Title: PESCA: A scalable platform for the development of cell-type-specific viral drivers

One sentence summary: Highly paralleled functional evaluation of enhancer activity in single cells generates new cell-type-specific tools with broad medical and scientific applications.

Authors:

Sinisa Hrvatin[1, *], Christopher P. Tzeng[1], M. Aurel Nagy[1], Hume Stroud[1], Charalampia Koutsioumpa[1, 2], Oren F. Wilcox[1], Eric C. Griffith[1], Michael E. Greenberg[1, *]

Affiliations:

1. Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, USA
2. BBS Program, Harvard Medical School, 220 Longwood Avenue, Boston, MA 02115, USA

*Co-corresponding authors:

Sinisa Hrvatin (sinisa_hrvatin@hms.harvard.edu),

Michael E. Greenberg (michael_greenberg@hms.harvard.edu)

Abstract:

Enhancers are the primary DNA regulatory elements that confer cell type specificity of gene expression, thus giving rise to the hundreds of different cell types that make up a complex organism. Recent studies characterizing individual enhancers have revealed their potential to direct heterologous gene expression in a highly cell-type-specific manner. However, it has not yet been possible to systematically identify such enhancers for each of the many cell types in an organism. We have developed PESCA, a scalable and generalizable method coupling ATAC and single-cell RNA sequencing protocols, to identify cell-type-specific enhancers that should enable genetic access and perturbation of gene function in virtually any mammalian cell type. Focusing on the highly heterogeneous mammalian cerebral cortex, we apply PESCA to find enhancers and generate viral reagents capable of accessing and manipulating a rare subset of somatostatin-expressing cortical interneurons with high specificity. This study demonstrates the utility of this readily generalizable platform for developing new cell-type-specific viral reagents, with significant implications for both basic and translational research.

Main text:

Introduction

Enhancers are DNA elements that regulate gene expression to produce the unique complement of proteins necessary to establish a specialized function for each cell type in an organism. Large scale efforts to build a definitive catalogue of cell types(*1–4*) based on their gene expression have only recently been met with attempts to map epigenomic regulatory landscapes as well(*5–7*), enabling a mechanistic understanding of the underlying gene expression that is critical for cell-type-specific development, identity, and unique function. Importantly, characterization of individual enhancers has revealed their potential to direct highly cell-type-specific gene expression in both endogenous and heterologous contexts(*8–11*), making them ideal for developing tools to access, study, and manipulate virtually any mammalian cell type.

Indeed, despite being able to transcriptionally catalogue distinct cell subpopulations, our efforts to study their function are fundamentally impeded by our limited ability to specifically access them. Nowhere is this more apparent than in the mammalian cerebral cortex, which is composed of a staggeringly intricate and interconnected ensemble of cell types that together enable the complex computations that the brain must make. Glutamatergic excitatory neurons propagate signal across neural circuits, while GABAergic inhibitory interneurons play an essential role in cortical signal processing by modulating neuronal activity, balancing excitability, and gating information(*12–14*). Although relatively sparse, interneurons are highly diverse; for example, somatostatin-expressing cortical interneurons comprise several anatomically, electrophysiologically, and molecularly defined cell types whose dysfunction is associated with neuropsychiatric and neurological disorders(*2*, *15*, *16*). Given the vast diversity of cell types in the brain, and the inability of our current transgenic tools to access most neuronal

70 cell types, enhancer-driven viral reagents have the potential to become the next generation of
71 cell-type-specific transgenic tools enabling, for the first time, facile, inexpensive, cross-species,
72 and targeted observation and functional study of neuronal cell-types and circuits.

73 Despite the potential of cell-type-specific enhancers to revolutionize neuroscience
74 research, cell-type-restricted gene regulatory elements (GREs) have not yet been systematically
75 identified. Moreover, functional evaluation of candidate GRE-driven viral vectors is currently
76 laborious, expensive, and low-throughput, typically relying on the production of individual viral
77 vectors and the assessment of expression across a limited number of cell types by *in situ*
78 hybridization or immunofluorescence. The lack of a high-throughput platform for rapid
79 identification and functional testing of enhancers is therefore a critical bottleneck impeding the
80 generation of new cell-type-specific viral reagents required to elucidate the function of each cell
81 type in a complex organism.

82 To address these issues, we developed a scalable Paralleled Enhancer Single Cell Assay
83 (PESCA) to identify and functionally assess the specificity of hundreds of GREs across the full
84 complement of cell types present in any tissue of interest. In the PESCA protocol, the expression
85 of a barcoded pool of AAV vectors harboring GREs is analyzed by single-nucleus RNA
86 sequencing (snRNA-seq) to evaluate the specificity of each constituent GRE across tens of
87 thousands of individual cells in the target tissue, through the use of an orthogonal cell-indexed
88 system of transcript barcoding (**Fig. 1a**).

89 We validated the efficacy of PESCA in the murine primary visual cortex by identifying
90 GREs that have the ability to confine AAV expression to a layer-restricted, electrophysiologically
91 distinct subset of somatostatin (SST)-expressing interneurons, and showed that these vectors can
92 be used to modulate neuronal activity in a cell-type-specific manner. We chose to focus on SST

93 neurons in the brain because this population is known to be diverse and to be composed of
94 relatively rare subpopulations(*2*, *15*, *17*), and thus might serve as a good test case for the efficacy
95 of PESCA. As described below, these results highlight the utility of PESCA for identifying an
96 enhancer that drives gene expression selectively in a subset of neurons and establish the use of
97 PESCA as a platform of broad interest to the research and gene therapy community, enabling the
98 generation of cell-type-specific AAVs for virtually any cell type.
99

100 **Results**

101 **GRE selection and library construction**

102 To identify candidate SST interneuron-restricted gene regulatory elements (GREs), we
103 carried out comparative epigenetic profiling of the three largest classes of cortical interneurons,
104 somatostatin (SST)-, vasoactive intestinal polypeptide (VIP)- and parvalbumin (PV)-expressing
105 cells. To this end, we employed the recently developed isolation of nuclei tagged in specific cell
106 types (INTACT)(*5*) method to isolate purified chromatin from of each of these cell types from
107 the cerebral cortex of adult (6-10-week-old) mice. Assay for transposase-accessible chromatin
108 using sequencing (ATAC-Seq)(*7*), which marks nucleosome-depleted gene regulatory regions
109 based on their enhanced accessibility to *in vitro* transposition by the Tn5 transposase, was then
110 used to identify genomic regions with enhanced accessibility in the SST (n = 279,221), PV (n =
111 275,631), and VIP (n = 258,646) chromatin samples. These datasets can be used as a resource to
112 identify putative gene regulatory elements as candidates for driving cell-type-specific gene
113 expression for the numerous SST, PV or VIP-expressing intraneuronal subtypes across diverse
114 cortical regions.

115 Among these putative gene regulatory regions, 16,386 (5.9%) were enriched or uniquely
116 present in SST cells relative to PV or VIP-expressing interneurons (**Fig. 1b, c**). To enrich for
117 GREs that thus could be useful reagents to study and manipulate interneurons across mammalian
118 species including in humans, we subsequently filtered the resulting list to exclude GREs with
119 poor mammalian sequence conservation (**Methods**, **Supplementary Fig. 1**). Remaining
120 elements were ranked based on cell-type-enrichment (**Methods**), with the top 287 SST-enriched
121 GREs selected for screening to identify enhancers that drive gene expression selectively in SST
122 interneurons of the primary visual cortex (**Fig. 1d**).
123 A PCR-based strategy was used to simultaneously amplify and barcode each GRE from
124 mouse genomic DNA (**Methods**). To minimize sequencing bias due to the choice of barcode
125 sequence, each GRE was paired with three unique barcode sequences. The resulting library of
126 861 GRE-barcode pairs was pooled and cloned into an AAV-based expression vector, with the
127 GRE element inserted 5' to a minimal promoter driving a GFP expression cassette and the GRE-
128 paired barcode sequences inserted into the 3' untranslated region (UTR) of the GRE-driven
129 transcript (**Methods, Fig. 2a, Supplementary Fig. 2)**. This configuration was chosen to
130 maximize the retrieval of the barcode sequence during single-cell RNA sequencing. The library
131 was packaged into AAV9, which exhibits broad neural tropism(*18*). The complexity of the
132 resulting rAAV-GRE library was then confirmed by Next Generation Sequencing, detecting 802
133 of the 861 barcodes (93.2%), corresponding to 285 of the 287 GREs (99.3%) (**Fig. 2b**).
134
135 **PESCA screen identifies GREs highly enriched for SST$^+$ interneurons**
136 To quantify the expression of each rAAV-GRE vector across the full complement of cell types in
137 the mouse visual cortex, we used a modified single-nucleus RNA-Seq (snRNA-Seq) protocol to

138 first determine the cellular identity of each nucleus and then quantify the abundance of the GRE-
139 paired barcodes in the transcriptome of nuclei assigned to each cell type. Two injections (800 nL
140 each) of the pooled AAV library (1×10^{13} viral genomes/mL) were first administered to the
141 primary visual cortex (V1) of two 6-week-old C57BL/6 mice. Twelve days following injection,
142 the injected cortical regions were dissected and processed to generate a suspension of nuclei for
143 snRNA-Seq using the inDrops platform(*19*, *20*) (**Methods**). A total of 32,335 nuclei were
144 subsequently analyzed across the two animals, recovering an average of 866 unique non-viral
145 transcripts per nucleus, representing 610 unique genes (**Supplementary Fig. 3a, b**).
146 Since droplet-based high-throughput snRNA-Seq samples the nuclear transcriptome with
147 low sensitivity(*19*), viral-derived transcripts were initially detected in only 3.9% of sampled
148 nuclei. We therefore designed a modified PCR-based approach to enrich for barcode-containing
149 viral transcripts, which yielded deep coverage of AAV-derived transcripts with simultaneous
150 shallow coverage of the non-viral transcriptome. PCR enrichment increased the viral transcript
151 recovery 382-fold in the sampled nuclei, to an average of 15.6 unique viral transcripts, 6.0
152 unique GRE-barcodes, and 5.7 unique GREs per cell **(Fig. 2b, Supplementary Fig. 3c)**. Using
153 this modified protocol, viral transcripts were identified across 86% of cells (**Supplementary Fig.
154 3d, e**), with a high correlation ($r = 0.9$, $p < 2.2\times10^{-16}$) observed between the abundance of each
155 barcoded AAV in the library and the number of cells infected by that AAV (**Supplementary
156 Fig. 3f**).
157 Nuclei were classified into ten cell types using graph-based clustering and expression of
158 known marker genes **(Methods, Fig. 2c, d, Supplementary Fig. 4**). The average expression of
159 each viral-derived barcoded transcript was analyzed across all ten cell types, and an enrichment
160 score was calculated from the ratio of expression in *Sst*$^+$ nuclei compared to all *Sst*$^-$ nuclei. As

161 expected, sets of three barcodes associated with the same GRE showed highly statistically
162 correlated enrichment scores ($r = 0.53 \pm 0.03$, $p < 2.2\times10^{-16}$) (**Fig. 2e, f, Supplementary Fig. 5**),
163 which were abolished when barcodes were randomly shuffled (shuffled $r = 0.002 \pm 0.06$; Wilcox
164 test between data and shuffled data, $p = 0.003$).
165 Having confirmed a robust, non-random correlation in enrichment scores among the three
166 barcodes associated with each GRE, we next computed a single expression value for each of the
167 287 viral drivers by aggregating expression data from barcodes associated with the same GRE,
168 and carried out differential gene expression analysis between *Sst*$^+$ and *Sst*$^-$ cells for each rAAV-
169 GRE. Differential gene expression analysis between *Sst*$^+$ and *Sst*$^-$ cells for each rAAV-GRE
170 revealed a marked overall enrichment of viral-derived transcripts in the *Sst*$^+$ subpopulation
171 (**Supplementary Fig. 6a**). Indeed, multiple viral drivers were identified that promoted highly
172 specific reporter expression in the *Sst*$^+$ subpopulation ($q < 0.01$, fold-change > 7, **Fig. 2g-i,**
173 **Supplementary Fig. 6b**).
174
175 ***In situ* characterization of rAAV-GRE reporter expression**
176 We next sought to validate the cell-type-specificity of the resulting hits using methods
177 that do not rely on single-cell sequencing-based approaches. To this end, we selected three of the
178 top five viral drivers (GRE12, GRE22, GRE44), as well as a control viral construct lacking the
179 GRE element (ΔGRE), for injection into V1 of adult transgenic Sst-Cre; Ai14 mice, in which
180 SST$^+$ cells express the red fluorescent marker tdTomato. Fluorescence analysis twelve days
181 following injection with rAAV-GRE12/22/44-GFP revealed strong yet sparse GFP labeling
182 centered around cortical layers IV and V (**Fig. 3a-c**). By contrast, the control rAAV-ΔGRE-GFP
183 showed a strikingly different pattern of GFP expression concentrated around the sites of

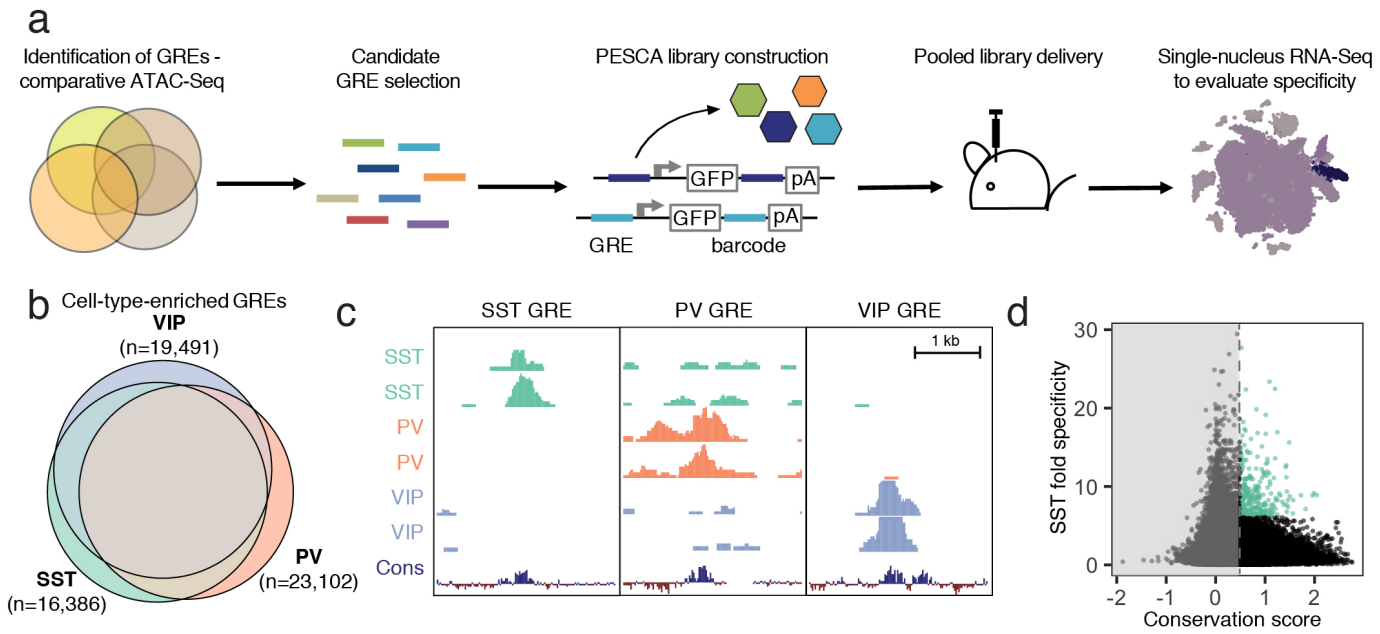
184 injection, with expression in a larger number of cells (**Fig. 3d**). Many rAAV-GRE12/22/44-GFP
185 virally infected cells were indeed SST-positive, marked by the high degree of overlapping GFP
186 and tdTomato expression: 90.7% ± 2.1% for rAAV-GRE12-GFP (170 cells, 4 animals); 72.9 ±
187 4.2% for rAAV-GRE22-GFP (1164 cells, 3 animals), and 95.8 ± 0.6% for rAAV-GRE44-GFP
188 (759 cells, 4 animals). (**Fig. 3e, f, Supplementary Fig. 7**). By contrast, we observed that only
189 27.2 ± 1.9% of GFP^+ cells following rAAV-ΔGRE-GFP infection were also positive for
190 tdTomato expression (2066 cells, 3 animals, **Fig. 3e, f**), indicating that the tested GREs serve to
191 effectively restrict AAV payload expression to SST^+ interneurons. It is notable that the GREs
192 seemingly not only promote expression in SST^+ cells but also reduce background expression in
193 SST^- cells, indicating the tested GREs confer both enhancer and insulator functionality.
194 Consistent with this hypothesis, the incorporation of the GREs into the rAAV both increased the
195 number of SST^+/GFP^+ cells (1.7-2-fold) and dramatically (3-32-fold) decreased the number of
196 SST^- cells that expressed GFP (**Fig. 3g, Supplementary Fig. 8**). To further investigate the
197 specificity of our viral drivers among cortical interneuron cell types we injected each construct
198 into Vip-Cre; $Ai14^+$ mice in which all VIP^+ cells express tdTomato, or used fluorescence
199 antibody staining to label PV-expressing cells (**Supplementary Fig. 9)**. Fluorescent signal
200 analysis indicated the percentage of GFP^+ cells that were either VIP^+ or PV^+ (rAAV-SST12-
201 GFP^+ [2.6 ± 2.6%], rAAV-GRE22-GFP^+ [3.5 ± 2.0%] and rAAV-GRE44-GFP^+ [6.0 ± 2.7%],
202 **Fig. 3h**). This confirms that among major interneuron cell classes, all three vectors are highly
203 SST-specific.
204 Because at least five subtypes of cortical SST^+ interneurons have been identified based
205 on the laminar distribution of their cell bodies and projections(*15*, *21*), we also investigated the
206 laminar distribution of GFP-expressing cells for the three Sst-enriched viral drivers. Intriguingly,

207 the majority of rAAV-GRE12-GFP$^+$ and rAAV-GRE44-GFP$^+$ SST$^+$ cells were found to reside in
208 layers IV and V, distinct from the distribution observed for the full SST$^+$ cell population in visual
209 cortex ($p = 1.3\times10^{-6}$, $p < 2.2\times10^{-16}$, respectively, Mann-Whitney *U* test, two-tailed, **Fig. 3i, j**),
210 raising the possibility that these constructs may preferentially label a specific subtype(s) of SST$^+$
211 interneuron. Consistent with this hypothesis, we observed that these two viral drivers mediated
212 reporter expression in only a relatively small fraction of all SST$^+$ cells within the region of
213 infection (44.5 ± 12.0% for rAAV-GRE12-GFP and 35.9 ± 6.2% for rAAV-GRE44-GFP)
214 compared to rAAV-GRE22-GFP (**Fig. 3k, Supplementary Fig. 10**). Together, these findings
215 suggest that PESCA may support the isolation of viral drivers capable of discriminating between
216 fine-grained cell-types within a given interneuron cell class.
217
218 **Electrophysiological characterization of rAAV-GRE-GFP-expressing SST subtypes**
219 In addition to variability in laminar distribution, different electrophysiological
220 phenotypes have also been observed in cortical SST interneurons(*22*, *23*). To determine whether
221 AAV-GRE reporters can be used to distinguish electrophysiologically distinct SST subtypes, we
222 injected rAAV-GRE44-GFP into the visual cortex of adult Sst-Cre; Ai14 mice and performed
223 whole-cell current-clamp recordings from double GFP and tdTomato positive neurons (rAAV-
224 GRE44-GFP$^+$), as well as immediately nearby tdTomato positive but GFP negative cells (rAAV-
225 GRE44-GFP$^-$). Our recordings indicate that both rAAV-GRE44-GFP$^+$ and rAAV-GRE44-GFP$^-$
226 neurons are adapting SST interneurons with high input resistances and features consistent with
227 deep layer cortical SST neurons(*22*, *24*) (**Fig. 4a, b**).
228 However, rAAV-GRE44-GFP$^+$ SST neurons are distinct in a several electrophysiological
229 parameters. Action potentials of rAAV-GRE44-GFP$^+$ SST neurons are significantly broader

230 than those of rAAV-GRE44-GFP$^-$ SST neurons (**Fig. 4c, d**), perhaps due to differences in
231 expression of certain channels in these subgroups of SST neurons, such as voltage-activated
232 potassium channels, and BK calcium-activated potassium channels(*25*, *26*). Furthermore,
233 rAAV-GRE44-GFP$^+$ SST neurons have a lower rheobase, and fire action potentials with a slower
234 rising phase, and at lower maximal frequencies compared to rAAV-GRE44-GFP$^-$ SST neurons
235 **(Fig. 4a, d, Supplementary Table 1)**. Our electrophysiology experiments further illuminate the
236 potential of PESCA to identify novel, functionally distinct subgroups of previously defined
237 interneuron types.
238
239 **Modulation of neuronal activity with rAAV-GREs**
240 Finally, we evaluated whether the identified viral drivers support sufficiently high and
241 persistent levels of payload expression to effectively modulate SST$^+$ cell physiology. Designer
242 receptors exclusively activated by designer drugs (DREADDs) are a commonly employed viral
243 payload to dynamically regulate neuronal activity in response to the synthetic ligand clozapine-
244 N-oxide (CNO)(*27*). We therefore injected the visual cortex of adult wild-type mice (6-8-week-
245 old) with rAAV-GRE12-Gq-DREADD-tdTomato and performed electrophysiological recordings
246 from tdTomato$^+$ cells of acute cortical slices in a whole-cell, current-clamp configuration two
247 weeks post-injection. All recordings from tdTomato$^+$ cells evoked with depolarizing current
248 steps showed striking sensitivity to CNO, as shown by significantly increased firing rates and
249 depolarized resting membrane potentials during bath application of CNO (**Fig. 4e-g**). To ensure
250 that increases in firing rate upon CNO application were specific to infected SST$^+$ neurons, we
251 performed recordings from nearby uninfected pyramidal neurons that were identified by
252 morphology and found that there was no statistically significant increase in firing rate upon CNO

253 application (**Fig. 4h-j**). These data demonstrate the ability of these reagents to robustly and
254 specifically modulate the activity of SST^+ cells in non-transgenic animals.
255
256 **Discussion**
257 The PESCA platform merges the principle of massively paralleled reporter assays (MPRA)(*28*,
258 *29*) with scRNA-seq and represents a significant advancement in current approaches to viral
259 vector design, as it enables the rapid screening of hundreds of viral permutations for enhanced
260 cell-type-specificity. In this study, we applied PESCA to screen putative enhancer elements for
261 drivers that robustly and specifically target a rare SST^+ population of GABAergic interneurons in
262 the mouse central nervous system. Given our understanding of the generality of enhancer
263 function across tissues, this approach could be readily applied to virtually any other neuronal or
264 non-neuronal cells types, diverse model organisms, tissues, and viral types. Moreover, PESCA is
265 not limited to GRE screening; the method can be easily adapted to assess the cell-type-specificity
266 of viral capsid variants or other mutable aspects of viral design. This study addresses the urgent
267 need for new tools to access, study, and manipulate specific cell types across complex tissues,
268 organ systems, and animal models, demonstrating the importance and broad utility of PESCA
269 beyond neuroscience research. Moreover, as the promise of gene therapy to treat and cure a
270 broad range of diseases is being realized, PESCA has the potential to pave the way for a new
271 generation of targeted gene therapy vehicles for diseases with cell-type-specific etiologies, such
272 as congenital blindness, deafness, cystic fibrosis, and spinal muscular atrophy.

273 **Figures and figure legends**

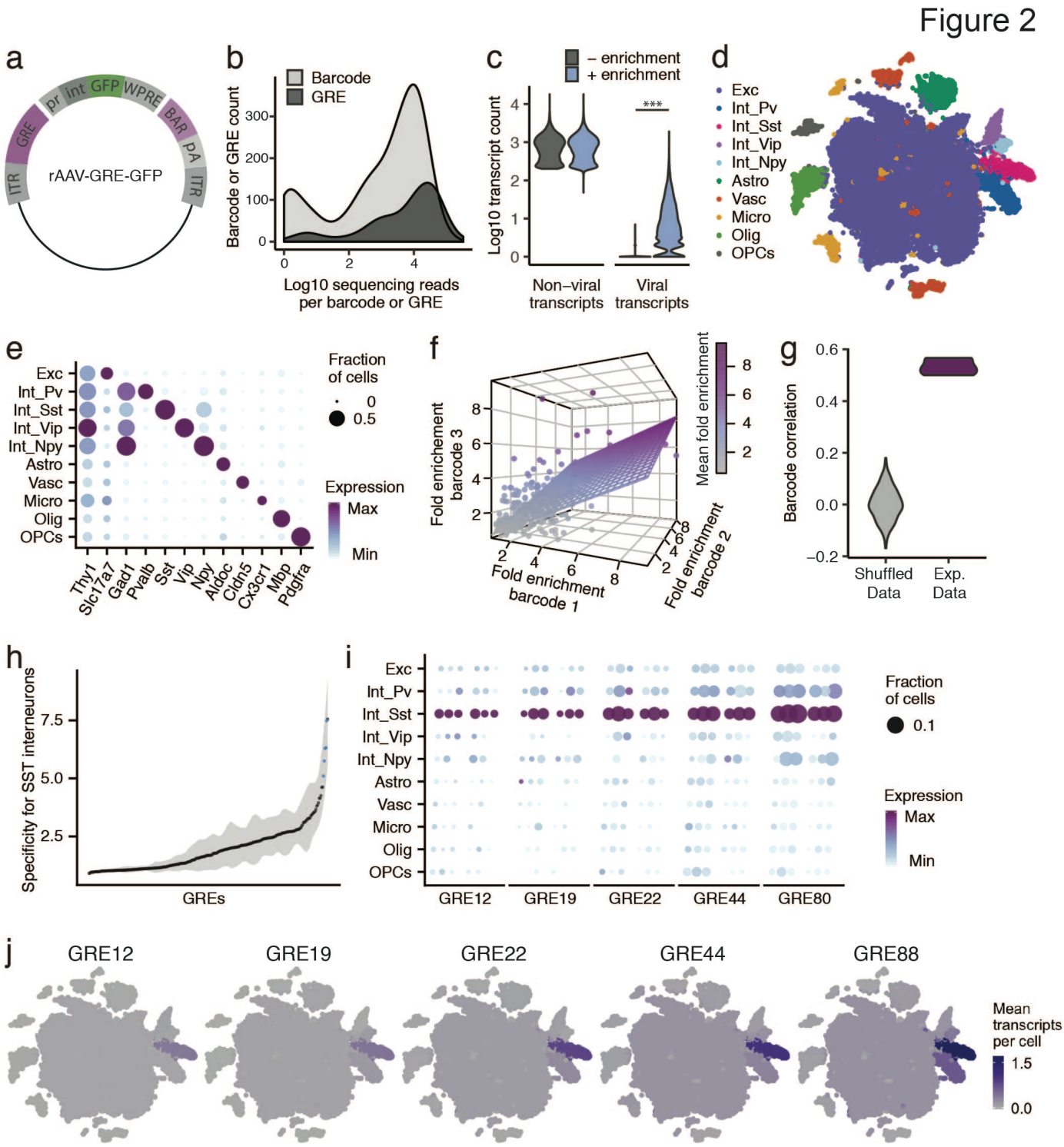


274 **Figure 1 – Experimental strategy and GRE selection**

275 (**a**) Paralleled Enhancer Single Cell Assay (PESCA). Comparative ATAC-Seq is used to
276 identify candidate GREs. A library of gene regulatory elements (GREs) is inserted upstream of a
277 minimal promoter-driven GFP. The viral barcode sequence is inserted in the 3'UTR, and the
278 vector packaged into rAAVs. Following *en masse* injection of the rAAV library, the specificity
279 of the constituent GREs for various cell types *in vivo* is determined by single-nucleus RNA
280 sequencing, measuring expression of the barcoded transcripts in tens of thousands of individual
281 cells in the target tissue. Finally, bioinformatic analysis determines the most cell-type-specific
282 barcode-associated rAAV-GRE-GFP constructs. pA = polyA tail.

283 (**b**) Area-proportional Venn diagram of the number of putative GREs identified by
284 ATAC-Seq of purified PV, SST, and VIP interneuron chromatin. Overlapping areas indicate

285 shared putative GREs. Non-overlapping areas represent GREs that are either unique or strongly
286 enriched in a single cell type.
287 (**c**) Representative ATAC-seq genome browser traces of a putative GRE enriched in SST-
288 , PV-, or VIP-interneurons. Sequence conservation across the Placental mammalian clade is also
289 shown.
290 (**d**) Putative GREs (n = 323,369) are plotted based on average sequence conservation
291 (phyloP, 60 placental mammals) and SST-specificity (ratio of the average ATAC-Seq signal
292 intensity between SST samples and non-SST samples). Dashed vertical line indicates the
293 minimal conservation value cutoff (0.5). Green coloring indicates the 287 most SST-specific
294 GREs selected for PESCA screening.



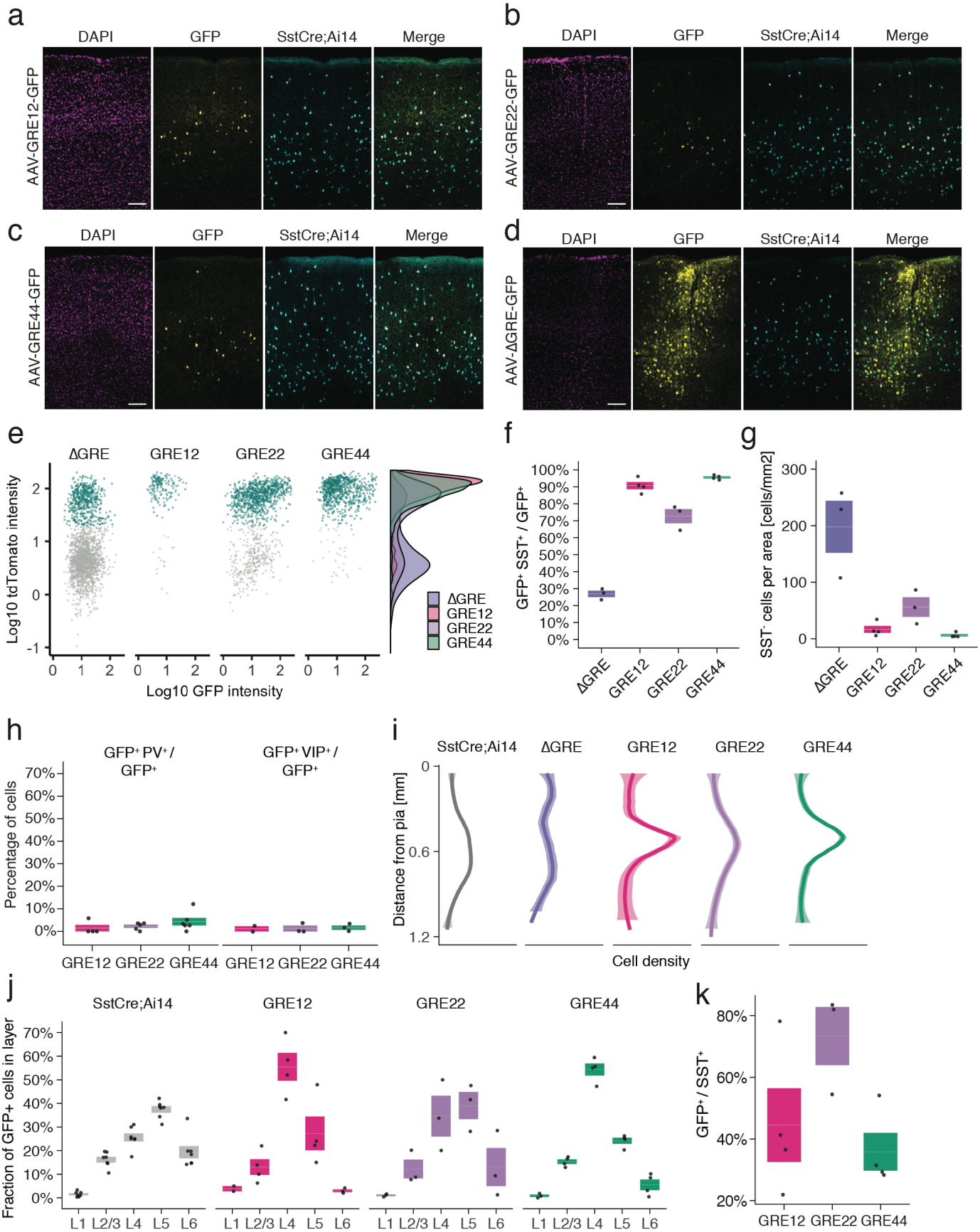
295 **Figure 2 – PESCA screen identifies GREs highly enriched for SST^{+} interneurons**

296 (**a**) PESCA library plasmid map. ITR, inverted terminal repeats; GRE, gene regulatory

297 element; pr, HBB minimal promoter; int, intron; GFP, green fluorescent protein; WPRE,

298 Woodchuck Hepatitis Virus post-transcriptional regulatory element; BAR, 10-mer sequence
299 barcode associated with each GRE; pA, polyadenylation signal.
300 (**b**) Library complexity plotted as distribution of the abundance of the 861 barcodes and
301 287 GREs in the AAV library. Barcodes and GREs were binned by number of sequencing reads
302 attributed to each barcode or GRE within the library.
303 (**c**) Transcript count per nucleus (n = 32,335 nuclei). Sequencing libraries were prepared
304 with or without PCR-enrichment for viral transcripts. PCR enrichment resulted in a 382-fold
305 increase in the number of recovered viral transcripts (p = 0) to an average of 15.6 unique viral
306 transcripts per nucleus. Displayed as Log10(Count+1).
307 (**d**) t-SNE plot of 32,335 nuclei from V1 cortex of two animals**.** Colors denote main cell
308 types: Exc (Excitatory neurons), Pv (PV Interneurons), Sst (SST Interneurons), Vip (VIP
309 interneurons), Npy (NPY Interneurons), Astro (Astrocytes), Vasc (Vascular-associated cells),
310 Micro (Microglia), Olig (Oligodendrocytes), OPCs (Oligodendrocyte precursor cells).
311 (**e**) Marker gene expression across cell types. Color denotes mean expression across all
312 nuclei normalized to the highest mean across cell types. Size represents the fraction of nuclei in
313 which the marker gene was detected.
314 (**f**) Three-dimensional dot plot with each dot representing one GRE (n = 287). The values
315 on each axis represent the SST fold-enrichment calculated for each GRE based on the three
316 barcodes paired with that GRE. Plane of correlation between the enrichment values calculated
317 from three sets of barcodes associated with 287 GREs ($r = 0.53 \pm 0.03$, $p < 2.2\times10^{-16}$). Color
318 gradient indicates the average enrichment between the three barcodes.

319 (**g**) Pairwise Pearson correlation between the enrichment values calculated from three sets
320 of barcodes associated with 287 GREs for experimental data (Exp. Data, $r = 0.53 \pm 0.03$, $p <$
321 2.2×10^{-16}) and after random shuffling of enrichment values (Shuffled Data, $r = 0 \pm 0.06$).
322 (**h**) GREs ranked by average expression specificity for SST interneurons. Shading
323 indicates the minimal and maximal specificity calculated by analyzing each of the three barcodes
324 associated with a GRE. Blue indicates the five top hits that also passed a statistical test for SST
325 interneuron enrichment (FDR-corrected $q < 0.01$).
326 (**i**) Expression of the top five hits: GRE12, GRE19, GRE22, GRE44, GRE80. For each
327 GRE, expression values are split into two animals, and, for each animal, into the three barcodes
328 associated with that GRE. Color denotes mean expression across all nuclei normalized to the
329 highest mean across cell types. Size represents the fraction of nuclei in which the marker gene
330 was detected.
331 (**j**) t-SNE plot of 32,335 nuclei from V1 cortex of two animals, showing expression of
332 GRE12, GRE19, GRE22, GRE44, and GRE80**.** Plot is pseudocolored based on the mean GRE
333 expression in each cell type.
334


a
DAPI
GFP
SstCre;Ai14
Merge
AAV-GRE12-GFP
b
AAV-GRE22-GFP
c
AAV-GRE44-GFP
d
AAV-ΔGRE-GFP
e
ΔGRE
GRE12
GRE22
GRE44
Log10 tdTomato intensity
Log10 GFP intensity
f
GFP+ SST+ / GFP+
g
SST- cells per area [cells/mm2]
h
GFP+ PV+ / GFP+
GFP+ VIP+ / GFP+
Percentage of cells
i
SstCre;Ai14
Distance from pia [mm]
Cell density
j
Fraction of GFP+ cells in layer
L1 L2/3 L4 L5 L6
k
GFP+ / SST+

335 **Figure 3 – *In situ* characterization of rAAV-GRE reporter expression**

336 (**a-d**) Fluorescent images from adult Sst-Cre; Ai14 mouse visual cortex twelve days
337 following injection with rAAV-GRE-GFP as indicated. Scale bars 100 µm.

338 (**e**) Identification of rAAV-GRE-GFP$^+$ cells that express tdTomato (SST$^+$). Each dot
339 represents a GFP$^+$ cell (n = 2066, 172, 1164, and 765, for AAV-[ΔGRE, GRE12, GRE22,
340 GRE44]-GFP, respectively). Cyan indicates tdTomato$^+$ (SST$^+$) cells. Distribution of cell
341 frequency across tdTomato intensity is plotted on the right for each construct.

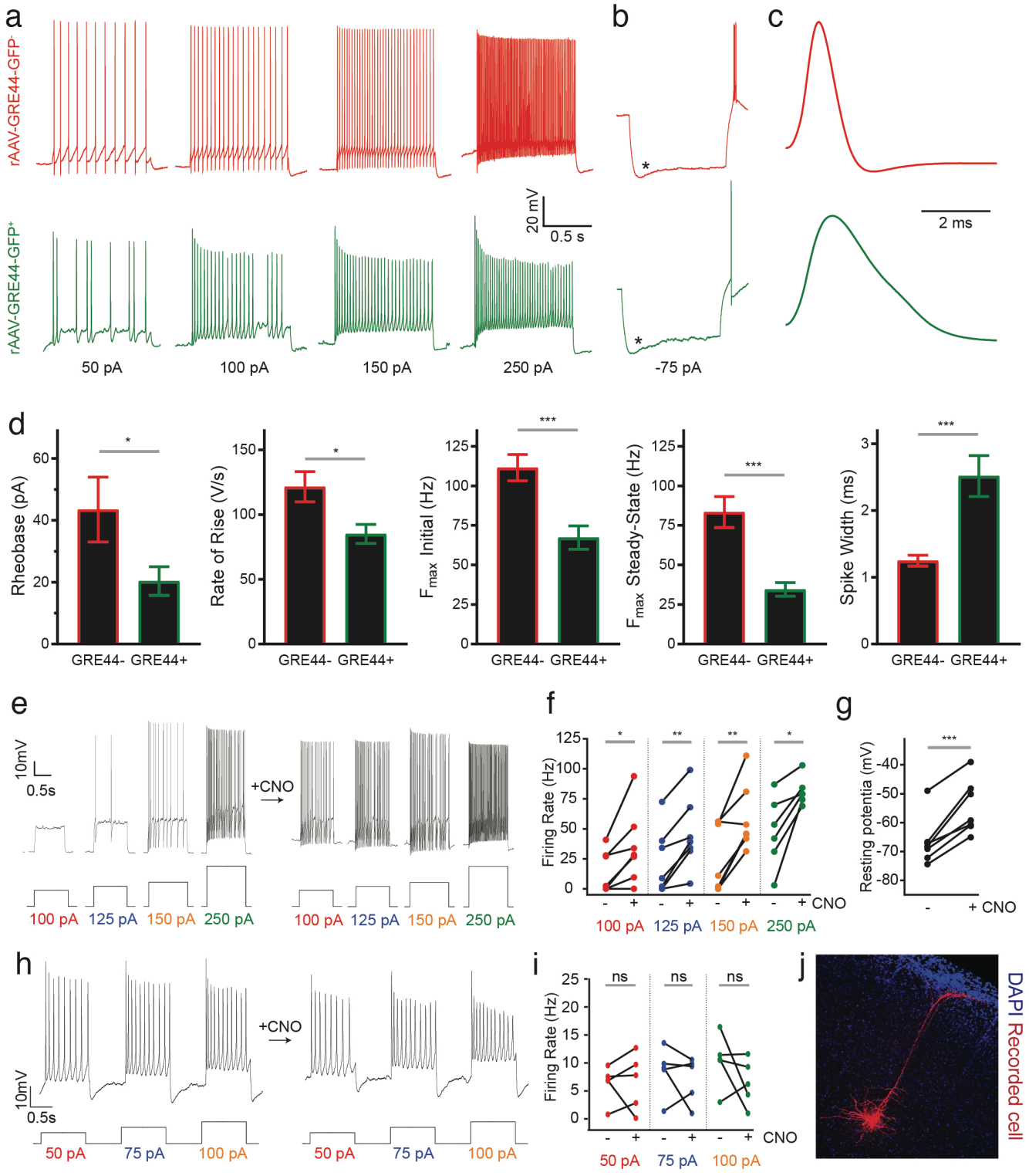
342 (**f**) Quantification of the fraction of GFP$^+$ cells that are SST$^+$. Each dot represents one
343 animal. Box plot represents mean ± standard error of the mean (s.e.m). Values are 27.2 ± 1.9%,
344 90.7 ± 2.1, 72.9 ± 4.2%, and 95.8 ± 0.6% for AAV-[ΔGRE, GRE12, GRE22, GRE44]-GFP,
345 respectively.

346 (**g**) Quantification of the number of GFP$^+$ SST$^-$ cells normalized for area of infection.
347 Each dot represents one animal. Box plot represents mean ± standard error of the mean (s.e.m).
348 Values are 198.0 ± 46.0, 16.4 ± 6.2, 56.0 ± 17.3 and 6.1± 2.1 cells/mm^2 for AAV-[ΔGRE,
349 GRE12, GRE22, GRE44]-GFP, respectively.

350 (**h**) Quantification of the fraction of GFP$^+$ cells that are PV$^+$ or VIP$^+$. Box plot represents
351 mean ± standard error of the mean (s.e.m). Fraction of AAV-GRE-GFP$^+$ cells that are PV$^+$ is 1.4
352 ± 1.4%, 2.2 ± 0.7, and 4.3 ± 1.7% for AAV-[GRE12, GRE22, GRE44]-GFP, respectively.
353 Similarly, the fraction of AAV-GRE-GFP$^+$ cells that are VIP$^+$ is 1.2 ± 1.2%, 1.3 ± 1.3%, and 1.7
354 ± 1.0% for AAV-[GRE12, GRE22, GRE44]-GFP$^+$ cells, respectively.

355 (**i**) Distribution of the location of GFP-expressing cells as function of distance from the
356 pia. Gray represents SST$^+$ cells (n = 2648); Colored lines represents GFP$^+$ SST$^+$ cells (n = 2066,

357 172, 1164, and 765, respectively, for AAV-[ΔGRE, GRE12, GRE22, GRE44]-GFP). Shading
358 represents the 95% confidence interval.
359 (**j**) Quantification of the fraction of GFP^+ cells that are present it each cortical layer. Each
360 dot represents one animal. Box plot represents mean ± standard error of the mean (s.e.m). Gray
361 represents all SST^+ cells, colored plots represent GFP^+ SST^+ cells respectively, for AAV-
362 [GRE12, GRE22, GRE44]-GFP).
363 (**k**) Quantification of the fraction of Sst-Cre; $Ai14^+$ cells within the infection area that are
364 GFP^+. Each dot represents one animal. Box plot represents mean ± standard error of the mean
365 (s.e.m). Values are 44.5 ± 12.0%, 73.4 ± 9.4%, and 35.9 ± 6.2% for AAV-[GRE12, GRE22,
366 GRE44]-GFP, respectively.



367 **Figure 4 – Electrophysiology of neurons expressing an rAAV-GRE-driven reporter and**

368 **modulation of neuronal activity with rAAV-GREs**

369 (**a**) Representative current-clamp recordings from SST neurons in the visual cortex of
370 Sst-Cre; Ai14 mice injected with rAAV-GRE44-GFP. *Top*: Representative traces from a cortical
371 SST neuron with Cre-dependent expression of red fluorescent protein (RFP), in response to
372 1000ms depolarizing current injections as indicated in black ("GRE44-"). *Bottom*: Traces from a
373 RFP+ SST neuron with GRE44-driven expression of GFP ("GRE44+"). GRE44- SST neurons
374 were only recorded in the immediate vicinity of GRE44+ SST neurons.
375 (**b**) Recordings from GRE44+ and GRE44- neurons in response to hyperpolarizing,
376 1000ms currents. Asterisks indicate the sag likely due to the hyperpolarization-activated current
377 I_h. Rebound action potentials following recovery from hyperpolarization, likely due to low-
378 threshold calcium spikes mediated by T-type calcium channels, were also present in cells of both
379 groups. Same scale as (a).
380 (**c**) Broader action potentials in GRE44+ SST neurons (bottom) compared to GRE44-
381 SST neurons (top). Same vertical scale as (a) and (b).
382 (**d**) Electrophysiological properties that differ between GRE44+ (n=16 cells from 5 mice)
383 and GRE44- (n=16 cells from 4 mice) SST neurons, including rheobase (minimal amount of
384 current necessary to elicit a spike), maximal rate of rise during the depolarizing phase of the
385 action potential, the initial and steady state firing frequencies (both measured at the maximal
386 current step before spike inactivation), and spike width (measured as the width at half-maximal
387 spike amplitude).
388 (**e**) Representative current-clamp recordings from AAV-GRE12-Gq-tdTomato$^+$ cells
389 before and during CNO application.

390 (**f**) Increased firing rates of AAV-GRE12-Gq-tdTomato$^+$ cells evoked by depolarizing
391 current injections upon bath application of CNO (3 animals, 6-7 cells). * $p < 0.05$; ** $p < 0.01$,
392 paired t-test, two-tailed.
393 (**g**) Robust depolarization of AAV-GRE12-Gq-tdTomato$^+$ cells upon bath application of
394 CNO (3 animals, 6-7 cells). *** $p < 0.001$, paired t-test, two-tailed.
395 (**h**) Representative recordings from nearby uninfected pyramidal neurons in the visual
396 cortex of mice that were injected with AAV-GRE-12-Gq-tdTomato$^+$, before (top) and during
397 CNO application (bottom).
398 (**i**) Firing rates of pyramidal neurons during CNO application remain unchanged (3
399 animals, 5 cells). ns, $p > 0.05$, paired t-test, two-tailed.
400 (**j**) Representative image of nearby uninfected pyramidal neuron that was recorded from
401
402
403
404
405
406
407
408
409
410
411
412

Acknowledgements

We thank D. Tom for help with image analysis, Drs. W. Renthal, J. Green, and members of the Greenberg lab for discussions, the HMS Single Cell Core for single-nucleus RNA-seq sample processing, Boston Children's Hospital Viral Core for AAV packaging, Dr. B. Bean for feedback on electrophysiology, and the Harvard NeuroDiscovery Center Enhanced Neuroimaging Core for imaging and image analysis. This work was supported by the National Institute of Health BRAIN Initiative grant R01 MH114081-01, NIH grant T32GM007753 to M.A.N. and NIH Training in the Molecular Biology of Neurodegeneration grant 5T32AG000222-23 to S.H.

Author Contributions

S.H. conceived the study and designed the experiments. H.S. performed ATAC-Seq. M.A.N. and S.H. analyzed ATAC-seq data. S.H. cloned the library and performed the PESCA screen. M.A.N. mapped sequencing data and viral barcodes. S.H. analyzed snRNA-seq data. S.H., C.K. and O.F.W. cloned individual viral constructs. O.F.W. performed stereotactic injections and tissue sectioning. S.H. performed imaging and image analysis. C.P.T. performed and analyzed electrophysiology recordings. M.E.G. and E.C.G. advised on all aspects of the study. S.H., C.P.T., M.A.N, E.C.G. and M.E.G. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

436 **Correspondence**

437 Correspondence and requests for materials should be addressed to S.H.

438 (sinisa_hrvatin@hms.harvard.edu) or M.E.G. (michael_greenberg@hms.harvard.edu).

439

440 **References**

441 1. A. Zeisel *et al.*, Cell types in the mouse cortex and hippocampus revealed by single-cell
442 RNA-seq. *Science*. **347**, 1138–1142 (2015).

443 2. B. Tasic *et al.*, Shared and distinct transcriptomic cell types across neocortical areas. *Nature*.
444 **563**, 72–78 (2018).

445 3. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism.
446 *Science*. **357**, 661–667 (2017).

447 4. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord
448 with split-pool barcoding. *Science*. **360**, 176–182 (2018).

449 5. A. Mo *et al.*, Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain.
450 *Neuron*. **86**, 1369–1384 (2015).

451 6. D. A. Cusanovich *et al.*, Multiplex single-cell profiling of chromatin accessibility by
452 combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

453 7. J. D. Buenrostro *et al.*, Single-cell chromatin accessibility reveals principles of regulatory
454 variation. *Nature*. **523**, 486–490 (2015).

455 8. J. Dimidschstein *et al.*, A viral strategy for targeting and manipulating interneurons across
456 vertebrate species. *Nat. Neurosci.* **19**, 1743–1749 (2016).

457 9. J. Juettner *et al.*, Targeting neuronal and glial cell types with synthetic promoter AAVs in
458 mice, non-human primates, and humans. *bioRxiv*, 434720 (2018).

459 10. L. T. Graybuck *et al.*, Prospective, brain-wide labeling of neuronal subclasses with enhancer-
460 driven AAVs. *bioRxiv*, 525014 (2019).

461 11. J. K. Mich *et al.*, Epigenetic landscape and AAV targeting of human neocortical cell classes.
462 *bioRxiv*, 555318 (2019).

463 12. A. Kepecs, G. Fishell, Interneuron cell types are fit to function. *Nature*. **505**, 318–326
464 (2014).

465 13. O. Marín, Interneuron dysfunction in psychiatric disorders. *Nat. Rev. Neurosci.* **13**, 107–120
466 (2012).

467 14. B. Rudy, G. Fishell, S. Lee, J. Hjerling-Leffler, Three groups of interneurons account for
468 nearly 100% of neocortical GABAergic neurons. *Dev. Neurobiol.* **71**, 45–61 (2011).

469 15. W. Muñoz, R. Tremblay, D. Levenstein, B. Rudy, Layer-specific modulation of neocortical
470 dendritic inhibition during active wakefulness. *Science*. **355**, 954–959 (2017).

471 16. X. Jiang *et al.*, Principles of connectivity among morphologically defined cell types in adult
472 neocortex. *Science*. **350**, aac9462 (2015).

473 17. B. Tasic *et al.*, Adult mouse cortical cell taxonomy revealed by single cell transcriptomics.
474 *Nat. Neurosci.* **19**, 335–346 (2016).

475 18. C. N. Cearley, J. H. Wolfe, Transduction characteristics of adeno-associated virus vectors
476 expressing cap serotypes 7, 8, 9, and Rh10 in the mouse brain. *Mol. Ther. J. Am. Soc. Gene*
477 *Ther.* **13**, 528–537 (2006).

478 19. A. M. Klein *et al.*, Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic
479 Stem Cells. *Cell*. **161**, 1187–1201 (2015).

480 20. R. Zilionis *et al.*, Single-cell barcoding and sequencing using droplet microfluidics. *Nat.*
481 *Protoc.* **12**, 44–73 (2017).

482 21. J. Urban-Ciecko, A. L. Barth, Somatostatin-expressing neurons in cortical networks. *Nat.*
483 *Rev. Neurosci.* **17**, 401–409 (2016).

484 22. Y. Ma, H. Hu, A. S. Berrebi, P. H. Mathers, A. Agmon, Distinct subtypes of somatostatin-
485 containing neocortical interneurons revealed in transgenic mice. *J. Neurosci. Off. J. Soc.*
486 *Neurosci.* **26**, 5069–5082 (2006).

487 23. R. Tremblay, S. Lee, B. Rudy, GABAergic Interneurons in the Neocortex: From Cellular
488 Properties to Circuits. *Neuron*. **91**, 260–292 (2016).

489 24. H. Xu, H.-Y. Jeong, R. Tremblay, B. Rudy, Neocortical somatostatin-expressing GABAergic
490 interneurons disinhibit the thalamorecipient layer 4. *Neuron*. **77**, 155–167 (2013).

491 25. B. P. Bean, The action potential in mammalian central neurons. *Nat. Rev. Neurosci.* **8**, 451–
492 465 (2007).

493 26. T. Kimm, Z. M. Khaliq, B. P. Bean, Differential Regulation of Action Potential Shape and
494 Burst-Frequency Firing by BK and Kv2 Channels in Substantia Nigra Dopaminergic
495 Neurons. *J. Neurosci. Off. J. Soc. Neurosci.* **35**, 16404–16417 (2015).

496 27. B. N. Armbruster, X. Li, M. H. Pausch, S. Herlitze, B. L. Roth, Evolving the lock to fit the
497 key to create a family of G protein-coupled receptors potently activated by an inert ligand.
498 *Proc. Natl. Acad. Sci.* **104**, 5163–5168 (2007).

499 28. R. P. Patwardhan *et al.*, Massively parallel functional dissection of mammalian enhancers in
500 vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

501 29. A. Melnikov *et al.*, Systematic dissection and optimization of inducible enhancers in human
502 cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).

503

Methods

Mice

Animal experiments were approved by the National Institute of Health and Harvard Medical School Institutional Animal Care and Use Committee, following ethical guidelines described in the US National Institutes of Health Guide for the Care and Use of Laboratory Animals. For INTACT we crossed Sst-IRES-Cre (The Jackson Laboratory Stock # 013044), Vip-IRES-Cre (The Jackson Laboratory Stock # 010908) and Pv-Cre (The Jackson Laboratory Stock # 017320) with SUN1-2xsfGFP-6xMYC (The Jackson Laboratory Stock # 021039) and used adult (6-12 wk old) male and female F1 progeny. For PESCA screening we used adult (6-10 wk) C57BL/6J (The Jackson Laboratory, Stock # 000664) mice. For confirmation of hits we crossed Sst-IRES-Cre (The Jackson Laboratory Stock # 013044), Vip-IRES-Cre (The Jackson Laboratory Stock # 031628) and Gad2-IRES-Cre (The Jackson Laboratory Stock # 028867) mice with Ai14 mice (The Jackson Laboratory Stock # 007914) and used adult (6-12 wk old) male and female F1 progeny. All mice were housed under a standard 12 hr light/dark cycle.

INTACT purification and *in vitro* transposition

INTACT employs a transgenic mouse that expresses a cell-type-specific Cre and a Cre-dependent SUN1-2xsfGFP-6xMYC (SUN1-GFP) fusion protein. Nuclear purifications were performed from whole cortex of adult mice as previously described using anti-GFP antibodies (Fisher G10362)(*1*, *2*). Isolated nuclei were gently resuspended in cold L1 buffer (50 mM Hepes pH 7.5, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.25% Triton X-100, 0.5% NP40, 10% Glycerol, protease inhibitors), and pelleted at 800g for 5 minutes at 4°C. DNA libraries were

prepared from the nuclei using the Nextera DNA Library Prep Kit (Illumina) according to manufacturer's protocols. The final libraries were purified using the Qiagen minelute kit and sequenced on a Nextseq 500 benchtop DNA sequencer (Illumina).

ATAC-seq mapping:

All ATAC-seq libraries were sequenced on the Nextseq 500 benchtop DNA sequencer (Illumina). Seventy-five base pair (bp) single-end reads were obtained for all datasets. ATAC-seq experiments were sequenced to a minimum depth of 20 million (M) reads. Reads for all samples were aligned to the mouse genome (GRCm38/mm10, December 2011) using default parameters for the Subread (subread-1.4.6-p3, (Liao et al., 2013)) alignment tool after quality trimming with Trimmomatic v0.33 (Bolger et al., 2014) with the following command: java -jar trimmomatic-0.33.jar SE -threads 1 -phred33 [FASTQ_FILE] ILLUMINACLIP:[ADAPTER_FILE]:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:20 MINLEN:45. Nextera adapters were trimmed out for ATAC-seq data. Duplicates were removed with samtools rmdup. To generate UCSC genome browser tracks for ATAC-seq visualization, BEDtools was used to convert output bam files to BED format with the bedtools bamtobed command. Published mm10 blacklisted regions (Consortium, 2012) were filtered out using the following command: bedops –not-element-of 1 [BLACKLIST_BED]. Filtered BED files were scaled to 20 M reads and converted to coverageBED format using the BEDtools genomecov command. bedGraphToBigWig (UCSC-tools) was used to generate bigWIG files for the UCSC genome browser.

ATAC-seq peak calling and quantification:

Two independent peak calling algorithms were employed to ensure robust, reproducible peak calls. First, tag directories were created using HOMER makeTagDirectory for each replicate, and peaks were called using default parameters for findPeaks with -style factor. MACS2 was also called using default parameters on each replicate. The summit files output by MACS2 were converted to bed format and each summit extended bidirectionally to achieve a total length of 300 bp. As the ATAC-seq peak calls would ultimately be used to narrow down potential regulatory elements for screening of a limited subset, we applied the overly stringent requirement that a peak be called by both approaches in a given replicate for its inclusion in the final peak list for that sample. Peaks identified in any sample in this way were aggregated to produce a final superset of 323,369 regulatory elements called as accessible in at least one cell type. The featurecounts package was used to obtain ATAC-seq read counts for each of these accessible putative GREs.

Identification of SST-enriched GREs

We used genomic coordinates of a superset of 323,369 genomic regions identified as a union of ATAC-Seq peaks across various cell types in the mouse cortex (manuscript in preparation) as a list of reference coordinates over which to quantify the ATAC-Seq signal from SST+, VIP+ and PV+ cells. A matrix was constructed representing the mean ATAC-Seq signal in SST+, VIP+ and PV+ cells for each of the 323,369 GREs and normalized such that the total ATAC-Seq signal from each cell population was scaled to 10^7. Fold-enrichment was calculated for each region/GRE as [(Signal in cell type A)+1] / [mean(signal in cell types B and C)+1]. GREs were subsequently ranked based on fold-enrichment score.

Identification of conserved GREs

To identify GREs whose sequence is highly conserved across mammals, we first needed to identify an appropriate conservation score to use as a threshold for high conservation. We reasoned that by analyzing the conservation of DNA sequences of the same length, but an arbitrary distance of 100,000 bases away from each identified GRE, we would generate a set of DNA sequences whose conservation can be used to determine this threshold.

To this end, conservation scores for GREs and corresponding GRE-distal sequences were calculated using the bigWigAverageOverBed command to determine the average PhyloP score of each sequence based on mm10.60way.phyloP60wayPlacental.bw PhyloP scores (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phyloP60way/)(*3*) . After plotting the conservation score (phyloP, 60 placental mammals) of 323,369 GRE-distal sequences, we determined the conservation score of the 95th percentile of this distribution (PhyloP score = 0.5) and chose it as a minimal conservation score needed to classify any GRE as conserved.

Viral barcode design

Viral barcode sequences were chosen to be at least 3 insertions, deletions, or substitutions apart from each other to minimize the effects of sequencing errors on the correct identification of each barcode. The R library "DNAbarcodes" and following functions were used:

initialPool = create.dnabarcodes(10, dist=3, heuristic="ashlock");

finalPool = create.dnabarcodes(10, pool = initialPool, metric="seqlev");

The result was a list of 1164 10-base barcodes that fit our initial criteria.

Amplification of GREs and barcoding

Genomic PCR:

PCR primers were designed using primer3 2.3.7.(*4*) such that a 150-400 bp flanking sequence was added to each side of the GRE. The forward primers contained a 5' overhang sequence for downstream in-Fusion (Clonetech) cloning into the AAV vector (5'-GCCGCACGCGTTTAAT). The reverse primers contained a 5' overhang sequence containing the recognition sites for AsiSI and SalI restriction enzymes(5'-GCGATCGCTTGTCGAC). Hot Start High-Fidelity Q5 polymerase (NEB) was used according to manufacturer's protocol with mouse genomic DNA as template.

Barcoding PCR:

The unpurified PCR products from the genomic PCR were used as templates for the barcoding PCR. A forward primer containing the sequence for downstream in-Fusion (Clonetech) cloning into the AAV vector (5'-CTGCGGCCGCACGCGTTTA) was used in all reactions. Reverse primers were constructed featuring (in the 5' → 3'direction): 1) a sequence for downstream in-Fusion (Clonetech) cloning into the AAV vector (5'-GCCGCTATCACAGATCTCTCGA), 2) a unique 10-base barcode sequence, and 3) sequence complementary with the AsiSI and SalI restriction enzyme recognition sites that were introduced during the first PCR (5'-GCGATCGCTTGTCGAC). Three different reverse primers were used for each of the GREs amplified during the genomic PCR. Hot Start High-Fidelity Q5 polymerase (NEB) was used according to the manufacturer's protocol.

<u>PESCA Library cloning</u>

All PCR reactions were pooled and the amplicons purified using Agencourt AMPure XP. The pAAV-mDlx-GFP-Fishell-1 was a gift from Gordon Fishell (Addgene plasmid # 83900). The

plasmid was digested with PacI and XhoI, leaving the ITRs and the polyA sequence. in-Fusion was used to shuttle the pool of GRE PCR products into the vector. Following transformation into High Efficiency NEB 5-alpha Competent E.coli and recovery, SalI and AsiSI were used to linearize the AAV vector containing the GREs. The expression cassette containing the human HBB promoter and intron followed by GFP and WPRE was isolated by PCR amplification from pAAV-mDlx-GFP-Fishell-1. The expression cassette was ligated with the linearized GRE-library-containing vector using T4 ligase and transformed into High Efficiency NEB 5-alpha Competent E.coli to yield the final library. 50 colonies were Sanger sequenced to determine the correct pairing between GRE and barcode and the correct arrangement of the AAV vector.

AAV preparation

The pooled PESCA library or individual AAV constructs (100 µg) were packed into AAV9 at the Boston Children's Hospital Viral Core. The titers (2-50 x 10^{13} genome copies/mL) were determined by qPCR. Next generation sequencing using the NextSeq 500 platform was used to determine the complexity of the pooled PESCA library (Fig. 2a).

V1 cortex injections

Animals were anesthetized with isofluorane (1–3% in air) and placed on a stereotactic instrument (Kopf) with a 37°C heated pad. The PESCA library (AAV9, 1.9 x 10^{13} genome copies/mL) was stereotactically injected in V1 (800 nL per site at 25 nL/min) using a sharp glass pipette (25-45 µm diameter) that was left in place for 5 min prior to and 10 min following injection to minimize

backflow. Two injections were performed per animal at coordinates 3.0 and 3.7 mm posterior, 2.5 mm lateral relative to bregma, and 0.6 mm ventral relative to the brain surface. Individual rAAV-GRE constructs were stereotactically injected at a titer of 1 x 10^{11} genome copies/mL. (250 nL per site at 25 nL/min). All injections were performed at two depths (0.4 and 0.7 mm ventral relative to the brain surface) to achieve broader infection across cortical layers. The injection coordinates relative to bregma were 3.0 or 3.7 mm posterior, 2.5 or -2.5 mm lateral.

Nuclear isolation

Single-nuclei suspensions were generated as described previously(*1*) , with minor modifications. V1 was dissected and placed into a Dounce with homogenization buffer (0.25 M sucrose, 25 mM KCl, 5 mM $MgCl_2$, 20 mM Tricine-KOH, pH 7.8, 1 mM DTT, 0.15 mM spermine, 0.5 mM spermidine, protease inhibitors). The sample was homogenized using a tight pestle with 10 stokes. IGEPAL solution (5%, Sigma) was added to a final concentration of 0.32%, and 5 additional strokes were performed. The homogenate was filtered through a 40-µm filter, and OptiPrep (Sigma) added to a final concentration of 25% iodixanol. The sample was layered onto an iodixanol gradient and centrifuged at 10,000g for 18 minutes as previously described[1,2] . Nuclei were collected between the 30% and 40% iodixanol layers and diluted to 80,000-100,000 nuclei/mL for encapsulation. All buffers contained 0.15% RNasin Plus RNase Inhibitor (Promega) and 0.04% BSA.

snRNA-Seq library preparation and sequencing

Single nuclei were captured and barcoded whole-transcriptome libraries prepared using the inDrops platform as previously described(*5*, *6*), collecting five libraries of approximately 3,000 nuclei from each animal. Briefly, single nuclei along with single primer-carrying hydrogels were captured into droplets using a microfluidic platform. Each hydrogel carried oligodT primers with a unique cell-barcode. Nuclei were lysed and the cell-barcode containing primers released from the hydrogel, initiating reverse transcription and barcoding of all cDNA in each droplet. Next, the emulsions were broken and cDNA across ~3000 nuclei pooled into the same library. The cDNA was amplified by second strand synthesis and *in vitro* transcription, generating an amplified RNA intermediate which was fragmented and reverse transcribed into an amplified cDNA library.

For enrichment of virally-derived transcripts, a fraction (3 μL) of the amplified RNA intermediate was reverse transcribed with random hexamers without prior fragmentation. PCR was next used to amplify virally derived transcripts. The forward primer was designed to introduce the R1 sequence and anneal to a sequence uniquely present 5' of the viral-barcode sequence present in the viral transcripts (5'- GCATCGATACCGAGCGC). The reverse primer was designed to anneal to a sequence present 5' of the cell-barcode (5'-GGGTGTCGGGTGCAG). The result of the PCR is preferential amplification of the viral-derived transcripts, while simultaneously retaining the cell-barcode sequence necessary to assign each transcript to a particular cell/nucleus. Following PCR amplification (18 cycles, Hot Start High-Fidelity Q5 polymerase) all the libraries were indexed, pooled, and sequenced on a Nextseq 500 benchtop DNA sequencer (Illumina).

inDrop sample mapping and viral barcode deconvolution by cell:

The published inDrops mapping pipeline (github.com/indrops/indrops) was used to assign reads to cells. To map viral sequences, a custom annotated transcriptome was generated using the indrops pipeline build_index command supplied with the following newly generated reference files: a custom genome with one additional contig comprising a shared 5' sequence (gcatcgataccgagcgcgcgatcgc), the given 10 bp barcode, and a shared 3' sequence (tcgagagatctgtgatagcggc) was appended to the GRCm38.dna_sm.primary_assembly.fa genome file for each cloned GRE. These sequences were also appended GRCm38.88.gtf gene annotation file, with all sequences assigned the same gene_id and gene_name, but unique transcript_id, transcript_name, and protein_id. After inDrops pipeline mapping and cell deconvolution, the pysam package was used to extract the 'XB' and 'XU' tags, which contain cell barcode and UMI sequences, respectively, from every read that mapped uniquely to any one of the custom viral contigs (i.e. requiring the read map to the 10 bp barcode with at most 1 mismatch) in the inDrops pipeline-output bam files. These barcode-UMI combinations were condensed to generate a final cell x GRE barcode UMI counts table for each sample.

Embedding and identification of cell types

Data from all nuclei (two animals, 5 libraries of ~3,000 nuclei per animal) were analyzed simultaneously. Viral-derived sequences were removed for the purposes of embedding clustering and cell type identification. The initial dataset contained 32,335 nuclei, with more than 200 unique non-viral transcripts (UMIs) assigned to each nucleus. The R software package Seurat(*7*, *8*) was used to cluster cells. First, the data were log-normalized and scaled to 10,000 transcripts per cell. Variable genes were identified using the FindVariableGenes() function. The following

parameters were used to set the minimum and maximum average expression and the minimum dispersion: x.low.cutoff=0.0125, x.high.cutoff=3, y.cutoff=0.5. Next, the data was scaled using the ScaleData() function, and principle component analysis (PCA) was carried out. The FindClusters() function using the top 30 principal components (PCs) and a resolution of 1.5 was used to determine the initial 29 clusters. Based on the expression of known marker genes we merged clusters that represented the same cell type. Our final list of cell types was: Excitatory neurons, PV Interneurons, SST Interneurons, VIP interneurons, NPY Interneurons, Astrocytes, Vascular-associated cells, Microglia, Oligodendrocytes, and Oligodendrocyte precursor cells.

Enrichment calculation

Viral vector expression for each of the 861 barcodes across the ten cell types was calculated by averaging the expression of barcoded transcripts across all the individual nuclei that were assigned to that cell type. The relative fold-enrichment in expression toward *Sst*+ cells was computed as the ratio of the mean expression in *Sst*+ cells and the mean expression in *Sst*- cells: (mean(*Sst*+ cells)+0.01)/ (mean(*Sst*- cells)+0.01).

Viral GRE expression for each of the 287 barcodes was calculated at the single-nucleus level as a sum of the expression of the three barcodes that were paired with that GRE. Average GRE-driven expression across the ten cell types was calculated by averaging the expression of the GRE transcripts across all the individual nuclei that were assigned to that cell type. The relative fold-enrichment in GRE expression toward *Sst*+ cells was determined as the ratio of the mean expression in *Sst*+ cells and the mean expression in *Sst*- cells: (mean(*Sst*+ cells)+0.01)/ (mean(*Sst*- cells)+0.01).

Differential gene expression

To identify which of the GRE-driven transcripts were statistically enriched in *Sst*+ vs. *Sst*- cells, we carried out differential gene expression analysis using the R package Monocle2(*9*) . The data were modeled and normalized using a negative binomial distribution, consistent with snRNA-seq experiments. The functions estimateSizeFactors(), estimateDispersions() and differentialGeneTest() were used to identify which of the GRE-derived transcripts were statistically enriched in *Sst*+ cells. GREs whose false discovery rate (FDR) was less than 0.01 were considered enriched.

Fluorescence microscopy

Sample preparation:

Mice were sacrificed and perfused with 4% PFA followed by PBS. The brain was dissected out of the skull and post-fixed with 4% PFA for 1-3 days at 4°C. The brain was mounted on the vibratome (Leica VT1000S) and coronally sectioned into 100 µm slices. Sections containing V1 were arrayed on glass slides and mounted using DAPI Fluoromount-G (Southern Biotech).

Sample imaging:

Sections containing V1 were imaged on a Leica SPE confocal microscope using an ACS APO 10x/0.30 CS objective (Harvard NeuroDiscovery Center). Tiled V1 cortical areas of ~1.2 mm by ~0.5 mm were imaged at a single optical section to avoid counting the same cell across multiple optical sections. Channels were imaged sequentially to avoid any optical crosstalk.

Immunostaining

To identify parvalbumin (PV)+ cells, coronal sections were washed three times with PBS containing 0.3% TritonX-100 (PBST) and blocked for 1 h at room temperature with PBST containing 5% donkey serum. Section were incubated overnight at 4°C with mouse anti-PVALB antibody 1:2000 (Milipore), washed again three times with PBST, and incubated for 1 h at room temperature with 1:500 donkey anti-mouse 647 secondary antibody (Life Technologies). After washing in PBST and PBS, samples were mounted onto glass slides using DAPI Fluoromount-G.

Quantification of the percentage of GFP+ cells that were SST+, VIP+, and PV+

Across all images, coordinates were registered for each GFP+ cell that could be visually discerned. An automated ImageJ script was developed to quantify the intensity of each acquired channel for a given GFP+ cell. We created a circular mask (radius = 5.7 μm) at each coordinate representing a GFP positive cell, background subtracted (rolling ball, radius = 72 μm) each channel, and quantified the mean signal of the masked area. To identify the threshold intensity used to classify each GFP+ cell as either SST+, VIP+ or PV+, we first determined the background signal in the channel representing SST, VIP or PV by selecting multiple points throughout the area visually identified as background. These background points were masked as small circular areas (radius = 5.7 μm), over which the mean background signal was quantified. The highest mean background signal for SST, VIP and PV was conservatively chosen as the threshold for classifying GFP+ cells as SST+, VIP+ or PV+, respectively.

Quantification of the distribution of cells as a function of distance from pia

A semiautomated ImageJ algorithm was developed to trace the pia in each image, generate a Euclidean Distance Map (EDM), and calculate the distance from the pia to each GFP+ cell.

Quantification of the percentage of SST+ cells that were GFP+

An automated algorithm was developed to identify SST+ cells after appropriate background subtraction, image thresholding, masking and filtering for all objects of appropriate size and circularity. The number of SST+ objects (cells) was then counted within a minimal polygonal area that encompassed all GFP+ cells in that image. The ratio of the number of GFP+ cells and SST+ cells within the area of infection (here identified as area with discernable GFP+ cells) was calculated.

Slice Preparation

Acute, coronal brain slices containing visual cortex of 250 – 300 μm thickness were prepared using a sapphire blade (Delaware Diamond Knives, Wilmington, DE) and a VT1000S vibratome (Leica, Deerfield, IL). Mice were anesthetized though inhalation of isoflurane, then decapitated. The head was immediately immersed in an ice-cold solution containing (in mM): 130 K-gluconate, 15 KCl, 0.05 EGTA, 20 HEPES, and 25 glucose (pH 7.4 with NaOH; Sigma). The brains were quickly dissected and cut in the same ice-cold, gluconate based solution while oxygenated with 95% O_2/5% CO_2. Slices then recovered at 32ºC for 20-30 minutes in oxygenated artificial cerebrospinal fluid (ACSF) in mM: 125 NaCl, 26 NaHCO3, 1.25 NaH2PO4, 2.5 KCl, 1.0 MgCl2, 2.0 CaCl2, and 25 glucose (Sigma), adjusted to 310-312 mOsm with water.

Electrophysiological Recordings

Using an Olympus BX51WI microscope equipped with a 60x water immersion objective, we used fluorescence illumination to identify rAAV-GRE44-GFP$^+$ (red and green) and rAAV-GRE44-GFP$^-$ (only red) SST neurons in the area of injection/ AAV infection (Fig. 4a-d). rAAV-GRE44-GFP$^-$ neurons were recorded if they were in the same field of view as rAAV-GRE44-GFP$^+$ neurons under 60x. For rAAV-GRE12-Gq-DREADD-tdTomato experiments (Fig. 4e-j), tdTomato$^+$ cells and morphologically identified pyramidal neurons in the same field of view under 60x were recorded. Whole-cell current clamp recordings of these neurons in coronal visual cortex slices of P50 to P80 wild-type mice were performed using borosilicate glass pipettes (3-6 MOhms, Sutter Instrument, Novato, CA) filled with an internal solution (in mM): 116 $KMeSO_3$, 6 KCl, 2 NaCl, 0.5 EGTA, 20 HEPES, 4 MgATP, 0.3 NaGTP, 10 $NaPO_4$ creatine (pH 7.25 with KOH; Sigma). Neurobiotin (1.5%) was occasionally included in the internal solution to allow for post-hoc morphological reconstruction of recorded cells. All experiments were performed at room temperature in oxygenated ACSF. Series resistance was compensated by at least 60% in a voltage-clamp configuration before switching to current-clamp ("I Clamp Normal"). After break-in, a systematic series of 1 second current injections ranging from -100 pA to 500 pA were applied to each cell using the User List function in the "Edit Waveform" tab of pClamp. After such baseline firing rates were calculated, CNO (2 μM, Sigma) was bath applied. An average of at least three trials for each current injection was calculated before and during CNO application.

Data Acquisition and Analysis

For electrophysiology, data acquisition of current-clamp experiments was performed using Clampex10.2, an Axopatch 200B amplifier, filtered at 2 kHz and digitized at 20 kHz with a DigiData 1440 data acquisition board (Molecular Devices, Sunnyvale, CA). Analysis of electrophysiological parameters was done using Clampfit (Molecular Devices, Sunnyvale, CA), Prism 7 (GraphPad Software, La Jolla, CA), Excel (Microsoft, Redmond, WA), and custom software written and generously shared by Dr. Bruce Bean in Igor Pro version 6.1.2.1 (WaveMetrics, Lake Oswego, OR). Membrane potentials in this study were not corrected for the liquid junction potential and are thus positively biased by 8 mV. For analysis of action potential waveform in Fig. 4a-d and Supplementary Table 1, the first action potential that appeared during a current injection equivalent to the rheobase was analyzed, as well as the first action potential of the subsequent two current injections. For example, if the rheobase were 20 pA, then all the parameters defined in the next section were also analyzed for the first action potential elicited with 20, 25, and 30 pA of injected current, and averaged.

Definition of Electrophysiological Parameters

AP Height (in millivolts): the difference between the peak of the action potential and the most negative voltage during the afterhyperpolarization immediately following the spike.

AP Peak (in millivolts): the most depolarized (positive) potential of the spike.

AP Trough (in millivolts): the most negative voltage reached during the afterhyperpolarization immediately following the spike.

F_{max} initial (in Hertz): the average of the reciprocal of the first three interstimulus intervals, measured at the maximal current step injected before spike inactivation.

F_{max} steady-state (in Hertz): the average of the reciprocal of the last three interstimulus intervals, measured at the maximal current step injected before spike inactivation.

Rate of rise (in volts per second): maximal voltage slope (*dV/dt*) during the upstroke (rising phase) of the action potential.

Rheobase (in picoamperes): the minimal 1000 ms current step (in increments of 5 pA) needed to elicit an action potential.

R_{in} (in megaohms, MΩ): input resistance, determined by using Ohm's law to measure the change in voltage in response to a -50 pA, 1000ms hyperpolarizing current at rest.

Spike adaptation ratio: the ratio of F_{max} steady-state to F_{max} initial

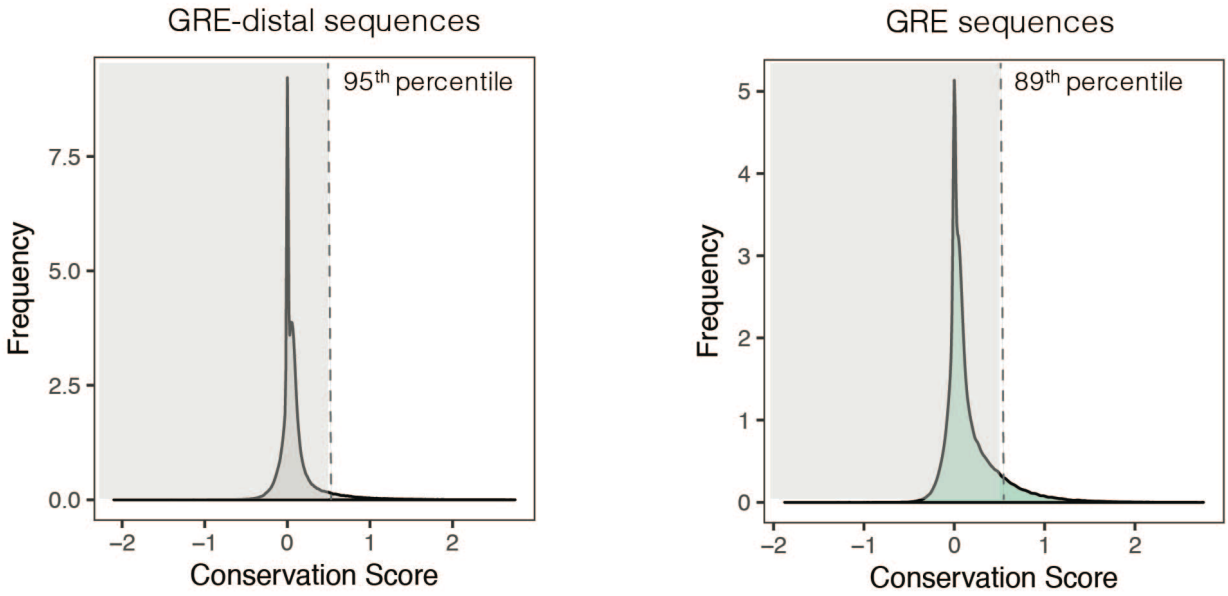
Spike width (in milliseconds, used interchangeably with spike half-width): the width at half-maximal spike height as defined above.

τ_m (in milliseconds): membrane time constant, determined by fitting a monoexponential curve to the voltage chance in response to a -50pA, 1000ms hyperpolarizing current at rest.

Threshold (in millivolts): the membrane potential at which dV/dt = 5 V/s.

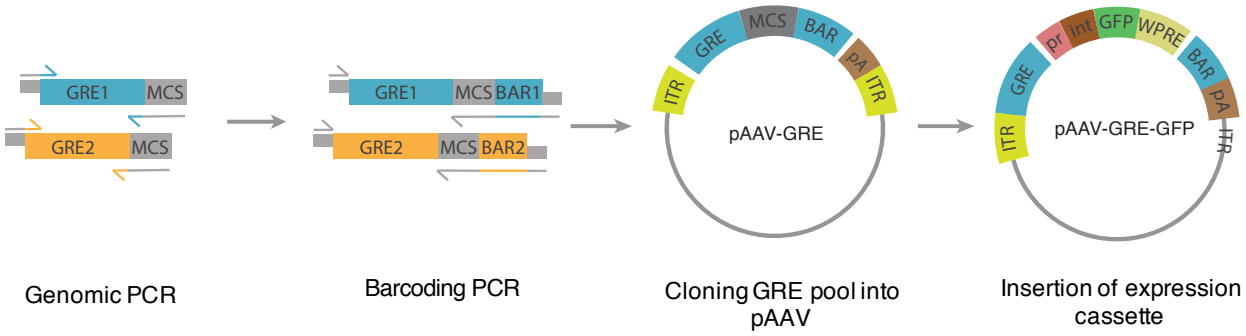
V_{rest} (in millivolts): resting membrane potential a few minutes after breaking in without any current injection.

Supplementary figures



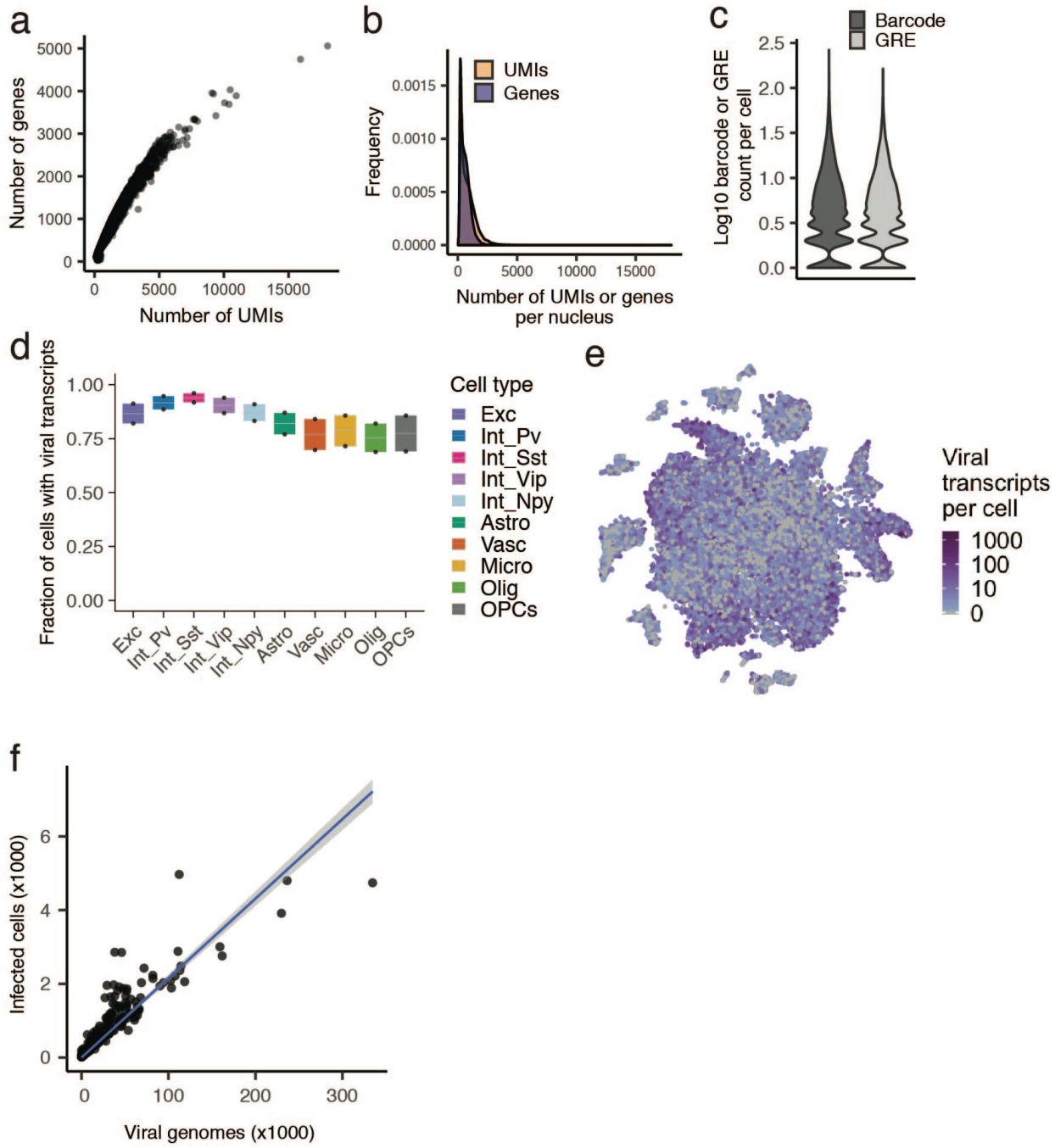
Supplementary figure 1

Identification of conserved GREs. Left: For each of the 323,369 genomic regions that were identified by ATAC-Seq as GREs in either SST^+, VIP^+ or PV^+ cells, a region of the same size was chosen exactly 100,000 bases away from the GRE. The mean sequence conservation score (phyloP, 60 placental mammals) for each of these GRE-distal regions was calculated and plotted. A vertical line at the conservation score of 0.5 indicates the 95th percentile of that distribution and was chosen as a minimal conservation score needed to consider a GRE sequence as conserved. Right: The mean sequence conservation score (phyloP, 60 placental mammals) for each of the 323,369 GREs was calculated and plotted. A vertical line indicates the minimal conservation score of 0.5. 36,215 GREs (11%) had a mean conservation greater than 0.5 and were deemed conserved.



Supplementary figure 2

(**a**) Schema for PESCA library construction. PCR is used to amplify GREs from the genomic DNA and to introduce appropriate restriction enzyme sites and, subsequently, a 10 bp barcode sequence. Each GRE is amplified three times using three different barcode sequences. The amplified GREs are pooled and cloned into an AAV vector. Restriction enzyme sites between the GRE and the barcode are used to insert an expression cassette consisting of a minimal promoter, intron, GFP and WPRE sequences. See methods for details.



Supplementary figure 3

(**a**) Dot plot of the number of unique molecular identifiers (UMIs) and the number of genes for each nucleus that was analyzed.

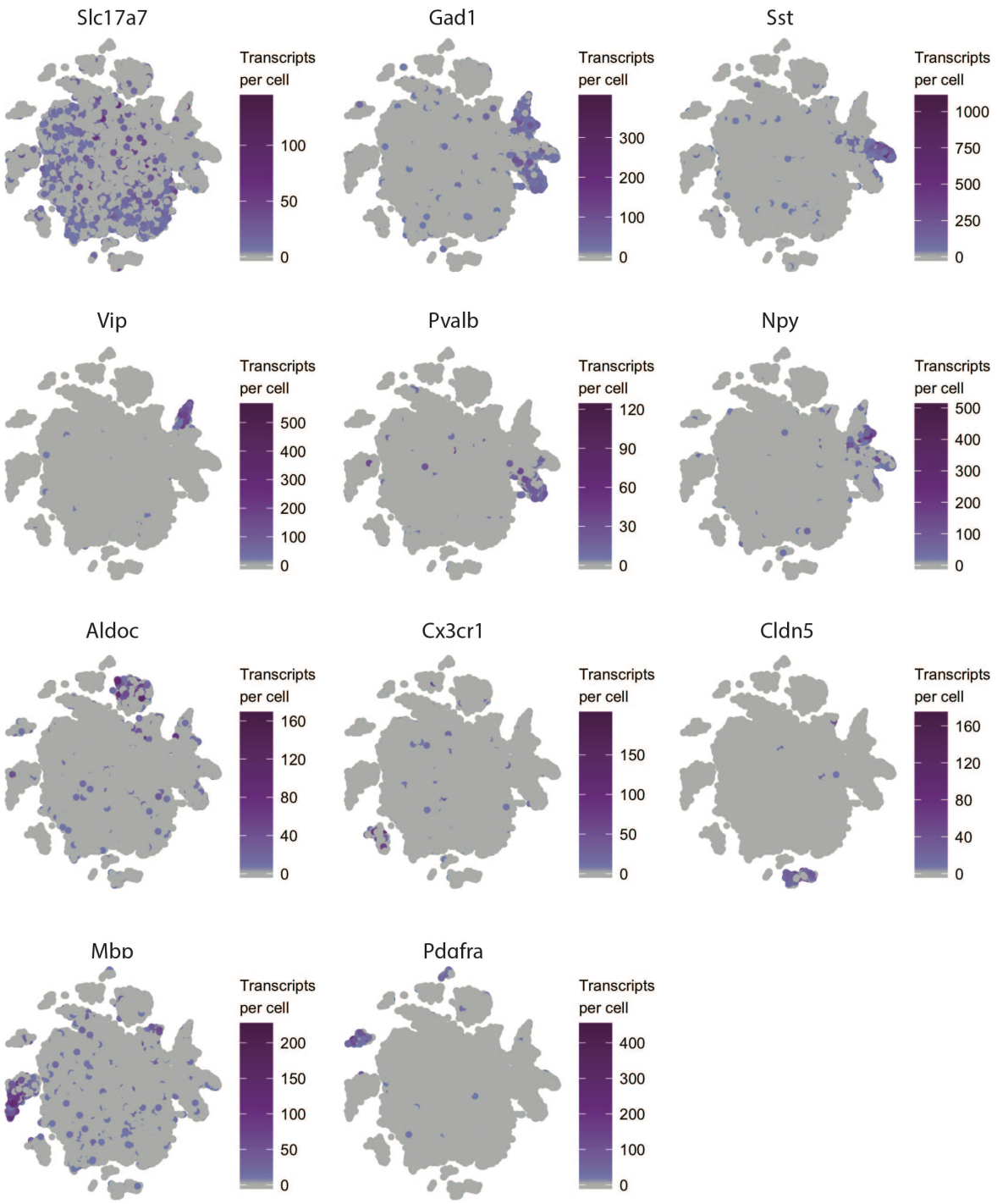
(**b**) Plot showing the density distribution of number of UMIs and genes per nucleus.

(**c**) Distribution of the number of unique barcodes and unique GREs detected per nucleus, displayed as Log10(Count+1).

(**d**) Quantification of the fraction of cells within each defined cell type in which we detected barcoded viral transcripts. Each dot represents one animal. Box plot represents mean ± standard error of the mean (s.e.m) for Exc (86.6 ± 4.6%), Int_Pv (91.6 ± 3.1%), Int_Sst (93.9 ± 2.1%), Int_Vip (90.4 ± 3.5%), Int_Npy (87.1 ± 3.8%), Astro (82.0 ± 5.0%), Vasc (76.9 ± 7.1%), Microglia (78.6 ± 7.1%), Olig (75.4 ± 6.6%), and OPCs (77.4 ± 8.3%).

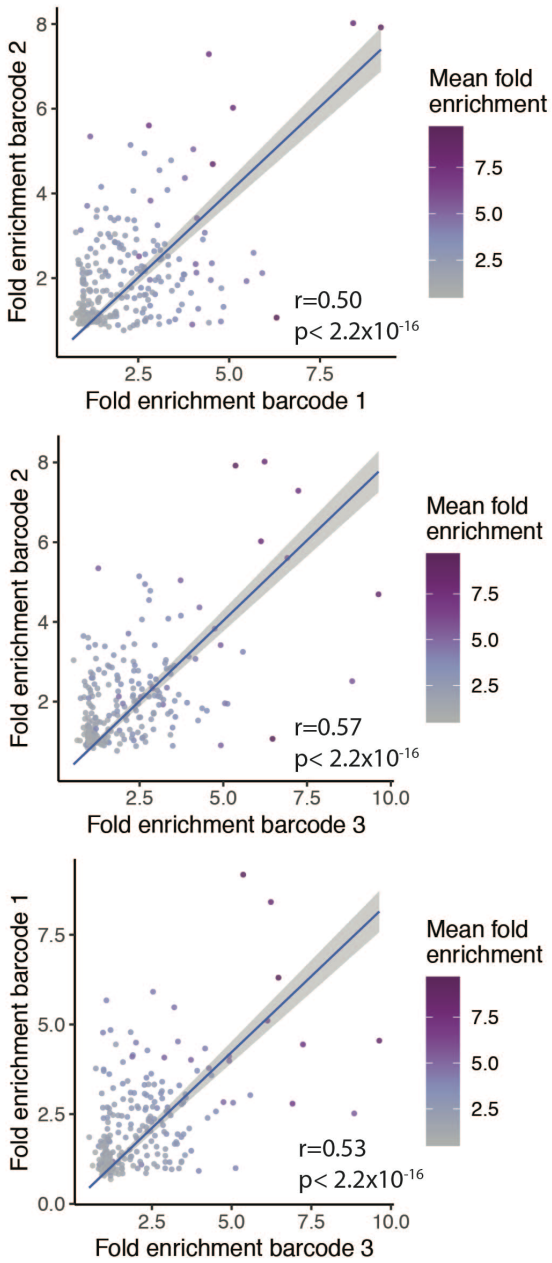
(**e**) t-SNE plot of 32,335 nuclei from V1 cortex of two injected animals**.** Color gradient denotes number of unique viral transcripts per nucleus displayed as Log10(Count+1).

(**f**) Dot plot of the number of viral genomes in the AAV library and the number of infected cells recovered after snRNA-Seq. Each dot represents one barcode (n = 861). Line of linear fit with 95% confidence intervals (shaded). Pearson correlation $r = 0.9$, $p < 2.2\times10^{-16}$.



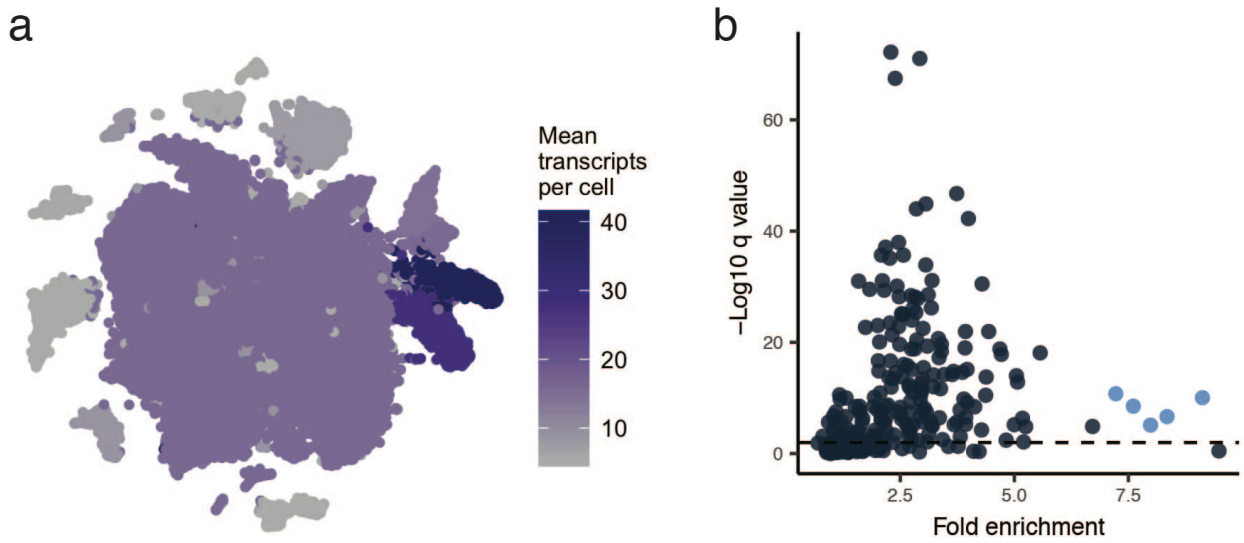
Supplementary figure 4

t-SNE plots of 32,335 nuclei from V1 cortex of two analyzed animals**.** Color gradient denotes number of unique transcripts per nucleus of the indicated cellular marker gene.



Supplementary figure 5

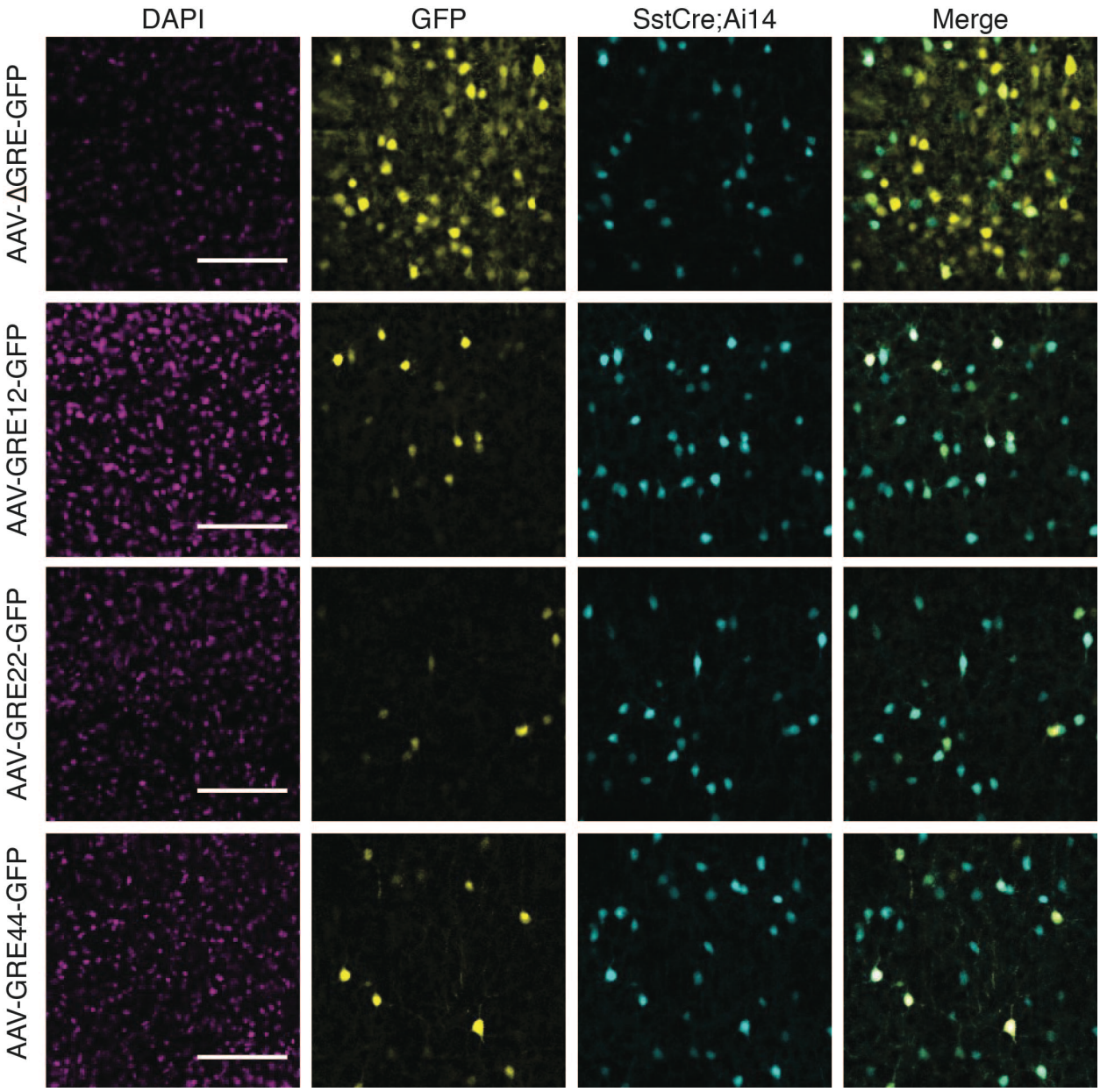
Dot plot of pairwise comparison between SST fold-enrichment values across three sets of barcodes. The values on each axis represent the SST fold-enrichment calculated for each GRE based on one of the three barcodes paired with that GRE. Blue line indicates linear fit with 95% confidence intervals (shaded). Correlation and p-values are indicated for each plot. Color gradient indicates the average enrichment between all three barcodes.



Supplementary figure 6

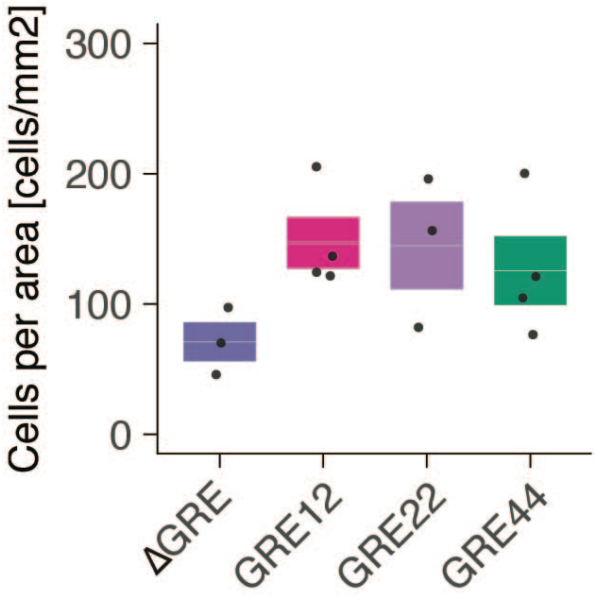
(**a**) t-SNE plot of 32,335 nuclei from V1 cortex of two analyzed animals showing the mean viral expression across all GREs**.** Plot is pseudocolored based on the mean expression in each cell type.

(**b**) Volcano plots for identified SST-enriched GREs (Fold-enrichment > 7 and FDR < 0.01). Colored dots represent the five SST-enriched GREs that were considered hits.



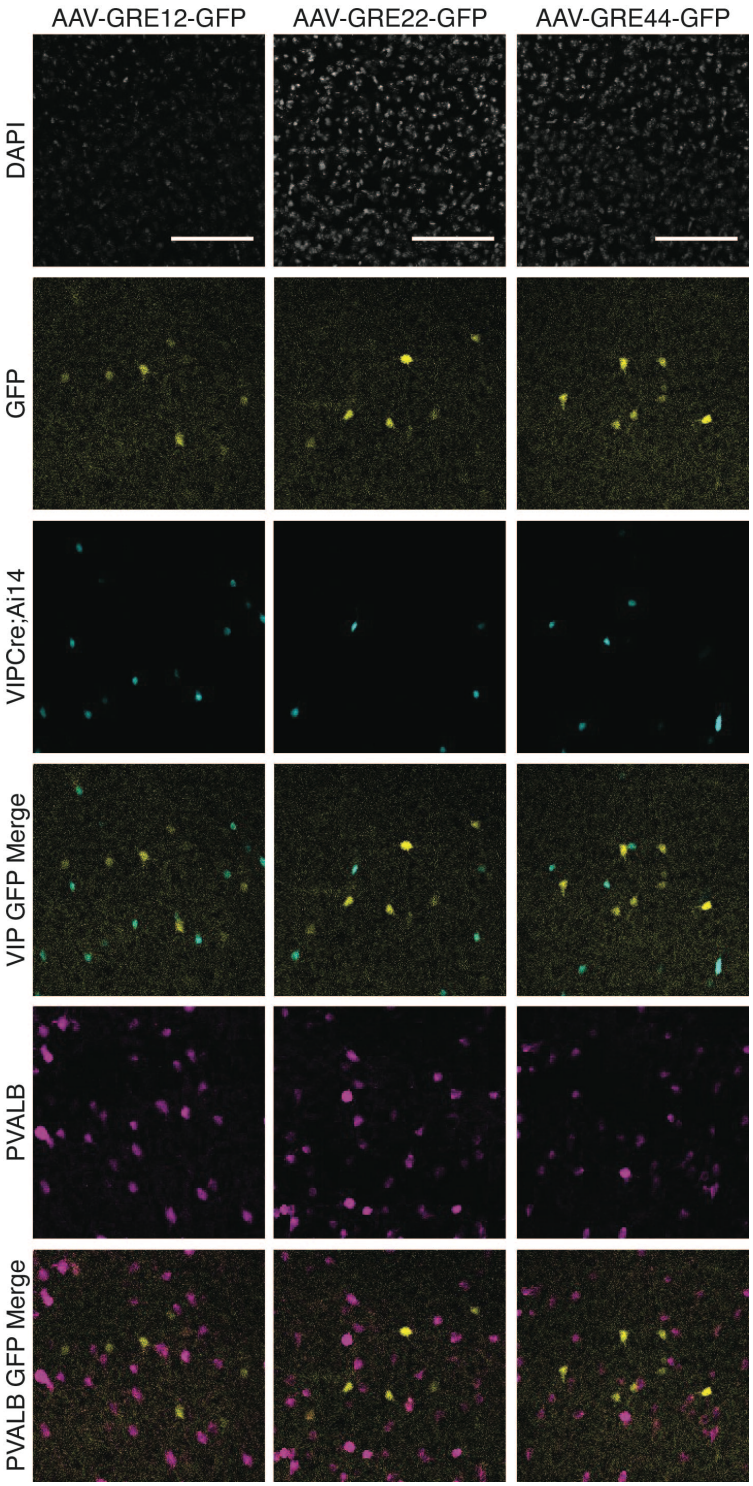
Supplementary Figure 7

Fluorescent images from adult Sst-Cre; Ai14 mouse visual cortex twelve days following injection with rAAV-GRE-GFP as indicated.



Supplementary Figure 8

Quantification of the number of GFP^+ SST^+ cells normalized for area of infection. Each dot represents one animal. Box plot represents mean ± standard error of the mean (s.e.m). Values are 73.0 ± 17.2, 146.9 ± 19.7, 144.8 ± 38.6 and 125.6± 26.4 cells/mm^2 for AAV-[ΔGRE, GRE12, GRE22, GRE44]-GFP, respectively.



Supplementary Figure 9

Fluorescent images from adult Vip-Cre; Ai14 mouse visual cortex immunostained for PVALB twelve days following injection with rAAV-GRE-GFP as indicated.

Supplementary Table 1
Electrophysiological Parameters of GRE44- and GRE44+ SST Neurons in Visual Cortex

	GRE44- (n=16)	**GRE44+** (n=16)	**p value (2-tailed unpaired t-test)**
V_{rest} (mV)	-62.4 ± 1.51	-60.6 ± 1.63	0.41
R_{in} (MΩ)	304 ± 54.8	391 ± 47.3	0.24
τ_m (ms)	14.2 ± 2.35	22.8 ± 4.32	0.094
Threshold (mV)	-45.6 ± 1.24	-48.1 ± 1.26	0.17
AP Peak (mV)	13 ± 3.38	11.6 ± 3.19	0.76
AP Trough (mV)	-63.6 ± 1.36	-63.7 ± 1.43	0.96
AP Height (mV)	76.5 ± 4.46	75.2 ± 4.06	0.83
Rate of Rise (V/s)	**122 ± 11.7**	**85 ± 7.34**	**0.013***
Rheobase (pA)	**43.5 ± 10.4**	**20.3 ± 4.64**	**0.044***
Spike Half-Width (ms)	**1.25 ± 0.0819**	**2.52 ± 0.307**	**0.0004*****
F_{max}, steady-state (Hz)	**83.4 ± 9.8**	**34.5 ± 4.32**	**0.0002*****
F_{max}, initial (Hz)	**111 ± 8.35**	**67.3 ± 7.48**	**0.0007*****
Spike adaptation ratio	0.763 ± 0.0839	0.561 ± 0.0699	0.08

Values are shown as mean ± SEM.

Supplementary Table 1. Electrophysiological Parameters of GRE44- and GRE44+ SST Neurons in Visual Cortex

References

1. A. Mo *et al.*, Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*. **86**, 1369–1384 (2015).

2. H. Stroud *et al.*, Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell*. **171**, 1151-1164.e16 (2017).

3. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

4. A. Untergasser *et al.*, Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).

5. A. M. Klein *et al.*, Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*. **161**, 1187–1201 (2015).

6. R. Zilionis *et al.*, Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).

7. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

8. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

9. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).