An arginine-rich nuclear localization signal (ArgiNLS) strategy for streamlined image segmentation of single-cells

Eric R. Szelenyi[1,2]*, Jovana S. Navarrete[1,2,3], Alexandria D. Murry[1,2], Yizhe Zhang[1,2], Kasey S. Girven[4], Lauren Kuo[1,5,9], Marcella M. Cline[1,7,10], Mollie X. Bernstein[1,7], Mariia Burdyniuk[8], Bryce Bowler[2], Nastacia L. Goodwin[1,2,3], Barbara Juarez[1,6,7,11], Larry S. Zweifel[1,6,7], and Sam A. Golden[1,2]*

[1]University of Washington Center of Excellence in Neurobiology of Addiction, Pain, and Emotion (NAPE), Seattle, WA, USA
[2]University of Washington, Department of Biological Structure, Seattle, WA, USA
[3]University of Washington, Graduate Program in Neuroscience, Seattle, WA, USA
[4]University of Washington, Department of Anesthesiology and Pain Medicine
[5]University of Washington Undergraduate Program in Biochemistry
[6]University of Washington, Department of Psychiatry and Behavioral Sciences, Seattle, WA, USA
[7]University of Washington, Department of Pharmacology, Seattle, WA, USA
[8]Carl Zeiss Microscopy LLC, White Plains, NY, USA

Current affiliation:
[9]Allen Institute for Cell Science, Seattle, WA, USA
[10]Cajal Neuroscience, Seattle, WA, USA
[11]University of Maryland School of Medicine, Department of Neurobiology, Baltimore, MD, USA

Co-corresponding authors: *Eric R Szelenyi; *Sam A. Golden

Email: szelenyi@uw.edu; sagolden@uw.edu

Author Contributions: E.R.S/S.A.G conceived of the project and E.R.S developed the genetic tag strategy. E.R.S performed all stereotaxic injections and confocal microscopy. E.R.S/L.K performed *in vitro* experiments and molecular cloning. Y.Z/B.B performed histology. E.R.S/J.N performed retro-orbital injections for all experiments. E.R.S/J.N performed epifluorescent imaging. J.N implemented QuPath for 2D cell segmentation testing. J.N/Y.Z/L.K performed manual annotations and F1 scoring for 2D cell segmentation. A.D.M performed iDISCO+ brain clearing and staining and LSFM imaging. E.R.S/A.D.M implemented ClearMap volumetric image processing pipeline. M.C.M/M.B performed behavioral battery experiments. B.J/K.S.G performed electrophysiology recordings and analysis. E.R.S performed all analyses. E.R.S/S.A.G wrote the manuscript.

Competing Interest Statement: E.R.S. and S.A.G. are inventors of a provisional patent related to ArgiNLS genetic technology (W149-0052USP1/49409.01US1). All other authors have no competing interests.

Classification: Biological Sciences/Neuroscience

Keywords: NLS; segmentation; volumetric; whole-brain; single-cell

`

Abstract

High-throughput volumetric fluorescent microscopy pipelines can spatially integrate whole-brain structure and function at the foundational level of single-cells. However, conventional fluorescent protein (FP) modifications used to discriminate single-cells possess limited efficacy or are detrimental to cellular health. Here, we introduce a synthetic and non-deleterious nuclear localization signal (NLS) tag strategy, called 'Arginine-rich NLS' (ArgiNLS), that optimizes genetic labeling and downstream image segmentation of single-cells by restricting FP localization near-exclusively in the nucleus through a poly-arginine mechanism. A single N-terminal ArgiNLS tag provides modular nuclear restriction consistently across spectrally separate FP variants. ArgiNLS performance in vivo displays functional conservation across major cortical cell classes, and in response to both local and systemic brain wide AAV administration. Crucially, the high signal-to-noise ratio afforded by ArgiNLS enhances ML-automated segmentation of single-cells due to rapid classifier training and enrichment of labeled cell detection within 2D brain sections or 3D volumetric whole-brain image datasets, derived from both staining-amplified and native signal. This genetic strategy provides a simple and flexible basis for precise image segmentation of genetically labeled single-cells at scale and paired with behavioral procedures.

Significance Statement

Quantifying labeled cells in fluorescent microscopy is a fundamental aspect of a modern biology. Critically, the use of short nuclear localization sequences (NLS) is a key genetic modification for discriminating single-cells labeled with fluorescent proteins (FPs). However, mainstay NLS approaches typically localize proteins to the nucleus with limited efficacy, while alternative non-NLS tag strategies can enhance efficacy at the cost of cellular health. Thus, quantitative cell counting using FP labels remains suboptimal or not compatible with health and behavior. Here, we present a novel genetic tagging strategy – named ArgiNLS – that flexibly and safely achieves FP nuclear restriction across the brain to facilitate machine learning-based segmentation of single-cells at scale, delivering a timely update to the behavioral neuroscientist's toolkit.

Main Text

Introduction

The structure of the mammalian brain is complex and spans multiple spatial scales, requiring the development and application of spatially integrated methods to reveal key relationships with its function. High-throughput volumetric fluorescent microscopy pipelines, facilitated by light-sheet fluorescence microscopy (LSFM)[1] and serial section block-face microscopy[2,3], have accelerated these efforts. Notably, they have democratized experimental access to spatially integrated intact whole-brain anatomy, at the resolution of single-cells and within computationally feasible datasets that are compatible with brain atlas mapping. For example, in combination with genetic fluorescent labeling, these techniques have facilitated the mapping of cellular activity[4–16], cell-type distributions[17–20], and endogenous genetic[21] and connectivity patterns[22].

However, while significant hardware-based advancements have been made in microscopy systems[2,23,24], and numerous software-based image analysis tools are readily available[5,13,14,25–29], there has been a relative lack of genetic tool development that maximizes single-cell detection and quantification through the optimization of fluorescent signal. As opposed to hardware- and software-based approaches, genetic approaches may offer improved cellular detection without the prohibitively expensive acquisition of cutting-edge microscopy and computational equipment. Past genetic-based approaches have relied on fluorescent protein (FP)

expression in its unmodified or modified forms[30], but the efficacy of these prior efforts have not been systematically examined at the level of single-cell discrimination across image datasets of various sizes. As a result, the quantitative accuracy of single-cell brain mapping measurements remains ambiguous both within and between studies.

Typically, single-cell discrimination is improved by targeted genetic modifications to FPs by (*i*) N- or C-terminal tags of a minimal nuclear localization signal (NLS) sequence from a naturally occurring protein like the simian virus 40 large T antigen (SV40nls)[31–33], or (*ii*) fusion with a full-length nuclear protein, like histone 2B[17,22,34]. Both approaches target FP localization to the cell nucleus to reduce extra-nuclear fluorescence. However, these approaches have limitations. Minimal NLS tag modifications only achieve partial nuclear localization, and therefore are sub-optimal for image segmentation. Conversely, full-length nuclear protein fusions guarantee nuclear restriction, but depend on overexpression of the fused nuclear gene, which can have significant deleterious effects on neuronal function and animal health[35].

Two major bottlenecks have slowed the development of new ectopic NLSs - including those identified in silico[36] or synthetic versions developed for genome-editing[37,38] – for use as FP fusion tags. First, NLS performance when ectopically tagged as an FP fusion is unpredictable. This is because the nuclear import efficiency of naturally occurring NLSs are influenced by many factors, including intrinsic NLS sequence affinity to nuclear transport receptors KPNA and KPNB[39], and their interplay with the amino acids surrounding the NLS in its native state[40,41]. Second, due to the barrier posed by challenging biological properties, there is not a high-throughput pipeline to systematically screen novel NLSs for whole brain mapping compatibility. Those properties include: (*i*) minimal extra-nuclear FP localization for high SNR, (*ii*) a simple and flexible sequence with minimal length and a single location of fused tag for unaltered FP function, (*iii*) modular compatibility with FP variants for spectral flexibility in imaging design, including multiplexing, (*iv*) equal functionality across cell-types for unbiased labeling of single-cells, and (*v*) non-deleterious effects of cellular and animal health. Thus, a genetic tag strategy that simultaneously synergizes and stereotypes FP-tagged NLS function while maintaining cellular health could satisfy these requirements.

Here, we present a new NLS genetic tag strategy named "Arginine-rich NLS" or "Argi-NLS", that improves nuclear localization of genetically encoded FPs and enhances image segmentation of single-cells through a poly-arginine driven nucleolar enrichment mechanism. This method does not alter neural physiology or mouse behavior after standard AAV expression windows and is therefore compatible with AAV-paired behavioral procedures. It is functionally modular when using spectrally separate FPs and possesses equal efficacy when expressed within genetically distinct cortical cell-types. Moreover, it significantly improves two- and three-dimensional image segmentation that decreases computational expense associated with ML-based segmentation.

Results

Design of a nuclear-restricting Arginine-rich NLS (ArgiNLS)

Arginine (R) is the most hydrophilic, basic natural amino acid residue with a pKa over 12, and possession of a positive charge within all biological environments. Because of these properties, synthetic polyarginine (poly-R) tags (i.e., a peptide sequence consisting of multiple, R repeats) serve a variety of experimental functions, including their use as a protein purification tag[42–45], cell-penetrating peptide signal, multi-DNA-binding domain linker, and nucleolar trafficking signal[42,46–48]. We hypothesized that a poly-R stretch coded in-between a NLS sequence and a downstream FP could act as an improved nuclear-localizing agent within neurons (**Fig. 1a**).

To explore this idea in vitro, we first tested the effects of a range of poly-R lengths on the nuclear localization of two known classical NLS (cNLS) types: (*i*) monopartite (MP), composed of a single cluster of 4-8 basic amino acids and consensus recognition motif of K (K/R) X (K/R) (X is any amino acid), and (*ii*) bipartite (BP), composed of two clusters of 2-3 positively charged amino acids, separated by a 9-12 amino acid linker region, and with a consensus sequence R/K(X)10-

12KRXX[39,49] (**Fig. 1b; Table S1**). We selected cNLS sequences derived from the widely applied and characterized MP SV40[50,51] or BP nucleoplasmin2 (NPM2)[52] genes, whose NLS protein sequences differ in length and net positive charge but share the same isoelectric point (pI) across poly-R length additions (**Fig. 1c-d**). We performed co-transfections of AAV-(SV40nls or NPM2nls)-poly-R-EGFP overexpression constructs with untagged mKate2 in mouse neuroblastoma 2A (N2A) cells. Nuclear or cellular fluorescence localization and brightness (see **Methods** for a detailed description quantification approach) in individual EGFP+/mKate2+ cells were measured using DAPI staining as a nuclear marker and untagged mKate2 expression to label entire cells.

Co-transfected EGFP+/mKate+ cells revealed a poly-R length-dependent increase in nucleolar localization for both cNLSs. NPM2nls, and not SV40nls, exhibited decreased extra-nuclear fluorescence regardless of poly-R length (**Fig. 1e**). Measured composite values across poly-R lengths was 77% (± 1.48) for SV40nls and 91% (± 0.85) for NPM2nls (data not shown), where NPM2nls significantly obtained stable, higher nuclear localization across all poly-R lengths (**Fig. 1f**). Additionally, unlike the linear decrease observed with SV40nls, nuclear brightness with NPM2nls peaked with poly-R7 and decayed with longer lengths (**Fig. 1g**).

We validated the optimal NPM2nls-R7 configuration in a separate heterologous mouse brain cell-line by comparing untagged performance for each cNLS type (**Fig. 1h-i**). The control SV40nls tag exhibited 63% of nuclear EGFP fluorescence that was significantly decreased to 55% with the poly-R7 addition, despite a nucleolar-like localization (**Fig. 1h-i**). NPM2nls-R7 enhanced nuclear localization above untagged levels from 70 to 95% (**Fig. 1i**). We next probed the specificity of poly-R7 on BP cNLS activity using a BP version of SV40nls (biSV40nls, **Fig. S1**). We observed a significant enhancement of nuclear-localized EGFP fluorescence with biSV40nls-R7 (93% ± 1.5) versus untagged biSV40nls (70% ± 7.3) (**Fig. S1b-c**).

Prior work has shown that a synthetic poly-R8 tag facilitates DNA-binding protein function when inserted as a linker sequence[47], suggesting that poly-R's can be used as a linker in other molecular contexts. Next, we tested the ability of the NPM2nls-R7 genetic tag – now termed ArgiNLS – to provide modularity with other FPs beyond EGFP (**Fig. 1j-l**). We attached ArgiNLS to a range of spectrally separate FPs that have diverging N-terminal sequences located within the first 10 amino acids, which could influence fused NLS tag activity (**Fig. 1j)**. Monomeric FPs included EGFP (Ex/Em: 488/507), oScarlet[53] (Ex/Em: 569/594), mKate2[54] (Ex/Em: 588/633), and miRFP670[55] (Ex/Em: 647/670) (**Fig. 1j; bottom**). In N2A cells, we measured nuclear restriction of native fluorescence for EGFP (93% ± 0.3), oScarlet (93% ± 1.0), mKate2 (93% ± 1.5), and miRFP670 (91% ± 0.3) (**Fig. 1k-l**).

Nuclear restriction of FP fluorescence *in vivo*

The *in vitro* results show enhanced ArgiNLS restriction of nuclear fluorescence. We next determined if the performance replicated *in vivo* using AAV overexpression in the mouse brain. Nuclear import of NLS-containing proteins – including SV40 and NPM2 - are controlled by nuclear transport receptors called importins (or karyopherins)[49]. Importins are dynamically expressed in multiple isoforms across cell-types and brain regions[56], providing a potential cause of variable NLS activity of NLSs in a cell-type and brain region-dependent manner. We used a conditional AAV co-expression strategy to quantify nuclear-localized EGFP fluorescence in two genetically distinct cortical cell classes, GABAergic or glutamatergic neurons. We reasoned that each class comprises a divergent set of multiple neuronal cell sub-types that have functionally and genetically distinct characteristics[20,57], and therefore are well suited for benchmarking the nuclear localizing efficacy of ArgiNLS- or conventional SV40nls-tagged EGFP.

We selected the barrel cortex – where GABAergic and glutamatergic cell-types are abundant[20,57] – for AAV co-injections using SV40nls or ArgiNLS tagged EGFP (AAV-EF1a-SV40nls- *or* ArgiNLS-EGFP), combined with a conditional Cre-dependent virus to express somatic tdTomato (AAV-CAG-FLEX-tdTomato) in GABAergic or glutamatergic cells within *vGat-Cre* or *vGlut1-Cre* animals, respectively (**Fig. 2a**). Confocal images were acquired of the injection site, and nuclear (as defined by DAPI+ co-stain) and somatic (as defined by tdTomato+ somatic

fluorescence) EGFP fluorescence was quantified at the level of single-cells across NLS tags and cell-types at both high (**Fig. 2b-c)** and low (**Fig. S2**) magnification.

Barrel cortex neurons infected with SV40nls-EGFP exhibited somatic EGFP concomitant with high extracellular background, putatively from SV40nls-EGFP localized in cellular processes (**Fig. 2b-c; left**). Conversely, ArgiNLS-EGFP did not exhibit extra-cellular EGFP background, which corresponded with nuclear-restricted EGFP and localization in presumptive nucleolar puncta (**Fig. 2b-c; right**). Sampled EGFP+/tdTomato+ single-cells used for nuclear/somatic EGFP fluorescence quantifications contained equal distributions of nuclear localization (corrected total nuclear fluorescence, CTNF; see **Methods**) levels across tags and cell-type (**Fig. 2d**). However, when CTNF was expressed as a percent cellular fluorescence (corrected total cellular fluorescence, CTCF; see **Methods**), nuclear localization variance was significantly different across tags; SV40nls showed high variance while ArgiNLS showed low variance (**Fig. 2e**). Despite differences in variance, tag type had a significant effect on nuclear localization, where ArgiNLS-EGFP contained significantly higher nuclear localization (GABAergic cells: 88% ± 0.9; glutamatergic cells: 87% ± 0.7) than SV40nls-EGFP (GABAergic cells: 65% ± 1.8; glutamatergic cells: 67% ± 2.0) (**Fig. 2f-h**).

We next determined how ArgiNLS impacts extracellular (EC) background fluorescence at the site of injection and within anatomically distal areas (**Fig. 2i-l**). First, we compared EC pixel intensity measurements taken from field of views (FOVs) positioned in-between EGFP-labeled cells of the injected barrel cortex (i.e., ipsilateral), and in-between DAPI+ cells of the uninjected (i.e., contralateral) barrel cortex, as a control measurement, from the same mice (**Fig. 2i**). Ipsilateral SV40nls FOVs contained a significant 4-fold increase in EC fluorescence compared to uninjected contralateral FOVs, that was also significantly higher than between-subject ipsilateral and contralateral ArgiNLS FOVs. EC fluorescence within ipsilateral ArgiNLS FOVs was unchanged from contralateral FOV levels from both SV40nls and ArgiNLS injected samples.

Neurons within the barrel cortex send diverse axonal projections throughout the brain[58]. ArgiNLS consistently eliminated non-nuclear labeling of axon fiber of passage in the corpus callosum (CC), as well as axonal terminations in the dorsolateral thalamus (dlTH), that was observed with SV40nls-EGFP injections (**Fig. 2j**). We then confirmed decreased CC axon labeling by ArgiNLS in separate quantifications made of EGFP sampled in high-magnification FOVs. To control for error in sample-to-sample variability of viral delivery, we restricted our analysis within the *vGlut1-Cre* injected sections and normalized EGFP+ fiber area to *vGlut1*:tdTomato+ fiber area (**Fig. 2k-l**). Accordingly, SV40nls-EGFP labeled 115% (± 13) of *vGlut1*:tdTomato+ fibers, indicating that extra-nuclear EGFP localization produced from constitutively expressed SV40nls-EGFP labels CC fibers within both *vGlut1(*+) and *vGlut1(-)* neurons of the barrel cortex (**Fig. 2l**). Conversely, ArgiNLS-EGFP labeled 3% (± 1.2) of *vGlut1*:tdTomato+ fibers.

Non-deleterious to neuronal and behavioral health

Full-length nuclear protein-FP fusions, such as H2B-EGFP[59], demonstrate high levels of nuclear-restricted fluorescence, but cause detrimental effects to cellular and animal health when overexpressed in the mouse brain[35]. Since ArgiNLS tagged FPs display similar nuclear-restricted fluorescence, we next determined if ArgiNLS-EGFP viral overexpression in the brain causes detrimental effects to baseline neuronal physiology or animal behavior (despite it lacking the fusion of an endogenous nuclear protein) (**Fig. 3**). We first examined baseline physiological properties of glutamatergic neurons within the barrel cortex of *vGlut1-Cre* animals. We used a local unilateral injection of Cre-dependent tdTomato virus to create control *vGlut1-Cre*:tdTomato(+) labeled neurons, and a contralateral co-injection of Cre-dependent tdTomato with ArgiNLS-EGFP viruses to create within-sample, experimental ArgiNLS-EGFP(+)/*vGlut1-Cre*:tdTomato(+) labeled neurons (**Fig. 3a**). Analysis of baseline physiological parameters indicated no effect of ArgiNLS-EGFP overexpression on resting membrane potential (**Fig. 3b**) and action potential generation across a range of injected currents (**Fig. 3c-d**).

Next, we used the PHP.eB viral serotype to examine the effect of systemic whole-brain ArgiNLS-EGFP or control virus expression – including a non-expressing negative Cre-dependent

tdTomato control virus, and positive SV40nls-EGFP and H2B-EGFP expressing controls - on behaviors including the open field task (OFT) and elevated plus maze (EPM) to study effects on acute exploratory drive, and a 60-hour locomotion monitoring assay to study chronic locomotion and sensorimotor performance across dark/light cycles (**Fig. 3g-m**). We found no differences in acute OFT performance across viral groups as measured by mean total distance (cm) traveled (**Fig. 3g**), or percent center time duration (**Fig. 3h**). Additionally, acute EPM performance as measured by open arm entry frequency (**Fig. 3i**), or open arm total duration (sec) (**Fig. 3j**) was the same across viral groups. Consecutive monitoring of homecage distance traveled (cm) across 3 nights and 2 days indicated a lack of effect from ArgiNLS-EGFP or SV40nls-EGFP systemic overexpression (**Fig**. **3k-m**). Conversely, H2B-EGFP-injected animals displayed dysregulated locomotor activity during the light phase. We confirmed these behavioral data using 2D histology and imaging of native fluorescence on a subset of animals (**Fig. S3**).

Enhanced compatibility with machine learning classification

Image classification of objects can be reliably automated and scaled through the creation of machine learning (ML) algorithms or "classifiers" that rapidly extract features based on pixel content to create image segmentations for quantification[60,61]. For fluorescently labeled images, the manual effort taken to compile training datasets is directly tied to the fluorescent signal's fidelity to label a target class. Thus, poorly defined labeling may require extensive training and user annotation. Since ArgiNLS enables consistent FP restriction in the nucleus of neurons, we next asked if this influences the creation – or performance - of a single-cell ML classifier, as compared to a classifier trained with cSV40nls tag labeling (**Fig. 4**).

We systemically labeled single-cells with SV40nls-EGFP or ArgiNLS-EGFP-expressing AAV.eB viruses through RO injections (**Fig. 4a**) and used DAPI counterstained images of coronal sections (2.11 and -0.52 from bregma) to benchmark single-cell ML classifier training and testing at varying anatomical locations throughout the brain (**Fig. 4b**). We observed that SV40nls displayed variable extra-nuclear fluorescence and background, and ArgiNLS conversely showed nuclear-restriction and an overall higher density of labeled single-cells (**Fig. 4c-d**).

We next created single-cell classifiers for SV40nls and ArgiNLS labels and benchmarked their performance using F1 scores: a machine learning metric that determines model accuracy based on a combination of precision (i.e., positive predictions) and recall (i.e., correct identification of positive class) measurements. We performed multiple training iterations consisting of two sample FOVs per session until maximum classifier performance, as measured by F1 scoring in FOVs with similar ground truth (GT) cell detections, was reached (**Fig. 4e**; **Fig. S4**). ArgiNLS (5 rounds, 10 input FOV) resulted in a substantial reduction in the total number of training iterations required to reach max classifier performance at both anatomical planes compared to SV40nls (14 rounds, 28 input FOV) (**Fig. 4f-h**). Specifically, for AP 2.11, SV40nls classification started with a 93% greater total amount of cell detections than GT, that was composed of 11% correct (24 ± 3.1) and 89% (222 ± 3.8) incorrect detections. Both correct and incorrect detections displayed an exponential-like scaling towards GT levels as function of training round number (**Fig. 4f; left**). Initial ArgiNLS classification consisted of 63% more total cell detections (184 ± 7.5; Correct: 87 ± 7.0; Incorrect: 96 ± 13.9) than GT, with correct cells matching 90% of GT (**Fig. 4g; left**), and incorrect cells progressively reduced over each training session.

Total cell detections for both classifiers at AP -0.52 was stable across all training iterations and corresponded closely to GT detection levels with training (**Fig. 4f-g; right**). Like the raw data, ArgiNLS classification outperformed SV40nls classification in both precision (**Fig. 4h, top**) and recall (**Fig. 4h, middle**). ArgiNLS significantly outperformed SV40nls in F-sore across nearly all training tile input FOV iterations (**Fig. 4h; bottom**).

Enriched brain-wide segmentation of single-cells

We next sought to determine ArgiNLS tag's impact on single-cell segmentation within a whole-brain mapping dataset using large-scale ML classification. We systemically delivered PHP.eB SV40nls-EGFP and ArgiNLS-EGFP expressing viruses and performed iDISCO+ whole-

mount clearing and EGFP staining to fluorescently label all the infected cells throughout the brain (**Fig. 5a, top; 5b**). Whole-brain samples were imaged with light sheet fluorescent microscopy (LSFM), and image datasets were run through a pipeline consisting of reference atlas registration, ML cellular segmentation, atlas segmentation of brain regions, and voxelized cell density analyses (**Fig. 5a, bottom**). Like our observations in 2D histology, SV40nls-EGFP staining produced higher image artifact and less cells counts than ArgiNLS-EGFP (**Fig. 5b, left**). We therefore trained separate ML classifiers using ilastik software to achieve optimal single-cell classification performance, separately for each tag (**Fig. S5a**). ArgiNLS reduced the model training requirements in 3D, where SV40nls classifier training required a total dataset of 44 training images and ArgiNLS required 20 training images (**Fig. S5b**).

Following extensive classifier validation (**SI Text**), we applied our whole-brain image processing pipeline to create maps of systemic AAV.eB infection. We found that ArgiNLS-EGFP significantly increases AAV.eB labeled whole-brain cell counts by 2.6-fold (SV40nls: 514,422 ± 90,388; ArgiNLS: 1,362,466 ± 268,889) (**Fig. 5c**). Additionally, we asked if the whole-brain count difference was anatomically distributed preferentially or evenly across gross areas of the brain. Statistically significant increases of ArgiNLS-labeled cell densities were found specifically in the isocortex (ISO) (SV40nls: 5526 ± 1156, ArgiNLS: 1798 ± 277) and olfactory area (OLF) (SV40nls: 1162 ± 211, ArgiNLS: 3753 ± 609), suggesting these two brain areas are major drivers of the differences in whole-brain cell counts.

To identify subregional differences in labeled cell densities, we used 1216 anatomical annotations provided by the Unified atlas[63]. First, we asked if the general distribution of labeled cell patterning across subregions differ amongst genetic tags (**Fig. 5e**). Linear regression analysis revealed a statistically significant positive correlation of ArgiNLS versus SV40nls labeled cell densities of subregions, suggesting an equivalent anatomical distribution of labeled cells across subregions. However, with the exception of 30 subregions, 98% contained greater cell densities when labeled with ArgiNLS (**Fig. 5e-f**).

Second, we screened for the identities of subregions significantly enriched in ArgiNLS-labeled cell density (**Fig. 5f bottom; 5g; Fig. S6-7; Dataset S1**). The most enriched gross regions – the ISO and OLF - contained a compilation of 60 (208/345; e.g., primary somatosensory area (S1)) and 23% (16/69; e.g., anterior olfactory area (AO), piriform cortex (Pir)) significant subregions, respectively (**Fig. 5g;** Supplemental Data Fig. 6). We found that ISO cortical enrichment was predominant in layers 2/3 to 5, with heavy concentration in areas related to sensory processing (**Fig. S7**). We additionally found significant enrichment of ArgiNLS-labeled cell density in subregions broadly located in every major division of the brain (HPF: 13/82 = 16%, e.g., dorsolateral entorhinal cortex (DLEnt), subiculum (SUB); CTXsp: 1/23 = 4%, e.g., intermediate endopiriform claustrum (IEn); CNU: 4/120 = 3%, olfactory tubercle (Tu); CB: 1/65 = 1.5%, e.g., lateral cerebellar nucleus parvicellular part (LatPC); TH: 6/94 = 6%, e.g., medial geniculate nucleus (MG); HY: 1/104 = 1%, medial mammillary nucleus (MM); MB: 8/109 = 7%, e.g., inferior colliculus (IC); and HB: 2/198% = 2%, e.g., Kolliker-fuse nucleus (KF)) (**Fig. 5g; Fig. S7**). We additionally performed an orthogonal, voxel-by-voxel analysis of differences in brain-wide cell density. Like our region-by-region analysis, significantly greater voxels were visualized across all major brain areas confirming wide-spread enrichment of labeled cell density by ArgiNLS (**Fig. 5h; Movie S1**).

ArgiNLS-optimized applications in volumetric brain cell counting

Spatially integrated features of brain structure and function can be studied through volumetric imaging paired with genetically-encoded fluorescent labels. For example, many transgenic lines are available for, among other purposes, guiding FP labeling amongst a constellation of cell-types of choice[64]. Further, AAV technology is advancing rapidly as seen by recent serotype development directing FP labeling in retrograde, transsynaptic, or systemic fashions[65]. Since ArgiNLS optimizes single-cell FP labeling and segmentation, we established its utility for cell counting applications when incorporated into a variety of molecular genetic tools.

We first compiled an AAV vector toolkit consisting of multiple ArgiNLS-tagged FPs, including bright green- (AausFP1[66] and mGreenLantern[67]), yellow- (mVenus ME[68]) and red-emitting (oScarlet[53]) variants within constitutive and conditional formats to offer experimental range (**Table S2**). We applied aqueous-based SHIELD clearing[69] and axial swept LSFM imaging[70] to demonstrate ease of ArgiNLS-assisted segmentation of single-cells and their total count quantification using 3D ML from native fluorescence in 3D volumes. We demonstrated ArgiNLS-assisted quantification by generating and injecting a retrograde AAV-EF1a-ArgiNLS-mKate2 virus into the basomedial amygdala (BMA) – an amygdalar structure involved in anxiety and fear modulation[71] - to create an anatomical input map (**Fig. 6a-c**). Co-injected EGFP-expressing virus confirmed BMA injection placement, and native retro-ArgNLS-mKate2+ single-cells were readily identifiable at local and long-range distances from the injection site (**Fig. 6a**). 3D ML classification of brain-wide ArgiNLS-mKate2+ cells quantified at total of $2.4e^5$ BMA inputs that were mainly distributed continuously throughout the ipsilateral cortical subplate including amongst multiple amygdalar subregions (Amy) and claustrum (Cl), as well as within the subiculum of the hippocampus (**Fig. 6b**). Major long-range BMA input cells resided in 5 nodes including (from anterior to posterior) the accessory olfactory bulb (AOB), infralimbic cortex (IL), anterior paraventricular nucleus (aPVT) of the thalamus, contralateral amygdala, and ventral premammillary nucleus of the hypothalamus (PMv). The aPVT represented one of the largest long-range and singular thalamic node containing $5.6e^3$ input cells, highlighting the existence of a putative aPVT-BMA circuit motif of unknown behavioral function.

Second, we generated and locally injected AAV.eB-CAG-FLEX-ArgiNLS-EGFP and AAV.eB-CAG-fDIO-ArgiNLS-mKate2 virus into the barrel cortex of *vGat-Cre* or *vGat-Flp* driver mice, respectively, to conditionally label and 3D segment GABAergic single-cells (**Fig. 6d-e**). High SNR-labeled single-cells were identically observed in Cre- or Flp-dependent conditions in 2D (**Fig. 6d-e, left**). In 3D, a clear distribution of single-cell-labeled viral spread was observed and devoid of extranuclear signal artifact in both Cre- and Flp-dependent conditions (**Fig. 6d-e, middle**), with 3D ML segmenting $1.8e^4$ and $2.5e^4$ GABAergic cells, respectively (**Fig. 6d-e, right**).

Finally, we generated and systemically delivered an AAV.eB-EF1a-ArgiNLS-oScarlet virus to demonstrate robust native fluorescence labeling of single-cells and their ease of segmentation of large cell amounts ubiquitously throughout the brain (**Fig. 6f; Movie S2**). To gauge sensitivity of detection levels enabled by ArgiNLS, we systemically delivered 3 amounts – $1e^{10}$ (1x), $1e^{11}$ (10x), and $2e^{11}$ (20x) - of viral GC payloads. 3D renderings of native signal displayed sharply contrasted single-cells distributed widely throughout the brain, and scaled with viral GC amount (**Fig. 6f, left**). 3D ML quantification of single hemispheres segmented $1.3e^5$, $1.8e^6$ (14-fold), and $3.4e^6$ (26-fold) single-cells for $1e^{10}$, $1e^{11}$, and $2e^{11}$ GC payloads, respectively, or 1 cell labeled for every 58,173 viral GCs delivered (**Fig. 6f, left;** Supplemental **Data Fig. 8a, left**). Expressing fold change values over the lowest $1e^{10}$ GC values indicated the utility of ArgiNLS-oScarlet to precisely label and segment 1.3-fold cell count increases for every 1-fold increase in viral GC delivered, within the 1 to 20-fold (i.e.,$1e^{10}$- $2e^{11}$) GC range tested (**Fig. S8a, right**). This total single-cell detection by viral GC relationship was maintained evenly throughout the brain as observed in cell counts across separate horizontal planes located in dorsal, central, and ventral anatomical positions (**Fig. 6f, right; Fig. S8b**).

Discussion

Volumetric imaging methods can be used to unveil spatially integrated representations – or maps – of brain structure and function at the direct level of single-cells. Pivotal to this approach is the ability of genetically-encoded fluorescent labels to discriminate single-cells within dense and complex brain tissue. Here, we developed and characterized a new nuclear-localizing genetic tag strategy named ArgiNLS that optimizes FP labeling of single-cells, without off-target effects on cellular physiology or animal behavior. By extension, ArgiNLS provides seamless compatibility and enrichment of ML-automated brain cell segmentation across small-to-large scale image datasets of various applications, demonstrating its use as an improved single-cell resolution brain mapping genetic tool.

Unlike other recently developed synthetic NLS approaches created for enhanced genome-editing[37,38,72], the ArgiNLS approach for image segmentation is based on a well-understood molecular foundation. This approach uses a poly-R-based rerouting mechanism to reduce extra-nuclear NLS-tagged FP localization by attracting it into the nucleolus. Optimal tag efficacy of ArgiNLS occurs when the BP cNLS sequence is combined with a specific poly-R length (R7) that titrates FP localization into a balanced nucleoplasmic/nucleolar composition with optimal native brightness. MP cNLS combinations fail to reduce extra-nuclear artifact in combinations up to 11 Rs, suggesting that the function of R7 synergism on ectopically-fused NLS activity requires sequence and/or length characteristics of a BP NLS. The function of this configuration within cells outside of the CNS and amongst varied species was not determined in our current study, and therefore context-specific modification may be required. However, as demonstrated in mouse brain here, the ArgiNLS strategy lays a robust framework for tailoring function through the parameterization of BP NLS sequence selection and/or poly-R lengths.

The nature of the poly-R7 component of ArgiNLS additionally provides functional fidelity across a wide variety of FPs, that can provide spectral flexibility for single-cell labeling, and potential nuclear localization of other non-FP proteins, through its dual role as a protein linker. This capability was demonstrated on an initial group of spectrally separate FPs. However, further validation amongst a larger FP selection and with non-FP proteins for applications outside of image segmentation is necessary to strengthen the generalization of this linker-like modularity. Additionally, while our current ArgiNLS strategy applies N terminus tags to FPs, we did not test linker-like capability or function when attached to the C terminus, thereby limiting its current principles of use to a single fusion location.

NLS performance of ArgiNLS *in vivo* demonstrates several advantages over current SV40nls tag or H2B fusion single-cell labeling approaches. First, ArgiNLS causes FP nuclear restriction that is equally robust whether an AAV is delivered locally or systemically to the brain. ArgiNLS eliminates background fluorescence-producing, extra-nuclear signal that is consistently observed with SV40nls. ArgiNLS elicits consistent nuclear-restricted FP labelling across major cortical cell classes. This property enables improved ML image classification of intact whole brain tissue, which is characterized by a reduction in model training requirements and coupled to increased precision and recall of segmentations. Unlike H2B fusion[35], these benefits were not associated with deleterious effects to neural physiology or animal behavior. Therefore, ArgiNLS is appropriate for genetic labeling experiments of single-cells that require concomitant behavioral procedures paired with AAV infection. Future time-course studies of *in vivo* expression durations longer than studied here (i.e., 3-4 weeks), will help understand the total range of temporal compatibility with cellular and animal health.

The in vivo use of ArgiNLS highlighted how the selection of genetic labelling strategy can significantly influence whole-brain single-cell segmentations, with an effective doubling of cell counts across subregions located in all major brain structures. The exact nature of this improvement is unclear; however, it is possibly related to depletion of extra-nuclear FP localization that enhances single-cell SNR. Additionally, since a fraction of wild-type FP molecules misfold or do not form a proper chromophore[73,74], poly-R driven interaction with nucleolar ribosomal RNA[46] and/or nucleolus-dependent changes of pH could provide a stabilizing influence in aggregation or folding of ArgiNLS-tagged FPs, leading to more single-cells detected.

Finally, ArgiNLS can be applied in various volumetric cell labeling applications to resolve cell counts seamlessly and rapidly with 3D ML segmentation. Our proof-of-concept experiments

begin to demonstrate ArgiNLS compatibility with, and enhancement of, anatomical input mapping, conditional segmentation of cell-types, and mapping a wide range of titrated cell counts within 3D brain volumes using various AAV vectors. However, ArgiNLS may also be incorporated into other molecular genetic tools that will provide improvements to single-cell counting applications including intersectional mapping[74] with or without multiplexing of spectrally separate ArgiNLS-tagged FPs, transsynaptic anatomical tracing[65], activity-dependent reporting[75], and transgenic reporter mouse lines for genomically-encoded expression[64]. Indeed, while we demonstrate improved single-cell labeling capacity of ArgiNLS over other tags, our *in vivo* characterization was solely based off AAV expression. The creation of future ArgiNLS-tagged expression vectors for gene delivery from other commonly used viral vectors[76], as well as gene-targeting vectors for the creation of reporter mouse lines, will help expand optimized single-cell labeling and segmentation amongst broad genetic delivery options.

In conclusion, the simple yet powerful ArgiNLS genetic tag strategy described here advances streamlined quantification of single-cells from small-to-large scale brain image datasets with enriched labeling capacity and enhanced ML capabilities.

Materials and Methods

Mice

Adult male and female mice (8-12 weeks old) were used for all experiments unless otherwise noted. Animals were group housed on a reverse 12-hr light-dark cycle (0700 OFF/1900 ON) and had access to water and food ad libitum. All experimental procedures were performed in accordance with University of Washington IACUC guidelines. Wild-type and transgenic mice lines were purchased from the following vendors: C57Bl6/J (JAX, 000664); CD-1 (Charles River, Crl:CD1(ICR)); vGlut1-Cre (Slc17a7-IRES2-Cre-D; JAX, 023527); vGat-Cre (Vgat-ires-cre knock-in (C57BL/6J); JAX, 028862); vGat-Flp (Slc32a1-IRES2-FlpO-D knock-in; JAX, 031331).

Fluorescent protein-expressing AAV Vector construction

Untagged and NLS-tagged fluorescent proteins were cloned through PCR amplification using standard primers with or without extension overhangs (Integrated DNA Technologies). The monopartite SV40nls coding tag sequence used was from amino acids 126-132 of large T antigen (Betapolyomavirus macacae; NCBI Reference sequence: NP_043127.1). Amino acids 156-172 of the bipartite mouse nucleoplasmin2 NLS (NCBI accession number: NP_001398932.1) was used for NPM2nls. Isolectric point (pI) and net chrarge (z) calculations of each poly-R7 modified or unmodified cNLS protein tag was determined through the open source Prot pi online tool (http://protpi.ch).

The following Addgene plasmids were used for PCR templates in AAV vector construction: pAAV-CAG-NLS-GFP (104061) for all EGFP variants; pAAV2-CAG-3xNLS-AausFP1 (191096); pcDNA3.1-mGreenLantern (161912); pEB1-mVenus ME (103989); pAAV-EF1a-oScarlet (137135); pAAV-RAM-d2TTA::TRE-NLS-mKate2-WPREpA (84474) ; pmiRFP670-N1 (79987); and pCS-H2B-EGFP (53744) . For constitutive expression AAV construction, fluorescent proteins and variants were subcloned into pAAV-Ef1a-DIO-Synaptophysin-GCaMP6s (105715) at XhoI/EcoRV restriction enzyme sites. For Cre-dependent expression AAV construction, all fluorescent proteins and ArgiNLS-tagged variants were subcloned into pAAV-CAG-FLEX-GFP (59331) at KpnI/XhoI restriction enzyme sites. For Flp-dependent expression AAV construction, ArgiNLS-mKate2 was subcloned into pAAV-CAG-fDIO-mNeonGreen (99133) at KpnI/XhoI restriction enzyme sites. ArgiNLS-GCaMP6s and ArgiNLS-EGFP variants containing less or more than 7 R's were gene synthesized and cloned into the pAAV-Ef1a-DIO-Synaptophysin-GCaMP6s vector at Genscript (Piscataway, NJ). All cloned insert sequences of final plasmids used in the manuscript were sequence verified by Sanger sequencing (Genewiz, Seattle, WA). Full AAV plasmid sequences and maps are made publicly available through Addgene.

Characterization of nuclear localization in vitro

Immortalized embryonic mouse hippocampal cells (Cedarlane; mHippoE-1; cat. CLU198)) or mouse Neuroblastoma 2A cells (ATCC; CCL-131) were cultured in DMEM (high glucose and L-glutamine; Gibco 11965–092), supplemented with 10% fetal bovine serum (Gibco; 1600) and 1% penicillin-streptomycin (Invitrogen; 15140122) at 37°C in 5% CO_2. Cells were sub-cultured following manufacturer's instructions. For transient DNA transfections, cells were seeded onto Poly-L-lysine-coated glass coverslips (Electron Microscopy Science, 72292-08) in a 24-well plate format. Once cells reached ~80% confluence, transient transfections were performed in duplicate with 250 ng of plasmid DNA per construct using Lipofectamine 3000 transfection reagent (ThermoFisher Scientific, L3000015). After 48 h, cells were fixed with formalin, counterstained with DAPI, and mounted onto glass slides. All NLS tagged variants were co-transfected with a spectrally separate untagged fluorescent protein to fluorescently label total cellular area in the following combinations: all NLS tagged EGFP constructs with AAV-EF1a-mKate2; all NLS tagged RFP constructs with p-EGFP-N1 (Addgene; 6085-1). To qualitatively verify Cre-dependent ArgiNLS variant expression, each plasmid was co-transfected with pCAG-Cre (Addgene; 13775), respectively (data not shown). To qualitatively verify Flp-dependent ArgiNLS-mKate2 expression, we PCR-amplified and subcloned FlpO from PGK-NLS-FlpO-bpA (a gift from Dr. Phillipe Soriano[77]) into ptCAV-12a[78] at AgeI/EcoRV restriction sites to create the FlpO-expressing ptCAV-FlpO-12a which was used for co-transfections (data not shown). Images of all transfection conditions were acquired with a Keyence BZ-X800E inverted fluorescence microscope using a 20x/0.75 NA objective (Nikon). The effects of experimental NLS and ArgiNLS tags on nuclear localization of fluorescence was quantified amongst single-cells (n=5 randomly selected cells/3 replicate cultures or otherwise noted) in a blinded fashion using ImageJ. Accordingly, manual segmentations of the nucleus as defined by DAPI fluorescence, and the whole cell as defined by untagged co-transfected plasmid fluorescence, were created in synchronized duplicate images of the target channel. Corrected total nuclear or whole cell fluorescence (CTNF or CTCF, respectively) were separately calculated using area, mean, and integrated density measurements within the following equation: CTF = Integrated density - (Area * Mean background). Percent nuclear localized fluorescence was then calculated for each cell as CTNF/CTCF * 100.

Adeno-associated virus production

Php.eB and retroAAV2 packaging plasmid construction

A single packaging plasmid to produce PHP.eB viruses, pDG_PHP.eB, was constructed by subcloning the PHP.eB capsid from pUCmini-iCAP-PHP.eB (Addgene, 103005) into pDG1 plasmid to replace the AAV1 capsid at ClaI/SwaI restriction enzyme sites. The final vector was produced at 1mg/mL by Nature Technologies (Lincoln, Nebraska). A single packaging plasmid for retroAAV2 was created by subcloning the retroAAV2 capsid[80] into the pDG plasmid in a manner identical to pDG_PHP.eB to create pDG_retroAAV2. The pDG_PHP.eB plasmid is publicly available through Addgene.

AAV generation

HEK293T cells were transfected with 25 µg AAV vector plasmid and 50 µg packaging vector (pDG_PhP.eB or pDG2_retroAAV2) per 15 cm plate. Two days after transfection, cells were harvested and subjected to three freeze–thaw cycles. The supernatant was transferred to a Beckman tube containing a 40% sucrose cushion and spun at 27,000 rpm overnight at 4°C. Pellets were resuspended in CsCl at a density of 1.37 g/ml and spun at 65000 rpm 4 hours at 4°C. 1 ml CsCl fractions were run on an agarose gel, and genome-containing fractions were selected and spun at 50000 rpm overnight at 4°C. The 1 ml fractions were collected again, and genome containing fractions were dialyzed overnight. The filtered solution was transferred to a Beckman tube containing a 40% sucrose cushion and spun at 27,000 rpm overnight at 4°C. The pellet (containing purified AAV) was resuspended in 100 µl 1× HBSS. Virus was aliquoted and stored at -80 ° C until use.

AAV genome copy (GC) titration

Two methods were employed to titrate viral DNA genome copies of AAV preparations: dPCR by Stanford's Gene Vector and Virus Core for initial constitutive and conditional viral preparations, and in-house SYBR green-based qRT-PCR for subsequent constitutive preparations. For dPCR, the Qiacuity 5-plex system was employed which uses random distribution of template in available partitions to detect and measure fluorescence emission from target amplicons. The virus is was first made DNA ready by adding 8ul of Biosearch Quickextract solution to 2 ul of sample and thermocycling at 65C for 6min and 98c for 2min. Extracted DNA was then combined at 100x, 10,000x, and 5,000,000x dilutions with Qiacuity ProbePCR mastermix and XFP primers (forward: agcaaagaccccaacgagaa; reverse: ggcggcggtcacgaact), and reaction mixes were subjected to a dPCR thermocycling program: 95 C for 2 min, and 50 cycles of 15 sec at 95 C and 30 sec of 60 C. Earlier determined titer of AAV-45 was included as internal control along with a No Template Control. The concentration 'copies/ul' is then used to calculate the genome copies per mL. For qRT-PCR, 5 ul of AAV at multiple dilutions (5000x, 25000x, 125,000x, 625,000x, and 3,125,000x) or linearized standard plasmid DNA ($1e^4$-$1e^8$ molecules) were combined into master mixes containing SsoAdvanced Universal SYBR Green (Bio-Rad, 1725271) and primer sets targeting either the WPRE3 sequence (forward: CTGGTTAGTTCTTGCCACGG; reverse: AATTGTCAGTGCCCAACAGC) for titration of constitutive viruses. 20 ul reaction mixes were then subjected to the following thermal cycling protocol using a Bio-Rad CFX Connect Real-Time PCR System: 2 min at 98 C, and 38 cycles of 15 sec at 98 C and 20 sec at 60 C. Viral genome copies/ul were calculated in Microsoft Excel based off of plasmid DNA standard curves that passed quality control of an R^2 of ~.99 and PCR efficiencies of ~100%.

Stereotaxic viral injections

All stereotaxic surgeries were performed under aseptic conditions. Briefly, animals were anesthetized with isoflurane prior to and during placement and surgery in a stereotaxic instrument (David Kopf Instruments). Scalp incisions were made to expose the skull surface and virus was injected intracranially using a 32 gauge 0.5 ul Neuros syringe (Hamilton Co.) at the following coordinates from bregma of the mouse brain: SSpBF: -1.67 anterioposterior, 2.75 mediolateral, -0.9 dorsoventral; BMA: 1.55 anterioposterior, 2.7 mediolateral, -5.2 dorsoventral. For NLS nuclear fluorescence comparisons in defined cell-types, 400 nl of AAV.eB-EF1a-ArgiNLS-EGFP ($7x10^{12}$ GC/ml) or AAV.eB-EF1a-SV40nls-EGFP ($7x10^{12}$ GC/ml) was co-injected with 100 nl of AAV.eB-CAG-FLEX-tdTomato (Addgene, 28306; $7x10^{13}$ GC/ml) into the SSpBF of homozygous vGlut1-Cre or vGat-Cre mice at a 0.1 ul/min delivery rate. For qualitative characterization of conditional ArgiNLS-tagged fluorescent proteins, 500 nl of AAV.eB-CAG-FLEX-ArgiNLS-EGFP ($1x10^{12}$ GC/ml) and AAV.eB-CAG-fDIO-ArgiNLS-mKate2 ($1x10^{12}$ GC/ml) virus was injected into the SSpBF of separate heterozygous vGat-Cre or vGat-Flp mice at a 0.1 ul/min delivery rate. For characterization of retrograde ArgiNLS-tagged fluorescent protein expression, 200 nl of retroAAV2-EF1a-ArgiNLS-mKate2 ($1.5x10^{13}$ GC/ml) virus was injected into the BMA of C57Bl6/J mice at a 0.02 ul/min delivery rate. All surgerized animals received 1 mg/ml ketoprofen analgesic and 500 ul saline subcutaneously followed by recovery in their homecage placed above a warm heating pad. All animals receiving stereotaxic injections were sacrificed for histological analysis at 2 weeks post-injection.

Characterization of ArgiNLS- EGFP nuclear localization in vivo

Local AAV-injected mice were terminally anesthetized with isofluorane and transcardially perfused with phosphate-buffered saline (PBS) followed by 10% neutral buffered formalin (Sigma, HT501128). Brains were extracted, post-fixed overnight in 4 C, and cryoprotected in 30% sucrose at 4 C. Cryoprotected brains were histologically sectioned in coronal orientation on a freezing cryostat (Leica Biosystems) at 50 um in 6 equal series. Sections were counterstained with DAPI (Millipore Sigma; 10236276001) and then mounted with VectaShield mounting media (Vector Laboratories; H-1400-10). Mounted sections were imaged using a Ziess LSM880 confocal microscope. Confocal images were captured with either 10x/0.45 NA (Zeiss) or 40x/1.2 NA (Zeiss) objectives. All scans were acquired in 16-bit with 2.05 ms pixel dwell time and line 8 averaging in a 1024 x 1024 dimension. 3-channel DAPI/EGFP/tdTomato images were captured

using the following excitation/emission filters: DAPI: 410-585 nm; EGFP: 493-564 nm; tdTomato: 566-691 nm. Consistent laser power and gain settings were used across all conditions. 40x scans used for quantification were acquired in Z-stacks consisting of 3 um optical steps across 15 um total tissue depth. Maximal z-projections were created in ImageJ and used for quantification. The effects of SV40nls and ArgiNLS tags on cellular localization of EGFP fluorescence was quantified in EGFP+/tdTomato+ single cells (n=5 randomly selected cells/3 vGlut1-Cre or vGat-cre animals) from 40x images in a blinded fashion using ImageJ in a manner similarly described for *in-vitro* quantification. Briefly, manual segmentations of the nucleus as defined by DAPI fluorescence, and the soma as defined by Cre-dependent tdTomato fluorescence, were created in synchronized duplicate images of the target EGFP channel. Since we observed different levels of extracellular background fluorescence at the injection site amongst SV40nls- and ArgiNLS-EGFP injections, we used DAPI-guided extracellular background fluorescence images and measurements from the uninjected contralateral side of the SSpBF for background subtraction. These values were also used to compare against the extracellular fluorescence background measurements taken within each injection site of all animals to perform ipsilateral versus contralateral extracellular background fluorescence comparisons. Axonal fluorescence was quantified from additional 20x image stack z maximal z-projections acquired in the corpus callosum from *vGlut1-Cre* samples. For quantification, we applied an adaptive manual threshold to isolate measurements to axons only using the Huang method in ImageJ on 608 x 1024 FOVs. Corrected total axonal fluorescence (CTAF) was then calculated for each channel as Integrated density – (Area * mean background). EGFP+ CTAF was internally normalized for each sample to *vGlut1-Cre*:tdTomato+ CTAF and expressed as a percent.

Retroorbital viral injections

Unless otherwise noted, $1x10^{11}$ GC of AAV.eB virus was diluted up to 100 ul with sterile saline and injected intravenously into the retro-orbital (RO) sinus of mice receiving brief isofluorane anesthesia. All RO-injected animals were sacrificed at 3 weeks post-injection for histological analysis.

ML classification performance in 2D images

Mice RO-injected with AAV.eB-EF1a-ArgiNLS-EGFP or AAV.eB-EF1a-Sv40nls-EGFP were terminally anesthetized with isofluorane and transcardially perfused and histologically processed as described above. Single-plane epifluorescent images of entire coronal planes at AP 2.11 and -0.52 were acquired in montages with a Keyence BZ-X800E inverted fluorescence microscope using a 10x/0.45 NA objective (Nikon). Consistent DAPI and EGFP channel exposure settings were applied across conditions. Montages were stitched using BZ-X800 Analyzer software (Keyence, v 1.1.1.8) and composite images were created in ImageJ and exported as TIFF images.

QuPath Classifier Training

Machine learning classifiers for ArgiNLS- and SV40nls-EGFP labels were created using random forest classification with QuPath[81] software applied to composite TIF images. We used 25 px x 25 px (625 px^2) tile FOVs labeled as ‘DAPI’ or ‘GFP’ as training input data. The object classification function was used to create classifiers with the following train object classifier parameters: object filter = all detections; classifier = random trees (RTrees), features = all measurements, classes = all classes, and training = unlocked annotations. Cell detection parameters consisted of: Setup parameters detection channel = DAPI; nucleus parameters background radius = 0 px; median filter radius = 0 px; sigma = 3 px; minimum area = 10 px^2; maximum area = 200 px^2; intensity threshold = 50 and split by shape; cell expansion = 0 px; general parameters = smooth boundaries and make measurements. The live update function was used to observe classifier training, and each training iteration included the addition of 2 training tiles. The first training iteration for each classifier included both a ‘DAPI’ and ‘GFP’-labeled training ROI. Total iterations to reach maximal performance for the ArgiNLS classifier was 10 and included a combination of 7

DAPI and 2 GFP training ROIs. Total iterations to reach maximal performance for the SV40nls classifier was 28 and included a combination of 22 DAPI and 6 GFP training ROIs.

Classifier Performance Quantification
Additional 100 px x 100 px (10,000 px^2) FOVs were created (4 per AP) to quantify classifier performance at every training iteration. FOV placement varied ranging from regions with small to large density of labeled cells. Labeled cells in each FOV were manually segmented separately by 3 expert raters, and classifier training was repeated, as described above. Both manual and automated segmentations were overlaid for each FOV in QuPath and manual only segmentations, classifier only segmentations, and co-segmentations were hand scored for each FOV at every training iteration. Data was recorded in Microsoft Excel and used to calculate precision, recall, and F1 scores.

3D volumetric characterization of ArgiNLS-EGFP performance
Two independent approaches were applied to systematically quantify and characterize single-cell labeling across the intact mouse brain with ArgiNLS-tagged FPs: 1) iDISCO+ solvent-based brain clearing and EGFP staining based on methods previously described[9,13], and 2) SHIELD aqueous-based brain clearing.

iDISCO+ whole-brain processing
Mice RO-injected with AAV.eB-EF1a-ArgiNLS-EGFP (n=4) or AAV.eB-EF1a-Sv40nls-EGFP (n=4) were sacrificed and brains were post-fixed as described above. We performed iDISCO+ beginning with a pretreatment of 1x PBS washes (3 x 30 min; RT) followed by dehydration using ascending incubations of methanol (MeOH; Sigma, 34860-4L-R) concentrations (1 h each step; RT rotating): 20%, 40%, 60%, 80%, 100%, and 100%. Samples were next delipidated with a 66% dichloromethane (DCM; Sigma, 270997-12X100ML)/34% MeOH solution (ON, RT rotating), and washed the following day in 100% MeOH (2 x 1 h, RT rotating). Brains were then bleached in a chilled solution consisting of 5% hydrogen peroxide (30% Hydrogen peroxide solution; Sigma, 21673-500ML)/methanol (ON,4 C). Bleached brains were then rehydrated using descending incubations of MeOH concentrations (1 h each step; RT): 80%, 60%, 40%, 20%, and finally washed in 1x PBS buffer containing 0.5% TritonX-100 (Sigma, T9284-1L) (2 x 1 h; RT rotating). To immunostain brains for EGFP, we first permeabilized pre-treated samples in a 1x PBS buffer containing 0.5% TritonX-100/1.4% glycine (Sigma, G7126-1KG)/20% dimethyl sulfoxide (DMSO; Sigma, D128-1) (2 days, 37 C rotating) followed by blocking of non-specific binding through incubation with a 1x PBS buffer containing 0.5% TritonX-100/6% normal donkey serum (NDS; Jackson ImmunoResearch, 017-000-121)/10% DMSO (2 d, 37 C rotating). EGFP primary antibody (chicken polyclonal; Aves, GFP1010) labeling was next performed by applying a 1:200 dilution in a 1x PBS buffer containing 0.5% Tween-20/10ug/mL heparin (Sigma, H3393-250KU)/ 3% NDS/5% DMSO (7 days, 37 C rotating). Excess antibody was then washed with 1x PBS buffer containing 0.5% Tween-20/10ug/mL heparin (4-5 times x 2 days, 37 C rotating). Secondary antibody (donkey anti-chicken Alexa Fluor-647-conjugated AffiniPure F(ab')2; Jackson ImmunoResearch; 703-606-155; Lot: 157044) staining was then performed by applying a 1:100 dilution in a 1x PBS buffer containing 0.5% Tween-20/10ug/mL Heparin/ 3% NDS/5% DMSO (7 days, 37 C rotating). Excess secondary antibody was washed with 1x PBS buffer containing 0.5% Tween-20/10ug/mL heparin (4-5 times x 3 days, 37 C rotating). Tissue clearing then began with dehydration using ascending incubations of MeOH concentrations (1 h each step; RT rotating): 20%, 40%, 60%, 80%, 100%, and 100% followed by delipidation with a 66% DCM/34% MeOH solution (3 h, RT rotating). Next, MeOH was washed with 100% DCM (2 x 15 min; RT rotating). Finally, samples were index matched for at least 1 day in DiBenzyl Ether (DBE; Sigma, 108014-1KG) prior to imaging.

iDISCO+-processed volumetric imaging
EGFP-stained and cleared whole-brains were imaged in DBE using an UltraMicroscope II light-sheet fluorescent microscope (LSFM) with Infinity Corrected Objective Lenses (Miltenyi Biotec).

Our microscope configuration consisted of a 1.1x/0.1 NA MI PLAN objective (LaVision BioTec), non-corrected dipping cap, and an Andor Zyla sCMOS camera. All samples were secured on a 3D-printed sample platform and positioned in horizontal orientation with the cortical surface facing up. 2-channel image acquisition was performed in 1 FOV tile for autofluorescence (Excitation: 488 nm laser, Emission: 535/43 bandpass filter) and EGFP stain (Excitation: 647 nm laser, Emission: 690/50 bandpass filter) in separate scans (scan order: z-x-y) at a near-isotropic pixel resolution of 5.91 um X x 5.91 um Y x 6 um Z. We applied the following fixed acquisition parameters across all samples using ImspectorPro software (Miltenyi Biotec; v 7.1.16): laser power = 20 (488 channel), 30 (647 channel); exposure = 100 ms (488 channel), 136 ms (647 channel); sheet NA = 0.16; sheet thickness = 3.89 um; sheet width = 100%; zoom = 1x; dynamic horizontal focus = 5 (647 channel only); dynamic horizontal focus processing = contrast adaptive; merge light-sheet = blend.

SHIELD whole-brain processing

We followed the SmartClear full active pipeline protocol (LifeCanvas Technologies, v5.05) for aqueous-based brain clearing and mounting. Accordingly, RO- or locally-injected mice were sacrificed and brains were post-fixed as described above. Brain samples were first incubated in freshly prepared SHIELD OFF solution, consisting of SHIELD buffer and epoxy solution (4 days, 4 C shaking), followed by SHIELD ON buffer (1 day, 37 C shaking). Samples were next equilibrated in Delipidation buffer (ON, RT shaking). Prior to active clearing, the SmartClear II Pro electrophoresis system (LifeCanvas Technologies) containing Delipidation buffer was equilibrated to 42 C and new membranes were installed. Samples were then transferred into mesh bags and placed into clearing chambers. Active clearing was performed on 4-sample batches using the Gentle Mode program (1500 mA/42 C; 1 day). After clearing, samples were removed from mesh bags and index matched first in 50% EasyIndex (1 day; 37 C slow shaking) followed by 100% EasyIndex (1day; 37 C slow shaking). Index matched samples were embedded on custom 3D-printed sample mounts in EasyIndex containing 2% w/v agarose (Sigma, A5030), and incubated post-embedding in 100% EasyIndex (ON, 37 C). Embedded samples were stored in 100% EasyIndex (RT) in the dark until imaging.

SHIELD-processed volumetric imaging of native fluorescence

Native signal from aqueous-based cleared brains were imaged horizontally using the SmartSPIM LSFM (LifeCanvas Technologies)[70] at 4 um isotropic pixel resolution. We imaged all samples in 488 and 563 channels to acquire native fluorescence from AAV expression from green- or red-emitting FPs, and background fluorescence in the channel without AAV expression for 3D rendering purposes. Laser power and acquisition settings were optimized individually towards each proof-of-concept sample, whereas settings for titrated AAV.eB-EF1a-ArgiNLS-oScarlet samples were held constant across conditions.

Whole-brain image processing

Brain-wide SV40nls versus ArgiNLS-tagged EGFP cell count comparisons

Volumetric LSFM-acquired images were processed for whole-brain template registration, automated cell segmentation, and regional segmentation as previously described[9] with modifications. First, we employed the Unified brain template and atlas[63] for registration and regional segmentation, respectively. Second, we applied ML classification of single-cells from SV40nls-EGFP or ArgiNLS-EGFP stains through the creation of separately trained classifiers using random forest classification algorithm through ilastik software[62]. Classifier Performance was quantified in FOVs in regions with small to large density of labeled cells as similarly described previously on non-ML segmented datasets[9,21]. Labeled cells in each FOV were manually segmented separately by 3 expert raters. Data from ImageJ processing was recorded in Microsoft Excel and used to calculate precision, recall, and F1 scores.

3D Visualization and ML

Zeiss arivis Pro (formerly, Vision4D) software was used to generate 3D renderings of brain samples throughout the manuscript. Volumetric segmentations of major brain areas on registered, mean voxelized and p-value map data was performed by importing the Unified atlas label stack as a separate image followed by label segmentation based on grayscale intensity level for each anatomical label using the Labeled Image Segmenter function. The names and hierarchy of the anatomical labels were separately imported using a custom python script. 3D ML classification of ArgiNLS-FP label in native state images was also performed in Zeiss arivis Pro using the Image-based ML feature that uses random forest classification. Classifier training was performed separately for each sample and consisted of two classes, background and single-cells, in which single-cell classes were stored as objects in order to flexibly quantify and color-code segmentations across volumetric spaces.

Electrophysiology

vGlut1-Cre mice injected with ArgiNLS-EGFP and Cre-dependent tdTomato virus were allowed 3-4 weeks recovery after surgery to allow for viral expression. All solutions were continuously bubbled with O_2/CO_2. Coronal (200 um) brain slices were prepared from 11-13 week old mice in a slush NMDG cutting solution (in mM: 92 NMDG, 2.5 KCl, 1.25 NaH_2PO_4, 30 $NaHCO_3$, 20 HEPES, 25 glucose, 2 thiourea, 5 Na-ascorbate, 3 Na-pyruvate, 0.5 $CaCl_2$, 10 $MgSO_4$, pH 7.3–7.4[82]. Slices recovered for ~12 min in the same solution in 32°C water bath. Slices were then transferred to a room temperature HEPES-aCSF solution (in mM: 92 NaCl, 2.5 KCl, 1.25 NaH_2PO_4, 30 $NaHCO_3$, 20 HEPES, 25 glucose, 2 thiourea, 5 Na-ascorbate, 3 Na-pyruvate, 2 $CaCl_2$, 2 $MgSO_4$). Slices recovered for an additional 60 min. Whole-cell patch-clamp recordings were made using an Axopatch 700B amplifier (Molecular Devices) with sampling at 10KHz and filtering at 1 kHz. EGFP+ and/or tdTomato+ cells were visually identified via fluorescence for patching with electrodes at 4–6 MΩ. Series resistance was monitored during all recordings with changes in resistance +10% qualifying the cell for exclusion. Recordings were made in aCSF (in mM: 126 NaCl, 2.5 KCl, 1.2 NaH_2PO_4, 1.2 $MgCl_2$ 11 D-glucose, 18 $NaHCO_3$, 2.4 $CaCl_2$) at 32°C continually perfused over slices at a rate of ~2 ml/min. Barrel cortex glutamate neurons were identified by fluorescence. Electrodes were filled with an internal solution containing (in mM): 130 potassium gluconate, 10 HEPES, 5 NaCl, 1 EGTA), 5 Mg2+/ATP, 0.5 Na+-GTP, pH 7.3, 280 mOsmol. Excitability I/V curves (-20-180 pA, 10 pA steps, 500 ms) were measured in I-Clamp mode after gaining whole cell access. All data were analyzed offline using Clampfit (Molecular Devices). Excitability was measured as the total number of events during each current step.

Behavioral health battery

To screen for behavioral effects from systemic AAV.eB expression from experimental viruses the following battery of behavioral tests were performed in chronological order on the same mice:

Open field testing:

Mice were placed in a large circular arena (120 cm diameter) in dim light and activity was recorded for a period of 10 minutes. Data were analyzed using Ethovision software, where time in center, time on edge, and total distance were calculated.

Elevated Plus maze (EPM):

The EPM is plus-shaped arena in which two maze arms are sheltered by enclosed walls, and two arms are open platforms. Each arm length is 15 cm, with a center area of 49 cm^2. Activity was recorded for a period of 10 min and analyzed using Ethovision software.

Locomotor activity:

Baseline locomotion was measured using locomotion chambers (Columbus instruments) that use infrared beam breaks to calculate ambulatory activity. Mice were singly housed in Allentown cages with reduced corncob bedding and provided with *ad libitum* access to food and water. Locomotion was monitored continuously over the course of 3 nights and 2 days.

Statistics

Raw data provided throughout the manuscript represent mean ± S.E.M. values. With the exception of voxel statistical tests (performed in ClearMAP software[13]), all other statistical

analyses and data graphing was performed with Prism 10 software (GraphPad). Alpha levels for statistical significance were set at 0.05 for p and q-values. Probability levels of statistical significance from group mean comparisons are indicated in figures with an asterisk labeling convention of: *p<0.05; **p<0.01; **p<0.01; ****p<0.001. Two group comparisons were analyzed using two-tailed, unpaired Student's t-test. One- or two-way ANOVA analysis of main effects were used to compare group means with one of two independent variables, respectively. Posthoc tests were chosen on an experiment-by-experiment basis and indicated in each figure legend. Region-by-region analysis of whole-brain datasets were performed using FDR-corrected multiple t-tests by the two-stage step-up method of Benjamini, Krieger, and Yekutieli[83] on uncollapsed levels of all subregions from major brain ontologies.

Acknowledgments

We would like to acknowledge support from the University of Washington Center of Excellence in Addiction Pain and Emotion (NAPE) Molecular Genetics Resource Core (P30DA048736). We would like to thank Valerie Tsai, Aida Chan, Axel Salazar, and Kevin Bai for their technical support. We thank the entire Golden and Heshmati Laboratories for their discussions and critiques of the manuscript.

References

1. Corsetti, S., Gunn-Moore, F. & Dholakia, K. Light sheet fluorescence microscopy for neuroscience. *J Neurosci Methods* **319**, 16–27 (2019).
2. Ragan, T. *et al.* Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nat Methods* **9**, 255–258 (2012).
3. Li, A. *et al.* Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science* **330**, 1404–1408 (2010).
4. Kim, Y. *et al.* Whole-Brain Mapping of Neuronal Activity in the Learned Helplessness Model of Depression. *Front Neural Circuits* **10**, 3 (2016).
5. Kim, Y. *et al.* Mapping social behavior-induced brain activation at cellular resolution in the mouse. *Cell Rep* **10**, 292–305 (2015).
6. Azevedo, H., Ferreira, M., Mascarello, A., Osten, P. & Guimarães, C. R. W. Brain-wide mapping of c-fos expression in the single prolonged stress model and the effects of pretreatment with ACH-000029 or prazosin. *Neurobiol Stress* **13**, 100226 (2020).
7. Cai, G. *et al.* Brain-wide mapping of c-Fos expression with fluorescence micro-optical sectioning tomography in a chronic sleep deprivation mouse model. *Neurobiol Stress* **20**, 100478 (2022).
8. Roland, A. V. *et al.* Alcohol Dependence Modifies Brain Networks Activated During Withdrawal and Reaccess: A c-Fos-Based Analysis in Mice. *Biol Psychiatry* S0006-3223(23)00046-X (2023) doi:10.1016/j.biopsych.2023.01.018.
9. Madangopal, R. *et al.* Incubation of palatable food craving is associated with brain-wide neuronal activation in mice. 2022.05.31.494210 Preprint at https://doi.org/10.1101/2022.05.31.494210 (2022).
10. DeNardo, L. A. *et al.* Temporal evolution of cortical ensembles promoting remote memory retrieval. *Nat Neurosci* **22**, 460–469 (2019).
11. Rijsketic, D. R. *et al.* UNRAVELing the synergistic effects of psilocybin and environment on brain-wide immediate early gene expression in mice. *Neuropsychopharmacology* (2023) doi:10.1038/s41386-023-01613-4.
12. Hansen, H. H. *et al.* Whole-brain activation signatures of weight-lowering drugs. *Mol Metab* **47**, 101171 (2021).

13. Renier, N. Mapping of Brain Activity by Automated Volume Analysis of Immediate Early Genes. 133.
14. Susaki, E. A. *et al.* Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis. *Cell* **157**, 726–739 (2014).
15. Yamaguchi, H. *et al.* Brain-wide mapping of neuronal architecture controlling torpor. 2023.03.03.531064 Preprint at https://doi.org/10.1101/2023.03.03.531064 (2023).
16. Roy, D. S. *et al.* Brain-wide mapping reveals that engrams for a single memory are distributed across multiple brain regions. *Nat Commun* **13**, 1799 (2022).
17. Kim, Y. *et al.* Brain-wide Maps Reveal Stereotyped Cell Type-based Cortical Architecture and Subcortical Sexual Dimorphism. *Cell* **171**, 456-469.e22 (2017).
18. Son, S. *et al.* Whole-Brain Wiring Diagram of Oxytocin System in Adult Mice. *J Neurosci* **42**, 5021–5033 (2022).
19. Habart, M., Lio, G., Soumier, A., Demily, C. & Sirigu, A. An optimized iDISCO+ protocol for tissue clearing and 3D analysis of oxytocin and vasopressin cell network in the developing mouse brain. *STAR Protoc* **4**, 101968 (2023).
20. Matho, K. S. *et al.* Genetic dissection of the glutamatergic neuron system in cerebral cortex. *Nature* **598**, 182–187 (2021).
21. Szelenyi, E. R. *et al.* Brain X chromosome inactivation is not random and can protect from paternally inherited neurodevelopmental disease. 458992 Preprint at https://doi.org/10.1101/458992 (2021).
22. Wang, Y. W., Hu, Y., Qi, J. T., Zhang, Z. & Luo, M. Q. Comparative Whole-Brain Mapping of Isoflurane and Ketamine-Activated Nuclei and Functional Networks. 2023.06.03.543576 Preprint at https://doi.org/10.1101/2023.06.03.543576 (2023).
23. Stelzer, E. H. K. *et al.* Light sheet fluorescence microscopy. *Nat Rev Methods Primers* **1**, 1–25 (2021).
24. Dean, K. M. *et al.* Isotropic imaging across spatial scales with axially swept light-sheet microscopy. *Nat Protoc* **17**, 2025–2053 (2022).
25. Claudi, F. *et al.* Visualizing anatomically registered data with brainrender. *eLife* **10**, e65751 (2021).
26. Niedworok, C. J. *et al.* aMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. *Nat Commun* **7**, 11879 (2016).
27. Tyson, A. L. *et al.* A deep learning algorithm for 3D cell detection in whole mouse brain image datasets. *PLOS Computational Biology* **17**, e1009074 (2021).
28. Tyson, A. L. *et al.* Accurate determination of marker location within whole-brain microscopy images. *Sci Rep* **12**, 867 (2022).
29. Negwer, M. *et al.* FriendlyClearMap: an optimized toolkit for mouse brain mapping and analysis. *Gigascience* **12**, giad035 (2022).
30. Choquet, D., Sainlos, M. & Sibarita, J.-B. Advanced imaging and labelling methods to decipher brain cell organization and function. *Nat Rev Neurosci* **22**, 237–255 (2021).
31. Lanford, R. E. & Butel, J. S. Construction and characterization of an SV40 mutant defective in nuclear transport of T antigen. *Cell* **37**, 801–813 (1984).
32. Goertsen, D. *et al.* AAV capsid variants with brain-wide transgene expression and decreased liver targeting after intravenous delivery in mouse and marmoset. *Nat Neurosci* **25**, 106–115 (2022).
33. Jang, M. J. *et al.* Spatial transcriptomics for profiling the tropism of viral vectors in tissues. *Nat Biotechnol* 1–15 (2023) doi:10.1038/s41587-022-01648-w.
34. Matsumoto, K. *et al.* Advanced CUBIC tissue clearing for whole-organ cell profiling. *Nat Protoc* **14**, 3506–3537 (2019).
35. Ito, S. *et al.* Loss of neuronal 3D chromatin organization causes transcriptional and behavioural deficits related to serotonergic dysfunction. *Nat Commun* **5**, 4450 (2014).
36. Bernhofer, M. *et al.* NLSdb—major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Research* **46**, D503–D508 (2018).
37. Luk, K. *et al.* Optimization of Nuclear Localization Signal Composition Improves CRISPR-Cas12a Editing Rates in Human Primary Cells. *GEN Biotechnology* **1**, 271–284 (2022).

38. Thumberger, T. *et al.* Boosting targeted genome editing using the hei-tag. *eLife* **11**, e70558 (2022).
39. Cautain, B., Hill, R., de Pedro, N. & Link, W. Components and regulation of nuclear transport processes. *FEBS J* **282**, 445–462 (2015).
40. Friedrich, B., Quensel, C., Sommer, T., Hartmann, E. & Köhler, M. Nuclear localization signal and protein context both mediate importin alpha specificity of nuclear import substrates. *Mol Cell Biol* **26**, 8697–8709 (2006).
41. Xiao, C. Y., Jans, P. & Jans, D. A. Negative charge at the protein kinase CK2 site enhances recognition of the SV40 large T-antigen NLS by importin: effect of conformation. *FEBS Lett* **440**, 297–301 (1998).
42. Bartnicki, F., Bonarek, P., Kowalska, E. & Strzalka, W. The Argi system: one-step purification of proteins tagged with arginine-rich cell-penetrating peptides. *Sci Rep* **7**, 2619 (2017).
43. Fuchs, S. M. & Raines, R. T. Polyarginine as a multifunctional fusion tag. *Protein Sci* **14**, 1538–1544 (2005).
44. Brewer, S. J. & Sassenfeld, H. M. The purification of recombinant proteins using C-terminal polyarginine fusions. *Trends in Biotechnology* **3**, 119–122 (1985).
45. Hristova, K. & Wimley, W. C. A Look at Arginine in Membranes. *J Membr Biol* **239**, 49–56 (2011).
46. Martin, R. M. *et al.* Principles of protein targeting to the nucleolus. *Nucleus* **6**, 314–325 (2015).
47. Schmidt, N., Mishra, A., Lai, G. H. & Wong, G. C. L. Arginine-rich cell-penetrating peptides. *FEBS Letters* **584**, 1806–1813 (2010).
48. Imanishi, M., Yan, W., Morisaki, T. & Sugiura, Y. An artificial six-zinc finger peptide with polyarginine linker: Selective binding to the discontinuous DNA sequences. *Biochemical and Biophysical Research Communications* **333**, 167–173 (2005).
49. Freitas, N. & Cunha, C. Mechanisms and Signals for the Nuclear Import of Proteins. *Curr Genomics* **10**, 550–557 (2009).
50. Lanford, R. E. & Butel, J. S. Construction and characterization of an SV40 mutant defective in nuclear transport of T antigen. *Cell* **37**, 801–813 (1984).
51. Ishii, N. *et al.* Analysis of a nuclear localization signal of simian virus 40 major capsid protein Vp1. *Journal of Virology* **70**, 1317 (1996).
52. Dingwall, C., Sharnick, S. V. & Laskey, R. A. A polypeptide domain that specifies migration of nucleoplasmin into the nucleus. *Cell* **30**, 449–458 (1982).
53. Fenno, L. E. *et al.* Comprehensive Dual- and Triple-Feature Intersectional Single-Vector Delivery of Diverse Functional Payloads to Cells of Behaving Mammals. *Neuron* **107**, 836-853.e11 (2020).
54. Shcherbo, D. *et al.* Far-red fluorescent tags for protein imaging in living tissues. *Biochem J* **418**, 567–574 (2009).
55. Shcherbakova, D. M. *et al.* Bright monomeric near-infrared fluorescent proteins as tags and biosensors for multiscale imaging. *Nat Commun* **7**, 12405 (2016).
56. Hosokawa, K., Nishi, M., Sakamoto, H., Tanaka, Y. & Kawata, M. Regional distribution of importin subtype mRNA expression in the nervous system: study of early postnatal and adult mouse. *Neuroscience* **157**, 864–877 (2008).
57. Huang, Z. J. & Paul, A. The diversity of GABAergic neurons and neural communication elements. *Nat Rev Neurosci* **20**, 563–572 (2019).
58. Zakiewicz, I. M., Bjaalie, J. G. & Leergaard, T. B. Brain-wide map of efferent projections from rat barrel cortex. *Front Neuroinform* **8**, 5 (2014).
59. Kanda, T., Sullivan, K. F. & Wahl, G. M. Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells. *Curr Biol* **8**, 377–385 (1998).
60. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* **18**, 100–106 (2021).
61. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv.org* https://arxiv.org/abs/1505.04597v1 (2015).

62. Berg, S. *et al.* ilastik: interactive machine learning for (bio)image analysis. *Nat Methods* **16**, 1226–1232 (2019).
63. Enhanced and unified anatomical labeling for a common mouse brain atlas | Nature Communications. https://www.nature.com/articles/s41467-019-13057-w.
64. Navabpour, S., Kwapis, J. L. & Jarome, T. J. A neuroscientist's guide to transgenic mice and other genetic tools. *Neurosci Biobehav Rev* **108**, 732–748 (2020).
65. Haggerty, D. L., Grecco, G. G., Reeves, K. C. & Atwood, B. Adeno-Associated Viral Vectors in Neuroscience Research. *Molecular Therapy - Methods & Clinical Development* **17**, 69–82 (2020).
66. Lambert, G. G. *et al.* Aequorea's secrets revealed: New fluorescent proteins with unique properties for bioimaging and biosensing. *PLOS Biology* **18**, e3000936 (2020).
67. mGreenLantern: a bright monomeric fluorescent protein with rapid expression and cell filling properties for neuronal imaging | PNAS. https://www.pnas.org/doi/10.1073/pnas.2000942117.
68. Cox, R. S., Dunlop, M. J. & Elowitz, M. B. A synthetic three-color scaffold for monitoring genetic regulation and noise. *Journal of Biological Engineering* **4**, 10 (2010).
69. Park, Y.-G. *et al.* Protection of tissue physicochemical properties using polyfunctional crosslinkers. *Nat Biotechnol* **37**, 73–83 (2019).
70. Dean, K. M. *et al.* Isotropic imaging across spatial scales with axially swept light-sheet microscopy. *Nat Protoc* **17**, 2025–2053 (2022).
71. Adhikari, A. *et al.* Basomedial amygdala mediates top-down control of anxiety and fear. *Nature* **527**, 179–185 (2015).
72. Hu, P., Zhao, X., Zhang, Q., Li, W. & Zu, Y. Comparison of Various Nuclear Localization Signal-Fused Cas9 Proteins and Cas9 mRNA for Genome Editing in Zebrafish. *G3 (Bethesda)* **8**, 823–831 (2018).
73. Stepanenko, O. V., Stepanenko, O. V., Kuznetsova, I. M., Verkhusha, V. V. & Turoverov, K. K. Chapter Four - Beta-Barrel Scaffold of Fluorescent Proteins: Folding, Stability and Role in Chromophore Formation. in *International Review of Cell and Molecular Biology* (ed. Jeon, K. W.) vol. 302 221–278 (Academic Press, 2013).
74. Cormack, B. P., Valdivia, R. H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38 (1996).
75. DeNardo, L. & Luo, L. Genetic strategies to access activated neurons. *Current Opinion in Neurobiology* **45**, 121–129 (2017).
76. Ghosh, S., Brown, A. M., Jenkins, C. & Campbell, K. Viral Vector Systems for Gene Therapy: A Comprehensive Literature Review of Progress and Biosafety Challenges. *Applied Biosafety* **25**, 7–18 (2020).
77. Raymond, C. S. & Soriano, P. High-Efficiency FLP and ΦC31 Site-Specific Recombination in Mammalian Cells. *PLOS ONE* **2**, e162 (2007).
78. Kremer, E. J., Boutin, S., Chillon, M. & Danos, O. Canine Adenovirus Vectors: an Alternative for Adenovirus-Mediated Gene Transfer. *J Virol* **74**, 505–512 (2000).
79. Grimm, D., Kern, A., Rittner, K. & Kleinschmidt, J. A. Novel tools for production and purification of recombinant adenoassociated virus vectors. *Hum Gene Ther* **9**, 2745–2760 (1998).
80. Tervo, D. G. R. *et al.* A Designer AAV Variant Permits Efficient Retrograde Access to Projection Neurons. *Neuron* **92**, 372–382 (2016).
81. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci Rep* **7**, 16878 (2017).
82. Ting, J. T., Daigle, T. L., Chen, Q. & Feng, G. Acute brain slice methods for adult and aging animals: application of targeted patch clamp analysis and optogenetics. *Methods Mol Biol* **1183**, 221–242 (2014).
83. Benjamini, Y., Krieger, A. M. & Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507 (2006).

Figures and Tables

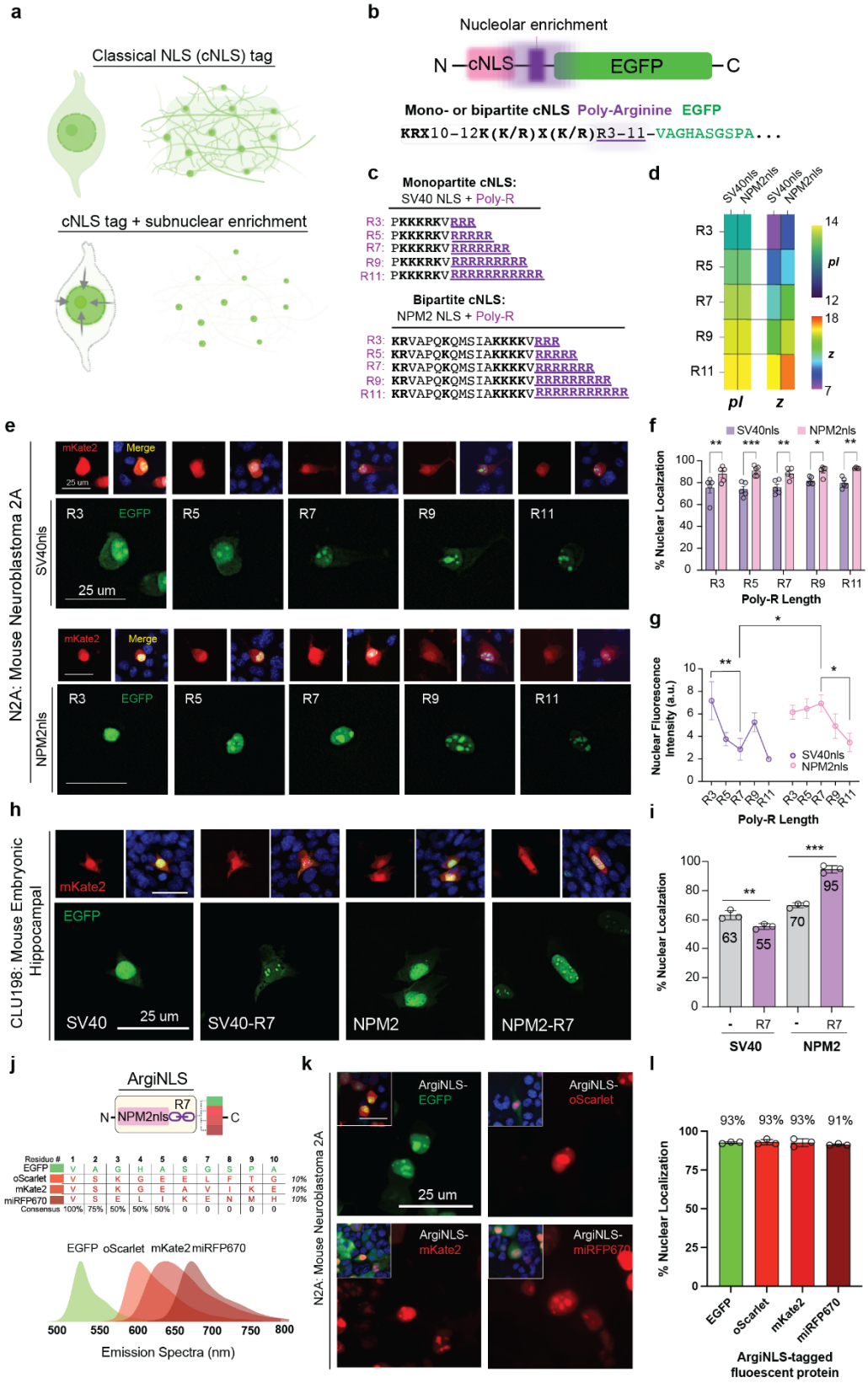


Figure 1. Strategy and *in vitro* characterization of an ariginine-rich NLS (ArgiNLS). a, Schematic depicting extracellular signal caused by classical NLS (cNLS) tag use, and depicting the theoretical effects of an optimized NLS tag that eliminates extra-nuclear signal artifact through a subnuclear enrichment strategy. **b**, Top: Configuration of cNLS, poly-arginine nucleolar enrichment tag, and EGFP in AAV overexpression vectors created for *in vitro* testing. Bottom: Corresponding amino acid composition of mono- or bipartite cNLS consensus sequences followed by poly-R stretch separating the coding sequence (CDS) of EGFP (first 10 amino acids shown). **c,** Amino acid sequences of SV40nls- and NPM2nls-poly-R tags. Basic amino acids are indicated in bold. **d**, Heat maps of isoelectric point (pI) and net positive charge (z) of each tag configuration. **e,** Representative images of transiently co-transfected single N2A cells displaying untagged mKate2 and SV40nls-poly-R (top) or NPM2nls-poly-R (bottom) tagged EGFP expression, spanning 3 to 11 R. **f,** Percent nuclear localization of each NLS across poly-R lengths (n=5/condition). **g**, Nuclear fluorescence intensity (a.u. = arbitrary units) across poly-R lengths for each NLS from same cells measured in **f**. **h**, Representative images of transient co-transfections with untagged mKate2 and each experimental EGFP-expressing construct within immortalized CLU198 embryonic hippocampal mouse cell line. **i,** Percent nuclear localization with/without R7 addition for each cNLS. **j**, Molecular schematic of ArgiNLS tag (NPM2nls-R7) and amino acid sequence comparison and consensus of the first 10 N-terminal residues to EGFP, oScarlet, mKate2, and miRFP670. **k**, Representative images of transient co-transfections with (from L to R and top to bottom) ArgiNLS-EGFP/mKate2, ArgiNLS-oScarlet/EGFP, ArgiNLS-mKate2/EGFP, and ArgiNLS-miRFP670/EGFP combinations within N2A mouse cell line. **l**, Percent nuclear localization of ArgiNLS-tagged FPs. **i** and **l** cell culture summary data is from 5 blinded and randomly selected cells quantified per transfection condition across 3 replicate cultures. The data in **f** was statistically analyzed using a two-way ANOVA followed by Holm-Sidak's multiple comparison post hoc test. The data in **g** was statistically analyzed using a two-way ANOVA followed by Holm-Sidak's multiple comparison post hoc tests for effects of poly-R length both within and across each cNLS. The data in **l** was statistically analyzed using a two-way (NLS (3) x +/- R7 (2)) ANOVA followed by Sidak post hoc test. *$p<0.05$; **$p<0.01$; ***$p<0.005$

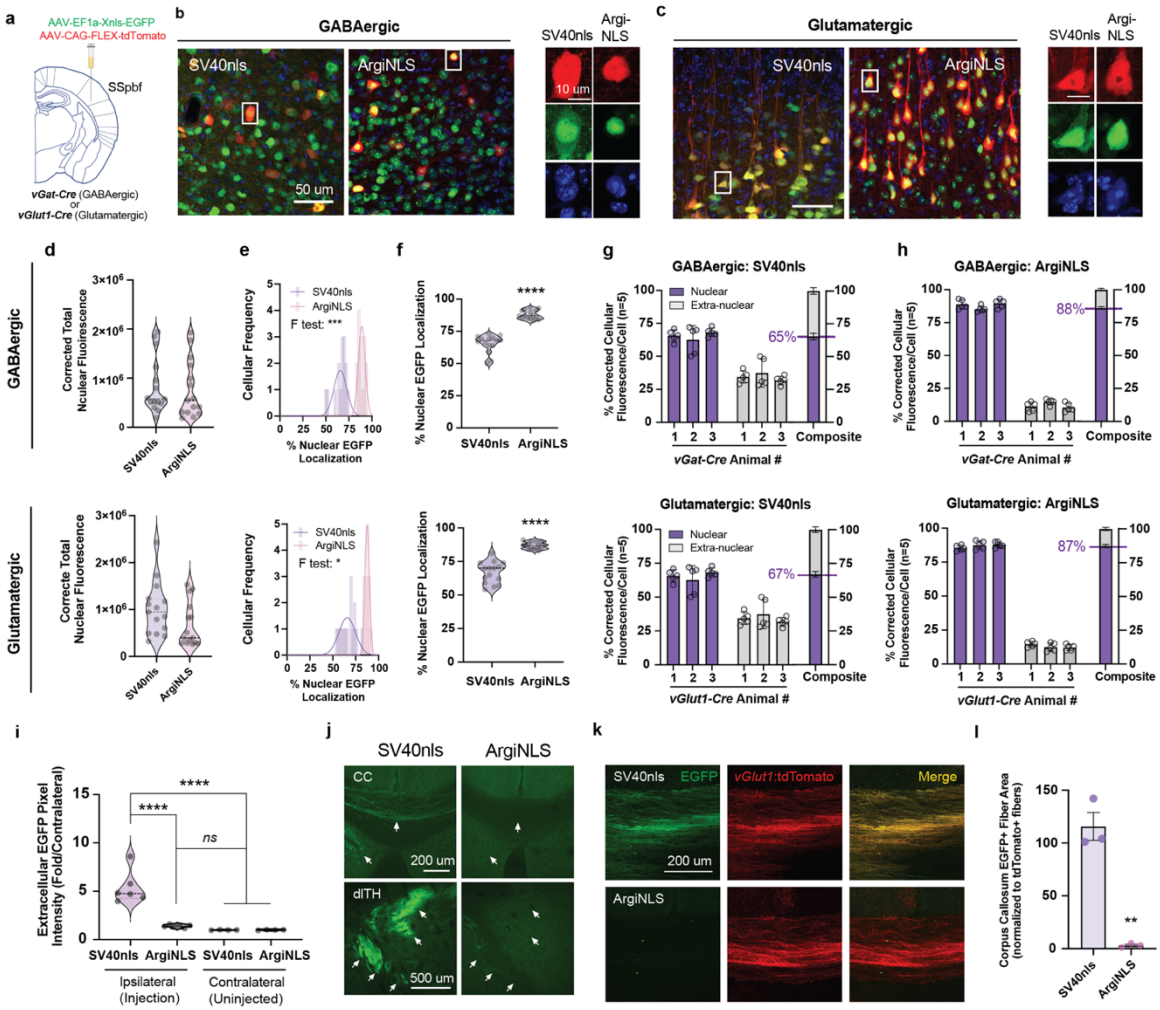


Figure 2. Optimized nuclear localization of fluorescence across major cortical cell-type classes *in vivo*. a, Schematic of stereotaxic AAV co-injection in the barrel field of primary somatosensory cortex (SSpBF) of *vGat-Cre* or *vGlut1-Cre* transgenic mice. **b-c**, Representative 40x confocal images of SV40nls-EGFP/FLEX-tdTomato and ArgiNLS-EGFP/ FLEX-tdTomato viral expression in (**b**) *vGat-Cre* or (**c**) *vGlut1-Cre* mice. Right: Cropped representative image of an example EGFP+/tdTomato+ cell**. d-h,** Quantification of NLS tag performance in GABAergic (top) or glutamatergic (bottom) EGFP+/tdTomato+ cells (n=random 5 cells/animal; n=3 animals per Cre-driver/tag). **d**, EGFP corrected total nuclear fluorescence (CTNF). **e**, Cellular frequency histogram of percent nuclear EGFP localization. **f**, Percent nuclear EGFP fluorescence comparison across tags. **g-h**, Nuclear and extra-nuclear EGFP fluorescence for all quantified cells of each animal across (**g**) SV40nls- or (**h**) ArgiNLS-labeled cells. The composite percent nuclear EGFP across all animals is indicated to the right. **i**, Total pixel intensity of extracellular (EC) EGFP signal ipsilateral to injection site, expressed as a fold change from non-injected contralateral side (n=3 EC quantifications/3 *vGlut1-Cre* mice**). j**, Representative images of extranuclear EGFP signal in axon bundles of passage in the corpus callosum (CC; top) or terminations in the dorsal thalamus (dTH; bottom). Arrows indicate location of labeling observed with SV40nls and are superimposed on ArgiNLS-EGFP. **k**, Representative 40x confocal images of SV40nls- or ArgiNLS-EGFP with *vGlut1:*tdTomato viral expression in fibers of the CC (n=3 EC quantifications/3 *vGlut1-Cre* mice). **l**, Quantification of EGFP+ fiber area in CC normalized to *vGlut1*:tdTomato+ fiber area. All bar graphs represent mean ± SEM. The data in **d, f,** and **l** was

statistically analyzed using an unpaired t-test. Group differences in population variance of data in **e** were statistically compared using an F-test. The data in **i** was statistically analyzed using a 1-way ANOVA followed by Tukey's multiple comparison post hoc test. Violin plots display median and interquartile range as dashed lines. *p<0.01; **p<0.05; **p<0.01; ****p<0.001

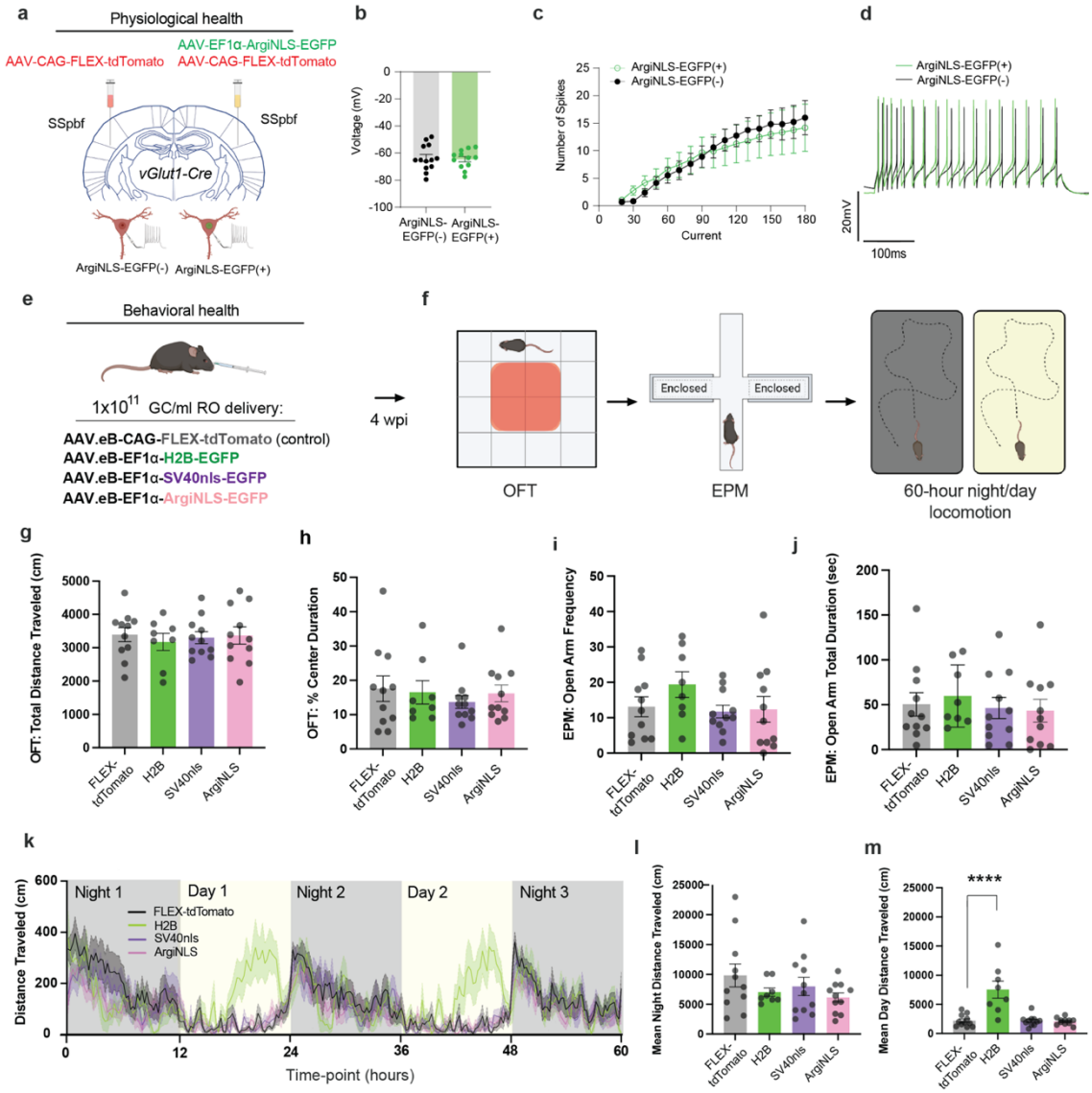


Figure 3. ArgiNLS-EGFP viral expression is not associated with physiological or behavioral deficits. a, Schematic illustration of bilateral stereotaxic AAV injection/co-injection in barrel cortices (SSpBF) of *vGlut1-Cre* transgenic mice, and electrophysiological recordings conducted on resulting *vGlut1-Cre*:tdTomato(+) neurons that are either ArgiNLS-EGFP(-) or ArgiNLS-EGFP(+). **b**, Resting membrane potential of ArgiNLS-EGFP(-) (n=13) or ArgiNLS-EGFP(+) (n=12) neurons. **c**, Current injection curve from ArgiNLS-EGFP(-) or ArgiNLS-EGFP(+) neurons. **d**, Example trace current injection at 100 pA. **e,** AAV.eB delivery and vector information schematic. Groups of mice were injected with 1 x 10^{11} GC of a negative control AAV.eB-CAG-FLEX-tdTomato (n=11), AAV.eB-EF1a-H2B-EGFP (n=8), AAV.eB-EF1a-SV40nls-EGFP (n=11), or AAV.eB-EF1a-ArgiNLS-EGFP (n=11). **f,** Baseline behavioral battery 4 weeks post viral injection (wpi). Mice were sequentially tested in the open field test (OFT), elevated plus maze (EPM), and 60-hour night/dark locomotion recording. **g**, Total distance traveled in the OFT. **h**, Percent time duration spent in the center of the OFT arena. **i**, Frequency of entries in the open arm of the EPM. **j**, Total duration of time spent in the open arm of the EPM. **k**, Locomotion XY plot showing total distance traveled across time of the 60-h recording. Day (yellow)/night (gray) phase of light cycle is underlaid the plot according to time of 60-h recording. **l**, Mean night distance traveled across the 60-h recording period. **m**, Mean day distance traveled across the 60-

h recording period. All bar graphs represent mean ± SEM. The data for each panel from **g-j** and **l-m** were statistically analyzed using an one-way ANOVA with Dunnet's multiple comparison posthoc test of each experimental group versus control FLEX-tdTomato group. ****$p<0.0001$

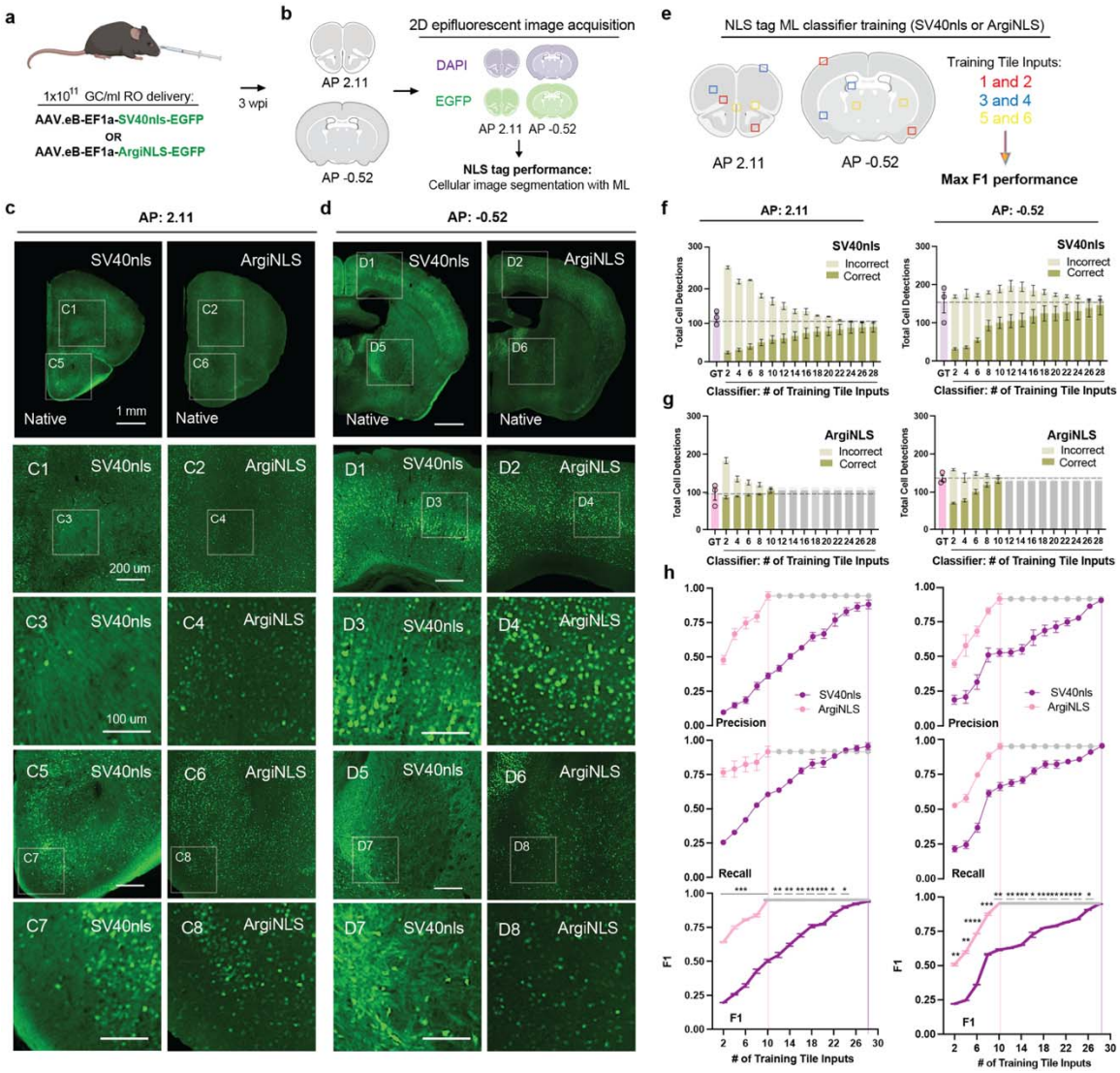


Figure 4. Enhanced performance of single-cell ML classification. a, Experimental schematic of systemic AAV.eB delivery of NLS-tagged EGFP. **b**, Microscopy acquisition schematic. **c-d**, Representative images of EGFP viral expression at matching (**c**) AP 2.11 or (**d**) -0.52 hemispheric coronal planes, across 2 FOVs at 3 magnification levels. **e**, Schematic of SV40nls and ArgiNLS classifier training workflow for single-cell segmentation of 2D images. **f-h**, Cell segmentation classifier benchmarking and performance metrics for (left) AP: 2.11 and (right) AP: -0.52 brain sections. Total cell detections for (**f**) SV40nls-EGFP or (**g**) ArgiNLS-EGFP for expert rater ground truth (GT; left bar) (n=3) and ML classifier across iterations of 2-increment additive training input rounds. Gray bars represent iterations after peak performance criteria was met. Dashed line indicates mean GT cell detections. **h**, Precision, recall, and F1 harmonic scoring of single-cell detection for SV40nls-EGFP and ArgiNLS-EGFP. Gray data points represent iterations after max F1 performance was met. All bar graphs represent mean ± SEM. The F1 data in **h** was statistically analyzed using a two-way repeated measures ANOVA followed by Holm-Sidak's post hoc test. *p<0.05; **p<0.01; **p<0.01; ****p<0.001

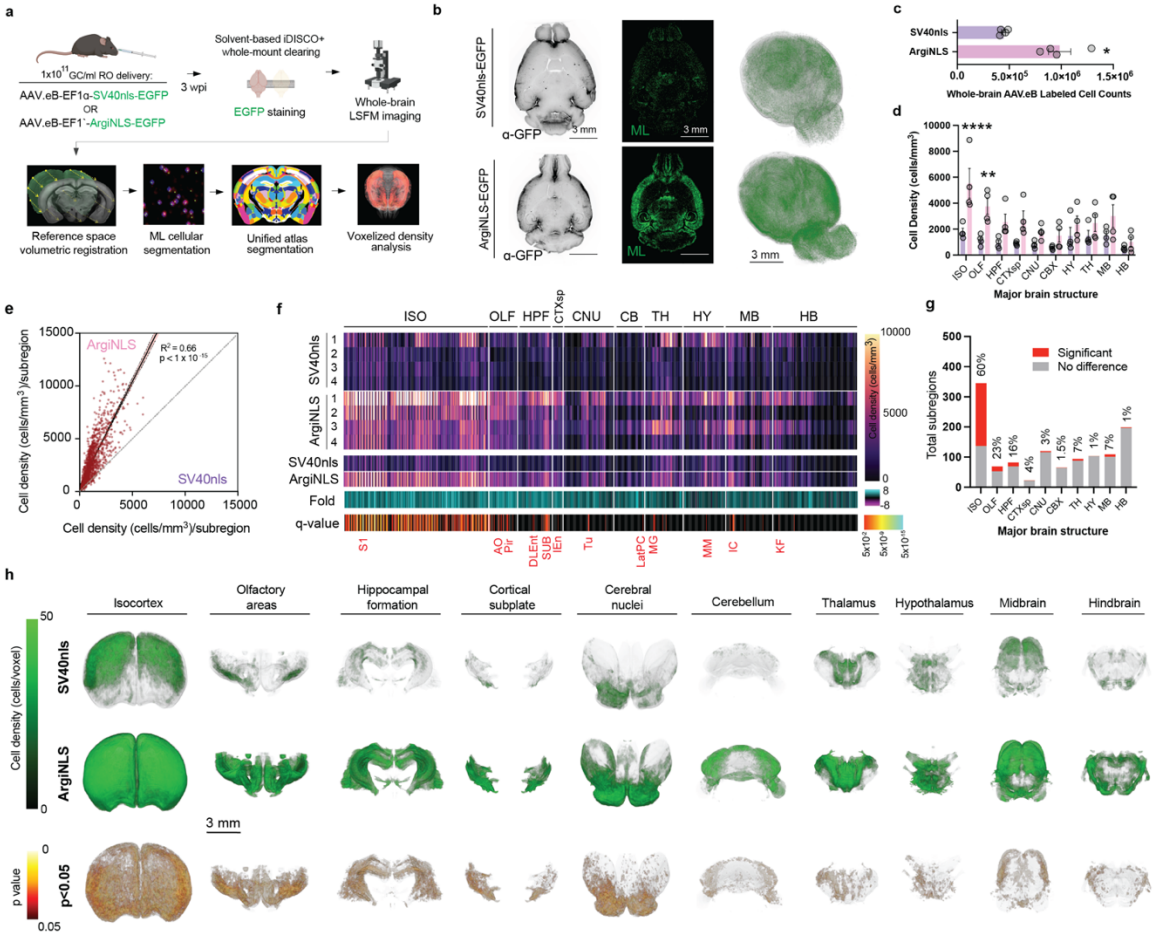
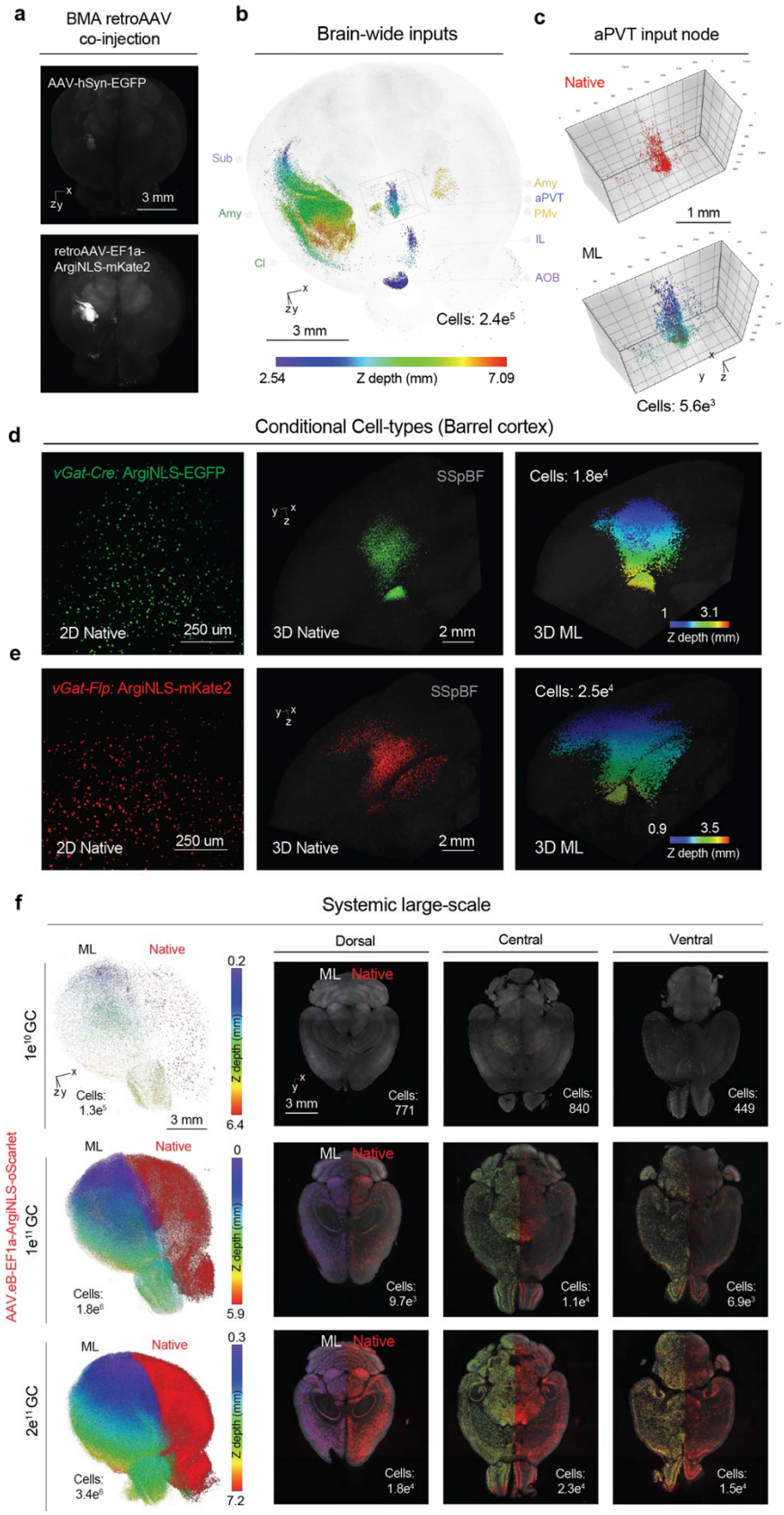


Figure 5. Enriched brain-wide segmentation of single-cells. **a**, Schematic of experimental and image processing workflow used to quantify systemic SV40nls- or ArgiNLS-EGFP-labeled cells across the whole mouse brain. **b,** Left: Raw LSFM image in horizontal orientation from a representative SV40nls-EGFP (top) or ArgiNLS-EGFP (bottom) expressing brain sample. Middle: corresponding ML-segmented cells within the same optical section. Right: Whole-brain 3D-renderings of voxelized cell density for each sample. **c,** Comparison of whole-brain AAV.eB-labeled cell counts. **d,** Comparison of cell density across major ontological brain structures. **e**, Scatterplot and linear regression of ArgiNLS (Y-axis) versus SV40nls (X-axis)-labeled mean cell density across all subdivisions of the brain. **f**, 2D heatmap plot of (from top to bottom): cell density across each individual sample, mean cell density, fold change over ArgiNLS, and statistical significance of mean cell density differences across all anatomical subdivisions by structure order. Major ontological brain structure positioning is indicated with white lines and written above. Example statistically significant subdivisions are listed in red. **g,** Bar graph of total statistically significant subdivisions versus no difference plotted for each major brain structure. Percent significant subregions is listed above each bar. **h**, Volumetric renderings of mean SV40nls-EGFP or ArgiNLS-EGFP voxelized cell densities (top) and the statistically significant voxels of the mean differences (bottom) across all major brain structures. All bar graphs represent mean ± SEM. The data for panel **c** was statistically analyzed using an unpaired Student's t-test. The data for panel **d** was statistically analyzed using a two-way ANOVA with Bonferroni post-hoc test for mean significant differences at each major region. The data for panel **e** was analyzed

with linear regression. The data for panel **f** was analyzed with FDR-corrected multiple t-tests using the two-stage step-up method of Benjamini, Krieger, and Yekutieli. *$p<0.05$; **$p<0.01$. Acronyms: ISO = isocortex; OLF = olfactory areas; HPF = hippocampal formation; CTXsp = cortical subplate; CNU = cerebral nuclei; CBX = cerebellum; HY = hypothalamus; TH = thalamus; MB = midbrain; HB = hindbrain; AO = accessory olfactory area; Pir: piriform cortex; DLEnt: dorsolateral entorhinal cortex; SUB: subiculum; Tu = olfactory tubercle; MG = medial geniculate; MM = medial mammillary area; IC = inferior colliculus; KF = Koelliker-Fuse subnucleus



Szelenyi et al., 2023 30

Figure 6. Optimized volumetric brain cell counting applications with ArgiNLS viral vectors. **a-c,** Example circuit mapping with ArgiNLS-mKate2-expressing retroAAV into the basomedia amygdala (BMA). **a,** Top: Whole-brain volumetric rendering of AAV-hSyn-EGFP native expression labeling the injection site in green channel. Bottom: Corresponding red channel image displaying native fluorescence of ArgiNLS-mKate2+ single-cell inputs. **b,** Volumetric rendering of single-cell brain-wide input segmentations and their total counts using 3D ML, gradient color coded by Z depth. Total BMA input cells are listed in the bottom right. Main anatomical areas of local and long-range inputs are listed to the left and right, respectively. Sub = subiculum; Amy = amygdala; Cl = claustrum; aPVT = anterior paraventricular nucleus of the thalamus; PMv = ventral premammillary nucleus; IL = infralimbic cortex; AOB = accessory olfactory bulb **c,** Single-cells (native fluorescence: top; ML segmentation: bottom) and their total count from the anterior PVT (aPVT) input node cropped and expanded from the outlined inlay in **b**. **d-e,** Conditional labeling of GABAergic single-cells in the barrel cortex using local (**d**) Cre- or (**e**) Flp-dependent AAV injections in *vGAT-Cre* or *vGAT-Flp* driver mice, respectively. Left: 2D confocal image stacks displaying native conditional single-cell labeling from separate example brain samples. Middle: Example volumetric renderings of native conditional single-cell labeling from separate brain samples. Right: Volumetric renderings of 3D ML-segmented single-cells and their total counts for each sample. **f,** Large-scale labeling of single-cells using systemic delivery of AAV.eB-Ef1a-ArgiNLS-oScarlet virus at $1e^{10}$, $1e^{11}$, or $2e^{11}$ GC payloads. Left: Whole-brain volumetric renderings of combined hemispheric native ArgiNLS-oScarlet fluorescence (left hemisphere) and 3D ML segmentation (right hemisphere) of single-cells from separate example brain samples per viral condition. Total hemispheric cell counts are listed in the bottom left. Right: 2D horizontal planes of each sample across (top to bottom) dorsal, central, and ventral positions. Images display native ArgiNLS-oScarlet fluorescence with overlaid 3D ML classification on the right hemisphere and total cell counts per plane listed in the bottom right.