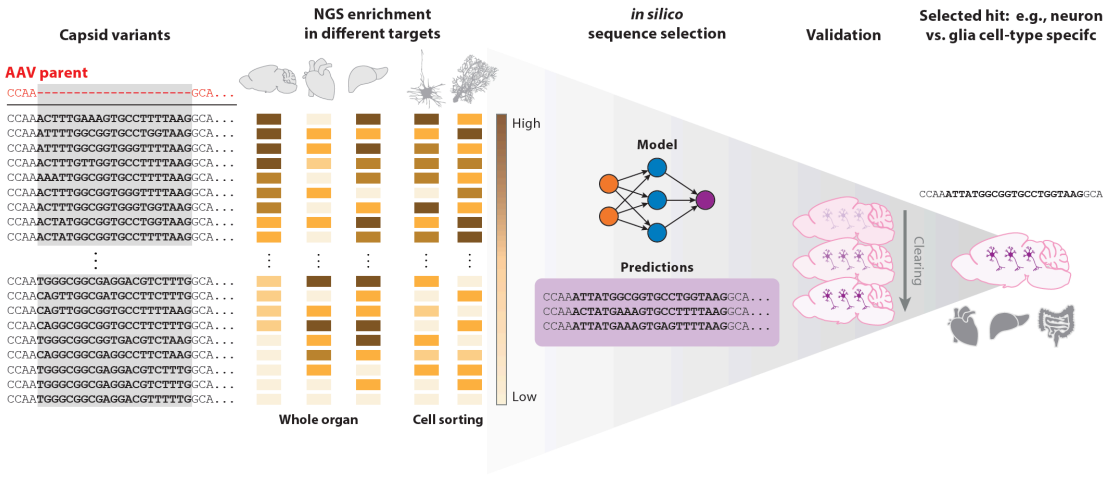


Figure 7.7 Optimized CREATE screening system using NGS to assess libraries of capsid variant enrichment in different tissues and cell types. Sequencing data can then be analyzed in silico to predict novel capsid variants with desired properties. The biodistribution and specificity of predicted capsid sequences can be tested broadly by qPCR and at the individual cell level by tissue clearing and imaging methods (271). Abbreviations: CREATE, Cre recombination–based AAV targeted evolution; NGS, next-generation sequencing; qPCR, quantitative polymerase chain reaction.

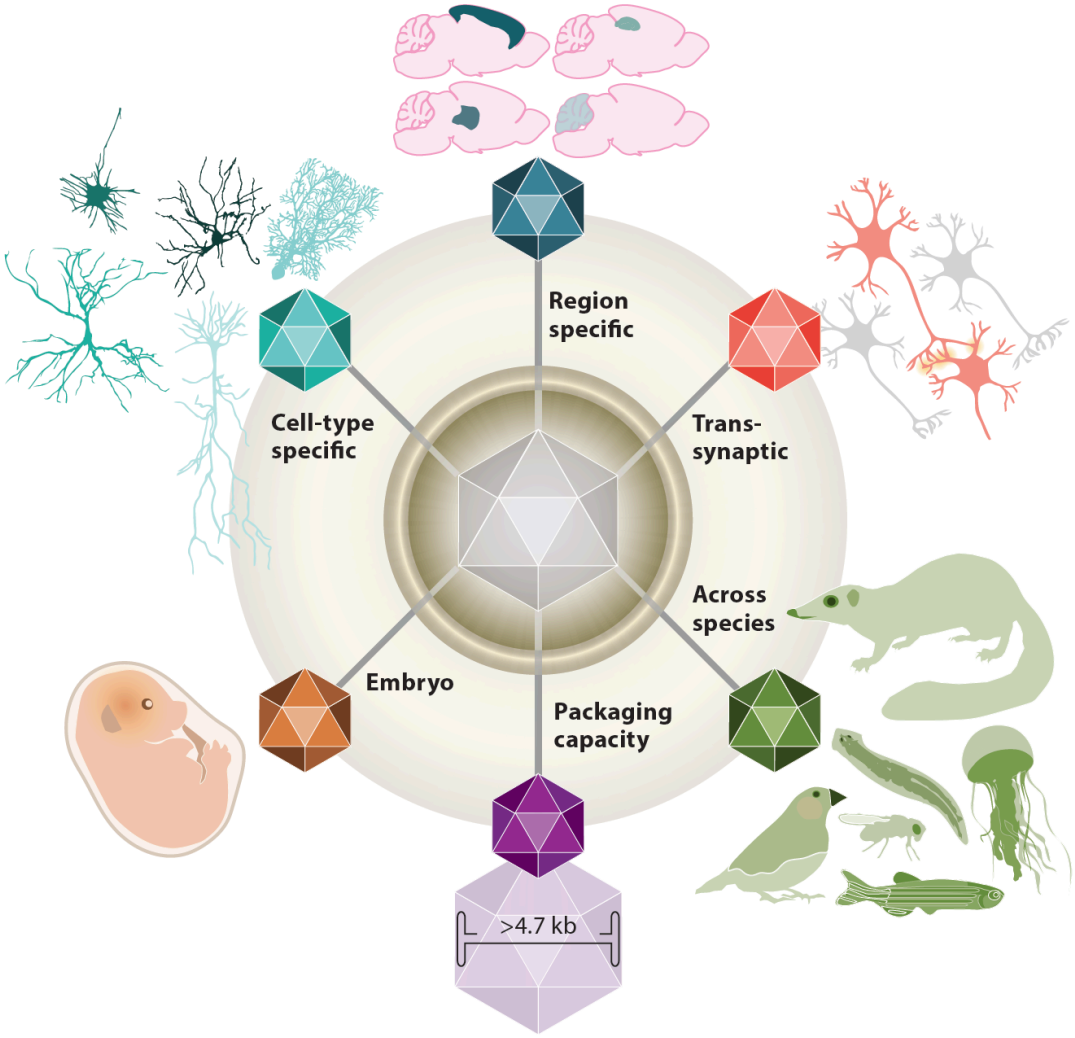


Figure 7.8 Outlook: future capsid engineering to achieve AAVs capable of region specific transduction, transsynaptic trafficking, gene delivery across species, packaging larger genomes, efficient transduction of the developing brain in utero, and transduction in a cell type–specific manner.

7.11 Supplemental tables

Supplemental Table 7.1. Delivery of rAAVs to key CNS and PNS targets

Target	Route of delivery	AAV serotype	References
Brain	Intracranial (direct)	rAAV1 rAAV2 rAAV5 rAAV9	Burger et al. 2004, Cearley et al. 2008, Cearley & Wolfe 2006, Davidson et al. 2000
	Intra-CSF (intracerebroventricular, intrathecal, intracisternal)	rAAV9 rAAV1 (neonates)	Bey et al. 2017, McLean et al. 2014, Passini et al. 2003
	Intranasal	rAAV2	Ma et al. 2016
	Intravenous	rAAV-PHP.B rAAV-PHP.eB rAAV9 rAAVrh8 rAAVrh10	Chan et al. 2017, Deverman et al. 2016, Foust et al. 2009, Yang et al. 2014a
Retina	Subretinal	rAAV8	Allocca et al. 2007
	Intravitreal	rAAV2 rAAV2 quad Y-F + T-V	Kay et al. 2013
	Intravenous	rAAV-PHP.B	Deverman et al. 2016
Spinal cord	Direct injection	rAAV1	Haenraets et al. 2017
	Intrathecal	rAAV6	Towne et al. 2009
	Intravenous	rAAV-PHP.B rAAV-PHP.S	Chan et al. 2017, Deverman et al. 2016
Dorsal root ganglia	Direct	rAAV5	Mason et al. 2010
	Intrasciatic	rAAV6	Iyer et al. 2014, Towne et al. 2009
	Subcutaneous	rAAV6	Towne et al. 2009
	Intramuscular	rAAV6	Towne et al. 2009
	Intrathecal	rAAV6	Towne et al. 2009
	Intraperitoneal	rAAV8 (neonates)	Foust et al. 2008
	Intravenous	rAAV-PHP.S rAAV8 (neonates)	Chan et al. 2017, Foust et al. 2008
Vagal nodose ganglia	Direct nodose/jugular complex	rAAV9	Chang et al. 2015, Williams et al. 2016
	Intravenous	rAAV-PHP.S	Challis et al. 2017, in review
Sympathetic chain ganglia	Intravenous	rAAV-PHP.S	Challis et al. 2017, in review
Motor neurons	Intravenous	rAAV-PHP.B	Deverman et al. 2016
Auditory nerve and inner hair cells	Scala media inoculation and cochlear delivery (direct)	rAAV8 rAAV-Anc80L65 (neonates)	Chien et al. 2016, Kilpatrick et al. 2011, Landegger et al. 2017
Heart and cardiac ganglion	Direct	rAAV9	Nussinovitch & Gepstein 2015
	Intravenous	rAAV9 rAAV-PHP.S	Chan et al. 2017, Vogt et al. 2015
Enteric nervous system (submucosal plexus/myenteric plexus)	Direct (into the wall of the descending colon)	rAAV6 (rats) rAAV9 (rats)	Benskey et al. 2015
	Intravenous	rAAV-PHP.S rAAV9 (neonates)	Chan et al. 2017, Gombash et al. 2014

Supplemental Table 7.2. Gene regulatory elements and recombination target sequences for controlled transgene expression in AAVs

Name	Properties	Size (bps)	References
Promoters			
CMV (cytomegalovirus early enhancer/promoter)	Ubiquitous	589–800	Gray et al. 2011a, Qin et al. 2010
Ubc (Ubiquitin C promoter)	Ubiquitous, weaker than CMV	403–1,177	Powell et al. 2015, Qin et al. 2010
CAGG/CAG (CMV enhancer, chicken β-actin promoter, and rabbit β-globin splice acceptor)	Ubiquitous	1,100–1,718	Chan et al. 2017, de Leeuw et al. 2016, Deverman et al. 2016, Niwa et al. 1991, Pignataro et al. 2017
CBh (modified miniature CAG promoter)	Ubiquitous	800	Gray et al. 2011a
Ef1α (human elongation factor 1 alpha promoter)	Ubiquitous	1,108	Sohal et al. 2009
EFS/EFFS (truncated Ef1α promoter)	Ubiquitous	253	Schambach et al. 2006
NSE (neuron-specific enolase promoter)	Broad, neuron-specific expression	1,800	Xu et al. 2001
hSyn1 (truncated human Synapsin I promoter)	Broad, neuron-specific expression	468	Kugler et al. 2003
MeCP2 (truncated methyl-CpG-binding protein-2)	Broad, primarily low-level neuronal expression	229	Gray et al. 2011a
BM88 (neural protein BM88 promoter)	Broad, neuron-specific expression	88	Papadodima et al. 2005, Pignataro et al. 2017
CHRNB2 (neuronal nicotinic receptor β promoter)	Neuron-specific expression	177	Bessis et al. 1995, Pignataro et al. 2017
CaMKIIα (Ca2+/calmodulin-dependent protein kinase II promoter)	Specific to glutamatergic neurons in the cortex and hippocampus	374–2,300	Dittgen et al. 2004, Hioki et al. 2007, Kuroda et al. 2008, Sohal et al. 2009
fSST (fugu somatostatin promoter)	Designed to target somatostatin (SST) inhibitory neuron, but instead provides expression in inhibitory interneurons broadly	2,597	Nathanson et al. 2009
mA93 (Riken gene A930038C07Rik promoter)	Expression in inhibitory neurons and glia	2,694	Nathanson et al. 2009
E1.1-NRSE (Artificial promoter construct with E1.1 binding sites and neuron restrictive silencing element)	Expression in SST and VIP, but not PV inhibitory neurons	214	Nathanson et al. 2009
MCH (mouse melanin-concentrated hormone promoter)	Expression in a subpopulation of the lateral hypothalamic neurons	830	van den Pol et al. 2004
MBP (myelin basic protein promoter)	Oligodendrocyte and Schwann cell–specific expression	1,943	Chan et al. 2017, Gow et al. 1992

hGFAP/gfaABC$_1$D (short human glial fibrillary acidic protein promoter)	Glial expression	681	Chan et al. 2017, Lee et al. 2008, Pignataro et al. 2017
hGFAPΔD (truncated hGFAP promoter)	Glial expression	476	Pignataro et al. 2017
mGFAP (truncated murine GFAP promoter)	Glial expression	543	Pignataro et al. 2017
rTH (rat tyrosine hydroxylase promoter)	Expression in dopaminergic neurons	2,500	Oh et al. 2009
mTH (mouse tyrosine hydroxylase promoter)	Efficient and specific expression in dopaminergic neurons in the SNc and VTA	2,571	Chan et al. 2017
PCP2 regulatory region (Ple155 MiniPromoter)	Expression in Purkinje cells of the cerebellum and retinal bipolar ON cells	1,652	Chan et al. 2017, de Leeuw et al. 2016
FEV regulatory region (Ple67 MiniPromoter)	Expression in serotonergic brain regions throughout most midbrain and hindbrain areas, including specific expression in the brain raphe nuclei and the retinal ganglion cell layer	2,202	Chan et al. 2017, de Leeuw et al. 2014, de Leeuw et al. 2016
DCX regulatory region (Ple53 MiniPromoter)	Expression in the retinal ganglion cell layer	3,310	de Leeuw et al. 2014
CCKBR regulatory region (Ple25 MiniPromoter)	Expression in the retinal ganglion cell layer	3,312	de Leeuw et al. 2014
CLDN5 regulatory regions (Ple34 MiniPromoter)	Expression in endothelial cells of brain blood vessels	3,845	de Leeuw et al. 2016
CLDN5 regulatory region (Ple261 MiniPromoter)	Expression in endothelial cells of the brain's blood vessels	2,963	de Leeuw et al. 2016
GPR88 regulatory region (Ple94 MiniPromoter)	Expression strongest in the striatum and in upper cortical layers	3,049	de Leeuw et al. 2016
SLC6A4 regulatory region (Ple198 MiniPromoter)	Expression strongest in thalamus	2,826	de Leeuw et al. 2016
C8ORF46 regulatory region (Ple251 MiniPromoter)	Expression strongest in the cortex and hippocampus	2,453	de Leeuw et al. 2016
NR2E1 regulatory region (Ple264 MiniPromoter)	Expression in retinal Müller glia	3,026	de Leeuw et al. 2016
S100B regulatory region (Ple266 MiniPromoter)	Expressed sporadically in the brain in a subset of GFAP+ astrocytes and Müller glia in the retina	2,982	de Leeuw et al. 2016
UGT8 regulatory region (Ple267 MiniPromoter)	Expression in globeruli-like structures in the olfactory bulb and in the cerebellar granule layer, with some Purkinje cells labeled	3,014	de Leeuw et al. 2016
TNNT1 regulatory region (Ple301 MiniPromoter)	Expression in muscle	1,209	de Leeuw et al. 2016
DCX regulatory region (Ple302 MiniPromoter)	Expression in ganglion cell layer in the retina	2,359	de Leeuw et al. 2016
NOV regulatory region (Ple303 MiniPromoter)	Expression strongest in cortical layers and hippocampus. Modest expression enrichment in horizontal cells of the retina	3,087	de Leeuw et al. 2016

OLIG1 regulatory region (Ple304/Ple305 MiniPromoter)	Expression in scattered cells in several brain regions, including the cortex and brainstem, resembling oligodendrocytes	2,596	de Leeuw et al. 2016
hs671 (small human enhancer, hs671, with a mouse minimal promoter Hsp68)	Widespread expression in adult mouse brain with stronger and more dense expression in cortex	2,129	de Leeuw et al. 2016
hs1218 (small human enhancer, hs1218, with a mouse minimal promoter Hsp68)	Widespread expression in adult mouse brain with high levels of expression in the midbrain region	2,338	de Leeuw et al. 2016
mDlx (mouse distal-less homeobox enhancer in front of a minimal promoter)	Selective expression within forebrain GABAergic interneurons	297	Dimidschstein et al. 2016
miRNA target site			
miR-1	Target for miRNA-1, a miRNA highly expressed in the heart and skeletal muscle (seq ATACATACTTCTTTACATTCCA)	22	Chan et al. 2017, Xie et al. 2011, Yang et al. 2014a
miR-122	Target for liver-specific miRNA-122 (seq ACAAACACCATTGTCACACTCCA)	23	Chan et al. 2017, Geisler et al. 2011, Qiao et al. 2011, Xie et al. 2011, Yang et al. 2014a
miR-124	Target for miRNA-124, a miRNA highly expressed in differentiated neurons (seq CCGTAAGTGGCGCACGGAAT)	20	Karali et al. 2011, Lagos-Quintana et al. 2002
miR-142-3p	Target for miRNA-142-3p, a miRNA specifically expressed in antigen-presenting cells (seq TCCATAAAGTAGGAAACACTACA)	23	Majowicz et al. 2013
miR-204	Target for miRNA-204, a miRNA that is strongly expressed in the retinal pigment epithelium (RPE) from as early as E10.5 to adulthood (seq TCCGTATCCTACTGTTTCCCTT)	22	Karali et al. 2011
miR-181	Target for miRNA-181, a miRNA expressed in the inner and middle retina (seq ACTCACCGACAGGTTGAA)	18	Kay et al. 2013
Recombinase recognition target sequence			
LoxP	Cre recombinase enzyme recognition target sequence	34	Fenno et al. 2014, Sauer & Henderson 1988
Lox2272	Modified LoxP site for Cre recombinase recognition	34	Fenno et al. 2014, Lee & Saito 1998

Lox71	Modified LoxP site for Cre recombinase recognition	34	Albert et al. 1995, Deverman et al. 2016
Lox66	Modified LoxP site for Cre recombinase recognition	34	Albert et al. 1995, Deverman et al. 2016
LoxN	Modified LoxP site for Cre recombinase recognition	34	Fenno et al. 2014, Livet et al. 2007
FRT	Flippase (flp) recombinase enzyme recognition target sequence	34	Fenno et al. 2014, Senecoff et al. 1988
F5	Modified FRT site for flp recombinase recognition	34	Fenno et al. 2014, Schlake & Bode 1994
F3	Modified FRT site for flp recombinase recognition	34	Fenno et al. 2014, Schlake & Bode 1994
Rox1	Dre enzyme recognition target sequence	32	Fenno et al. 2014, Sauer & McDermott 2004
Rox2	Modified Rox site for Dre recombinase recognition	32	Fenno et al. 2014

Supplemental Table 7.3. Natural and engineered AAVs optimal for specific applications in the nervous system

Application	Sero-type	Source	Notes	References
Direct injection: general use	AAV1	Natural serotype	Commonly used for direct injections. Good CNS transduction efficiency and spread	Burger et al. 2004
	AAV5	Natural serotype	Commonly used for direct injections. Good CNS transduction efficiency	Burger et al. 2004, Davidson et al. 2000
Direct injection: restricted spread	AAV2	Natural serotype	Best-studied AAV serotype. Limited spread from injection site	Burger et al. 2004, Davidson et al. 2000
Direct injections: high expression/high spread	AAVrh. 10	Natural serotype	High CNS transduction and spread	Cearley & Wolfe 2006
	AAV9	Natural serotype	High CNS transduction and spread	Cearley & Wolfe 2006
Direct injections: retrograde trafficking	AAV2-retro	Engineered AAV serotype: 7-aa insertion into loop 8 of AAV2	Efficient retrograde trafficking from axon terminals	Tervo et al. 2016

Direct injections: anterograde trafficking	AAV1	Natural serotype	Weak trans-synaptic anterograde trafficking. Cre used for signal amplification	Zingg et al. 2017
	AAV9	Natural serotype	Weak trans-synaptic anterograde trafficking. Cre used for signal amplification	Zingg et al. 2017
Systemic delivery: CNS (passage across the BBB)	AAV9	Natural serotype	Weak CNS transduction when delivered in adults mostly enriched in astrocytes than neurons. Efficient CNS transduction when delivered in neonates	Foust et al. 2009, Zhang et al. 2011
	AAVrh.8	Natural serotype	Weak CNS transduction when delivered in adults mostly enriched in astrocytes than neurons. Efficient CNS transduction when delivered in neonates	Yang et al. 2014a, Zhang et al. 2011
	AAVrh.10	Natural serotype	Weak CNS transduction when delivered in adults mostly enriched in astrocytes than neurons. Efficient CNS transduction when delivered in neonates	Yang et al. 2014a, Zhang et al. 2011
	AAV-AS	Engineered AAV serotype: insertion of 19-mer poly-alanine peptide in N-term of VP2 capsid protein	Enhanced CNS transduction. Six- and 15-fold more efficient than AAV9 in spinal cord and cerebrum	Choudhury et al. 2016
	AAV-PHP.B	Engineered AAV serotype: 7-aa insertion into loop 8 of AAV9	Enhanced CNS transduction. Forty- to 90-fold more efficient than AAV9 across many CNS regions	Deverman et al. 2016
	AAV-PHP.A	Engineered AAV serotype: 7-aa insertion into loop 8 of AAV9	Enhanced transduction of CNS astrocytes. Three- to 8-fold more efficient than AAV9 depending on the specific region	Deverman et al. 2016
	AAV-PHP.eB	Engineered AAV serotype: 2-aa replacement in loop 8 or AAV-PHP.B	Enhanced CNS transduction. >2.5-fold more efficient transduction compared with AAV-PHP.B	Chan et al. 2017
Systemic delivery: outside the brain	AAV9	Natural serotype	Intravenous delivery for transduction in motor neurons (neonatal), PNS neurons, and cardiomyocytes	Duque et al. 2009, Vogt et al. 2015
	AAV-PHP.S	Engineered AAV serotype: 7-aa insertion into loop 8 of AAV9	Enhanced transduction in DRGs, cardiac ganglion, enteric nervous system as well as expression in many organs, including the liver, lungs, heart, and stomach	Chan et al. 2017
Ependyma	AAV4	Natural serotype	Injected into the lateral ventricles of neonatal and adult animals	Liu et al. 2005

Chapter 8

THE JELLYFISH *CASSIOPEA* EXHIBITS A SLEEP-LIKE STATE

A version of this chapter has been published as (317)

8.1 Introduction

Do all animals sleep? Sleep has been observed in many vertebrates, and there is a growing body of evidence for sleep-like states in arthropods and nematodes (318-322). Here we show that sleep is also present in Cnidaria (323-325), an earlier branching metazoan lineage. Cnidaria, along with Ctenophora, are the first metazoan phyla to evolve tissue-level organization and differentiated cell types, such as neurons and muscle (326-332). In Cnidaria, neurons are organized into a non-centralized radially symmetric nerve net (327, 330, 331, 333, 334) that nevertheless shares fundamental properties with the vertebrate nervous system: action potentials, synaptic transmission, neuropeptides, and neurotransmitters (327, 333-337). It was reported that cnidarian soft corals (338) and box jellyfish (339, 340) exhibit periods of quiescence, a pre-requisite for sleep-like states, prompting us to ask if sleep is present in Cnidaria. Within Cnidaria, the upside-down jellyfish *Cassiopea spp.* displays a quantifiable pulsing behavior, allowing us to perform long-term behavioral tracking. Monitoring *Cassiopea* pulsing activity for consecutive days and nights revealed behavioral quiescence at night that is rapidly reversible, and a delayed response to stimulation in the quiescent state. When deprived of nighttime quiescence, *Cassiopea* exhibited decreased activity and reduced responsiveness to a sensory stimulus during the subsequent day, consistent with homeostatic regulation of the quiescent state. Together these results indicate that *Cassiopea* has a sleep-like state, supporting the hypothesis that sleep arose early in the metazoan lineage, prior to the emergence of a centralized nervous system.

Three behavioral characteristics define a sleep state (323, 324, 341): (1) behavioral quiescence, a period of decreased activity; (2) reduced responsiveness to stimuli during the

quiescent state; and (3) homeostatic regulation of the quiescent state. Both behavioral quiescence and reduced responsiveness must be rapidly reversible to differentiate sleep-like states from other immobile states (e.g. paralysis or coma) and reduced responsiveness distinguishes sleep from quiet wakefulness. Homeostatic regulation results in a rebound response, i.e. a compensatory period of increased sleep following sleep deprivation. Here we asked whether the cnidarian jellyfish *Cassiopea* exhibits these behavioral characteristics.

8.2 Results

Cassiopea are found throughout the tropics in shallow ocean waters and mudflats [**Figure 8.1**; (342, 343)]. They rarely swim and rather remain stationary with their bell on a surface, hence their name, the upside-down jellyfish [**Figure 8.1B**; **Supplementary Figure 8.1A**; (342, 343)]. *Cassiopea*, like coral and sea anemones, have a photosynthetic obligate endosymbiote, *Symbiodinium* (**Figure 8.1C**). *Cassiopea* continuously pulse by relaxing and contracting their bell at a rate of about 1 pulse/second (**Figure 8.1D**). This pulsing behavior generates fluid currents that facilitate vital processes such as filter feeding, circulation of metabolites, expulsion of byproducts, and gamete dispersion (343, 344). The pulsing behavior is controlled by light and gravity sensing organs called rhopalia [**Figure 8.1C**; (330)]. This stationary pulsing behavior makes *Cassiopea* a suitable jellyfish for behavioral tracking.

To track behavior in *Cassiopea*, we designed an imaging system (**Supplementary Figure 8.1C-F**) for counting pulses of individual jellyfish over successive cycles of day and night, defined as a 12-hour period when the light is on or off, respectively. As *Cassiopea* pulse, the relaxation and contraction of the bell causes a corresponding change in average pixel intensity, which was measured for each frame of the recording, producing a pulse-trace (**Figure 8.1D**). Pulse events were counted using the peak of the pulse-trace, and the inter-pulse interval (IPI) was calculated as the time between the peaks (**Figure 8.1D**; **Supplementary Figure 8.2**).

We observed that *Cassiopea* pulse less at night than during the day (**Figure 8.2**). To quantify this difference in pulsing frequency, we tracked the pulsing behavior of 23 jellyfish over 6 consecutive days and nights (**Figure 8.2C**). We define *activity* as the total number of pulses in the first 20 minutes of each hour. While individual jellyfish showed different basal activity levels (**Figure 8.2C**), all showed a large decrease in mean activity (~32%) at night (781 ± 199 pulses/20 min, mean ± s.d.) compared to the day (1155 ± 315 pulses/20 min, mean ± s.d.; **Figure 8.2C,E**). To determine if fast and slow pulsing jellyfish change their activity to a similar degree, we normalized activity of individual jellyfish by their mean day activity. Despite variations in basal activity, the relative change from day to night was similar between jellyfish (**Figure 8.2D**). Jellyfish activity decreased throughout the first 3-6 hours of the night, with the lowest activity occurring 6-12 hours after the day to night transition. Pulsing activity peaked upon feeding, occurring on the 4^{th} hour of each day (**Figure 8.2C,D**). To ensure that day feeding does not cause the day-night behavioral difference, we tracked the activity of 16 jellyfish over three consecutive days and nights without feeding and observed results consistent with those including feeding (**Figure 8.2F,G**; **Supplementary Figure 8.3D**). These results demonstrate that *Cassiopea* have a quiescent state during the night. To test the reversibility of this nighttime quiescent state we introduced a food stimulus at night, which transiently increased activity to daytime levels (**Supplementary Figure 8.3E**). The nighttime quiescent state in *Cassiopea* is thus rapidly reversible, consistent with a sleep-like behavior.

To better understand the nighttime quiescence, we compared day and night pulse-traces of individual jellyfish. The day and night pulse-traces of one representative jellyfish are shown in **2.2A**. During the night, the IPI is typically longer than during the day (**Figure 8.2A,B**; **Figure 8. 7A**). Two features contribute to this lengthening of the IPI: (1) the mode of the IPI distribution is longer at night than during the day, and (2) night pulsing is more often interrupted by pauses of variable length. These pauses are seen as a tail in the IPI frequency distribution (**Figure 8.2B**: 95^{th} percentile of night IPI frequency distribution (gray) is 13.9 s). Such long pauses are rarely seen during the day (**Figure 8.2B**: 95^{th} percentile of day IPI frequency distribution (yellow) is 2.5 s). This pause behavior may be

analogous to long rest bouts observed in *Drosophila* and zebrafish, which are suggested to be periods of deep quiescence with reduced responsiveness to stimuli (318, 345).

To test whether *Cassiopea* exhibit reduced responsiveness to stimuli during their nighttime-quiescent state, we designed an experiment to deliver a consistent arousing stimulus to the jellyfish. We observed in our nursery that *Cassiopea* prefer staying on solid surfaces as is found in nature. If *Cassiopea* are released into the water column, they quickly reorient and move to the bottom of the tank. We used placement into the water column as a stimulus to compare responsiveness during the night versus the day. *Cassiopea* were put inside a short PVC pipe with a screen bottom (**Figure 8.3A**). This was lifted to a fixed height, held for 5 min to allow the jellyfish to acclimate, and then rapidly lowered, which placed the jellyfish free-floating into the water column. We then scored the time it took for the jellyfish to first pulse and the time to reach the screen bottom (**Figure 8.3A**; **Methods**). At night, the jellyfish showed an increase in the time to first pulse and the time to reach bottom, compared to day (time to first pulse day: 2.1 ± 0.9 s versus night: 5.9 ± 4.0 s, and the time to reach bottom day: 8.6 ± 2.9 s versus night: 12.0 ± 3.2 s, mean ± s.d.; n = 23 animals) (**Figure 8.3B,C**). This increased latency in response to stimulus indicates that *Cassiopea* have reduced responsiveness to stimulus during the night.

To determine if the increased latency at night is rapidly reversible, a second drop was initiated within 30 s of the first drop, that is, after the jellyfish have been aroused. Reversibility was tested during both the day and night for 23 jellyfish. During the night, there is a large decrease in the time to first pulse and time to reach the bottom, after the second drop when compared to the first drop (**Figure 3D,E**). During the day and night, the time to first pulse and time to bottom after the second drop were indistinguishable, demonstrating that after perturbation, animals have similar arousal levels during the day and night. These results indicate that *Cassiopea* have rapidly reversible reduced responsiveness to a stimulus during the night.

To test whether *Cassiopea* nighttime quiescence is homeostatically regulated, we deprived jellyfish of behavioral quiescence for either 6 or 12 hours using a mechanical stimulus

(**Figure 8.4**). The stimulus consisted of a brief (10 s) pulse of water every 20 min, which caused a transient increase in pulsing activity. This increase in pulsing activity lasts for approximately 5 min after the 10 s pulse of water. Thus, the perturbation disrupts quiescence for approximately 25% of the perturbation period (either 6 hours or 12 hours). When the perturbation was performed during the last 6 hours of the night (**Figure 8.4A**), we observed a significant decrease in activity (~12%) during the first 4 hours of the following day relative to the pre-perturbation day (mean of first 4 hours of pre-perturbation day: 1146 ± 232 pulses/20 min compared to post-perturbation day: 1008 ± 210 pulses/20 min, mean ± s.d.; n = 30 animals; **Figure 8.4C**). This period of decreased activity is due to both decreased pulsing frequency (increased mode of IPI-length) and increased pause length (increase in the IPI-length 95th percentile) (**Supplementary Figure 8.4B,C**). This result is consistent with an increased sleep-drive after sleep deprivation. After a single day of decreased activity, the jellyfish return to baseline levels of day and night activity. Similar results were observed after an entire night of perturbation (12 hours; **Figure 8.4D**), with a large decrease in activity (~17%) throughout the following day (mean of 12 hours of pre-perturbation day: 1361 ± 254 pulses/20 min compared to post-perturbation day: 1132 ± 263 pulses/20 min, mean ± s.d.; n = 16 animals; **Figure 8.4F**). The decrease in activity caused by the 12-hour perturbation was larger than that of the 6-hour perturbation, indicating that the amount of sleep rebound is dependent on the level of sleep deprivation. During periods of decreased activity after either the 6-hour or 12-hour perturbation, we also observed increased response latency to a sensory stimulus (**Supplementary Figure 8.4A**), indicating a sleep-like state.

If the reduced activity following nighttime perturbation is due to sleep deprivation rather than muscle fatigue, applying the perturbation during the day, when *Cassiopea* are much less quiescent, should not result in reduced activity. To distinguish between sleep deprivation and muscle fatigue, we performed the 6- or 12-hour mechanical stimulus experiments during the day (**Figure 8.4B,E**). We observed no significant difference between pre- and post- perturbation activity levels (**Figure 8.4C,F**), indicating that the rebound response is specific to deprivation of nighttime quiescence. Taken together, these

results demonstrate that *Cassiopea* have a nighttime-quiescent state that is homeostatically controlled.

In many animals sleep is regulated by both homeostatic and circadian systems (346), but this is not always the case (320, 322-324, 347). For instance, the nematode *C. elegans* exhibits a developmentally regulated sleep state, and adult *C. elegans* show a non-circadian stress-induced-sleep state (320, 322, 348). A fully functioning circadian system is also not essential for sleep to occur; animals with null mutations of circadian rhythm genes still sleep, though sleep timing is altered (347). To test if nighttime quiescence in *Cassiopea* is regulated by a circadian rhythm, we first entrained the jellyfish for one week in a normal 12:12-hour light/dark cycle, and then shifted them to constant lighting conditions for 36 hours. We tested low- (~0.5 Photosynthetic Photon Flux [PPF]), mid- (~100 PPF), and full-intensity (~200 PPF) light, as well as dark (**Supplementary Figure 8.4D,E**). If jellyfish activity is regulated by a circadian rhythm, cycling activity should persist in the absence of entraining stimuli, such as light. We observed no circadian oscillation of jellyfish activity under any of the constant light conditions (**Supplementary Figure 8.4D**). However, we do observe circadian oscillation of activity in constant dark conditions (**Supplementary Figure 8.4E**). This result suggests that the quiescent state may be under circadian regulation.

Cassiopea display the key behavioral characteristics of a sleep-like state: a reversible quiescent state with reduced responsiveness to stimuli and both homeostatic and possibly circadian regulation. To our knowledge, our finding is the first example of a sleep-like state in an organism with a diffuse nerve net (324, 325), suggesting that this behavioral state arose prior to the evolution of a centralized nervous system. Though at least 600 million years of evolution separate cnidarians from bilaterians (326-328, 330-333, 349), many aspects of the nervous system are conserved, including neuropeptides and neurotransmitters (327, 333-337). One such conserved molecule, melatonin (350), promotes sleep in diurnal vertebrates, including zebrafish (351) and humans (352), and induces quiescence in invertebrates (353). We observed that melatonin induces a reversible decrease in activity in *Cassiopea* during the day in a concentration-dependent manner (**Supplementary Figure**

8.4F-H), suggesting that melatonin has a conserved quiescence-inducing effect in *Cassiopea*. Pyrilamine, a histamine H1 receptor antagonist that induces sleep in vertebrates(354), also induces concentration-dependent quiescence in *Cassiopea* (**Supplementary Figure 8.4F**). These results suggest that at least some mechanisms involved in vertebrate sleep may be conserved in *Cassiopea*.

8.3 Discussion

Although future studies are required to test whether other cnidarians sleep, field studies showing behavioral quiescence, diel vertical migration, and swimming speeds that vary with diel period (339, 340) suggest that a sleep-like state may not be specific to *Cassiopea*. A cnidarian sleep-like state could result from either divergent or convergent evolution. The observation of behaviorally and mechanistically conserved sleep-like states across the animal kingdom (323, 324) strongly supports the possibility for an early rooted sleep state rather than many instances of convergent evolution. It has been hypothesized that sleep has multiple functions, including synaptic homeostasis, regulation of neurotransmitters, repair of cellular damage, removal of toxins, memory consolidation, and energy conservation (324), although the ancestral role and selective advantage of sleep remains elusive. Our discovery of a sleep-like state in an ancient metazoan phylum suggests that the ancestral role of sleep is rooted in basic requirements that are conserved across the animal kingdom. The ancestral function of sleep may be revealed by further study of early branching metazoa.

8.4 Experimental methods

Experimental model and subjects details

Cassiopea spp. medusae used in this study were originally collected from the Florida Keys. For the majority of the experiments, a collection of multiple *Cassiopea* species were used (**Supplementary Figure 8.1A,B**). For the experiments shown in **Supplementary Figure 8.4A,E,F** a young (2-4 months old) clonal population of medusa

were used (*Cassiopea xamachana*). This clonal polyp line was generated in Monica Medina's lab at Pennsylvania State University.

Cassiopea were reared in artificial seawater (ASW, Instant Ocean, 30-34 ppt) at pH 8.1-8.3, 26-28°C with a 12-hour day/night cycle. During the day, 450 and 250 W light sources were used to generate 200-300 PPF (Photosynthetic Photon Flux, a measurement of light power between 400 and 700 nm). To limit waste buildup, the *Cassiopea* aquarium was equipped with a refugium (*Chaetomorpha* algae aquaculture), a protein skimmer (Vertex Omega Skimmer), carbon dosing bio-pellets (Bulk Reef Supply), activated carbon in a media reactor (Bulk Reef Supply), and a UV sterilizer (Emperor Aquatics 25 W). Waste products were kept at or below the following levels: 0.1 ppm ammonia, 5 ppb phosphorus, 0 ppm nitrite, and 0 ppm nitrate.

Cassiopea were fed daily with brine shrimp (*Artremia nauplii*, Brine Shrimp Direct) enriched with *Nannochloropsis* algae (Reed Mariculture), and they were fed oyster roe once per week (Reed Mariculture). *Cassiopea* were group housed in a 60 gallon holding tank. Animals were randomly assigned to experimental groups. Medusae between 3-6 cm in diameter were used for experiments.

Cassiopea **Genotyping**

Cassiopea is a genus with many species that have not been classified. All of our experiments were performed with *Cassiopea spp.* of a range of sizes, ages, sex and morphologies (**Supplementary Figure 8.1A,B**). To assess the diversity of *Cassiopea spp.* within our population we genotyped several animals by amplification and sequencing of the Mitochondrial cytochrome *c* oxidase I (COI). Genomic DNA extractions were performed as described (355). Jellyfish fragments, about 2 mm of tissue from the tentacles, were placed in 400 µL DNA extraction buffer (50% w/v guanidinium isothiocyanate; 50 mM Tris pH 7.6; 10 µM EDTA; 4.2% w/v sarkosyl; 2.1% v/v β-mercaptoethanol). Samples were incubated at 72°C for 10 min, centrifuged at 16,000 g for 5 min, and the resulting supernatant mixed with an equal volume of isopropanol and incubated at –20°C

overnight. The DNA was precipitated by centrifugation at 16,000 g for 15 min and the DNA pellet washed in 70% ethanol and resuspended and stored in water.

Amplification of COI was performed using primers designed by Folmer *et al.* (356), which amplify a ~710 base pair fragment of COI across the broadest array of invertebrates. COI primers:

LCO1490 forward primer: 5'-ggtcaacaaatcataaagatattgg-3'

HC02198 reverse primer: 5'-taaacttcagggtgaccaaaaaatca-3'

Amplifications were performed under the following PCR conditions: 2 min at 92°C, 30 cycles of 94°C for 30 s, 55°C for 30 s and 72°C for 45 s, with a final 72°C extension for 7 min. Amplification products were then TOPO-cloned using OneTaq (NEB) and sequenced.

Multiple sequence alignment of *Cassiopea spp.* COI sequences were generated using Clustal Omega software. Sequences were aligned with each other (see **Supplementary Figure 8.1B**), and to the previously identified cryptic species *Cassiopea ornata*, *Cassiopea andromeda*, and *Cassiopea frondosa* (342). The level of identity between these sequences is presented in **Supplementary Figure 8.1B**. Of the 15 *Cassiopea spp.* sequenced there were 8 identical COI sequences and 7 COI sequences with 45-90% identity.

***Cassiopea* behavioral tracking.**

Individual jellyfish were placed into 700 mL square clear plastic containers (cubbies), with white sand bottoms, in 35 L (10 gallon) glass tanks (**Supplementary Figure 8.1C-F**). Eight containers can fit in each tank, so eight jellyfish can be simultaneously recorded per tank. Tanks were housed inside Sterilite utility cabinets (65 cm W x 48 cm L x 176 cm H) with a door to eliminate ambient light in the recording setup. During the 12-hour day (lights on) tanks were illuminated with 24-inch florescent lamps, each containing four florescent bulbs that provide a combination of wavelengths optimized for photosynthesis in water: two 24 W, 6000 K Mid-day lights, and two 24 W Actinic lights (Giesemann), which combined provided 200-300 PPF. During the 12-hour night (lights

off) low-intensity red-LEDs were used to illuminate jellyfish to enable visualization. For all jellyfish recordings we used Unibrain 501b cameras above the tank running Firei software capturing at 15 frames per second. Camera aperture and Firei settings were adjusted to increase the contrast between jellyfish and background. Recordings were saved directly onto hard drives.

Jellyfish were acclimated in the recording tank in their cubbies for 2-3 days before starting recordings. 24-hour recordings were taken for successive days (7 am – 7 pm) and nights (7 pm – 7 am), unless otherwise indicated. *Cassiopea* were fed each day at 10:30 am, 3.5 hours after the lights turn on. Each jellyfish received 5 mL of 16 g/L brine shrimp. For each circadian rhythm experiment a different light condition was left on for 36-hours: dark conditions, low-intensity light conditions (an array of white-LED lights, 0-0.5 PPF), mid-intensity light conditions (two 24 W, 6000 K Mid-day lights, 75-150 PPF), or full light conditions (two 24 W, 6000 K Mid-day lights, and two 24 W Actinic lights, 200-300 PPF). For 6-hour and 12-hour rebound experiments the mechanical stimulus was applied for 10 s every 20 min.

All analysis was done using open-source packages in the SciPy ecosystem (357-359). To monitor jellyfish activity, pulsing information was extracted from the individual frames of each recording. Approximately 648,000 frames were collected every 12 hours. To quantify pulsing activity, we processed the first 18,000 frames of every hour (20 min). As *Cassiopea* pulse, the relaxation and contraction of the bell causes a corresponding change in average pixel intensity. To measure this change in average pixel intensity we drew a rectangular region of interest (ROI) around each jellyfish (**Figure 8.1D**; **Supplementary Figure 8.1F**). A user manually selected a ROI around each of the eight jellyfish in the first and last of the 18,000 frames. This was done so that the selected ROI accounts for any movement of the jellyfish. To control for noise from oscillations in ambient lighting, we perform background subtraction using a similarly sized ROI containing no jellyfish.

We analyzed pixel intensity data, and identified pulse events and inter-pulse intervals (IPI) in a four-step process. Step 1: Gaussian smoothing of the mean intensity over time

to eliminate high frequency oscillations (**Supplementary Figure 8.2A**). This smoothed trace was used to account for large movements in the mean intensity due to jellyfish translational movement within the selected ROI. Step 2: Normalization of the mean intensity values with the max mean intensity and the smoothed mean intensity:

$$T^n = \frac{T^n_{raw} - T^n_{smooth}}{T_{max} - T^n_{smooth}},$$

where T_{raw} is the raw intensity trace, T_{smooth} is the smoothed trace generated in Step 1, T_{max} is maximum intensity across the raw trace, and n is the index of each frame of the recording. Step 3: find the indices (time) of local maxima and minima in the normalized trace. Because of noise in the pulsing trace there is a high rate of false positives when finding local maxima and minima (**Supplementary Figure 8.2B**). We have used a set of criteria to identify a true pulse event from the local maxima and local minima. Step 4: identifying pulses from local maxima and minima (**Supplementary Figure 8.2C**). A local maximum can be defined as a pulse peak if it meets two criteria. First, it must be above a set threshold (to eliminate local maxima due to noise in pause regions of the pulse trace). Second, it must be above a set distance from the next local maxima (to prevent double counting of a single pulse). The standard deviation of the Gaussian smoothing, the threshold level, and the minimum distance between pulses can all be changed from one jellyfish to another. For all data analysis these parameter values were optimized to quantify pulsing events for each animal.

We calculated the total number of pulses and the IPI for each 20-min time bin. With some jellyfish the difference in pixel intensity from the contracted to non-contracted state was not big enough to easily identify pulsing above the noise. These jellyfish were excluded from analysis. During the 20-min recordings jellyfish would occasionally move out of the selected ROI. We would then exclude that 20-min recording for that jellyfish from the analysis. In compiling data to generate activity versus time plots we excluded jellyfish that we could not analyze for more than three 20-min recordings during a 12-hour day or night period.

For the arousal assay we designed an experiment to systematically test this sensory responsiveness. *Cassiopea* respond to being placed in the water column by rapidly orienting themselves and moving towards a stable surface. For the experimental system, *Cassiopea* were placed inside a 20 cm tall, 12 cm diameter, PVC pipe with a 53 μm filter screen bottom, called a *Cassiopea* dropper (CD). The experiment consists of four steps, as seen in the four panels in **Figure 8.3A**. Step 1, the jellyfish were placed on the screen bottom of the CD, which was positioned two cm below the water surface (h_L) and were acclimated for five min. At night jellyfish took less than five min to return to quiescence after being placed in the CD. Step 2, the CD was then "dropped" to a set depth (18 cm from the surface, h_D). This action leaves the jellyfish free-floating, two cm below the water surface. Step 3, the time to first pulse was measured. Step 4, the time to reach bottom was measured. To determine if the nighttime arousal latency is reversible, a second drop experiment was performed within 30 s of the initial drop. The CD was returned to two cm below the water surface, but instead of waiting for five min, steps 2 and 3 were performed immediately. Time to first pulse and time to bottom are not completely independent measures, though there is also not a perfect correlation. A jellyfish could pulse quickly but be delayed in reaching the bottom due to, for example, inactivity after the first pulse.

***Cassiopea* staining and imaging.**

Actin was stained using Alexa Flour 488-Phalloidin (ThermoFisher A12379). Jellyfish were anesthetized in ice-cold 0.8 mM menthol/ASW, and then fixed in 4% formaldehyde on ice for 45 min. Fixed jellyfish were permeabilized in 0.5% Triton/PBS for 2 hours and blocked using 3% BSA for 1 hour. They were then incubated in 1:100 Phalloidin solution in 0.5% Triton/PBS, for 18-24 hours in the dark at 4°C (360). Stained jellyfish were mounted in refractive index matching solution (361) and imaged using a LSM 780 confocal microscope (Zeiss).

Quantification and statistical analysis

The following statistical tests were used: two-sided paired Student's *t*-tests, two-sided unpaired Student's *t*-tests, and two-way ANOVA with Bonferroni posttest. We performed D'Agostino's omnibus K^2 normality test on all data sets to assess whether or not to reject the null hypothesis that all values were sampled from a population that follows a Gaussian distribution. For paired values, we tested if the pairs were sampled from a population where the difference between pairs follows a Gaussian distribution. Experimental groups that were statistically compared were tested for equal variance. The normality tests showed that all datasets were approximately Gaussian distributed with the exception of the time to first pulse arousal data. The time to first pulse data also showed grounds for rejecting the null hypothesis that there was equal variance between experimental groups. Tests of the log transformed time to first pulse data showed that the transformed data was approximately Gaussian distributed with equal variance between experimental groups, validating the use of standard two-way ANOVA and unpaired t-tests on the transformed data. Statistical tests were performed using either statistical functions from the SciPy ecosystem or GraphPad Prism (version 6.04 for Windows, GraphPad Software, San Diego California USA). No statistical methods were used to predetermine sample size. For these experiments we performed at least two laboratory replicates within our recording setup, which is limited to 8 jellyfish. Investigators were not blinded to allocation during experiments and outcome assessment. No specific method for randomization was used.

Data and software availability

Code used for tracking jellyfish activity and analysis are available at https://github.com/GradinaruLab/Jellyfish.

8.5 Figures

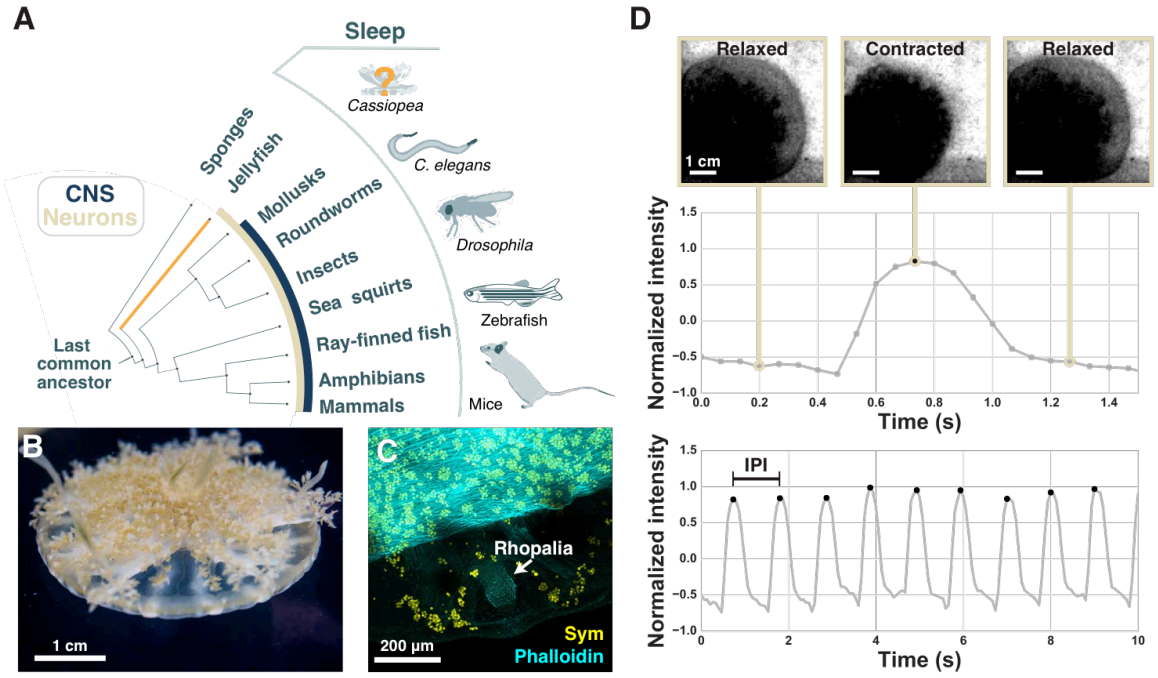


Figure 8.1. The pulsing behavior of the upside-down jellyfish, *Cassiopea spp.*, is trackable. (**A**) Phylogenetic tree schematic highlighting animals in which sleep behavior has been described, the presence of neurons (tan), and the emergence of a centralized nervous system (dark blue). See boxed key. (**B**) An image of *Cassiopea*. (**C**) Higher magnification view of *Cassiopea* with labeled actin-rich muscle (phalloidin stain; cyan), autofluorescent *Symbiodinium* (yellow), and a rhopalia, the sensory organ that controls pulsing, which is free of *Symbiodinium*. (**D**) As *Cassiopea* pulse the relaxation and contraction of the bell causes a corresponding change in average pixel intensity. Pulsing behavior was tracked by measuring this change in pixel intensity within the region of interest. (top) Representative frames and corresponding normalized pixel intensities for one pulse event. The local maxima in the pulse-trace was used to count pulse events. (bottom) A 10-second recording of one jellyfish shows multiple pulsing events. The inter-pulse interval (IPI) was calculated as the time between the maxima.

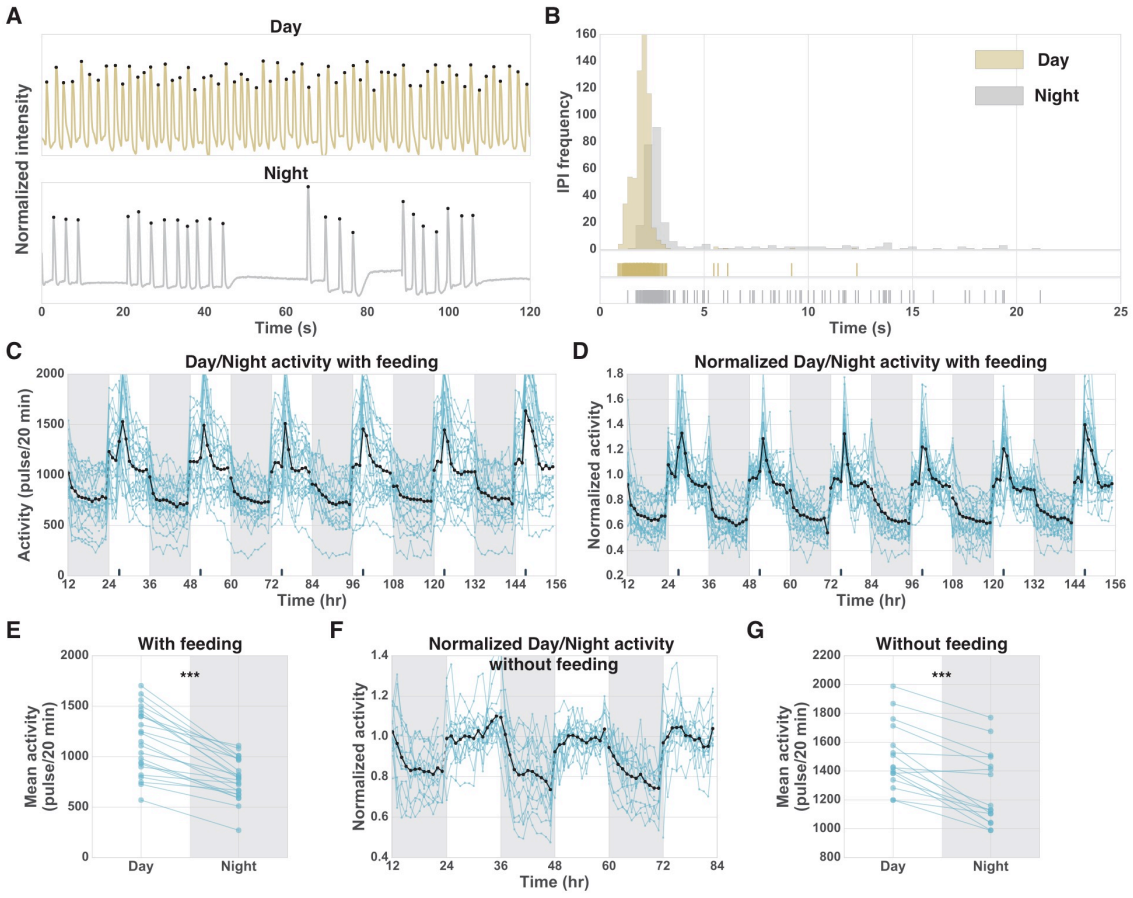


Figure 8.2. Continuous tracking of *Cassiopea* reveals pulsing quiescence at night. (**A**) Pulsing-traces for individual jellyfish during day and night over 120 s. (**B**) The distribution of IPI length for a 12-hour day and a 12-hour night for the same jellyfish shown in **A**. Tick marks below the distribution show each IPI length during the day and night. This highlights the long-pause events, which are more common at night (**Supplementary Figure 8.3A**). (**C-G**) Each blue line corresponds to a single jellyfish. The black line indicates the mean activity of all jellyfish. Dark gray shading indicates night periods. Dark tick marks on the x-axis indicate time of feeding. (**C**) Baseline activity (pulses/20 min) of 23 jellyfish tracked for six days from four laboratory replicates. (**D**) Normalized baseline activity for jellyfish shown in **C**, where each jellyfish is normalized by their mean day activity. (**E**) Mean day activity versus mean night activity for each jellyfish over the six-day experiment shown in **C**. Two-sided paired *t*-test, day versus night, $P = 6\text{x}10^{-9}$. (**F**) Normalized baseline activity without feeding of 16 jellyfish tracked over three days from two laboratory replicates,

where each jellyfish is normalized by its mean day activity. (**G**) Mean day activity versus mean night activity for each jellyfish over the three-day experiment shown in **F**. Two-sided paired *t*-test, day versus night, $P = 10^{-5}$. ***$P < 10^{-3}$.

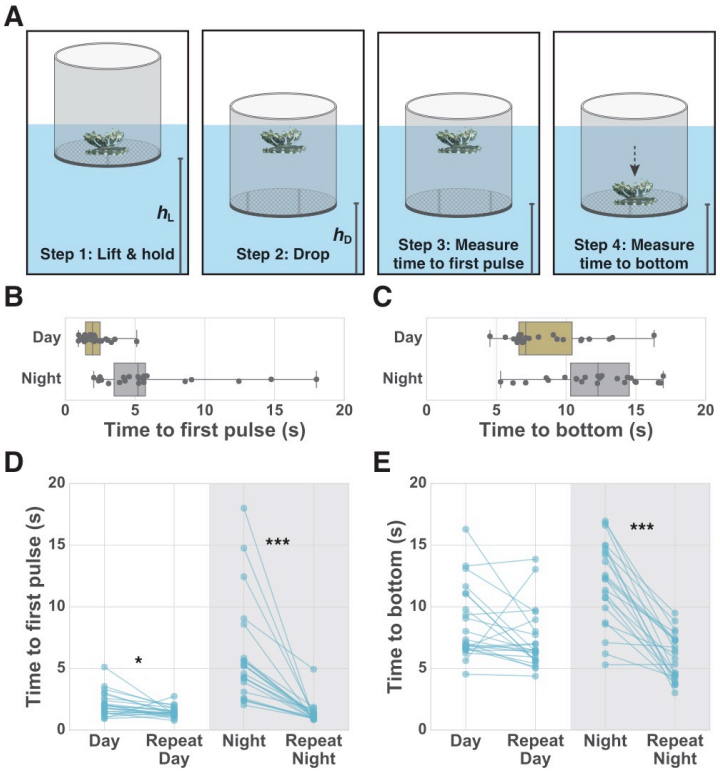


Figure 8.3. *Cassiopea* show reduced responsiveness to a sensory stimulus at night. (**A**) Schematic of experiment to test sensory responsiveness. Jellyfish were lifted and held at a fixed height (h_L) and then dropped to a fixed height (h_D). h_L and h_D were kept constant throughout experiments. Boxplots of time to first pulse after drop (**B**) for 23 jellyfish and time to reach bottom after drop (**C**) for 23 jellyfish during the day and night. Dots represent individual jellyfish collected from two laboratory replicates. Two-sided unpaired *t*-test, day versus night, (**B**) $P < 10^{-4}$ and (**C**) $P = 5\text{x}10^{-4}$. (**D**) Time to first pulse after initial drop and after perturbation for both day and night for 23 jellyfish. (**E**) Time to reach bottom after initial drop and after perturbation for both day and night for 23 jellyfish. Two-way analysis of variance (ANOVA) for data shown in **D** and **E**, followed by post-hoc comparisons between experimental groups using B2onferroni posttest (*$P<5\text{x}10^{-2}$, ***$P<10^{-3}$). For the time to first pulse, two-sided unpaired *t*-test (**B**) and two-way ANOVA (**D**) were performed after log-transformation (**8.4 Methods**).

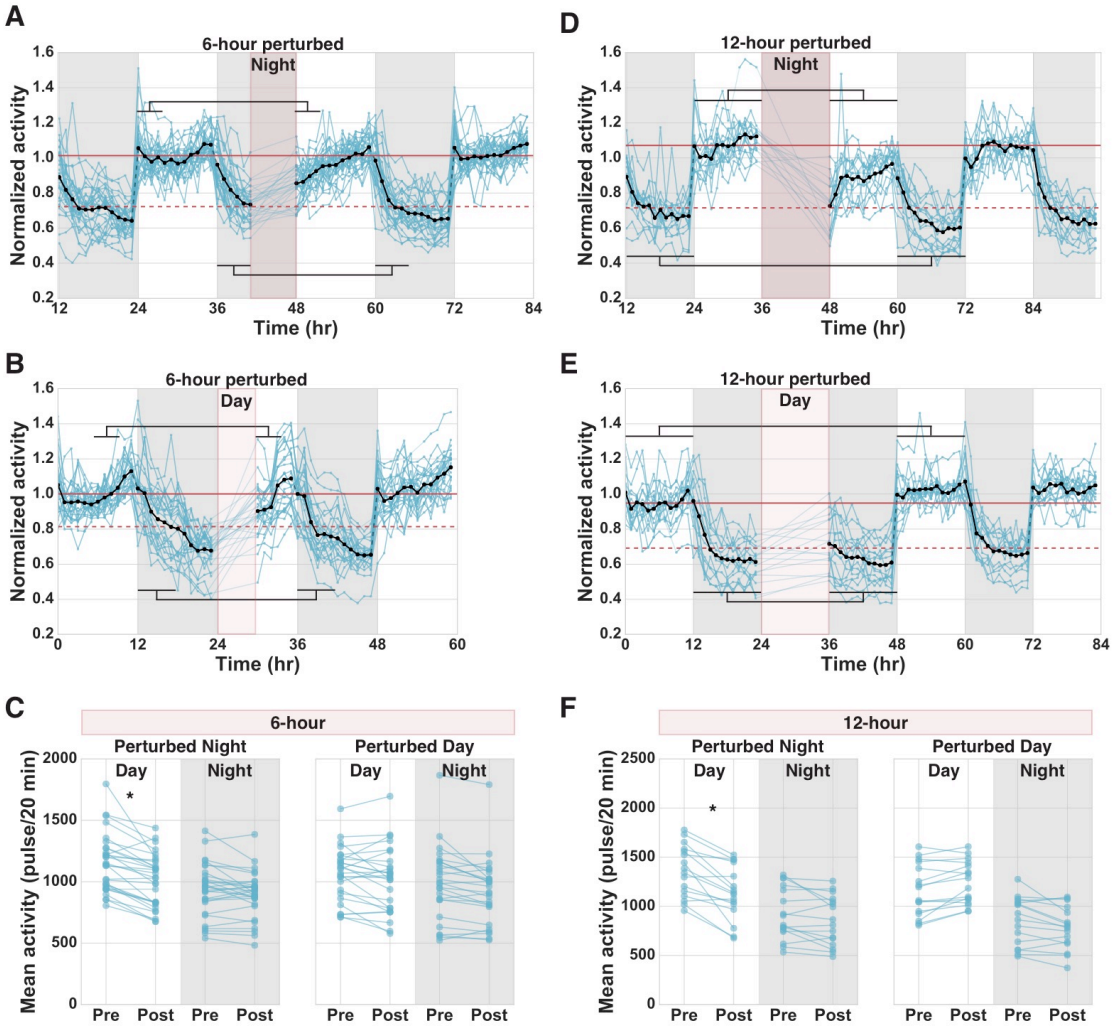
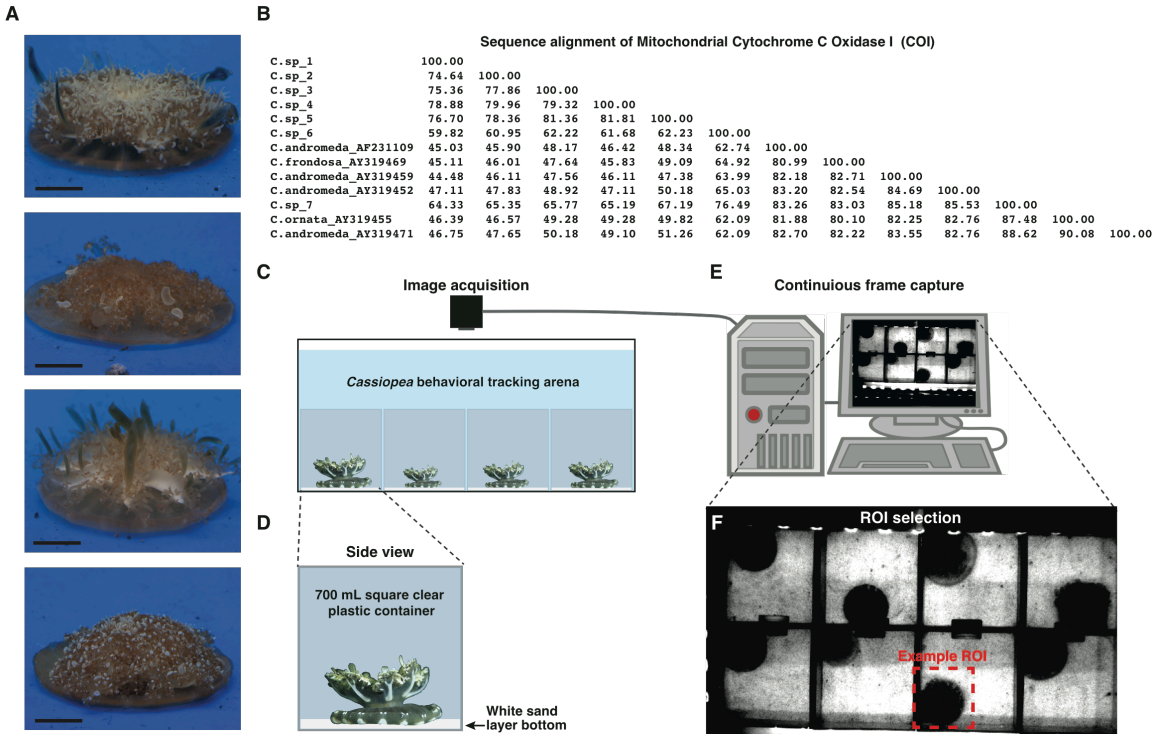


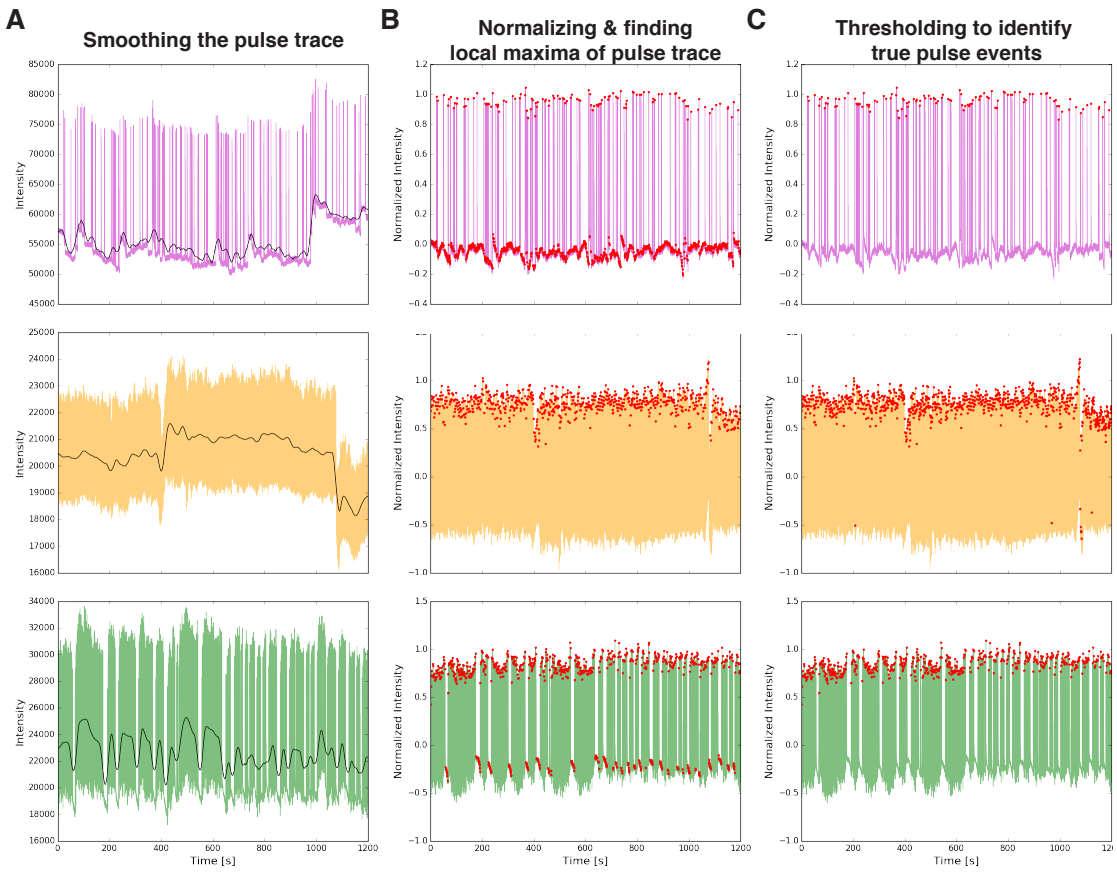
Figure 8.4. Homeostatic rebound in *Cassiopea*. Each blue line corresponds to a single jellyfish. The black line indicates the mean activity of all jellyfish. Dark gray shading indicates night periods. Maroon shading indicates perturbation periods with 10 s water pulses every 20 min. Jellyfish were exposed to different perturbation lengths (6 or 12 hours) at different times (day or night). The normalized activity of all jellyfish tracked over multiple days is plotted. Maroon horizontal lines show the mean activity of pre-perturbation day (solid) and pre-perturbation night (dashed). (**A**) Perturbation of 30 jellyfish for the last 6 hours of the night. (**B**) Perturbation of 26 jellyfish for the first 6 hours of the day. (**C**) Mean day and night activity pre- and post-perturbation for experiments shown in **A** and **B**. (**D**) Perturbation of 16 jellyfish for an entire 12-hour night. (**E**)

Perturbation of 16 jellyfish for an entire 12-hour day. (**F**) Mean day and night activity pre- and post-perturbation for experiments shown in **D** and **E**. Black-horizontal lines in **A**, **B**, **D**, and **E** indicate the windows of time used for calculating pre- and post-perturbation means shown in **C** and **F** for both the night (bottom lines) and day (top lines). For the 6-hour experiments we compared the first 4 hours of the post-perturbation day to the equivalent time pre-perturbation, and also compared the first 6 hours of post-perturbation night to the equivalent time pre-perturbation. For the 12-hour experiments we compared the full 12-hour days and nights pre- and post-perturbation. Two-way ANOVA followed by post-hoc comparisons between experimental groups using Bonferroni posttest ($*P<5x10^{-2}$). Both day and night 6-hour perturbation experiments include data from four laboratory replicates. Both day and night 12-hour perturbation experiments include data from two laboratory replicates.

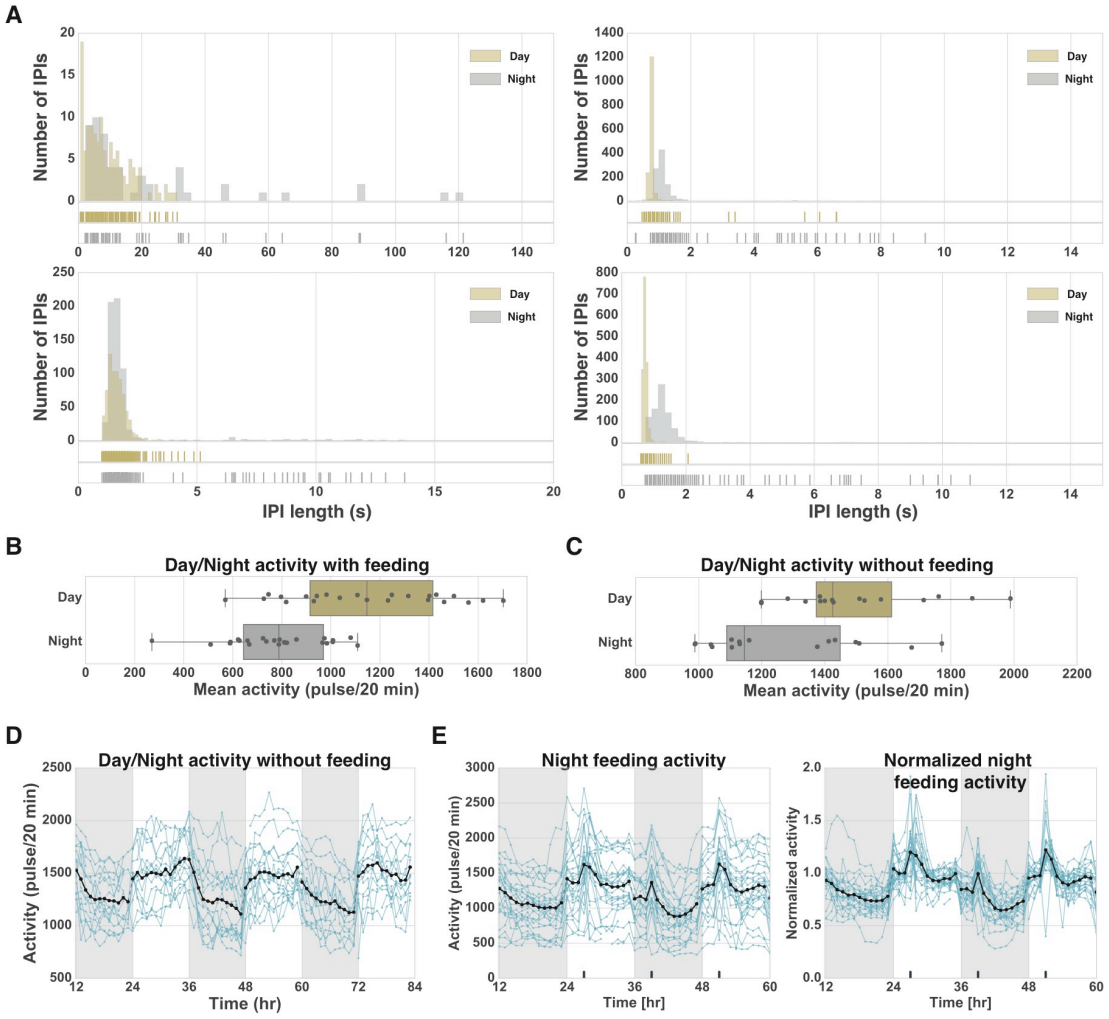
8.6 Supplemental figures



Supplementary Figure 8.1. *Cassiopea spp.* diversity and behavioral tracking system. (**A**) Images of four *Cassiopea spp.* with different morphology (scale bar 1 cm). This is representative of the range of morphologies used in the experiments. (**B**) Percent amino acid identity matrix comparing mitochondrial cytochrome C oxidase I (COI) amino acid sequences of seven *Cassiopea spp*. used in this study (C.sp_1 – C. sp_7) with six previously described *Cassiopea spp.* (Taxon_GeneBank number). (**C**) For the behavioral tracking system jellyfish were placed in behavioral tracking arenas with cameras recording from above. (**D**) Each jellyfish was placed in a clear, plastic container with white sand layering the bottom. The white sand provides contrast, allowing better behavioral tracking. (**E**) Images were captured at a rate of 15 frames per second and saved directly onto solid-state hard drives. (**F**) A region of interest (ROI) around each jellyfish was selected for downstream processing.

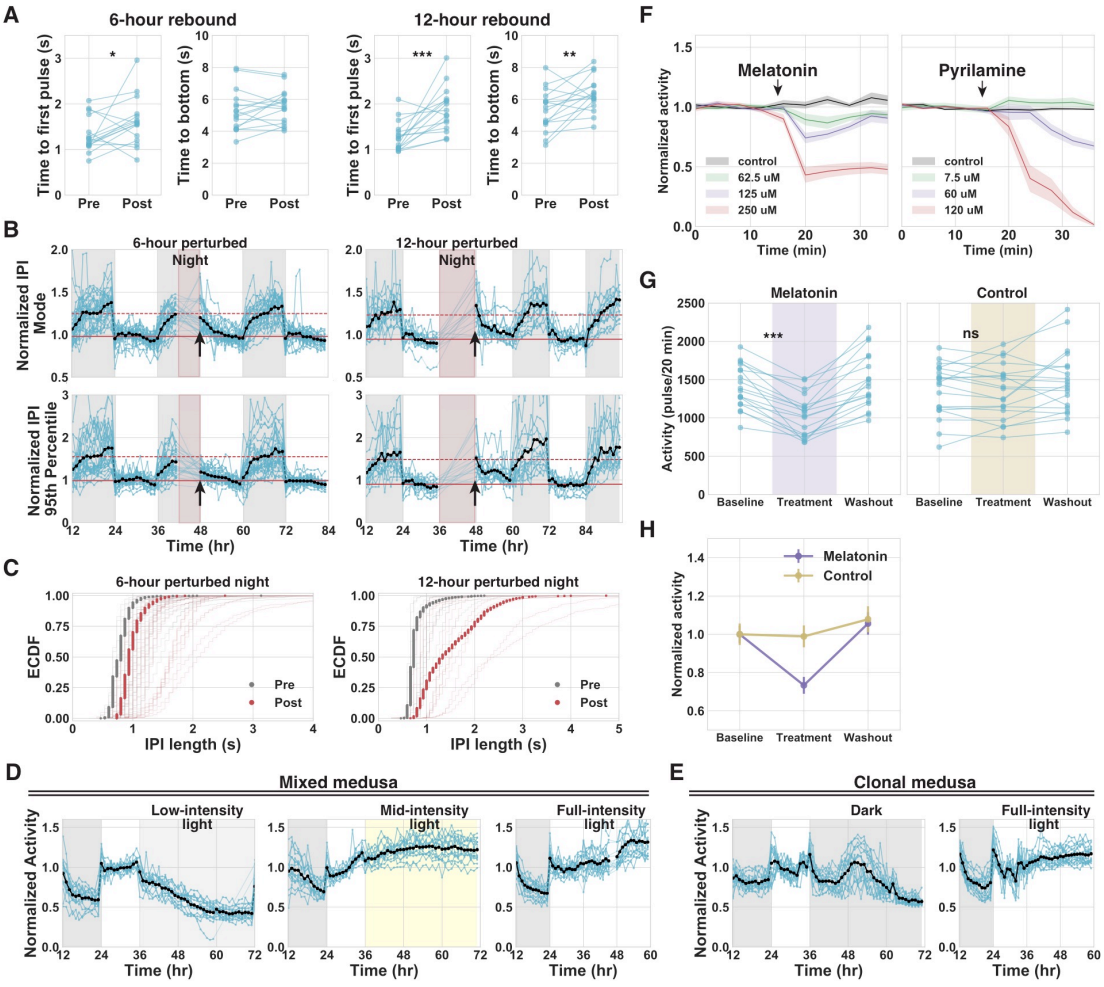


Supplementary Figure 8.2. Processing the jellyfish pulse-trace data to count pulse events. Each color represents data from a different jellyfish (pink, orange, and green). (**A**) Smoothing the pulse-trace for normalization. Black line represents the smoothed trace for a 20 min recording. (**B**) Normalized pulsing traces for three different jellyfish with local maxima indicated by red dots. Many local maxima are detected within pauses in activity due to noise (small fluctuations in intensity), which are removed by thresholding. (**C**) Thresholding to identify local maxima at pulsing peaks. Pulsing peaks are indicated by red dots. For more details see the '*Cassiopea* behavioral tracking' section of the **Methods**.



Supplementary Figure 8.3. *Cassiopea* pulsing quiescence at night. (**A**) Distribution of IPI length for four *Cassiopea* during the day (yellow) and night (gray) showing each IPI event. Tick marks below the distributions show each IPI length during the day (yellow) and night (gray). The ticks highlight the long-pauses that are more common at night for all jellyfish (**Data S1**). Box plot of *Cassiopea* day and night pulsing activity with feeding (**B**)**,** and without feeding (**C**). Each dot represents a single jellyfish, mean activity is calculated over 6 (feeding, **B**) or 3 (without feeding, **C**) days and nights. For **D** and **E** each blue line corresponds to a single jellyfish. The black line indicates the mean activity of all jellyfish. Dark gray shading indicates night periods. (**D**) Day and night activity of *Cassiopea* without feeding. Baseline activity (pulses/20 min) without feeding of 16 jellyfish tracked over three days. (**E**) Feeding induced arousal rapidly reverses the night quiescent state. Dark tick

marks on x-axis indicate time of feeding. Activity (pulses/20 min) and normalized activity of 30 jellyfish tracked over two day/nights from six laboratory replicates. Jellyfish were fed 4 hours into each day and 4 hours into the second night.



Supplementary Figure 8.4. Regulation of quiescence in *Cassiopea*. Each blue line corresponds to a single jellyfish. The black line indicates the mean activity of all jellyfish. Dark gray shading indicates night periods. (**A**) Sensory responsiveness was tested during periods of decreased activity before (pre) and after (post) either the 6-hour or 12-hour perturbation periods (10 s water pulses every 20 min) using the assay described in **Figure 2.3**. Time to first pulse after drop and time to reach bottom after drop were measured during the day pre or post perturbation. After perturbation (post), an increased response latency was observed. Two-sided paired *t*-test, pre versus post, $*P<5x10^{-2}$, $**P<10^{-2}$, $***P<10^{-3}$. (**B**) Maroon horizontal lines show the mean activity of pre-perturbation day (solid) and pre-perturbation night (dashed). Maroon shading indicates perturbation periods with 10 s water pulses every 20 min. In these experiments jellyfish were exposed to

different perturbation lengths (either 6 or 12 hours) during the night. Plotted here is the normalized mode and 95th percentile of the IPI length for all jellyfish tracked over multiple days. Perturbation of either 30 jellyfish for the last 6 hours of the night or 16 jellyfish for an entire 12-hour night. For both the 6-hour and 12-hour perturbation there is an increase in the mode and 95th percentile of the IPI length after perturbation (black arrowhead). (**C**) Empirical cumulative distribution function (ECDF) of daytime IPI length for all jellyfish pre (gray) and post (maroon) perturbation (thin lines, single jellyfish; dots, all jellyfish). Jellyfish exhibited increased IPI lengths after perturbation compared to before perturbation. These results suggest that the increased quiescence observed in **Figure 2.4** results from both a decreased frequency of pulsing and an increase in the length of pause events. (**D-E**) Monitoring activity with different light or dark conditions suggests that nighttime quiescence may be under circadian regulation. (**D**) Prolonged light exposure of *Cassiopea* shows no circadian cycling. 16 jellyfish were exposed to either 36-hours of continuous low-intensity light (light-gray shading) from hour 36 to hour 72, 36-hours of continuous mid-intensity light (yellow shading) from hour 36 to hour 72, or 36-hours of continuous full-intensity light from hour 24 to hour 60. Each experiment represents two laboratory replicates using a mixed population of *Cassiopea spp*. (**E**) Prolonged exposure to dark conditions of jellyfish shows circadian cycling when using a clonal population of medusa (*Cassiopea xamachana*), see **Methods**. 16 jellyfish were exposed to dark conditions from hour 36 to hour 72 or full-intensity light from hour 24 to hour 60. With this clonal population of jellyfish, circadian cycling of behavior is only observed for constant dark conditions and not constant full-intensity light conditions, consistent with results seen in the mixed population of *Cassiopea* shown in (**D**). (**F-H)** *Cassiopea* exhibit a decrease in activity in response to melatonin and pyrilamine exposure during the day. (**F**) Treatment with either pyrilamine or melatonin effects pulsing activity. The colored lines represent different concentrations of compounds tested. Activity was monitored before and after treatment. Time of treatment is indicated by a black arrow. Both melatonin and pyrilamine induce a concentration-dependent decrease in pulsing activity. (**G**) Activity of 18 *Cassiopea* exposed to 125 μM melatonin solubilized in ethanol compared to 19 *Cassiopea* treated with ethanol vehicle control from four laboratory replicates. *Cassiopea* were

monitored for 20 min before (baseline), during (treatment), and after (washout) either melatonin or vehicle treatment. Two-sided paired *t*-test, before/during melatonin treatment: $P = 4x10^{-7}$, and before/during vehicle treatment: $P = 7x10^{-1}$. ***$P<10^{-3}$, ns not significant (ns) $P>5x10^{-2}$. (**H**) Comparison of the normalized mean activity between the melatonin and control treatment. Error-bars represent the standard error of the mean.

BIBLIOGRAPHY

1. McIsaac RS, Bedbrook CN, & Arnold FH (2015) Recent advances in engineering microbial rhodopsins for optogenetics. *Current opinion in structural biology* 33:8-15.
2. Zhang F*, et al.* (2011) The microbial opsin family of optogenetic tools. *Cell* 147(7):1446-1457.
3. Gradinaru V, Mogri M, Thompson KR, Henderson JM, & Deisseroth K (2009) Optical deconstruction of parkinsonian neural circuitry. *Science* 324(5925):354-359.
4. Kralj JM, Douglass AD, Hochbaum DR, Maclaurin D, & Cohen AE (2011) Optical recording of action potentials in mammalian neurons using a microbial rhodopsin. *Nature methods* 9(1):90-95.
5. Gong Y, Wagner MJ, Zhong Li J, & Schnitzer MJ (2014) Imaging neural spiking in brain tissue using FRET-opsin protein voltage sensors. *Nature communications* 5:3674.
6. Ernst OP*, et al.* (2014) Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. *Chemical reviews* 114(1):126-163.
7. Spudich JL, Yang CS, Jung KH, & Spudich EN (2000) Retinylidene proteins: structures and functions from archaea to humans. *Annual review of cell and developmental biology* 16:365-392.
8. Nagel G*, et al.* (2003) Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* 100(24):13940-13945.
9. Nagel G*, et al.* (2002) Channelrhodopsin-1: a light-gated proton channel in green algae. *Science* 296(5577):2395-2398.
10. Boyden ES, Zhang F, Bamberg E, Nagel G, & Deisseroth K (2005) Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience* 8(9):1263-1268.
11. Mattis J*, et al.* (2011) Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nature methods* 9(2):159-172.
12. Zhang F*, et al.* (2007) Multimodal fast optical interrogation of neural circuitry. *Nature* 446(7136):633-639.
13. Kralj JM, Hochbaum DR, Douglass AD, & Cohen AE (2011) Electrical spiking in Escherichia coli probed with a fluorescent voltage-indicating protein. *Science* 333(6040):345-348.
14. Hochbaum DR*, et al.* (2014) All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nature methods* 11(8):825-833.
15. Flytzanis NC*, et al.* (2014) Archaerhodopsin variants with enhanced voltage-sensitive fluorescence in mammalian and Caenorhabditis elegans neurons. *Nature communications* 5:4894.
16. Gong Y, Li JZ, & Schnitzer MJ (2013) Enhanced Archaerhodopsin Fluorescent Protein Voltage Indicators. *PloS one* 8(6):e66959.
17. Deisseroth K & Hegemann P (2017) The form and function of channelrhodopsin. *Science* 357(6356).
18. Bamann C, Kirsch T, Nagel G, & Bamberg E (2008) Spectral characteristics of the photocycle of channelrhodopsin-2 and its implication for channel function. *Journal of molecular biology* 375(3):686-694.
19. Zhang F*, et al.* (2010) Optogenetic interrogation of neural circuits: technology for probing mammalian brain structures. *Nature protocols* 5(3):439-456.
20. Lin JY (2011) A user's guide to channelrhodopsin variants: features, limitations and future developments. *Experimental physiology* 96(1):19-25.
21. Looger LL & Griesbeck O (2012) Genetically encoded neural activity indicators. *Current opinion in neurobiology* 22(1):18-23.

22. Looger LL (2011) Running in reverse: rhodopsins sense voltage. *Nature methods* 9(1):43-44.
23. Kato HE*, et al.* (2012) Crystal structure of the channelrhodopsin light-gated cation channel. *Nature* 482(7385):369-374.
24. Melaccio F, Ferre N, & Olivucci M (2012) Quantum chemical modeling of rhodopsin mutants displaying switchable colors. *Physical chemistry chemical physics : PCCP* 14(36):12485-12495.
25. Engqvist MK*, et al.* (2015) Directed evolution of Gloeobacter violaceus rhodopsin spectral properties. *Journal of molecular biology* 427(1):205-220.
26. Wang W*, et al.* (2012) Tuning the electronic absorption of protein-embedded all-trans-retinal. *Science* 338(6112):1340-1343.
27. Lin JY, Knutsen PM, Muller A, Kleinfeld D, & Tsien RY (2013) ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nature neuroscience* 16(10):1499-1508.
28. Zhang F*, et al.* (2008) Red-shifted optogenetic excitation: a tool for fast neural control derived from Volvox carteri. *Nature neuroscience* 11(6):631-633.
29. Lin JY, Lin MZ, Steinbach P, & Tsien RY (2009) Characterization of engineered channelrhodopsin variants with improved properties and kinetics. *Biophysical journal* 96(5):1803-1814.
30. Prigge M*, et al.* (2012) Color-tuned channelrhodopsins for multiwavelength optogenetics. *The Journal of biological chemistry* 287(38):31804-31812.
31. Sineshchekov OA, Govorunova EG, Wang J, Li H, & Spudich JL (2013) Intramolecular proton transfer in channelrhodopsins. *Biophysical journal* 104(4):807-817.
32. Hou SY*, et al.* (2012) Diversity of Chlamydomonas channelrhodopsins. *Photochemistry and photobiology* 88(1):119-128.
33. Valderrama-Rincon JD*, et al.* (2012) An engineered eukaryotic protein glycosylation pathway in Escherichia coli. *Nature chemical biology* 8(5):434-436.
34. Zhou X, Sundholm D, Wesolowski TA, & Kaila VR (2014) Spectral tuning of rhodopsin and visual cone pigments. *Journal of the American Chemical Society* 136(7):2723-2726.
35. Berndt A, Lee SY, Ramakrishnan C, & Deisseroth K (2014) Structure-guided transformation of channelrhodopsin into a light-activated chloride channel. *Science* 344(6182):420-424.
36. Wietek J*, et al.* (2014) Conversion of channelrhodopsin into a light-gated chloride channel. *Science* 344(6182):409-412.
37. Yellen G (2002) The voltage-gated potassium channels and their relatives. *Nature* 419(6902):35-42.
38. Berndt A, Yizhar O, Gunaydin LA, Hegemann P, & Deisseroth K (2009) Bi-stable neural state switches. *Nature neuroscience* 12(2):229-234.
39. Lanyi JK & Schobert B (2003) Mechanism of proton transport in bacteriorhodopsin from crystallographic structures of the K, L, M1, M2, and M2' intermediates of the photocycle. *Journal of molecular biology* 328(2):439-450.
40. Chuong AS*, et al.* (2014) Noninvasive optical inhibition with a red-shifted microbial rhodopsin. *Nature neuroscience* 17(8):1123-1129.
41. Klapoetke NC*, et al.* (2014) Independent optical excitation of distinct neural populations. *Nature methods* 11(3):338-346.
42. Nagel G*, et al.* (2005) Light activation of channelrhodopsin-2 in excitable cells of Caenorhabditis elegans triggers rapid behavioral responses. *Current biology : CB* 15(24):2279-2284.
43. Gradinaru V*, et al.* (2010) Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* 141(1):154-165.

44. Bedbrook CN*, et al.* (2017) Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *Proceedings of the National Academy of Sciences of the United States of America*.
45. Bedbrook CN, Yang KK, Rice AJ, Gradinaru V, & Arnold FH (2017) Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS computational biology* 13(10):e1005786.
46. Maclaurin D, Venkatachalam V, Lee H, & Cohen AE (2013) Mechanism of voltage-sensitive fluorescence in a microbial rhodopsin. *Proceedings of the National Academy of Sciences of the United States of America* 110(15):5939-5944.
47. McIsaac RS*, et al.* (2014) Directed evolution of a far-red fluorescent rhodopsin. *Proceedings of the National Academy of Sciences of the United States of America* 111(36):13034-13039.
48. Wagner NL, Greco JA, Ranaghan MJ, & Birge RR (2013) Directed evolution of bacteriorhodopsin for applications in bioelectronics. *Journal of the Royal Society, Interface* 10(84):20130197.
49. Kolodner P, Lukashev EP, Ching YC, & Rousseau DL (1996) Electric-field-induced Schiff-base deprotonation in D85N mutant bacteriorhodopsin. *Proceedings of the National Academy of Sciences of the United States of America* 93(21):11618-11621.
50. Zou P*, et al.* (2014) Bright and fast multicoloured voltage reporters via electrochromic FRET. *Nature communications* 5:4625.
51. Mutoh H, Akemann W, & Knopfel T (2012) Genetically engineered fluorescent voltage reporters. *ACS chemical neuroscience* 3(8):585-592.
52. Jin L*, et al.* (2012) Single action potentials and subthreshold electrical events imaged in neurons with a fluorescent protein voltage probe. *Neuron* 75(5):779-785.
53. St-Pierre F*, et al.* (2014) High-fidelity optical reporting of neuronal electrical activity with an ultrafast fluorescent voltage sensor. *Nature neuroscience* 17(6):884-889.
54. Gobel W, Kampa BM, & Helmchen F (2007) Imaging cellular network dynamics in three dimensions using fast 3D laser scanning. *Nature methods* 4(1):73-79.
55. Cao G*, et al.* (2013) Genetically targeted optical electrophysiology in intact neural circuits. *Cell* 154(4):904-913.
56. Chen TW*, et al.* (2013) Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 499(7458):295-300.
57. Peters AJ, Chen SX, & Komiyama T (2014) Emergence of reproducible spatiotemporal activity during motor learning. *Nature* 510(7504):263-267.
58. Quirin S, Jackson J, Peterka DS, & Yuste R (2014) Simultaneous imaging of neural activity in three dimensions. *Frontiers in neural circuits* 8:29.
59. Hochbaum DR*, et al.* (2014) All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nature methods*.
60. Tian L*, et al.* (2009) Imaging neural activity in worms, flies and mice with improved GCaMP calcium indicators. *Nature methods* 6(12):875-881.
61. Akerboom J*, et al.* (2012) Optimization of a GCaMP calcium indicator for neural activity imaging. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32(40):13819-13840.
62. Mattis J*, et al.* (2012) Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nature methods* 9(2):159-172.
63. Chow BY*, et al.* (2010) High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature* 463(7277):98-102.
64. Kralj JM, Douglass AD, Hochbaum DR, Maclaurin D, & Cohen AE (2012) Optical recording of action potentials in mammalian neurons using a microbial rhodopsin. *Nature methods* 9(1):90-95.

65. Engqvist MK*, et al.* (2014) Directed Evolution of Gloeobacter violaceus Rhodopsin Spectral Properties. *Journal of molecular biology*.
66. McIsaac RS*, et al.* (2014) Directed Evolution of a Far-Red Fluorescent Rhodopsin. *Proceedings of the National Academy of Sciences*.
67. Chalasani SH*, et al.* (2007) Dissecting a circuit for olfactory behaviour in Caenorhabditis elegans. *Nature* 450(7166):63-70.
68. Kato S, Xu Y, Cho CE, Abbott LF, & Bargmann CI (2014) Temporal responses of C. elegans chemosensory neurons are preserved in behavioral dynamics. *Neuron* 81(3):616-628.
69. Nickell WT, Pun RY, Bargmann CI, & Kleene SJ (2002) Single ionic channels of two Caenorhabditis elegans chemosensory neurons in native membrane. *The Journal of membrane biology* 189(1):55-66.
70. Gao S & Zhen M (2011) Action potentials drive body wall muscle contractions in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* 108(6):2557-2562.
71. Looger LL (2012) Running in reverse: rhodopsins sense voltage. *Nature methods* 9(1):43-44.
72. Grishkevich V, Hashimshony T, & Yanai I (2011) Core promoter T-blocks correlate with gene expression levels in C. elegans. *Genome research* 21(5):707-717.
73. Mello C & Fire A (1995) DNA transformation. *Methods in cell biology* 48:451-482.
74. Okazaki A, Sudo Y, & Takagi S (2012) Optical silencing of C. elegans cells with arch proton pump. *PloS one* 7(5):e35370.
75. Chronis N, Zimmer M, & Bargmann CI (2007) Microfluidics for in vivo imaging of neuronal and behavioral activity in Caenorhabditis elegans. *Nature methods* 4(9):727-731.
76. Cho JY & Sternberg PW (2014) Multilevel modulation of a sensory motor circuit during C. elegans sleep and arousal. *Cell* 156(1-2):249-260.
77. Enami N*, et al.* (2006) Crystal structures of archaerhodopsin-1 and -2: Common structural motif in archaeal light-driven proton pumps. *Journal of molecular biology* 358(3):675-685.
78. Bedbrook CN*, et al.* (2015) Genetically Encoded Spy Peptide Fusion System to Detect Plasma Membrane-Localized Proteins In Vivo. *Chemistry & biology* 22(8):1108-1121.
79. Chalfie M, Tu Y, Euskirchen G, Ward WW, & Prasher DC (1994) Green fluorescent protein as a marker for gene expression. *Science* 263(5148):802-805.
80. Tsien RY (1998) The green fluorescent protein. *Annual review of biochemistry* 67:509-544.
81. Goedhart J*, et al.* (2012) Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%. *Nature communications* 3:751.
82. Patterson GH & Lippincott-Schwartz J (2002) A photoactivatable GFP for selective photolabeling of proteins and cells. *Science* 297(5588):1873-1877.
83. Piatkevich KD, Subach FV, & Verkhusha VV (2013) Far-red light photoactivatable near-infrared fluorescent proteins engineered from a bacterial phytochrome. *Nature communications* 4:2153.
84. Nienhaus K & Nienhaus GU (2014) Fluorescent proteins for live-cell imaging with super-resolution. *Chemical Society reviews* 43(4):1088-1106.
85. Zhou XX & Lin MZ (2013) Photoswitchable fluorescent proteins: ten years of colorful chemistry and exciting applications. *Current opinion in chemical biology* 17(4):682-690.
86. Dean KM & Palmer AE (2014) Advances in fluorescence labeling strategies for dynamic cellular imaging. *Nature chemical biology* 10(7):512-523.
87. Los GV*, et al.* (2008) HaloTag: a novel protein labeling technology for cell imaging and protein analysis. *ACS chemical biology* 3(6):373-382.
88. Gautier A*, et al.* (2008) An engineered protein tag for multiprotein labeling in living cells. *Chemistry & biology* 15(2):128-136.

89. Uttamapinant C*, et al.* (2010) A fluorophore ligase for site-specific protein labeling inside living cells. *Proceedings of the National Academy of Sciences of the United States of America* 107(24):10914-10919.
90. Juillerat A*, et al.* (2003) Directed evolution of O6-alkylguanine-DNA alkyltransferase for efficient labeling of fusion proteins with small molecules in vivo. *Chemistry & biology* 10(4):313-317.
91. Keppler A*, et al.* (2003) A general method for the covalent labeling of fusion proteins with small molecules in vivo. *Nature biotechnology* 21(1):86-89.
92. Lukinavicius G*, et al.* (2013) A near-infrared fluorophore for live-cell super-resolution microscopy of cellular proteins. *Nature chemistry* 5(2):132-139.
93. Gronemeyer T, Chidley C, Juillerat A, Heinis C, & Johnsson K (2006) Directed evolution of O6-alkylguanine-DNA alkyltransferase for applications in protein labeling. *Protein engineering, design & selection : PEDS* 19(7):309-316.
94. Zakeri B*, et al.* (2012) Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proceedings of the National Academy of Sciences of the United States of America* 109(12):E690-697.
95. Zhang WB, Sun F, Tirrell DA, & Arnold FH (2013) Controlling macromolecular topology with genetically encoded SpyTag-SpyCatcher chemistry. *Journal of the American Chemical Society* 135(37):13988-13997.
96. Shemiakina, II*, et al.* (2012) A monomeric red fluorescent protein with low cytotoxicity. *Nature communications* 3:1204.
97. Glauser DA*, et al.* (2011) Heat avoidance is regulated by transient receptor potential (TRP) channels and a neuropeptide signaling pathway in Caenorhabditis elegans. *Genetics* 188(1):91-103.
98. Schwabe T, Neuert H, & Clandinin TR (2013) A network of cadherin-mediated interactions polarizes growth cones to determine targeting specificity. *Cell* 154(2):351-364.
99. Prober DA*, et al.* (2008) Zebrafish TRPA1 channels are required for chemosensation but not for thermosensation or mechanosensory hair cell function. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 28(40):10102-10110.
100. Gupta VK, You Y, Gupta VB, Klistorner A, & Graham SL (2013) TrkB Receptor Signalling: Implications in Neurodegenerative, Psychiatric and Proliferative Disorders. *International journal of molecular sciences* 14(5):10122-10142.
101. Kohl J*, et al.* (2014) Ultrafast tissue staining with chemical tags. *Proceedings of the National Academy of Sciences of the United States of America* 111(36):E3805-3814.
102. McMurray MA & Thorner J (2008) Septin stability and recycling during dynamic structural transitions in cell division and development. *Current biology : CB* 18(16):1203-1208.
103. Kimble J & Hirsh D (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. *Developmental biology* 70(2):396-417.
104. Jorgensen EM (GABA. in *WormBook*, ed Community TCeR (WormBook).
105. Hausser M (2014) Optogenetics: the age of light. *Nature methods* 11(10):1012-1014.
106. Malenka RC & Bear MF (2004) LTP and LTD: an embarrassment of riches. *Neuron* 44(1):5-21.
107. Cong L*, et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* 339(6121):819-823.
108. Chung K*, et al.* (2013) Structural and molecular interrogation of intact biological systems. *Nature* 497(7449):332-337.
109. Yang B*, et al.* (2014) Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell* 158(4):945-958.
110. Kille S*, et al.* (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS synthetic biology* 2(2):83-92.

111. Sun F, Zhang WB, Mahdavi A, Arnold FH, & Tirrell DA (2014) Synthesis of bioactive protein hydrogels by genetically encoded SpyTag-SpyCatcher chemistry. *Proceedings of the National Academy of Sciences of the United States of America* 111(31):11269-11274.
112. Brenner S (1974) The genetics of Caenorhabditis elegans. *Genetics* 77(1):71-94.
113. Hodgkin J, Horvitz HR, & Brenner S (1979) Nondisjunction Mutants of the Nematode CAENORHABDITIS ELEGANS. *Genetics* 91(1):67-94.
114. Bedbrook CN*, et al.* (2017) Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *Proceedings of the National Academy of Sciences of the United States of America* 114(13):E2624-E2633.
115. Overington JP, Al-Lazikani B, & Hopkins AL (2006) How many drug targets are there? *Nature reviews. Drug discovery* 5(12):993-996.
116. Urban DJ & Roth BL (2015) DREADDs (designer receptors exclusively activated by designer drugs): chemogenetic tools with therapeutic utility. *Annual review of pharmacology and toxicology* 55:399-417.
117. Yizhar O, Fenno LE, Davidson TJ, Mogri M, & Deisseroth K (2011) Optogenetics in neural systems. *Neuron* 71(1):9-34.
118. Andrell J & Tate CG (2013) Overexpression of membrane proteins in mammalian cells for structural studies. *Molecular membrane biology* 30(1):52-63.
119. Lluis MW, Godfroy JI, 3rd, & Yin H (2013) Protein engineering methods applied to membrane protein targets. *Protein engineering, design & selection : PEDS* 26(2):91-100.
120. Cymer F, von Heijne G, & White SH (2015) Mechanisms of integral membrane protein insertion and folding. *Journal of molecular biology* 427(5):999-1022.
121. Chapple JP & Cheetham ME (2003) The chaperone environment at the cytoplasmic face of the endoplasmic reticulum can modulate rhodopsin processing and inclusion formation. *The Journal of biological chemistry* 278(21):19087-19094.
122. Conn PM & Ulloa-Aguirre A (2010) Trafficking of G-protein-coupled receptors to the plasma membrane: insights for pharmacoperone drugs. *Trends in endocrinology and metabolism: TEM* 21(3):190-197.
123. Voigt CA, Martinez C, Wang ZG, Mayo SL, & Arnold FH (2002) Protein building blocks preserved by recombination. *Nature structural biology* 9(7):553-558.
124. Suzuki T*, et al.* (2003) Archaeal-type rhodopsins in Chlamydomonas: model structure and intracellular localization. *Biochemical and biophysical research communications* 301(3):711-717.
125. Sineshchekov OA, Jung KH, & Spudich JL (2002) Two rhodopsins mediate phototaxis to low- and high-intensity light in Chlamydomonas reinhardtii. *Proceedings of the National Academy of Sciences of the United States of America* 99(13):8689-8694.
126. Schneider F, Grimm C, & Hegemann P (2015) Biophysics of Channelrhodopsin. *Annual review of biophysics* 44:167-186.
127. Ishizuka T, Kakuda M, Araki R, & Yawo H (2006) Kinetic evaluation of photosensitivity in genetically engineered neurons expressing green algae light-gated channels. *Neuroscience research* 54(2):85-94.
128. Scott DJ, Kummer L, Tremmel D, & Pluckthun A (2013) Stabilizing membrane proteins through protein engineering. *Current opinion in chemical biology* 17(3):427-435.
129. Sarkar CA*, et al.* (2008) Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. *Proceedings of the National Academy of Sciences of the United States of America* 105(39):14808-14813.
130. Newstead S, Kim H, von Heijne G, Iwata S, & Drew D (2007) High-throughput fluorescent-based optimization of eukaryotic membrane protein overexpression and purification in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* 104(35):13936-13941.

131. Trudeau DL, Smith MA, & Arnold FH (2013) Innovation by homologous recombination. *Current opinion in chemical biology* 17(6):902-909.
132. Romero PA & Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nature reviews. Molecular cell biology* 10(12):866-876.
133. Drummond DA, Silberg JJ, Meyer MM, Wilke CO, & Arnold FH (2005) On the conservative nature of intragenic recombination. *Proceedings of the National Academy of Sciences of the United States of America* 102(15):5380-5385.
134. Endelman JB, Silberg JJ, Wang ZG, & Arnold FH (2004) Site-directed protein recombination as a shortest-path problem. *Protein engineering, design & selection : PEDS* 17(7):589-594.
135. Smith MA, Romero PA, Wu T, Brustad EM, & Arnold FH (2013) Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein science* 22(2):231-238.
136. Meyer MM, Hochrein L, & Arnold FH (2006) Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein engineering, design & selection : PEDS* 19(12):563-570.
137. Crameri A, Raillard SA, Bermudez E, & Stemmer WP (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391(6664):288-291.
138. Otey CR*, et al.* (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS biology* 4(5):e112.
139. Li Y*, et al.* (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nature biotechnology* 25(9):1051-1056.
140. Romero PA*, et al.* (2012) SCHEMA-designed variants of human Arginase I and II reveal sequence elements important to stability and catalysis. *ACS synthetic biology* 1(6):221-228.
141. Heinzelman P*, et al.* (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *Proceedings of the National Academy of Sciences of the United States of America* 106(14):5610-5615.
142. Kianianmomeni A, Stehfest K, Nematollahi G, Hegemann P, & Hallmann A (2009) Channelrhodopsins of Volvox carteri are photochromic proteins that are specifically expressed in somatic cells under control of light, temperature, and the sex inducer. *Plant physiology* 151(1):347-366.
143. Ernst OP*, et al.* (2008) Photoactivation of channelrhodopsin. *The Journal of biological chemistry* 283(3):1637-1643.
144. Govorunova EG, Spudich EN, Lane CE, Sineshchekov OA, & Spudich JL (2011) New channelrhodopsin with a red-shifted spectrum and rapid kinetics from Mesostigma viride. *mBio* 2(3):e00115-00111.
145. Govorunova EG, Sineshchekov OA, Li H, Janz R, & Spudich JL (2013) Characterization of a highly efficient blue-shifted channelrhodopsin from the marine alga Platymonas subcordiformis. *The Journal of biological chemistry* 288(41):29911-29922.
146. Krause A & Golovin D (2014) Submodular function maximization. *Tractability: Practical Approaches to Hard Problems.*, (Cambridge University Press).
147. Wagner S, Bader ML, Drew D, & de Gier JW (2006) Rationalizing membrane protein overexpression. *Trends in biotechnology* 24(8):364-371.
148. Smith MA, Bedbrook CN, Wu T, & Arnold FH (2013) Hypocrea jecorina cellobiohydrolase I stabilizing mutations identified using noncontiguous recombination. *ACS synthetic biology* 2(12):690-696.
149. Heinzelman P*, et al.* (2009) SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *The Journal of biological chemistry* 284(39):26229-26233.
150. Chauhan JS, Rao A, & Raghava GP (2013) In silico platform for prediction of N-, O- and C-glycosites in eukaryotic protein sequences. *PloS one* 8(6):e67008.

151. Smith MA & Arnold FH (2014) Designing libraries of chimeric proteins using SCHEMA recombination and RASPP. *Methods in molecular biology* 1179:335-343.
152. Smith MA & Arnold FH (2014) Noncontiguous SCHEMA protein recombination. *Methods in molecular biology* 1179:345-352.
153. Carpenter AE*, et al.* (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology* 7(10):R100.
154. Oliphant TE (2007) Python for Scientific Computing. *Computing in Science and Engineering* 9(3):10-20.
155. Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering* 9(3):90-95.
156. Walt SVD, Colbert SC, & Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. *Computing in Science and Engineering* 13(2):22-30.
157. Lee MC, Miller EA, Goldberg J, Orci L, & Schekman R (2004) Bi-directional protein transport between the ER and Golgi. *Annual review of cell and developmental biology* 20:87-123.
158. Klenk C, Ehrenmann J, Schutz M, & Pluckthun A (2016) A generic selection system for improved expression and thermostability of G protein-coupled receptors by directed evolution. *Scientific reports* 6:21294.
159. Fleming KG (2014) Energetics of membrane protein folding. *Annual review of biophysics* 43:233-255.
160. Elazar A*, et al.* (2016) Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *eLife* 5.
161. White SH & Wimley WC (1999) Membrane protein folding and stability: physical principles. *Annual review of biophysics and biomolecular structure* 28:319-365.
162. von Heijne G (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* 341(6241):456-458.
163. Duong MT, Jaszewski TM, Fleming KG, & MacKenzie KR (2007) Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. *Journal of molecular biology* 371(2):422-434.
164. Habibi N, Mohd Hashim SZ, Norouzi A, & Samian MR (2014) A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. *BMC bioinformatics* 15:134.
165. Wilkinson DL & Harrison RG (1991) Predicting the solubility of recombinant proteins in Escherichia coli. *Bio/technology* 9(5):443-448.
166. Chang CC*, et al.* (2016) Periscope: quantitative prediction of soluble protein expression in the periplasm of Escherichia coli. *Scientific reports* 6:21844.
167. Wang H*, et al.* (2016) Crysalis: an integrated server for computational analysis and design of protein crystallization. *Scientific reports* 6:21383.
168. Radivojac P*, et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nature methods* 10(3):221-227.
169. Rasmussen CE & Williams CKI (2006) Gaussian Processes for Machine Learning (the MIT Press, Cambridge, MA).
170. Romero PA, Krause A, & Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences of the United States of America* 110(3):E193-201.
171. Srinivas N, Krause A, Kakade SM, & Seeger M (2010) Gaussian process optimization in the bandit setting: No regret and experimental design. *The 27 th International Conference on Machine Learning*.
172. Kyte J & Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology* 157(1):105-132.

173. Walt S, Colbert SC, & Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. *Computing in Science and Engineering* 13:22-30.
174. Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science and Engineering* 9:90-95.
175. Oliphant TE (2007) Python for Scientific Computing. *Computing in Science and Engineering* 9:10-20.
176. Pedregosa F*, et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825-2830.
177. Berndt A*, et al.* (2016) Structural foundations of optogenetics: Determinants of channelrhodopsin ion selectivity. *Proceedings of the National Academy of Sciences of the United States of America* 113(4):822-829.
178. Gunaydin LA*, et al.* (2010) Ultrafast optogenetic control. *Nature neuroscience* 13(3):387-392.
179. Chan KY*, et al.* (2017) Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nature neuroscience* 20(8):1172-1179.
180. Deverman BE*, et al.* (2016) Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nature biotechnology* 34(2):204-209.
181. Weitzman MD & Linden RM (2011) Adeno-associated virus biology. *Methods in molecular biology* 807:1-23.
182. Samulski RJ, Chang LS, & Shenk T (1989) Helper-free stocks of recombinant adeno-associated viruses: normal integration does not require viral gene expression. *Journal of virology* 63(9):3822-3828.
183. Samulski RJ & Muzyczka N (2014) AAV-Mediated Gene Therapy for Research and Therapeutic Purposes. *Annual review of virology* 1(1):427-451.
184. Hastie E & Samulski RJ (2015) Adeno-associated virus at 50: a golden anniversary of discovery, research, and gene therapy success--a personal perspective. *Human gene therapy* 26(5):257-265.
185. Alisky JM*, et al.* (2000) Transduction of murine cerebellar neurons with recombinant FIV and AAV5 vectors. *Neuroreport* 11(12):2669-2673.
186. Burger C*, et al.* (2004) Recombinant AAV viral vectors pseudotyped with viral capsids from serotypes 1, 2, and 5 display differential efficiency and cell tropism after delivery to different regions of the central nervous system. *Molecular therapy : the journal of the American Society of Gene Therapy* 10(2):302-317.
187. Davidson BL*, et al.* (2000) Recombinant adeno-associated virus type 2, 4, and 5 vectors: transduction of variant cell types and regions in the mammalian central nervous system. *Proceedings of the National Academy of Sciences of the United States of America* 97(7):3428-3432.
188. Lawlor PA, Bland RJ, Mouravlev A, Young D, & During MJ (2009) Efficient gene delivery and selective transduction of glial cells in the mammalian brain by AAV serotypes isolated from nonhuman primates. *Molecular therapy : the journal of the American Society of Gene Therapy* 17(10):1692-1702.
189. Van der Perren A*, et al.* (2011) Efficient and stable transduction of dopaminergic neurons in rat substantia nigra by rAAV 2/1, 2/2, 2/5, 2/6.2, 2/7, 2/8 and 2/9. *Gene therapy* 18(5):517-527.
190. McCown TJ, Xiao X, Li J, Breese GR, & Samulski RJ (1996) Differential and persistent expression patterns of CNS gene transfer by an adeno-associated virus (AAV) vector. *Brain research* 713(1-2):99-107.
191. Große S*, et al.* (2017) Relevance of assembly-activating protein for Adeno-associated virus vector production and capsid protein stability in mammalian and insect cells. *Journal of virology*.
192. Heidenreich M & Zhang F (2016) Applications of CRISPR-Cas systems in neuroscience. *Nature reviews. Neuroscience* 17(1):36-44.

193. Suzuki K, *et al.* (2016) In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* 540(7631):144-149.
194. Grieger JC, Choi VW, & Samulski RJ (2006) Production and characterization of adeno-associated viral vectors. *Nature protocols* 1(3):1412-1428.
195. Lock M, *et al.* (2010) Rapid, simple, and versatile manufacturing of recombinant adeno-associated viral vectors at scale. *Human gene therapy* 21(10):1259-1271.
196. Bey K, *et al.* (2017) Efficient CNS targeting in adult mice by intrathecal infusion of single-stranded AAV9-GFP for gene therapy of neurological disorders. *Gene therapy* 24(5):325-332.
197. Chakrabarty P, *et al.* (2013) Capsid serotype and timing of injection determines AAV transduction in the neonatal mice brain. *PloS one* 8(6):e67680.
198. McLean JR, *et al.* (2014) Widespread neuron-specific transgene expression in brain and spinal cord following synapsin promoter-driven AAV9 neonatal intracerebroventricular injection. *Neuroscience letters* 576:73-78.
199. Passini MA, *et al.* (2003) Intraventricular brain injection of adeno-associated virus type 1 (AAV1) in neonatal mice results in complementary patterns of neuronal transduction to AAV2 and total long-term correction of storage lesions in the brains of beta-glucuronidase-deficient mice. *Journal of virology* 77(12):7034-7040.
200. Samaranch L, *et al.* (2012) Adeno-associated virus serotype 9 transduction in the central nervous system of nonhuman primates. *Human gene therapy* 23(4):382-389.
201. Samaranch L, *et al.* (2013) Strong cortical and spinal cord transduction after AAV7 and AAV9 delivery into the cerebrospinal fluid of nonhuman primates. *Human gene therapy* 24(5):526-532.
202. Schuster DJ, *et al.* (2014) Biodistribution of adeno-associated virus serotype 9 (AAV9) vector after intrathecal and intravenous delivery in mouse. *Frontiers in neuroanatomy* 8:42.
203. Snyder BR, *et al.* (2011) Comparison of adeno-associated viral vector serotypes for spinal cord and motor neuron gene delivery. *Human gene therapy* 22(9):1129-1135.
204. Yang B, *et al.* (2014) Global CNS transduction of adult mice by intravenously delivered rAAVrh.8 and rAAVrh.10 and nonhuman primates by rAAVrh.10. *Molecular therapy : the journal of the American Society of Gene Therapy* 22(7):1299-1309.
205. Foust KD, *et al.* (2009) Intravascular AAV9 preferentially targets neonatal neurons and adult astrocytes. *Nature biotechnology* 27(1):59-65.
206. Zhang H, *et al.* (2011) Several rAAV vectors efficiently cross the blood-brain barrier and transduce neurons and astrocytes in the neonatal mouse central nervous system. *Molecular therapy : the journal of the American Society of Gene Therapy* 19(8):1440-1448.
207. Fu H, *et al.* (2003) Self-complementary adeno-associated virus serotype 2 vector: global distribution and broad dispersion of AAV-mediated transgene expression in mouse brain. *Molecular therapy : the journal of the American Society of Gene Therapy* 8(6):911-917.
208. McCarty DM, DiRosario J, Gulaid K, Muenzer J, & Fu H (2009) Mannitol-facilitated CNS entry of rAAV2 vector significantly delayed the neurological disease progression in MPS IIIB mice. *Gene therapy* 16(11):1340-1352.
209. Wang S, Olumolade OO, Sun T, Samiotaki G, & Konofagou EE (2015) Noninvasive, neuron-specific gene therapy can be facilitated by focused ultrasound and recombinant adeno-associated virus. *Gene therapy* 22(1):104-110.
210. Gray SJ, *et al.* (2010) Directed evolution of a novel adeno-associated virus (AAV) vector that crosses the seizure-compromised blood-brain barrier (BBB). *Molecular therapy : the journal of the American Society of Gene Therapy* 18(3):570-578.
211. Choudhury SR, *et al.* (2016) Widespread Central Nervous System Gene Transfer and Silencing After Systemic Delivery of Novel AAV-AS Vector. *Molecular therapy : the journal of the American Society of Gene Therapy* 24(4):726-735.

212. Chang RB, Strochlic DE, Williams EK, Umans BD, & Liberles SD (2015) Vagal Sensory Neuron Subtypes that Differentially Control Breathing. *Cell* 161(3):622-633.
213. Williams EK, *et al.* (2016) Sensory Neurons that Detect Stretch and Nutrients in the Digestive System. *Cell* 166(1):209-221.
214. Iyer SM, *et al.* (2014) Virally mediated optogenetic excitation and inhibition of pain in freely moving nontransgenic mice. *Nature biotechnology* 32(3):274-278.
215. Samad OA, *et al.* (2013) Virus-mediated shRNA Knockdown of Nav1.3 in Rat Dorsal Root Ganglion Attenuates Nerve Injury-induced Neuropathic Pain. *Molecular therapy : the journal of the American Society of Gene Therapy* 21(1):49-56.
216. Towne C, Pertin M, Beggah AT, Aebischer P, & Decosterd I (2009) Recombinant adeno-associated virus serotype 6 (rAAV2/6)-mediated gene transfer to nociceptive neurons through different routes of delivery. *Molecular pain* 5:52.
217. Francois A, *et al.* (2017) A Brainstem-Spinal Cord Inhibitory Circuit for Mechanical Pain Modulation by GABA and Enkephalins. *Neuron* 93(4):822-839 e826.
218. Nussinovitch U & Gepstein L (2015) Optogenetics for in vivo cardiac pacing and resynchronization therapies. *Nature biotechnology* 33(7):750-754.
219. Vogt CC, *et al.* (2015) Systemic gene transfer enables optogenetic pacing of mouse hearts. *Cardiovascular research* 106(2):338-343.
220. Duque S, *et al.* (2009) Intravenous administration of self-complementary AAV9 enables transgene delivery to adult motor neurons. *Molecular therapy : the journal of the American Society of Gene Therapy* 17(7):1187-1196.
221. Gombash SE, *et al.* (2014) Intravenous AAV9 efficiently transduces myenteric neurons in neonate and juvenile mice. *Frontiers in molecular neuroscience* 7:81.
222. Montgomery KL, Iyer SM, Christensen AJ, Deisseroth K, & Delp SL (2016) Beyond the brain: Optogenetic control in the spinal cord and peripheral nervous system. *Science translational medicine* 8(337):337rv335.
223. Gray SJ, *et al.* (2011) Preclinical differences of intravascular AAV9 delivery to neurons and glia: a comparative study of adult mice and nonhuman primates. *Molecular therapy : the journal of the American Society of Gene Therapy* 19(6):1058-1069.
224. Pulicherla N, *et al.* (2011) Engineering liver-detargeted AAV9 vectors for cardiac and musculoskeletal gene transfer. *Molecular therapy : the journal of the American Society of Gene Therapy* 19(6):1070-1078.
225. Gray SJ, *et al.* (2011) Optimizing promoters for recombinant adeno-associated virus-mediated gene expression in the peripheral and central nervous system using self-complementary vectors. *Human gene therapy* 22(9):1143-1153.
226. Pignataro D, *et al.* (2017) Adeno-Associated Viral Vectors Serotype 8 for Cell-Specific Delivery of Therapeutic Genes in the Central Nervous System. *Frontiers in neuroanatomy* 11:2.
227. Portales-Casamar E, *et al.* (2010) A regulatory toolbox of MiniPromoters to drive selective expression in the brain. *Proceedings of the National Academy of Sciences of the United States of America* 107(38):16589-16594.
228. Nathanson JL, *et al.* (2009) Short Promoters in Viral Vectors Drive Selective Expression in Mammalian Inhibitory Neurons, but do not Restrict Activity to Specific Inhibitory Cell-Types. *Frontiers in neural circuits* 3:19.
229. de Leeuw CN, *et al.* (2014) Targeted CNS Delivery Using Human MiniPromoters and Demonstrated Compatibility with Adeno-Associated Viral Vectors. *Molecular therapy. Methods & clinical development* 1:5.
230. de Leeuw CN, *et al.* (2016) rAAV-compatible MiniPromoters for restricted expression in the brain and eye. *Molecular brain* 9(1):52.
231. Lee Y, Messing A, Su M, & Brenner M (2008) GFAP promoter elements required for region-specific and astrocyte-specific expression. *Glia* 56(5):481-493.

232. Gow A, Friedrich VL, Jr., & Lazzarini RA (1992) Myelin basic protein gene contains separate enhancers for oligodendrocyte and Schwann cell expression. *The Journal of cell biology* 119(3):605-616.
233. Kugler S, Kilic E, & Bahr M (2003) Human synapsin 1 gene promoter confers highly neuron-specific long-term transgene expression from an adenoviral vector in the adult rat brain depending on the transduced area. *Gene therapy* 10(4):337-347.
234. Lee AT, Vogt D, Rubenstein JL, & Sohal VS (2014) A class of GABAergic neurons in the prefrontal cortex sends long-range projections to the nucleus accumbens and elicits acute avoidance behavior. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 34(35):11519-11525.
235. Dimidschstein J, *et al.* (2016) A viral strategy for targeting and manipulating interneurons across vertebrate species. *Nature neuroscience* 19(12):1743-1749.
236. Shima Y, *et al.* (2016) A Mammalian enhancer trap resource for discovering and manipulating neuronal cell types. *eLife* 5:e13503.
237. Gossen M, *et al.* (1995) Transcriptional activation by tetracyclines in mammalian cells. *Science* 268(5218):1766-1769.
238. Agha-Mohammadi S, *et al.* (2004) Second-generation tetracycline-regulatable promoter: repositioned tet operator elements optimize transactivator synergy while shorter minimal promoter offers tight basal leakiness. *The journal of gene medicine* 6(7):817-828.
239. Chenuaud P, *et al.* (2004) Optimal design of a single recombinant adeno-associated virus derived from serotypes 1 and 2 to achieve more tightly regulated transgene expression from nonhuman primate muscle. *Molecular therapy : the journal of the American Society of Gene Therapy* 9(3):410-418.
240. Liu X, *et al.* (2012) Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* 484(7394):381-385.
241. Ye L, *et al.* (2016) Wiring and Molecular Features of Prefrontal Ensembles Representing Distinct Experiences. *Cell* 165(7):1776-1788.
242. Kozomara A & Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research* 39(Database issue):D152-157.
243. Lagos-Quintana M, *et al.* (2002) Identification of tissue-specific microRNAs from mouse. *Current biology : CB* 12(9):735-739.
244. Xie J, *et al.* (2011) MicroRNA-regulated, systemically delivered rAAV9: a step closer to CNS-restricted transgene expression. *Molecular therapy : the journal of the American Society of Gene Therapy* 19(3):526-535.
245. Brown BD & Naldini L (2009) Exploiting and antagonizing microRNA regulation for therapeutic and experimental applications. *Nature reviews. Genetics* 10(8):578-585.
246. Karali M, *et al.* (2011) MicroRNA-restricted transgene expression in the retina. *PloS one* 6(7):e22166.
247. Fenno LE, *et al.* (2014) Targeting cells with single vectors using multiple-feature Boolean logic. *Nature methods* 11(7):763-772.
248. Fenno LE, Mattis J, Ramakrishnan C, & Deisseroth K (2017) A Guide to Creating and Testing New INTRSECT Constructs. *Current protocols in neuroscience* 80:4 39 31-34 39 24.
249. Xie Q, *et al.* (2002) The atomic structure of adeno-associated virus (AAV-2), a vector for human gene therapy. *Proceedings of the National Academy of Sciences of the United States of America* 99(16):10405-10410.
250. Nonnenmacher M & Weber T (2012) Intracellular transport of recombinant adeno-associated virus vectors. *Gene therapy* 19(6):649-658.
251. Cearley CN & Wolfe JH (2006) Transduction characteristics of adeno-associated virus vectors expressing cap serotypes 7, 8, 9, and Rh10 in the mouse brain. *Molecular therapy : the journal of the American Society of Gene Therapy* 13(3):528-537.

252. Cearley CN*, et al.* (2008) Expanded repertoire of AAV vector serotypes mediate unique patterns of transduction in mouse brain. *Molecular therapy : the journal of the American Society of Gene Therapy* 16(10):1710-1718.
253. Taymans JM*, et al.* (2007) Comparative analysis of adeno-associated viral vector serotypes 1, 2, 5, 7, and 8 in mouse brain. *Human gene therapy* 18(3):195-206.
254. Klimczak RR, Koerber JT, Dalkara D, Flannery JG, & Schaffer DV (2009) A novel adeno-associated viral variant for efficient and selective intravitreal transduction of rat Muller cells. *PloS one* 4(10):e7467.
255. Koerber JT*, et al.* (2009) Molecular evolution of adeno-associated virus for enhanced glial gene delivery. *Molecular therapy : the journal of the American Society of Gene Therapy* 17(12):2088-2095.
256. Yla-Herttuala S (2012) Endgame: glybera finally recommended for approval as the first gene therapy drug in the European union. *Molecular therapy : the journal of the American Society of Gene Therapy* 20(10):1831-1832.
257. Gao G*, et al.* (2004) Clades of Adeno-associated viruses are widely disseminated in human tissues. *Journal of virology* 78(12):6381-6388.
258. Bello A*, et al.* (2009) Isolation and evaluation of novel adeno-associated virus sequences from porcine tissues. *Gene therapy* 16(11):1320-1328.
259. Farkas SL*, et al.* (2004) A parvovirus isolated from royal python (Python regius) is a member of the genus Dependovirus. *The Journal of general virology* 85(Pt 3):555-561.
260. Murlidharan G, Samulski RJ, & Asokan A (2014) Biology of adeno-associated viral vectors in the central nervous system. *Frontiers in molecular neuroscience* 7:76.
261. Vance MA, Mitchell A, Samulski RJ, & Hashad D (2015) AAV Biology, Infectivity and Therapeutic Use from Bench to Clinic. *Gene Therapy - Principles and Challenges*, (InTech).
262. Grimm D*, et al.* (2008) In vitro and in vivo gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *Journal of virology* 82(12):5887-5911.
263. Koerber JT, Jang JH, & Schaffer DV (2008) DNA shuffling of adeno-associated virus yields functionally diverse viral progeny. *Molecular therapy : the journal of the American Society of Gene Therapy* 16(10):1703-1709.
264. Maheshri N, Koerber JT, Kaspar BK, & Schaffer DV (2006) Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nature biotechnology* 24(2):198-204.
265. Xie Q*, et al.* (2017) The 2.8 A Electron Microscopy Structure of Adeno-Associated Virus-DJ Bound by a Heparinoid Pentasaccharide. *Molecular therapy. Methods & clinical development* 5:1-12.
266. Zhong L*, et al.* (2008) Next generation of adeno-associated virus 2 vectors: point mutations in tyrosines lead to high-efficiency transduction at lower doses. *Proceedings of the National Academy of Sciences of the United States of America* 105(22):7827-7832.
267. Muller OJ*, et al.* (2003) Random peptide libraries displayed on adeno-associated virus to select for targeted gene therapy vectors. *Nature biotechnology* 21(9):1040-1046.
268. Tervo DG*, et al.* (2016) A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. *Neuron* 92(2):372-382.
269. Adachi K, Enoki T, Kawano Y, Veraz M, & Nakai H (2014) Drawing a high-resolution functional map of adeno-associated virus capsid by massively parallel sequencing. *Nature communications* 5:3075.
270. Kotterman MA & Schaffer DV (2014) Engineering adeno-associated viruses for clinical gene therapy. *Nature reviews. Genetics* 15(7):445-451.
271. Treweek JB & Gradinaru V (2016) Extracting structural and functional features of widely distributed biological circuits with single cell resolution via tissue clearing and delivery vectors. *Current opinion in biotechnology* 40:193-207.

272. Treweek JB*, et al.* (2015) Whole-body tissue stabilization and selective extractions via tissue-hydrogel hybrids for high-resolution intact circuit mapping and phenotyping. *Nature protocols* 10(11):1860-1896.
273. Korbelin J*, et al.* (2016) Pulmonary Targeting of Adeno-associated Viral Vectors by Next-generation Sequencing-guided Screening of Random Capsid Displayed Peptide Libraries. *Molecular therapy : the journal of the American Society of Gene Therapy* 24(6):1050-1061.
274. Korbelin J*, et al.* (2017) Optimization of design and production strategies for novel adeno-associated viral display peptide libraries. *Gene therapy* 24(8):470-481.
275. Allen WE*, et al.* (2017) Global Representations of Goal-Directed Behavior in Distinct Cell Types of Mouse Neocortex. *Neuron* 94(4):891-907 e896.
276. Dana H*, et al.* (2016) Sensitive red protein calcium indicators for imaging neural activity. *eLife* 5.
277. Hillier D*, et al.* (2017) Causal evidence for retina-dependent and -independent visual motion computations in mouse cortex. *Nature neuroscience* 20(7):960-968.
278. Morabito G*, et al.* (2017) AAV-PHP.B-Mediated Global-Scale Expression in the Mouse Nervous System Enables GBA1 Gene Therapy for Wide Protection from Synucleinopathy. *Molecular therapy : the journal of the American Society of Gene Therapy*.
279. Cai D, Cohen KB, Luo T, Lichtman JW, & Sanes JR (2013) Improved tools for the Brainbow toolbox. *Nature methods* 10(6):540-547.
280. Lin MZ & Schnitzer MJ (2016) Genetically encoded indicators of neuronal activity. *Nature neuroscience* 19(9):1142-1153.
281. Economo MN*, et al.* (2016) A platform for brain-wide imaging and reconstruction of individual neurons. *eLife* 5:e10566.
282. Kim SY*, et al.* (2013) Diverging neural pathways assemble a behavioural state from separable features in anxiety. *Nature* 496(7444):219-223.
283. Xiao C*, et al.* (2016) Cholinergic Mesopontine Signals Govern Locomotion and Reward through Dissociable Midbrain Pathways. *Neuron* 90(2):333-347.
284. Castle MJ, Gershenson ZT, Giles AR, Holzbaur EL, & Wolfe JH (2014) Adeno-associated virus serotypes 1, 8, and 9 share conserved mechanisms for anterograde and retrograde axonal transport. *Human gene therapy* 25(8):705-720.
285. Castle MJ, Perlson E, Holzbaur EL, & Wolfe JH (2014) Long-distance axonal transport of AAV9 is driven by dynein and kinesin-2 and is trafficked in a highly motile Rab7-positive compartment. *Molecular therapy : the journal of the American Society of Gene Therapy* 22(3):554-566.
286. Rothermel M, Brunert D, Zabawa C, Diaz-Quesada M, & Wachowiak M (2013) Transgene expression in target-defined neuron populations mediated by retrograde infection with adeno-associated viral vectors. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33(38):15195-15206.
287. Salegio EA*, et al.* (2013) Axonal transport of adeno-associated viral vectors is serotype-dependent. *Gene therapy* 20(3):348-352.
288. Junyent F & Kremer EJ (2015) CAV-2--why a canine virus is a neurobiologist's best friend. *Current opinion in pharmacology* 24:86-93.
289. Danielson NB*, et al.* (2017) In Vivo Imaging of Dentate Gyrus Mossy Cells in Behaving Mice. *Neuron* 93(3):552-559 e554.
290. Haenraets K*, et al.* (2017) Spinal nociceptive circuit analysis with recombinant adeno-associated viruses: the impact of serotypes and promoters. *Journal of neurochemistry*.
291. Wagner MJ, Kim TH, Savall J, Schnitzer MJ, & Luo L (2017) Cerebellar granule cells encode the expectation of reward. *Nature* 544(7648):96-100.
292. Zhao Z*, et al.* (2017) A Central Catecholaminergic Circuit Controls Blood Glucose Levels during Stress. *Neuron* 95(1):138-152 e135.

293. Zingg B, *et al.* (2017) AAV-Mediated Anterograde Transsynaptic Tagging: Mapping Corticocollicular Input-Defined Neural Pathways for Defense Behaviors. *Neuron* 93(1):33-47.
294. Lerner TN, *et al.* (2015) Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. *Cell* 162(3):635-647.
295. Menegas W, *et al.* (2015) Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *eLife* 4:e10032.
296. Rompani SB, *et al.* (2017) Different Modes of Visual Integration in the Lateral Geniculate Nucleus Revealed by Single-Cell-Initiated Transsynaptic Tracing. *Neuron* 93(4):767-776 e766.
297. Schwarz LA, *et al.* (2015) Viral-genetic tracing of the input-output organization of a central noradrenaline circuit. *Nature* 524(7563):88-92.
298. Wall NR, Wickersham IR, Cetin A, De La Parra M, & Callaway EM (2010) Monosynaptic circuit tracing in vivo through Cre-dependent targeting and complementation of modified rabies virus. *Proceedings of the National Academy of Sciences of the United States of America* 107(50):21848-21853.
299. Wertz A, *et al.* (2015) PRESYNAPTIC NETWORKS. Single-cell-initiated monosynaptic tracing reveals layer-specific cortical network modules. *Science* 349(6243):70-74.
300. Osakada F & Callaway EM (2013) Design and generation of recombinant rabies virus vectors. *Nature protocols* 8(8):1583-1601.
301. Callaway EM & Luo L (2015) Monosynaptic Circuit Tracing with Glycoprotein-Deleted Rabies Viruses. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 35(24):8979-8985.
302. Reardon TR, *et al.* (2016) Rabies Virus CVS-N2c(DeltaG) Strain Enhances Retrograde Synaptic Transfer and Neuronal Viability. *Neuron* 89(4):711-724.
303. Ciabatti E, Gonzalez-Rueda A, Mariotti L, Morgese F, & Tripodi M (2017) Life-Long Genetic and Functional Access to Neural Circuits Using Self-Inactivating Rabies Virus. *Cell* 170(2):382-392 e314.
304. Beier KT, *et al.* (2011) Anterograde or retrograde transsynaptic labeling of CNS neurons with vesicular stomatitis virus vectors. *Proceedings of the National Academy of Sciences of the United States of America* 108(37):15414-15419.
305. Lo L & Anderson DJ (2011) A Cre-dependent, anterograde transsynaptic viral tracer for mapping output pathways of genetically marked neurons. *Neuron* 72(6):938-950.
306. Krebs HA (1975) The August Krogh Principle: "For many problems there is an animal on which it can be most conveniently studied". *The Journal of experimental zoology* 194(1):221-226.
307. Mundell NA, *et al.* (2015) Vesicular stomatitis virus enables gene transfer and transsynaptic tracing in a wide range of organisms. *The Journal of comparative neurology* 523(11):1639-1663.
308. Drouin LM, *et al.* (2016) Cryo-electron Microscopy Reconstruction and Stability Studies of the Wild Type and the R432A Variant of Adeno-associated Virus Type 2 Reveal that Capsid Structural Stability Is a Major Factor in Genome Packaging. *Journal of virology* 90(19):8542-8551.
309. Platt RJ, *et al.* (2014) CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* 159(2):440-455.
310. Acland GM, *et al.* (2001) Gene therapy restores vision in a canine model of childhood blindness. *Nature genetics* 28(1):92-95.
311. Kaplitt MG, *et al.* (2007) Safety and tolerability of gene therapy with an adeno-associated virus (AAV) borne GAD gene for Parkinson's disease: an open label, phase I trial. *Lancet* 369(9579):2097-2105.
312. MacLaren RE, *et al.* (2014) Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. *Lancet* 383(9923):1129-1137.
313. Passini MA, *et al.* (2010) CNS-targeted gene therapy improves survival and motor function in a mouse model of spinal muscular atrophy. *The Journal of clinical investigation* 120(4):1253-1264.

314. Pasca AM*, et al.* (2015) Functional cortical neurons and astrocytes from human pluripotent stem cells in 3D culture. *Nature methods* 12(7):671-678.
315. Quadrato G*, et al.* (2017) Cell diversity and network dynamics in photosensitive human brain organoids. *Nature* 545(7652):48-53.
316. Kaplitt MG*, et al.* (1994) Long-term gene expression and phenotypic correction using adeno-associated virus vectors in the mammalian brain. *Nature genetics* 8(2):148-154.
317. Nath RD*, et al.* (2017) The Jellyfish Cassiopea Exhibits a Sleep-like State. *Current biology : CB* 27(19):2984-2990 e2983.
318. Hendricks JC*, et al.* (2000) Rest in Drosophila is a sleep-like state. *Neuron* 25(1):129-138.
319. Shaw PJ, Cirelli C, Greenspan RJ, & Tononi G (2000) Correlates of sleep and waking in Drosophila melanogaster. *Science* 287(5459):1834-1837.
320. Hill AJ, Mansfield R, Lopez JM, Raizen DM, & Van Buskirk C (2014) Cellular stress induces a protective sleep-like state in C. elegans. *Current biology : CB* 24(20):2399-2405.
321. Raizen DM*, et al.* (2008) Lethargus is a Caenorhabditis elegans sleep-like state. *Nature* 451(7178):569-572.
322. Trojanowski NF & Raizen DM (2016) Call it Worm Sleep. *Trends Neurosci* 39(2):54-62.
323. Allada R & Siegel JM (2008) Unearthing the phylogenetic roots of sleep. *Current biology : CB* 18(15):R670-R679.
324. Joiner WJ (2016) Unraveling the Evolutionary Determinants of Sleep. *Current biology : CB* 26(20):R1073-R1087.
325. Kirszenblat L & van Swinderen B (2015) The Yin and Yang of Sleep and Attention. *Trends Neurosci* 38(12):776-786.
326. Arendt D, Tosches MA, & Marlow H (2016) From nerve net to nerve ring, nerve cord and brain--evolution of the nervous system. *Nature reviews. Neuroscience* 17(1):61-72.
327. Bosch TCG*, et al.* (2017) Back to the Basics: Cnidarians Start to Fire. *Trends in Neurosciences* 40(2):92-105.
328. Dunn CW, Giribet G, Edgecombe GD, & Hejnol A (2014) Animal Phylogeny and Its Evolutionary Implications. *Annu Rev Ecol Evol S* 45:371-+.
329. Elphick MR, Mirabeau O, & Larhammar D (2018) Evolution of neuropeptide signalling systems. *J Exp Biol* 221(Pt 3).
330. Hejnol A & Rentzsch F (2015) Neural nets. *Current Biology* 25(18):R782-R786.
331. Katsuki T & Greenspan RJ (2013) Jellyfish nervous systems. *Current Biology* 23(14):R592-R594.
332. Kelava I, Rentzsch F, & Technau U (2015) Evolution of eumetazoan nervous systems: insights from cnidarians. *Philos T R Soc B* 370(1684).
333. Grimmelikhuijzen CJ & Westfall JA (1995) The nervous systems of cnidarians. *EXS* 72:7-24.
334. Satterlie RA (2011) Do jellyfish have central nervous systems? *Journal of Experimental Biology* 214(8):1215-1223.
335. Dupre C & Yuste R (2017) Non-overlapping Neural Networks in Hydra vulgaris. *Current biology : CB* 27(8):1085-1097.
336. Grimmelikhuijzen CJP, Williamson M, & Hansen GN (2002) Neuropeptides in cnidarians. *Can J Zool* 80(10):1690-1702.
337. Watanabe H, Fujisawa T, & Holstein TW (2009) Cnidarians and the evolutionary origin of the nervous system. *Dev Growth Differ* 51(3):167-183.
338. Kremien M, Shavit U, Mass T, & Genin A (2013) Benefit of pulsation in soft corals. *Proceedings of the National Academy of Sciences of the United States of America* 110(22):8978-8983.
339. Garm A, Bielecki J, Petie R, & Nilsson DE (2012) Opposite Patterns of Diurnal Activity in the Box Jellyfish Tripedalia cystophora and Copula sivickisi. *Biol Bull-Us* 222(1):35-45.
340. Seymour JE, Carrette TJ, & Sutherland PA (2004) Do box jellyfish sleep at night? *Med J Aust* 181(11-12):707.

341. Campbell SS & Tobler I (1984) Animal sleep: a review of sleep duration across phylogeny. *Neurosci Biobehav Rev* 8(3):269-300.
342. Holland BS, Dawson MN, Crow GL, & Hofmann DK (2004) Global phylogeography of Cassiopea (Scyphozoa : Rhizostomeae): molecular evidence for cryptic species and multiple invasions of the Hawaiian Islands. *Mar Biol* 145(6):1119-1128.
343. Jantzen CW, C.; Rasheed, M.; El-Zibdah, M.; Richter, C. (2010) Enhanced pore water nutrient fluxes by the upside-down jellyfish Cassiopea sp. in a Red Sea coral reef. *Mar Ecol Prog Ser* 411:117–125.
344. Santhanakrishnan A, Dollinger M, Hamlet CL, Colin SP, & Miller LA (2012) Flow structure and transport characteristics of feeding and exchange currents generated by upside-down Cassiopea jellyfish. *Journal of Experimental Biology* 215(14):2369-2381.
345. Zhdanova IV (2011) Sleep and its regulation in zebrafish. *Rev Neurosci* 22(1):27-36.
346. Borbely AA & Achermann P (1999) Sleep homeostasis and models of sleep regulation. *J Biol Rhythms* 14(6):557-568.
347. Zimmerman JE, Naidoo N, Raizen DM, & Pack AI (2008) Conservation of sleep: insights from non-mammalian model systems. *Trends Neurosci* 31(7):371-376.
348. Nath RD, Chow ES, Wang H, Schwarz EM, & Sternberg PW (2016) C. elegans Stress-Induced Sleep Emerges from the Collective Action of Multiple Neuropeptides. *Current biology : CB* 26(18):2446-2455.
349. Erwin DH & Davidson EH (2002) The last common bilaterian ancestor. *Development* 129(13):3021-3032.
350. Peres R*, et al.* (2014) Developmental and light-entrained expression of melatonin and its relationship to the circadian clock in the sea anemone Nematostella vectensis. *Evodevo* 5:26.
351. Zhdanova IV, Wang SY, Leclair OU, & Danilova NP (2001) Melatonin promotes sleep-like state in zebrafish. *Brain research* 903(1-2):263-268.
352. Brzezinski A*, et al.* (2005) Effects of exogenous melatonin on sleep: a meta-analysis. *Sleep Med Rev* 9(1):41-50.
353. Tosches MA, Bucher D, Vopalensky P, & Arendt D (2014) Melatonin signaling controls circadian swimming behavior in marine zooplankton. *Cell* 159(1):46-57.
354. Gandhi AV, Mosser EA, Oikonomou G, & Prober DA (2015) Melatonin Is Required for the Circadian Regulation of Sleep. *Neuron* 85(6):1193-1199.
355. Stat MP, X.; Cowie, R.; Gates, R.D. (2010) Specificity in communities of Symbiodinium in corals from Johnston Atoll. *Mar. Ecol. Prog. Ser.* 386:83–9.
356. Folmer O, Black M, Hoeh W, Lutz R, & Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol* 3(5):294-299.
357. van der Walt S, Colbert SC, & Varoquaux G (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Comput Sci Eng* 13(2):22-30.
358. Hunter JD (2007) Matplotlib: A 2D graphics environment. *Comput Sci Eng* 9(3):90-95.
359. Oliphant TE (2007) Python for scientific computing. *Comput Sci Eng* 9(3):10-20.
360. Abrams MJ, Basinger T, Yuan W, Guo CL, & Goentoro L (2015) Self-repairing symmetry in jellyfish through mechanically driven reorganization. *Proceedings of the National Academy of Sciences of the United States of America* 112(26):E3365-E3373.
361. Treweek JB*, et al.* (2015) Whole-body tissue stabilization and selective extractions via tissue-hydrogel hybrids for high-resolution intact circuit mapping and phenotyping. *Nat Protoc* 10(11):1860-1896.