



(51) International Patent Classification:

G16C 20/70 (2019.01) G16B 40/20 (2019.01)

(21) International Application Number:

PCT/US2020/038528

(22) International Filing Date:

18 June 2020 (18.06.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/921,452 18 June 2019 (18.06.2019) US

(71) Applicant: CARNEGIE MELLON UNIVERSITY

[US/US]; 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213 (US).

(72) Inventors: PFENNING, Andreas P.; 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213 (US). RAMA-

MURTHY, Easwaran; 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213 (US). LAWLER, Alyssa; 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213 (US).

(74) Agent: KAYTOR, Elizabeth N. et al.; Fish & Richardson

P.C., P.O. Box 1022, Minneapolis, Minnesota 55440-1022 (US).

(81) Designated States (unless otherwise indicated, for every

kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every

kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- with international search report (Art. 21(3))
- with sequence listing part of description (Rule 5.2(a))

(54) Title: SPECIFIC NUCLEAR-ANCHORED INDEPENDENT LABELING SYSTEM

(57) Abstract: Materials for generating data representing a synthetic genetic sequence configured for labeling at least one cell type by causing expression of a marker in the at least one cell type are provided herein. Also provided herein are methods and materials for labeling and isolating particular cell types from mixed cell populations are provided herein.



SPECIFIC NUCLEAR-ANCHORED INDEPENDENT LABELING SYSTEM

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims benefit of priority from U.S. Provisional Application Serial No. 62/921,452, filed June 18, 2019. The disclosure of the prior application is considered part of (and is incorporated by reference in) the disclosure of this application.

TECHNICAL FIELD

This document relates to materials and methods for labeling and subsequently isolating particular cell types from mixed cell populations.

BACKGROUND

The brain and other tissues consist of heterogeneous populations of cells with specialized properties and functions. Genetically and epigenetically distinguishable cell types are likely to play different roles and exhibit different responses in disease and other conditions. The vast majority of genomic assays, however, are conducted on bulk tissue and not individual cell types, making the results of such assays difficult to interpret and disentangle for cell type-specific mechanisms. Improved techniques for nuclei labeling and isolation are needed to solve this problem.

SUMMARY

This document is based, at least in part, on the development of compositions and methods for labeling and isolating specific populations of cells. The methods are referred to herein as SNAIL (Specific Nuclear-Anchored Independent Labeling) and cSNAIL (Cre-Specific Nuclear-Anchored Independent Labeling). cSNAIL and SNAIL provide improvements over current methods for nuclear isolation methods, as they are easier, more time-efficient, and more cost-effective, without any loss of selection. Moreover, cSNAIL and SNAIL have the added benefits of being compatible with multiplexing and other transgenic models, extending to new cell types, and being transferrable across species. The availability of this technology increases the practicality of cell type-specific genomics, and allows these approaches to be used in other mammals, including humans.

In a first aspect, this document features a nucleic acid construct containing (a) a sequence encoding a tagged Sun1 fusion polypeptide, where the tagged Sun1 fusion

polypeptide comprises, consists of, or consists essentially of (i) an N-truncated fragment of a Sun1 gene, where the fragment encodes at least a portion of a portion of a Sun1 protein, and where the portion of the Sun1 protein is 400 to 600 amino acids in length, and (ii) a sequence encoding a tag polypeptide, and (b) a promoter sequence specific for a selected cell type, where the promoter sequence is operably linked to the sequence encoding the tagged Sun1 fusion polypeptide, and is effective to drive expression of the sequence encoding the tagged Sun1 fusion polypeptide in the selected cell type. (By “consists essentially of” is meant that a nucleic acid or a polypeptide contains specified components, and can contain additional sequences that do not materially affect the basic and novel characteristics of the nucleic acid or polypeptide.) The N-truncated fragment of the Sun1 gene can encode a polypeptide with at least 95% sequence identity to amino acids 208-757 of the mouse Sun1 protein having the amino acid sequence set forth in SEQ ID NO:1. The tag polypeptide can be a fluorescent polypeptide (e.g., a green fluorescent protein (GFP), such as a GFP with an amino acid sequence that at least 95% identical to the superfolder GFP sequence set forth in SEQ ID NO:2). The construct can further include virus sequences (e.g., adeno-associated virus (AAV) sequences or lentivirus sequences). The promoter sequence can be specific for parvalbumin positive (PV+) neurons, such the sequence encoding the tagged Sun1 fusion polypeptide is expressed in PV+ neurons but not expressed in at least 90% of parvalbumin negative (PV-) neurons.

In another aspect, this document features a nucleic acid construct containing (a) a sequence encoding a tagged Sun1 fusion polypeptide, where the tagged Sun1 fusion polypeptide comprises, consists of, or consists essentially of (i) an N-truncated fragment of a Sun1 gene, where the fragment encodes a portion of a Sun1 protein, and where the portion of the Sun1 protein is 400 to 600 amino acids in length, and (ii) a sequence encoding a tag polypeptide, (b) a first lox sequence flanking the 5' end of the sequence encoding the tagged Sun1 fusion polypeptide and a second lox sequence flanking the 3' end of the sequence encoding the tagged Sun1 fusion polypeptide, and (c) a promoter sequence downstream of the lox sequence flanking the 3' end of the sequence encoding the tagged Sun1 fusion polypeptide, such that the sequence encoding the tagged Sun1 fusion polypeptide is in reverse orientation with respect to the promoter. The N-truncated fragment of the Sun1 gene can encode a polypeptide with at least 95% sequence identity to amino acids 208-757 of the mouse Sun1 protein having the amino acid sequence set

forth in SEQ ID NO:1. The tag polypeptide can be a fluorescent polypeptide (e.g., a GFP, such as a GFP having an amino acid sequence that is at least 95% identical to the superfolder GFP sequence set forth in SEQ ID NO:2). The construct can further include virus sequences (e.g., AAV sequences or lentivirus sequences).

This document also features virus particles (e.g., AAV or lentivirus particles) containing one or more nucleic acid constructs as described herein.

In another aspect, this document features a method for labeling a selected cell type within a population of different cell types, where the method includes introducing into the population of different cell types a nucleic acid construct containing (a) a sequence encoding a tagged Sun1 fusion polypeptide, where the tagged Sun1 fusion polypeptide comprises, consists of, or consists essentially of (i) an N-truncated fragment of a Sun1 gene, where the fragment encodes at least a portion of a portion of a Sun1 protein, and where the portion of the Sun1 protein is 400 to 600 amino acids in length, and (ii) a sequence encoding a tag polypeptide, and (b) a promoter sequence specific for a selected cell type, where the promoter sequence is operably linked to the sequence encoding the tagged Sun1 fusion polypeptide, and is effective to drive expression of the sequence encoding the tagged Sun1 fusion polypeptide in the selected cell type, and wherein the tagged Sun1 fusion polypeptide is expressed in and thereby labels the selected cell type. The N-truncated fragment of the Sun1 gene can encode a polypeptide with at least 95% sequence identity to amino acids 208-757 of the mouse Sun1 protein having the amino acid sequence set forth in SEQ ID NO:1. The tag polypeptide can be a fluorescent polypeptide (e.g., a GFP, such as a GFP having an amino acid sequence that is at least 95% identical to the superfolder GFP sequence set forth in SEQ ID NO:2). The construct can further include virus sequences (e.g., AAV sequences or lentivirus sequences). The promoter sequence can be specific for PV+ neurons, such that the sequence encoding the tagged Sun1 fusion polypeptide is expressed in PV+ neurons but not expressed in at least 90% of PV- neurons.

In another aspect, this document features a method for labeling a selected cell type within a population of different cell types, where the method includes (a) introducing into the population of different cell types (i) a first nucleic acid construct containing (1) a sequence encoding a tagged Sun1 fusion polypeptide and (2) a sequence encoding a tag polypeptide, where the tagged Sun1 fusion polypeptide comprises, consists of, or consists essentially of an N-truncated fragment of a Sun1 gene, where the fragment encodes a

portion of a Sun1 protein, and wherein the portion of the Sun1 protein is 400 to 600 amino acids in length, and (2) a first lox sequence flanking the 5' end of the sequence encoding the tagged Sun1 fusion polypeptide and a second lox sequence flanking the 3' end of the sequence encoding the tagged Sun1 fusion polypeptide, and (3) a promoter sequence downstream of the lox sequence flanking the 3' end of the sequence encoding the tagged Sun1 fusion polypeptide, such that the sequence encoding the tagged Sun1 fusion polypeptide is in reverse orientation with respect to the promoter, and (ii) a second nucleic acid construct comprising a sequence encoding a Cre recombinase operably linked to a promoter sequence specific for the selected cell type, and (b) incubating the population of different cell types under conditions in which the Cre recombinase is expressed in the selected cell type, where the expressed Cre recombinase excises the sequence encoding the tagged Sun1 fusion polypeptide from the nucleic acid construct and irreversibly re-inserts the sequence encoding the tagged Sun1 fusion polypeptide in the correct orientation for expression in the selected cell type. The N-truncated fragment of the Sun1 gene can encode a polypeptide with at least 95% sequence identity to amino acids 208-757 of the mouse Sun1 protein having the amino acid sequence set forth in SEQ ID NO:1. The tag polypeptide can be a fluorescent polypeptide (e.g., a GFP, such as a GFP having an amino acid sequence that is at least 95% identical to the superfolder GFP sequence set forth in SEQ ID NO:2). The construct can further include virus sequences (e.g., AAV sequences or lentivirus sequences).

In still another aspect, this document features a method for generating, by a data processing system, data representing a synthetic genetic sequence configured for labeling at least one cell type by causing expression of a marker in the at least one cell type. The method can include: receiving training data associating at least one feature of a genetic sequence of the at least one cell type with expression of a hallmark of the at least one cell type; training, based on the training data, a model configured to generate data representing a synthetic genetic sequence; receiving input data comprising the at least one feature; and generating the synthetic genetic sequence in response to receiving the input data comprising the at least one feature. The model can include a generative adversarial network (GAN). The GAN can be a deep convolutional GAN (DCGAN). The GAN can include: a generator network configured to receive latent random noise data as the input data and generate the synthetic genetic sequence; and a discriminator network configured to generate a probability value representing whether the input sequence is drawn from the

synthetic genetic sequence from the generator network or from a distribution of natural genetic sequences. The method can include training the generator network and the discriminator network of the GAN by alternating the input data between synthetic sequence derived from the latent noise data input to the generator network and natural genetic sequence data. The method can include: receiving input data representing endogenous genomic sequences underlying previously profiled open chromatin regions of at least two cell types; and updating the discriminator network for distinguishing between the at least two cell types. The at least two cell types can include parvalbumin positive (PV+) and parvalbumin negative (PV-) neurons, where the discriminator network is configured to distinguish between the PV+ and PV- neurons based on the input data. The method can include optimizing a feature of the synthetic genetic sequence by applying a class label to both the generator network and the discriminator network. The class label can be configured to force a first probability value for a first input type and a second probability value for a second input type that is different from the first input type. The feature can represent an enhancer of an activity in a cell type. The method can further include generating a nucleic acid sequence comprising the synthetic genetic sequence. The nucleic acid sequence can include the synthetic genetic sequence operably linked to a nucleotide sequence encoding a marker (e.g., a tagged Sun1 fusion polypeptide). The nucleic acid sequence can further include a virus nucleic acid sequence (e.g., an AAV or lentivirus sequence). The method can include: receiving results data representing a delivery of a nucleic acid including the synthetic genetic sequence to an organism, the results data representing a successful labeling of the at least one cell type or an unsuccessful labeling of the at least one cell type; updating the model using the results data; and generating an updated synthetic genetic sequence based on the updated model. The method can further include: receiving training data including the hallmark representing marker positive results marker negative results, or both; extracting at least one feature corresponding to the marker positive result or to the marker negative result; and adding the at least one feature to the model. The at least one feature can be a set of k-mer or gapped k-mer counts, and wherein extracting the at least one feature comprises scanning the sequence to determine the set of k-mer counts that form the sequence. The model can include a support vector machine that includes: a feature space; and a support vector representing a classification border in the feature space, where the method includes adding, to the support vector of the support vector machine, a given set

of k-mer counts within a predefined distance of the classification border in the feature space of the support vector machine. The model can include a neural network. The neural network can be a convolutional neural network. The neural network can include one or more weight values each associated with a feature of the synthetic genetic sequence. The feature can include a set of k-mer counts of the genetic sequence. The feature can represent a transcription factor binding motif. The synthetic genetic sequence can be configured to distinguish between parvalbumin positive (PV+) and parvalbumin negative (PV-) neurons; PV+ and excitatory (EXC) neurons; or PV+ and vasoactive intestinal peptide-expressing (VIP+) neurons.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention pertains. Although methods and materials similar or equivalent to those described herein can be used to practice the invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

The details of one or more embodiments of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

FIGS. 1A-1D demonstrate validation of the cSNAIL probe system in a non-limiting example of a neuron subtype in mice, with imaging and molecular data. **FIG. 1A** is a schematic of the Cre-dependent mechanism of the cSNAIL viral probe. The Sun1GFP coding sequence between the LoxP sites is expressed in the presence of Cre (left), but not in the absence of Cre (right). **FIGS. 1B and 1C** are images of brain tissue (M1 region) from a PV^{Cre}/ai14 mouse injected with the cSNAIL Sun1GFP virus. Viral labeling was indicated by green fluorescent protein (GFP), in direct comparison with transgenically labeled Cre⁺ (PV⁺) cells (originally in red). **FIG. 1D** is an image showing nuclei in suspension after brain tissue homogenization. Nuclei were stained blue with 4',6-diamidino-2-phenylindole (DAPI) stain, and virally tagged PV⁺ nuclei were co-labeled

with GFP. The GFP+ cells adhered to magnetic beads (straight arrow), while GFP- cells did not (curved arrows), allowing for separation of the two groups. **FIGS. 1E and 1F** are schematics showing that when compared with the INTACT (Isolation of Nuclei Tagged in Specific Cell Types) method, Assay for Transposase-Accessible Chromatin (ATAC-seq) data from cSNAIL probes yielded similar cell type-specific open chromatin profiles.

FIG. 2A is a flow diagram of a process for generating synthetic enhancers by a machine learning model.

FIG. 2B is a schematic of a conditional generative adversarial network (GAN) for generating synthetic enhancers to label cell types using SNAIL.

FIG. 3 is a series of graphs plotting the performance of gapped k-mer support vector machine (SVM) and convolutional neural network (CNN) models on predicting cell type-specific regulatory activity of genomic sequences in PV+ neurons relative to PV- neurons, excitatory neurons, and VIP+ neurons.

FIG. 4A is a set of images showing validation of two SNAIL probes for parvalbumin positive neurons, alongside a non-specific control virus. The images show SUN1GFP labeled cells and parvalbumin-immunoreactive neurons (*Swant PV27/Alexa Fluor 594*) within mouse cortical tissue. The scale bars measure 100 μ m. **FIGS. 4B and 4C** show that open chromatin signatures derived from SNAIL-isolated PV+ cells recapitulate known PV+ features from cSNAIL and single nucleus ATAC-seq data. **FIG 4C** depicts hypergeometric enrichment significance of the overlap between cell type-specific SNAIL open chromatin regions and cluster-specific snATAC-seq open chromatin regions (data from data.nemoarchive.org/biccn/assay/chromatin/cemba/scell/processed/analysis/EckerRen_Mouse_MOp_methylation_ATAC/).

FIG. 5A is a schematic depicting the structure of the pAAV-Ef1a-DIO-SUN1GFP-WPRE-pA construct. **FIGS. 5B and 5C** depict the structures of two PV SNAIL pAAV constructs.

FIGS. 6A-6C show that cSNAIL targets Cre+ nuclei with high precision. **FIG. 6A** is an image showing that SUN1GFP expression from cSNAIL correctly localized to the nuclear envelope in tissue. Specifically, SUN1GFP (yielding green fluorescence) from cSNAIL was expressed in a Cre+ cell in cortical tissue from a Pvalb-2A-Cre/Ai14 mouse that expresses tdTomato as a reporter of Cre (yielding magenta fluorescence). **FIG. 6B** is an image showing that SUN1GFP expression from cSNAIL was sufficient to bind anti-

GFP-coated magnetic beads for affinity purification. **FIG. 6C** is a series of images showing that SUN1GFP expression from cSNAIL was restricted to Cre⁺ cells in the cortex, striatum, and GPe. These images show SUN1GFP expression from cSNAIL in a Pvalb-2A-Cre/Ai14 mouse. The tissue slices were also stained for NeuN, a pan-neuronal marker with blue fluorescence. cSNAIL virus specificity and efficiency within each brain region are indicated below the images. Standard errors were calculated across four independent images. The total SUN1GFP⁺ and/or Cre⁺ cells in the analysis were cortex N = 1088 cells, striatum N = 107 cells, and GPe N = 773 cells. The expected neuron transduction metrics were reported elsewhere (Chan et al., *Nat. Neurosci.* 20:1172-1179, 2017). ND = no data.

FIGS. 7A-7C show that cSNAIL-isolated nuclei from Pvalb-2A-Cre mice recapitulate PV⁺ transcriptomic and epigenomic signals from external data. **FIG. 7A** is a genome browser visualization of ATAC-seq and RNA-seq signal at the Pvalb locus from pooled p-value tracks. cSNAIL data is labeled in black and is compared with publicly available data (Mo et al., *Neuron* 86:1369-1384, 2015) labeled in grey. **FIG. 7B** shows enrichment of mouse cortical snATAC-seq cluster markers (LFC > 0.25 and p < 0.05 relative to all other clusters) within cell type-specific ATAC-seq peaks in the data by the hypergeometric test (data and method from Signac vignette, available online at github.com/timoast/signac). **FIG. 7C** shows expression of cortical cSNAIL PV⁺ gene signatures per cell of external mouse cortex snRNA-seq (data from portal.nemoarchive.org).

FIGS. 8A-8F show molecular signatures of PV⁺ neurons across brain regions. **FIG. 8A** shows the numbers of differentially expressed genes (DESeq2 $p_{\text{adj}} < 0.01$ and $|\text{LFC}| > 1$) between cSNAIL-isolated PV⁺ and PV⁻ fractions from the cortex, striatum, and GPe of sham animals. **FIG. 8B** shows normalized read counts from example marker transcripts in each sample. DESeq2 normalized counts were converted to the proportion that they represent of the total read count for that transcript within the given brain region. **FIG. 8C** shows representative pathway enrichments for pan-PV⁺ and pan-PV⁻ marker genes by g:Profiler (Peterson et al., *Nucleic Acids Res* 47:W191-W198, 2019). The size of the bubble represents the negative log of the adjusted p-value after multiple hypotheses correction. **FIG. 8D** shows the numbers of cell type-enriched ATAC-seq peaks ($p_{\text{adj}} < 0.01$ and $|\text{LFC}| > 1$) in PV⁺ populations from each brain region. **FIG. 8E** shows DESeq2 normalized counts of ATAC-seq reads within 5kb of the transcription start site of marker

genes. Again, counts were normalized to the total counts for that transcript in the cortex, striatum, or GPe. **FIG. 8F** shows examples of motifs enriched in PV+ specific or PV-specific ATAC-seq peaks relative to all peaks in that cell type by AME (McLeay and Bailey, *BMC Bioinformatics* 11:165, 2010). Each brain region was assessed separately, and these motifs were significantly enriched in PV+ or PV- sequences across all brain regions.

FIGS. 9A-9D show that divergent gene expression in dopamine depleted mice is largely restricted to GPe PV+ neurons and implicates the HIF pathway. **FIG. 9A** is a table showing the numbers of genes that have differential expression with dopamine depletion (DD) in each cell type. **FIG. 9B.** is a volcano plot of DD-affected genes in GPe PV+ neurons. Genes meeting the significance threshold of $p_{adj} < 0.05$ (above the dotted line) are listed in order of significance. **FIG. 9C** shows that DD is associated with a specific increase in Hif2a (Eras1) transcription and its target genes in GPe PV+ neurons. The animals are referred to as S1 and S2 for sham animals and DD1-4 for 6-OHDA treated animals in order of depletion severity. **FIG. 9D** is a graph plotting the proportion of GPe PV+ cells that express Hif2a protein, demonstrating that there was a significant increase between the images from healthy tissue and DD tissue (t-test).

FIGS. 10A-10C show that open chromatin changes associated with DD showed cell type-specific HIF motif enrichment. **FIG. 10A** is a table showing the numbers of ATAC-seq regions in each cell type that were differentially open in DD animals compared to sham animals in a categorical analysis. **FIG. 10B** shows HIF family transcription factor binding motifs as defined in JASPAR 2018 (Khan et al., *Nucleic Acids Res* 46:260-266, 2018). **FIG. 10C** is a series of graphs plotting cell type-specific HIF family motif enrichments among sets of sequences underlying ATAC-seq peak summits that increase in accessibility in DD animals.

FIGS. 11A-11B show the quantification of dopamine depletion per hemisphere. **FIG. 11A** is a graph plotting the levels of striatal tyrosine hydroxylase for each hemisphere of four 6-OHDA lesioned animals and two sham animals. The p-values reflect a significant difference in populations by the standard t-test. **FIG. 11B** includes examples of the quantified images, indicated by the arrows in **FIG. 11A**.

FIGS. 12A-12D show that the ATAC-seq data was high quality. **FIGS. 12A and 12B** show that ATAC-seq samples tended to cluster with other samples of the same tissue and cell type by principal component analysis and hierarchical clustering of genome-wide

open chromatin profiling. **FIG. 12C** is a graph demonstrating that the data exhibited the characteristic periodicity in fragment length distributions of high quality ATAC-seq data, reflecting nucleosome positioning. The plot shown is a representative example from one GPe PV+ sample. **FIG. 12D** shows that the ATAC-seq signal was enriched at the transcription start sites (TSS), indicative of high signal-to-noise. The displayed data are a representative example from one GPe PV+ sample.

FIGS. 13A and 13B show additional properties of the mouse cortex snRNA-seq data. **FIG. 13A** shows cluster annotations per cell of the snRNA-seq data with UMAP embedding. **FIG. 13B** shows that *Pvalb* gene expression is highest within the *Pvalb* interneuron cluster (circled).

FIGS. 14A and 14B show that PV+ and PV- marker genes from cSNAIL were expressed in relevant cell clusters from scRNA-seq. DropViz was used to define scRNA-seq clusters in the frontal cortex, striatum, and globus pallidus. Per cell expression of each indicated gene is shown by the size of the points. Where meta-cell expression of a gene was greater than 70% of the maximum expression in that meta-cell, the cluster is labeled by its cluster number.

FIGS. 15A and 15B show RNA-seq sample similarities. **FIG. 15A** is a graph showing that the first two principal components separated samples by brain region and cell type. **FIG. 15B** shows that in hierarchical clustering, samples mainly clustered by brain region and then cell type.

FIG. 16 is a series of images showing that with dopamine depletion, the proportion of GPe PV+ neurons that express Hif2a protein increases. Example images from healthy and DD GPe tissue are shown. Some instances of PV+Hif2a+ cells are highlighted by white arrows. The scale bars measure 100µm.

FIG. 17 shows the overlap between DD-affected ATAC-seq peaks of different cell types. The width of each outer segment of the circle shows how many DD-affected peaks were recovered in that cell type. Bidirectional arcs between two cell types signify shared DD-affected peaks where the width of the connection represents the number of shared peaks. The data were plotted with circize (Gu et al., *Bioinformatics* 30:2811-2812, 2014).

FIG. 18 is a summary of enriched annotations in DD-affected ATAC-seq peak sets. Enrichments were determined using GREAT for each differential set of peaks against a background of all peaks in that cell type. The arrow of a term indicates that it

was enriched in DD-increasing peaks (up arrow) or DD-decreasing peaks (down arrow).

FIG. 19 is a graph plotting TSS accessibility trends with differential gene expression in GPe PV+ neurons (t-test).

DETAILED DESCRIPTION

Genomic assays of specific cell types first require identifying and separating the cells of interest from the rest of the tissue. In brain tissue, for example, this can be especially difficult due to the high diversity of cell types that are tightly packed together and connected with long, fragile projections. In the adult brain, it is impractical to separate whole cells in sufficient quantities because these projections can break off, spilling their contents into the media. Instead, researchers have turned to sorting the nuclei of cell types of interest. One method for isolation of nuclei without using fixatives (which are incompatible with some genomic assays) is the INTACT method (Mo et al., *supra*), which requires a double transgenic mouse. The first transgene expresses Cre recombinase in a cell type-specific manner, and the second expresses the Sun1GFP fusion protein only in the presence of Cre recombinase. An INTACT mouse therefore expresses Sun1GFP exclusively in the cells of interest.

Sun1 is a mammalian nuclear envelope protein that can serve as a robust tool for cell labeling and isolation from homogenized tissues, due to its localization. The Sun1 coding sequence can be fused with sequences encoding two copies of super-folder GFP, which reliably folds into the correct conformation even in protein fusions. The result of Sun1GFP expression is GFP that is tightly anchored to the nuclear envelope. Tissue (e.g., brain tissue) from transgenic mice expressing Sun1GFP in one or more particular cell types (e.g., particular neuron types) can be homogenized into a suspension of single nuclei, and GFP+ nuclei can be separated from GFP- nuclei using affinity purification in which magnetic beads coated with an anti-GFP antibody bind GFP+ nuclei, but not GFP- nuclei. The beads with GFP+ nuclei can then be pulled away from the GFP- nuclei with a magnet and used for further studies.

The INTACT method has been instrumental for studying specific subsets of neurons in isolation. Its reliance on double transgenic mice has limited its use, however, because transgenic breeding can be highly variable and resource-intensive, and because transgenic strains are often only feasible for cell types that can be characterized by a single marker locus, leaving many populations inaccessible. The current technology is

limited mainly to the laboratory *Mus musculus* species, and the difficult nature of creating new transgenic organisms means that it is impractical to extend the framework to non-model organisms. In addition, multiple transgenic breeding makes it problematic to use cell type-specific mouse lines in conjunction with other transgenic traits, including disease models or strains that label an additional cell population of interest.

As described herein, viral delivery of cell type-specific labels can provide much more flexibility across cell types and species, as compared to transgene expression. In general, the SNAIL and cSNAIL materials and methods provided herein utilize an affinity purification strategy that is similar to the INTACT method for immunopurification of nuclei, but without the need for double transgenic animals. In some cases, instead of transgenesis, viruses can be used to deliver a simple, compact labeling gene that is sufficient to achieve success. When this non-limiting embodiment is paired with a virus [e.g., Adeno-associated virus (AAV) or lentivirus] delivery method, the probes can be introduced into any appropriate mammals (e.g., mice, rats, sheep, pigs, dogs, or non-human primates) through intravenous injection, direct injection into the brain, or any other appropriate method. This delivery procedure is a much more time and resource-efficient way to tag cell populations when compared with the intricacies of transgenic breeding. The method also drastically minimizes the number of collateral animals that are bred for an experiment but cannot be used due to undesirable genotypes.

The probes provided herein include, for example, Cre-dependent (cSNAIL) viral probes that are immediately compatible with existing strains of Cre transgenic mice, as well as Cre-independent (SNAIL) viral probes, which are developed individually for different cell types and, as described more fully below, accomplish complete independence from animal strain. Both probe types contain a newly designed sequence that includes an N-truncated exon-only version of the Sun1GFP fusion protein. This sequence is 1/16th the size of the INTACT sequence (2,367 bp), and was sufficient to achieve successful nuclear envelope GFP labeling for affinity purification even *in vivo* (e.g., in the brain), which was rather surprising. The smaller size enables much more flexibility in gene delivery, including compatibility with AAV virus.

The cell type specificity of the cSNAIL viral probe is conferred through its dependence on DNA rearrangement by Cre recombinase (**FIG. 1A**), which is only present in certain cells. Specifically, a double-floxed inverted orientation (DIO) system (Atasoy et al., *J. Neurosci.* 28:7025-7030, 2008) is used, in which the Sun1GFP sequence is

delivered in the reverse orientation, such that it cannot be transcribed. When the DNA encounters Cre recombinase protein, Cre can excise the Sun1GFP sequence and irreversibly re-insert it in the correct orientation for expression. This system is more tightly regulated than the lox-flanked stop codon system that is used to confer Cre dependence in the INTACT transgenic mouse strain (Atasoy et al., *supra*).

For SNAIL viral probes, a cell type-specific regulatory element is required to drive targeted gene expression in the desired cell type of interest. As described in the Examples below, a regulatory element was designed to label one specific cell type – parvalbumin positive (PV+) inhibitory interneurons – using *in silico screening*. Profiles of cell type-specific open chromatin were used for this neuron subtype to train machine learning models that link genome sequence to parvalbumin-specific regulatory activity. In particular, three gapped k-mer support vector machines were built to discriminate between the comparative regulatory sequence activity in PV+ vs. PV- neurons, PV+ vs. excitatory neurons, and PV+ vs. vasoactive intestinal peptide-expressing (VIP+) neurons. In addition to gapped k-mer support vector machines, another example embodiment trained similar convolutional neural networks (CNNs), which are the state-of-the-art model for relating raw genomic sequence to regulatory function. All of these models were used to screen PV+ neuron-specific regions of open chromatin for their ability to independently drive cell type-specific gene expression. A top candidate was chosen based on the predicted cell type-specificity, its length, and its proximity to cell type-specific genes. That candidate regulatory sequence was cloned into a plasmid vector in which it could drive the expression of Sun1GFP, and the vector was transduced into the mouse brain using AAV-PHP.eB (Chan et al., *Nat. Neurosci.* 20:1172-1179, 2017).

This document provides nucleic acid constructs that contain sequences encoding fusion polypeptides that include a Sun1 portion and a tag portion. The Sun1 portion enables the fusion polypeptide to localize to the nuclear membrane, while the tag portion enables cells and/or nuclei containing the expressed fusion polypeptide to be identified and isolated away from cells and/or nuclei that do not express the fusion polypeptide. The nucleic acids provided herein can further include other sequences, such as sequences to control expression of the fusion polypeptide (e.g., sequences to promote expression in particular types of cells, and/or sequences to prevent expression unless Cre is present).

The terms “nucleic acid” and “polynucleotide” can be used interchangeably, and refer to both RNA and DNA, including cDNA, genomic DNA, synthetic (e.g., chemically

synthesized) DNA, and DNA (or RNA) containing nucleic acid analogs. Polynucleotides can have any three-dimensional structure. A nucleic acid can be double-stranded or single-stranded (i.e., a sense strand or an antisense single strand). Non-limiting examples of polynucleotides include genes, gene fragments, exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers, as well as nucleic acid analogs.

As used herein, “isolated,” when in reference to a nucleic acid, refers to a nucleic acid that is separated from other nucleic acids that are present in a genome, including nucleic acids that normally flank one or both sides of the nucleic acid in the genome. The term “isolated” as used herein with respect to nucleic acids also includes any non-naturally-occurring sequence, since such non-naturally-occurring sequences are not found in nature and do not have immediately contiguous sequences in a naturally-occurring genome.

An isolated nucleic acid can be, for example, a DNA molecule, provided one of the nucleic acid sequences normally found immediately flanking that DNA molecule in a naturally-occurring genome is removed or absent. Thus, an isolated nucleic acid includes, without limitation, a DNA molecule that exists as a separate molecule (e.g., a chemically synthesized nucleic acid, or a cDNA or genomic DNA fragment produced by PCR or restriction endonuclease treatment) independent of other sequences, as well as DNA that is incorporated into a vector, an autonomously replicating plasmid, a virus (e.g., a pararetrovirus, a retrovirus, lentivirus, adenovirus, or herpes virus), or the genomic DNA of a prokaryote or eukaryote. In addition, an isolated nucleic acid can include a recombinant nucleic acid such as a DNA molecule that is part of a hybrid or fusion nucleic acid. A nucleic acid existing among hundreds to millions of other nucleic acids within, for example, cDNA libraries or genomic libraries, or gel slices containing a genomic DNA restriction digest, is not to be considered an isolated nucleic acid.

A nucleic acid can be made by, for example, chemical synthesis or polymerase chain reaction (PCR). PCR refers to a procedure or technique in which target nucleic acids are amplified. PCR can be used to amplify specific sequences from DNA as well as RNA, including sequences from total genomic DNA or total cellular RNA. Various PCR methods are described, for example, in PCR Primer: A Laboratory Manual, Dieffenbach

and Dveksler, eds., Cold Spring Harbor Laboratory Press, 1995. Generally, sequence information from the ends of the region of interest or beyond is employed to design oligonucleotide primers that are identical or similar in sequence to opposite strands of the template to be amplified. Various PCR strategies also are available by which site-specific nucleotide sequence modifications can be introduced into a template nucleic acid.

Isolated nucleic acids also can be obtained by mutagenesis. For example, a naturally occurring nucleic acid sequence can be mutated using standard techniques, including oligonucleotide-directed mutagenesis and site-directed mutagenesis through PCR. See, Short Protocols in Molecular Biology, Chapter 8, Green Publishing Associates and John Wiley & Sons, edited by Ausubel *et al.*, 1992.

Recombinant nucleic acid constructs (e.g., vectors) containing sequences encoding the tagged Sun1 fusion polypeptides also are provided herein. A “vector” is a replicon, such as a plasmid, phage, or cosmid, into which another DNA segment may be inserted so as to bring about the replication of the inserted segment. Vector backbones include, for example, plasmids, viruses, artificial chromosomes, bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), and phage artificial chromosomes (PACs), as well as RNA vectors, and linear or circular DNA or RNA molecules that include chromosomal, non-chromosomal, semi-synthetic, or synthetic nucleic acids. Vectors include those capable of autonomous replication (episomal vectors) and/or expression of nucleic acids to which they are linked (expression vectors). Generally, a vector is capable of replication when associated with the proper control elements. The term “vector” includes cloning and expression vectors, as well as viral vectors and integrating vectors. An “expression vector” is a vector that includes one or more expression control sequences to control and regulate the transcription and/or translation of another DNA sequence. Suitable expression vectors include, without limitation, plasmids and viral vectors derived from, for example, bacteriophage, baculoviruses, tobacco mosaic virus, herpes viruses, cytomegalovirus, retroviruses, vaccinia viruses, adenoviruses, and adeno-associated viruses. Numerous vectors and expression systems are commercially available.

Viral vectors include, without limitation, retrovirus, adenovirus, parvovirus (e.g., adeno associated viruses), coronavirus, negative strand RNA viruses such as orthomyxovirus (e.g., influenza virus), rhabdovirus (e.g., rabies and vesicular stomatitis virus), paramyxovirus (e.g., measles and Sendai), positive strand RNA viruses such as picornavirus and alphavirus, and double-stranded DNA viruses including adenovirus,

herpesvirus (e.g., Herpes Simplex virus types 1 and 2, Epstein-Barr virus, cytomegalovirus), and poxvirus (e.g., vaccinia, fowlpox and canarypox). Other viruses include Norwalk virus, togavirus, flavivirus, reoviruses, papovavirus, hepadnavirus, and hepatitis virus, for example. Examples of retroviruses include: avian leukosis-sarcoma, mammalian C-type, B-type viruses, D type viruses, HTLV-BLV group, lentivirus, spumavirus (Coffin, "Retroviridae: The viruses and their replication," in Fundamental Virology, Third Edition, B. N. Fields, et al., eds., Lippincott-Raven Publishers, Philadelphia, 1996).

The terms "regulatory region," "control element," and "expression control sequence" refer to nucleotide sequences that influence transcription or translation initiation and rate, and stability and/or mobility of the transcript or polypeptide product. Regulatory regions include, without limitation, promoter sequences, enhancer sequences, response elements, protein recognition sites, inducible elements, promoter control elements, protein binding sequences, 5' and 3' untranslated regions (UTRs), transcriptional start sites, termination sequences, polyadenylation sequences, introns, and other regulatory regions that can reside within coding sequences, such as secretory signals, Nuclear Localization Sequences (NLS) and protease cleavage sites.

As used herein, "operably linked" means incorporated into a genetic construct so that expression control sequences effectively control expression of a coding sequence of interest. A coding sequence is "operably linked" and "under the control" of expression control sequences in a cell when RNA polymerase is able to transcribe the coding sequence into RNA, which if an mRNA, then can be translated into the protein encoded by the coding sequence. Thus, a regulatory region can modulate, e.g., regulate, facilitate or drive, transcription in the plant cell, plant, or plant tissue in which it is desired to express a modified target nucleic acid.

A promoter is an expression control sequence composed of a region of a DNA molecule, typically (but not always) within 100 nucleotides upstream of the point at which transcription starts (generally near the initiation site for RNA polymerase II). Promoters are involved in recognition and binding of RNA polymerase and other proteins to initiate and modulate transcription. To bring a coding sequence under the control of a promoter, it typically is necessary to position the translation initiation site of the translational reading frame of the polypeptide between one and about fifty nucleotides downstream of the promoter. A promoter can, however, be positioned as much as about 5,000 nucleotides upstream of the translation start site, or about 2,000 nucleotides upstream of the

transcription start site. A promoter typically includes at least a core (basal) promoter. A promoter also may include at least one control element such as an upstream element. Such elements include upstream activation regions (UARs) and, optionally, other DNA sequences that affect transcription of a polynucleotide such as a synthetic upstream element.

The choice of promoters to be included depends upon several factors, including, but not limited to, efficiency, selectability, inducibility, desired expression level, and cell or tissue specificity. For example, tissue-, organ- and cell-specific promoters that confer transcription only or predominantly in a particular tissue, organ, and cell type, respectively, can be used. In some cases, a promoter that is active in a particular type of cell (e.g., a particular type of neuron) can be used. Such promoters – or other expression control sequences – can be identified using methods such as those described herein, and can be used in the SNAIL constructs provided herein. Alternatively, constitutive promoters can promote transcription of an operably linked nucleic acid in essentially any tissue of an organism. Such promoters can be used in the cSNAIL constructs provided herein. Other classes of promoters include, without limitation, inducible promoters that confer transcription in response to external stimuli such as chemical agents, developmental stimuli, or environmental stimuli.

A 5' untranslated region (UTR) is transcribed, but is not translated, and lies between the start site of the transcript and the translation initiation codon and may include the +1 nucleotide. A 3' UTR can be positioned between the translation termination codon and the end of the transcript. UTRs can have particular functions such as increasing mRNA message stability or translation attenuation. Examples of 3' UTRs include, but are not limited to, polyadenylation signals and transcription termination sequences. A polyadenylation region at the 3'-end of a coding region can also be operably linked to a coding sequence.

Recombinant nucleic acid constructs can include a polynucleotide sequence inserted into a vector suitable for transformation of cells (e.g., plant cells or animal cells). Recombinant vectors can be made using, for example, standard recombinant DNA techniques (*see, e.g.,* Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual, 2nd ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY). In some cases, a recombinant nucleic acid sequence as described herein can integrate into the genome of a cell via illegitimate (i.e., random, non-homologous, non site-specific) recombination, or a

recombinant nucleic acid sequence as described herein can be adapted to integrate into the genome of a cell via homologous recombination. Nucleic acid sequences adapted for integration via homologous recombination are flanked on both sides with sequences that are similar or identical to endogenous target nucleotide sequences, which facilitates integration of the recombinant nucleic acid at the particular site(s) in the genome containing the endogenous target nucleotide sequences. Nucleic acid sequences adapted for integration via homologous recombination also can include a recognition site for a sequence-specific nuclease. Alternatively, the recognition site for a sequence-specific nuclease can be located in the genome of the cell to be transformed.

The nucleic acid constructs described herein typically contain a sequence encoding a tagged Sun1 fusion polypeptide. The term “polypeptide” as used herein refers to a compound of two or more subunit amino acids, regardless of post-translational modification (e.g., phosphorylation or glycosylation). The subunits may be linked by peptide bonds or other bonds such as, for example, ester or ether bonds. The term “amino acid” refers to either natural and/or unnatural or synthetic amino acids, including D/L optical isomers.

By “isolated” or “purified” with respect to a polypeptide it is meant that the polypeptide is separated to some extent from cellular components with which it is normally found in nature (e.g., other polypeptides, lipids, carbohydrates, and nucleic acids). An purified polypeptide can yield a single major band on a non-reducing polyacrylamide gel. A purified polypeptide can be at least about 75% pure (e.g., at least 80%, 85%, 90%, 95%, 97%, 98%, 99%, or 100% pure). Purified polypeptides can be obtained by, for example, extraction from a natural source, by chemical synthesis, or by recombinant production in a host cell or transgenic plant, and can be purified using, for example, affinity chromatography, immunoprecipitation, size exclusion chromatography, and ion exchange chromatography. The extent of purification can be measured using any appropriate method, including, without limitation, column chromatography, polyacrylamide gel electrophoresis, or high-performance liquid chromatography.

The Sun1 portion of the fusion polypeptides provided herein (and encoded by the nucleic acids provided here) can include a fragment of a mouse Sun1 protein. A representative Sun1 protein sequence is provided in SEQ ID NO: 1:

Mus musculus Sun1 (NCBI Ref. NM_001256118.1)

MDFSRLHTYTPPQCVPENTGYTYALSSSYSSDALDFETEHKLEPVFDSPRMSRRS
LRLVTTASYSSGDSQAIDSHISTSRAATPAKGRETRTVKQRRSASKPAFSINHLSGK
GLSSSTSHDSSCSLRSATVLRHPVLDESLIREQTKVDHFWGLDDDGDLKGGNKA
ATQGGNGELAAEVASSNGYTCRDCRMLSARTDALTAHSAIHGTTSRVYSRDRTLK
PRKAASGTFWWLGSGWYQFVTLISWLVNLFLLTRCLRNICKVFVLLLPLLLLGA
GVSLWGQGNFFSLLPVLNWTAMQPTQRVDDSKGMHRPGPLPPSPPPKVDHKAS
QWPQESDMGQKVASLSAQCHNHDERLAELTVLLQKLQIRVDQVDDGREGLSLW
VKNVVGQHLQEMGTIEPPDAKTDFMTFHHDEVRSLNLEDVLRKLTEKSEAIQK
ELEETKLKAGSRDEEQPLLDREVQHLELELNLLKSQLSDWQHLLKTSCEQAGARIQE
TVQLMFSEDQQGGSLEWLLEKLSSRFVSKDELQVLLHDLELKLLQNITHHITVTG
QAPTSEAIVSAVNQAGISGITEAQAHIIVNNALKLYSQDKTGMVDFALESAGGGSIL
STRCSETYETKTALLSLFGVPLWYFSQSPRVVIQPDYPGNCWAFKGSQGYLVVR
LSMKIYPTTFTMEHIPKTLSPGTGNISSAPKDFAVYGLETEYQEEGQPLGRFTYDQE
GDSLQMFHTLERPDQAFQIVELRVLSNWGHPEYTCLYRFRVHGEPIQ (SEQ ID
NO:1)

The Sun1 fragment included in a fusion polypeptide as provided herein can have a length between about 20 amino acids and about 750 amino acids (e.g., about 20 to 50 amino acids, about 50 to about 100 amino acids, about 100 to about 200 amino acids, about 200 to about 400 amino acids, about 400 to about 600 amino acids, or about 600 to about 750 amino acids). In some cases, for example, the Sun1 portion of a tagged Sun1 fusion polypeptide can include about 400 to about 600 consecutive amino acids from SEQ ID NO:1. In some cases, the Sun1 portion of the tagged Sun1 fusion polypeptide can have at least 90% sequence identity (e.g., at least 92%, 93%, 95%, 97%, 98%, or 99% sequence identity) to a fragment of SEQ ID NO:1 that is about 20 to about 800 amino acids in length (e.g., about 20 to 50, about 50 to about 100, about 100 to about 200, about 200 to about 400, about 400 to about 600, or about 600 to about 750 amino acids in length). In some cases, as described in the Examples, a Sun1 fragment can include amino acids 208 to 757 of SEQ ID NO:1, or can have a sequence that is at least 90% (e.g., at least 92%, 93%, 95%, 97%, 98%, or 99%) identical to amino acids 208 to 757 of SEQ ID NO:1. In some cases, a Sun1 fragment can include a portion of the fragment containing amino acids 208 to 757 of SEQ ID NO:1 (e.g., a portion that is 20 to 50, 50 to 100, 100 to 300, or 300 to 500 amino acids in length), or a Sun1 fragment can have a sequence that is at least 90% (e.g., at least 92%, 93%, 95%, 97%, 98%, or 99%) identical to a portion of the fragment containing amino acids 208 to 757 of SEQ ID NO:1.

The percent sequence identity between a particular amino acid sequence and a sequence referenced by a particular sequence identification number is determined as follows. First, an amino acid sequence is compared to the sequence set forth in a particular query sequence using the BLAST 2 Sequences (B12seq) program from the stand-alone version of BLASTZ containing BLASTP version 2.0.14. This stand-alone version of BLASTZ can be obtained online at fr.com/blast or at ncbi.nlm.nih.gov. Instructions explaining how to use the B12seq program can be found in the readme file accompanying BLASTZ. B12seq performs a comparison between two sequences using the BLASTN or BLASTP algorithm. BLASTN is used to compare nucleic acid sequences, while BLASTP is used to compare amino acid sequences. To compare two amino acid sequences, the options of B12seq are set as follows: -i is set to a file containing the first amino acid sequence to be compared (e.g., C:\seq1.txt); -j is set to a file containing the second amino acid sequence to be compared (e.g., C:\seq2.txt); -p is set to blastp; -o is set to any desired file name (e.g., C:\output.txt); and all other options are left at their default setting. For example, the following command can be used to generate an output file containing a comparison between two amino acid sequences: C:\B12seq -i c:\seq1.txt -j c:\seq2.txt -p blastp -o c:\output.txt. If the two compared sequences share homology, then the designated output file will present those regions of homology as aligned sequences. If the two compared sequences do not share homology, then the designated output file will not present aligned sequences.

Once aligned, the number of matches is determined by counting the number of positions where an identical nucleotide or amino acid residue is presented in both sequences. The percent sequence identity is determined by dividing the number of matches either by the length of the sequence set forth in the identified sequence or by an articulated length (e.g., 100 consecutive amino acid residues from a sequence set forth in an identified sequence), followed by multiplying the resulting value by 100. For example, the reference mouse Sun1 amino acid sequence set forth in SEQ ID NO:1 is 757 residues in length. An amino acid sequence that has 740 matches when aligned with the reference sequence is 97.8 percent identical to the reference sequence (i.e., $740/757 \times 100 = 97.8$). It is noted that the percent sequence identity value is rounded to the nearest tenth. For example, 75.11, 75.12, 75.13, and 75.14 are rounded down to 75.1, while 75.15, 75.16, 75.17, 75.18, and 75.19 are rounded up to 75.2. It also is noted that the length value will always be an integer.

In addition to the Sun1 portion, the fusion polypeptides provided herein can include a tag sequence designed to facilitate manipulation or detection (e.g., localization, purification, and/or isolation) of the expressed polypeptide. Any suitable tag can be included. In some cases, a fluorescent tag can be used (e.g., GFP, yellow fluorescent protein, or mCherry). A representative amino acid sequence for a superfolder GFP is provided in SEQ ID NO:2:

MSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTCLKFICTTGKLPV
PWPTLVTTLTYGVCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTR
AEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFSHNVYITADKQKNGIKAN
FKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHM
VLLEFVTAAGITHGMDELYK (SEQ ID NO:2)

In some cases, a tagged Sun1 fusion polypeptide can include an amino acid sequence that is at least 90% (e.g., at least 92%, 93%, 95%, 97%, 98%, or 99%) identical to the sequence set forth in SEQ ID NO:2. Other suitable tags that can be included in the fusion polypeptides provided herein include, without limitation, glutathione S-transferase (GST), polyhistidine, c-myc, hemagglutinin, or Flag™ tag (Kodak, New Haven, CT). The tag sequence can be located within the Sun1 portion of the fusion polypeptide, or can be at or adjacent to the carboxyl or amino terminus of the Sun1 portion of the fusion polypeptide.

In addition to the sequence encoding the tagged Sun1 fusion polypeptide, the cSNAIL constructs provided herein can include a promoter sequence (e.g., a constitutive promoter, an organ specific promoter, or a promoter that is activated by neural activity or a particular disease state such as oxidative stress or cancer, such as an IEG-sensitive promoter as disclosed by Sorensen et al., *eLife* 5:e13918, 2016) and loxP sites flanking either side of the sequence encoding the fusion polypeptide, such that one of the loxP sites separates the sequence encoding the fusion polypeptide from the promoter. As depicted in **FIGS. 1A and 5A**, for example, the sequence encoding the tagged Sun1 fusion polypeptide can be inverted between the loxP sequences, such that the fusion polypeptide is not expressed until after the coding sequence is excised by a Cre recombinase and reinserted in the “correct” orientation.

The SNAIL constructs provided herein (see, e.g., **FIGS. 5B and 5C**) lack loxP sequences, and contain promoter and/or enhancer sequences that drive expression of the tagged Sun1 fusion polypeptide in a specific cell type, thus allowing for identification and isolation of cells of the specific, selected type. In some applications of SNAIL, synthetic enhancer sequences can be constructed purely through computational optimization, and

can be paired with SNAIL to generate more specific and efficient cell type labels than endogenous genomic sequences prioritized using the CNNs and SVMs. In addition to other generative models, deep convolutional generative adversarial networks (DCGANs) that have been applied successfully for deep generative design of realistic faces, objects, and art can be used. Sequence classification tasks share many parallels with image classification. CNNs are the state-of-the-art model for both tasks, and DCGANs, which utilize CNNs under the hood for discrimination, are therefore an ideal choice for generating realistic sequences. Further, conditional DCGANs, which are a variant of GANs, can be utilized to generate sequences on the condition that they confer specific activity in a given cell type of interest versus other surrounding cell types. These can be trained on any neural cell type for which high resolution open chromatin or enhancer activity measurements exist. The generated sequences can then be cloned into the SNAIL vector to screen and generate high quality AAV-delivered cell type labels.

Generally, as previously discussed, machine learning models can be used for generating data representing a synthetic genetic sequence configured for labeling at least one cell type by causing expression of a marker in the at least one cell type. The models can include GANs, such as DCGANs, CNNs, support vector machines (SVMs), and so forth. The machine learning models are configured for generating the synthetic genetic sequences that are candidates for the SNAIL process (e.g., as viral probes).

FIG. 2A shows an example process 250 for training and executing the machine learning model to generate the synthetic genetic sequences. The process 250 includes receiving (252) training data associating at least one feature of a genetic sequence of the at least one cell type with expression of a hallmark of the at least one cell type. The process 250 includes training (254), based on the training data, a model configured to generate data representing a synthetic genetic sequence. The process 250 includes receiving (256) input data including the at least one feature of the genetic sequence. The process 250 includes generating (258) the synthetic genetic sequence in response to receiving the input data comprising the at least one feature.

In some implementations, the machine learning model includes a GAN. The GAN generally includes a generator network configured to receive latent random noise data as the input data and generate the synthetic genetic sequence and a discriminator network configured to generate a probability value representing whether the input sequence is drawn from the synthetic genetic sequence from the generator network or from a

distribution of natural genetic sequences. Turning briefly to **FIG. 2B**, an example machine learning model 200 is shown for generating data representing a synthetic genetic sequence configured for labeling at least one cell type by causing expression of a marker in the at least one cell type.

Generally, the machine learning model 200 is configured to generate data that models the SNAIL constructs (e.g., viral probes) discussed throughout this specification. The model 200 is a particular model including a DCGAN that has a generator G, a sequence S, and a discriminator D. However, as previously described, while the DCGAN is shown for illustrative purposes, other machine learning models are possible, such as support vector machines (SVNs), neural networks (NN) such as CNNs, and so forth. The models for generating regulatory sequences is not specific for SVM and CNNs, but also can utilize generative components to the CNNs that relate genomic sequences to regulatory activity and use them to construct synthetic enhancer sequences. Machine learning models, such as model 200, are configured to construct synthetic enhancer sequences that are computationally optimized to drive cell type-specific labeling when used in an AAV construct, more so than the endogenous genomic sequences that were used to train the CNN classifiers.

Returning to **FIG. 2A**, in some implementations, the machine learning model is configured for generating a nucleic acid sequence comprising the synthetic genetic sequence. In this example, the nucleic acid sequence comprises the synthetic genetic sequence operably linked to a nucleotide sequence encoding a marker. For example, the marker can include a Sun1GFP fusion polypeptide. In another example, the nucleic acid sequence further comprises a virus nucleic acid sequence. Here, the virus can include an adeno-associated virus (AAV).

The machine learning model is configured for receiving results data representing a delivery of a nucleic acid including the synthetic genetic sequence to an organism. The results data represents a successful labeling of the at least one cell type or an unsuccessful labeling of the at least one cell type. To train the model, the model is updated using the results data. The model is then configured to generate an updated synthetic genetic sequence based on the updated model. Updating the model can also be done by receiving training data including the hallmark representing marker positive results marker negative results, or both. The at least one feature corresponding to the marker positive result or to the marker negative result is extracted. The at least one feature is then added to the model.

In some implementations, the feature is a set of k-mer or gapped k-mer counts. Extracting the at least one feature can include scanning the sequence to determine the set of k-mer counts that form the sequence.

Other models besides the GAN of FIG. 2B can be used. For example, a support vector machine includes a feature space a support vector representing a classification border in the feature space. The model is executed to add, to the support vector of the support vector machine, a given set of k-mer counts within a predefined distance of the classification border in the feature space of the support vector machine.

In another example, the machine learning model includes a neural network. For example, the machine learning model can include a convolutional neural network. The neural network is loaded with one or more weight values (also called activation values) each associated with a feature of the synthetic genetic sequence. For example, a feature of the neural network can include a set of k-mer counts of the genetic sequence. In some implementations, the feature represents a transcription factor binding motif.

In any of the machine learning models described in this specification, the synthetic genetic sequence that is generated from the model is configured to distinguish between cell types. The particular cell types are based on the training data used. For example, the genetic sequence that is generated from the model can be configured to distinguish between parvalbumin positive (PV+) and parvalbumin negative (PV-) cells (e.g., the generated sequence can lead to expression in at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of all PV+ neurons that are transduced by the virus and have the chance to express the machinery, but would not lead to expression in at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of PV- cells). For example, the genetic sequence that is generated from the model can be configured to distinguish between PV+ and excitatory (EXC) neurons (e.g., the generated sequence can lead to expression in at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of all PV+ neurons that are transduced by the virus and have the chance to express the machinery, but would not lead to expression in at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of EXC neurons). For example, the generated genetic sequence that is generated from the model can be configured to distinguish between PV+ and vasoactive intestinal peptide-expressing (VIP+) neurons (e.g., the generated sequence can lead to expression in at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of all PV+ neurons that are transduced by the virus and have the chance to express the

machinery, but would not lead to expression in at least 70%, at least 80%, at least 90%, at least 95%, or at least 99% of VIP+ neurons). However, other cell types can be distinguished by the generated genetic sequence based on the training data used. These examples are non-exhaustive and illustrate possible examples from hundreds or thousands of cell types. The regulatory sequence features learned may extend beyond cell types as well as to cellular states, such as neurons that in the process or responding to neural activity or cells that are responding to DNA damage.

As shown in **FIG. 2B**, the DCGAN for labeling at least one cell type by causing expression of a marker in the at least one cell type includes two neural network modules, a generator network G and a discriminator network D. The generator G receives latent random noise input (z) and outputs a synthetic sequence. The discriminator network is configured to generate probability value representing a probability that an input sequence is drawn from a distribution of real sequences. The parameters of both G and D can be tuned together using alternating phases of training with real sequences from existing data sources and synthetic sequences generated from G. As the models are trained, G generally generates synthetic sequences that are closer to the real data distribution than before. The discriminator is more likely to classify these generated sequences as genuine.

The generator is training G to generate realistic sequence data. For example, a GAN could be applied to generate synthetic enhancers that are likely highly active in PV+ or PV- neurons by training on real data drawn from endogenous genomic sequences underlying previously profiled open chromatin regions in PV+ and PV- neurons. Further, to optimize specific properties of the sequence, a conditional GAN can be used where an additional class label (y) is input to both G and D. For example, to generate sequences specific to PV+ neurons relative to PV- neurons, a GAN could be trained on real sequences active in both PV+ and PV-neurons, but the generation can be forced to be specific to PV+ neurons by setting y to be 1 for all sequences active in PV+ neurons and 0 for all sequences active in PV- neurons. This enables discriminative generation of sequences, where the optimized property is enhancer activity in a cell type relative to surrounding or background cell types.

Generally, using the GAN includes training the generator network and the discriminator network of the GAN by alternating the input data between synthetic sequence derived from the latent noise data input to the generator network and natural genetic sequence data. The GAN receives input data representing endogenous genomic

sequences underlying previously profiled open chromatin regions of at least two cell types and then updates the discriminator network for distinguishing between the at least two cell types. In an example, the at least two cell types include parvalbumin positive (PV+) and parvalbumin negative (PV-) neurons; however, other cell types can be used, and are described throughout this specification. In this example, the discriminator network is configured to distinguish between the PV+ and PV- neurons based on the input data, but other cell types can be used as well, such as excitatory (EXC) neurons, vasoactive intestinal peptide-expressing (VIP+) neurons, and so forth.

The GAN is configured for optimizing a feature of the synthetic genetic sequence by applying a class label to both the generator network and the discriminator network. The class label is configured to force a first probability value for a first input type and a second probability value for a second input type that is different from the first input type. In some implementations, the feature represents an enhancer of an activity in a cell type.

The model 200 is configured to receive (212) the noise and condition input together to generate a compressed representation of the output sequence. The compressed representation is unpacked (210) using a deconvolution process. Fully connected layers can be used in the model 200 to generate a synthetic sequence. The model 200 is configured to generate (208) sequence binary encoding to input to the discriminator. The discriminator is configured to determine whether the sequence it receives is a synthetic sequence or a naturally occurring sequence. The discriminator includes a one dimensional (1D) convolution layer configured to scan (206) the received sequences for patterns. Max pooling can be used (204) to down-sample the input sequence. This convolution and max pooling process can be iterated as needed. The class label (e.g., condition) and compressed output from multiple fully connected layers are output (202) into an output unit.

For SNAIL viral probes, a cell type-specific regulatory element is required to drive targeted gene expression in the desired cell type of interest. In one embodiment, a regulatory element can be configured to label one specific cell type, parvalbumin positive (PV+) inhibitory interneurons. In an experiment, profiles of cell type-specific open chromatin were used for this neuron subtype to train machine learning models that link genome sequence to parvalbumin-specific regulatory activity. Specifically, three gapped k-mer support vector machines were built that discriminate between the comparative regulatory sequence activity in PV+ vs. PV- neurons, PV+ vs. excitatory neurons, and

PV+ vs. vasoactive intestinal peptide-expressing (VIP+) neurons. In addition to gapped k-mer support vector machines, another example embodiment trained similar convolutional neural networks (CNNs) which are the state-of-the-art model for relating raw genomic sequence to regulatory function. All of these models were used to screen PV+ neuron-specific regions of open chromatin for their ability to independently drive cell type-specific gene expression. The resulting data 300 are shown in graphs 302, 304, 306, and 308 of **FIG. 3** and subsequently described. A top candidate is chosen based on the predicted cell type-specificity, its length, and its proximity to cell type-specific genes. That candidate regulatory sequence was cloned into a plasmid vector in which it drives the expression of Sun1GFP and the vector was transduced into the mouse brain using AAV-PHP.eB.

This document also provides methods for using the constructs described herein to label and isolate nuclei and/or cells of a select type. The methods can include, for example providing a cSNAIL or SNAIL nucleic acid construct as described herein, introducing the construct into a population of different cell types, and culturing or incubating the cells under conditions in which a tagged Sun1 fusion polypeptide encoded by the cSNAIL or SNAIL construct is expressed in the selected cell type (and is not expressed in cells that are not of the selected type), thereby labeling the selected cell type. When a cSNAIL nucleic acid construct is used, the methods also include introducing a nucleic acid construct encoding a Cre recombinase into the population of different cell types, where the sequence encoding the Cre recombinase is operably linked to a promoter that is specific for the selected cell type. In addition, when a cSNAIL construct is used, the conditions under which the cell population is cultured or incubated permit the Cre recombinase to be expressed in the selected cell type, thus leading to excision of the sequence encoding the tagged Sun1 fusion polypeptide from its nucleic acid construct and re-insertion of the tagged Sun1 fusion polypeptide coding sequence in the correct orientation for expression. It is to be noted that in some cases, the population of cells is within a mammal.

Any appropriate method can be used to introduce the SNAIL construct or the cSNAIL and Cre recombinase constructs into a population of cells. As used herein, “transformed” and “transfected” encompass the introduction of one or more nucleic acid molecules (e.g., one or more expression vectors) into a cell by any of a number of techniques. Suitable methods for transforming and transfecting host cells can be found,

for example, in Sambrook et al., Molecular Cloning: A Laboratory Manual (2nd edition), Cold Spring Harbor Laboratory, New York (1989). For example, calcium phosphate precipitation, electroporation, heat shock, lipofection, microinjection, and virus-mediated nucleic acid transfer can be used to introduce nucleic acid molecules into cells. In addition, naked DNA can be delivered directly to cells *in vivo* (see, e.g., U.S. Patent Nos. 5,580,859 and 5,589,466). The isolated nucleic acid molecule transformed into a host cell can be integrated into the genome of the cell or maintained in an episomal state. Thus, host cells can be stably or transiently transfected with a construct containing an isolated nucleic acid molecule provided herein. In some cases, one or more nucleic acid constructs can be incorporated into virus particles (e.g., AAV or lentivirus particles), and the virus particles can be delivered to cells in order to transfer their nucleic acid contents to the cells.

The invention will be further described in the following examples, which do not limit the scope of the invention described in the claims.

EXAMPLES

Example 1

A cSNAIL viral probe was injected into PVcre mice, and Sun1GFP expression was selectively activated in Cre⁺ cells due to viral expression of Sun1GFP in PV⁺ neurons (**FIGS. 1B and 1C**). The success of this probe was quantified against the traditional double-transgenic strategy using different fluorophores in the same mouse. In the primary motor cortex (M1) of PVcre mice, the viral probe had an efficiency of 80.5% and a specificity of 94.1% (N = 1088 cells). In SSTcre mice, the cSNAIL probe labeled SST neurons in M1 with an efficiency of 67.9% and a specificity of 94.7% (N = 3359 cells). The efficiency was expected to be substantially less than 100% because it was limited by viral transduction rates, as AAV PHP.eB reportedly transduces 55-70% of neurons in the cortex (Chan et al., *supra*). The data for these neuron subtypes were in concordance or more optimistic than this characterization. Using cSNAIL technology, PV⁺ and PV⁻ were isolated neurons from the mouse cortex, open chromatin assays (Buenrostro et al., *Curr. Protoc. Mol. Biol.* 109:21.29.1-21.29.9, 2015; and Buenrostro et al., *Nat. Methods* 10:1213-1218, 2013) were performed to identify active gene regulatory elements in these populations (**FIGS. 1E and 1F**). The resulting data were high quality and exhibited characteristic PV⁺ or PV⁻ gene regulatory activity as previously described

using INTACT. This demonstrated that the probes and methods provided herein are capable of generating comparable data as current state-of-the art technology.

A SNAIL viral probe for nuclei isolation of PV+ neurons also was validated. When a successful candidate sequence is cloned upstream of the modified Sun1GFP gene in the virus construct, it can only activate fluorescence expression in cells that can recognize and activate the promoter sequence. Thus, while a SNAIL viral probe will transduce cells indiscriminately, it will only express Sun1GFP in the targeted cell population.

To prioritize PV+ neuron-specific enhancer candidates, predictions from the SVM and CNN models were used. Three pairwise classifiers of each type (SVM and CNN) were built to discriminate sequences of open chromatin specific to PV+ versus PV- cell, excitatory neurons (EXC), and vasoactive intestinal peptide-expressing (VIP) neurons. The training data for the PV+ vs. PV- models were based on ATAC-seq collected using cSNAIL in Pvalb-2A-Cre transgenic mice. The training data for the two other models were based on published ATAC-seq conducted with the double-transgenic INTACT method (Mo et al., *supra*). All ATAC-seq data were processed using the ENCODE ATAC-seq pipeline (github.com/kundajelab/atac_dnase_pipelines), and reproducible peaks were called on 2-3 replicates of each cell type. Peak regions were tested for differential accessibility between two cell types using the negative binomial model in DESeq2 (Love et al., *Genome Biol* 15:1-21). Differential peaks (adjusted $p < 0.01$ and $|\log_2\text{FoldDifference}|$) were summit centered and unified to 500 bp in length. Where more than one peak summit was called within a differential region and if these summits were further than 100 bp apart, both were included in the training data. These regions were filtered to remove promoters, because promoters tend to be less cell type-specific than enhancers and may have different sequence properties. Regions that overlapped with super enhancers were also removed, as they are too long for viral constructs and may exhibit unique sequence features. To be able to test for over-fitting of the models, the data were divided by chromosome into separate training (chromosomes 3, 4, 5, 6, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, and X), validation (chromosomes 8 and 9), and test sets (chromosomes 1 and 2).

The SVM models were built using the center weighted gapped k-mer kernel (Ghandi et al., *PLoS Comp Bio* 10:e1003711, 2014; code available at github.com/Dongwon-Lee/lsgkm), option -t 4. The models with the best performance

were constructed using the with the following parameters, PV+ vs. PV- model: -l 8 -d 2 -c 0.4 -k 6, PV+ vs. EXC model: -l 8 -d 2 -c 0.4 -k 6, PV+ vs. VIP model: -l 7 -d 1 -c 0.3 -k 6. After the training period, the models were evaluated on the held out validation set of enhancers with known cell type-specific activity. All three SVMs achieved a validation set AUROC > 0.9 and AUPRC > 0.83, shown in graphs 302 and 304 (**FIG. 3**). Of the top 100 most confident PV+ neuron-specific predictions in each model, 97-100 were true positives. Model performance on the PV+ and excitatory ATAC-seq peaks (Mo et al., *supra*) also was evaluated. On this independent data, the PV+ vs. PV- classifier achieved an accuracy of 77% and an F1 of 0.76. All three CNN models also achieved strong performance on the training set, as well as a held-out validation chromosome (AUROC > 0.94, AUPRC > 0.74), shown in graphs 306 and 308 (**FIG. 3**).

Using scores across all models and manual curation, four highly consistent PV+ neuron-specific regulatory sequences were identified as being most likely to induce PV+ specific Sun1GFP expression in an independent viral probe. These candidate sequences were incorporated into separate pAAV vectors, operably linked to the expression of a Sun1GFP fusion gene. A non-specific control vector also was designed with the same sequence except for the promoter region, which was replaced with a constitutive Efla promoter. After packaging into virus, these probes were assessed for their ability to drive PV+ specific expression of Sun1GFP in AAV by comparing Sun1GFP to PV neurons identified with antibodies against parvalbumin. Two of four viruses were successful at restricting Sun1GFP expression to PV+ neurons, hereafter referred to as Rank58 and Rank72 (**FIG. 4A**). Using affinity purification and ATAC-seq, it also was determined that the Rank58 and Rank72 viruses had chromatin landscapes that were similar to PV+ signatures collected with cSNAIL (**FIG. 4B**) or single nucleus ATAC-seq (**FIG. 4C**), indicating shared identity.

The materials and methods described herein have the striking benefit of strain independence. Because of this, the materials and methods can be used in wild type mice or any organism for which AAV can be used as a delivery method. Additionally, they can be used in conjunction with transgenic strains, such as disease models, which generally cannot be interbred with INTACT transgenic mice.

Thus, applications of cSNAIL and SNAIL are wide-ranging. For example, cSNAIL viral probes are an immediate alternative to double transgenic mice for cell isolation. These can be used with any existing cell type-specific Cre mouse strain that

researchers may already be using. SNAIL probes provide an extremely powerful tool for cell isolation in any strain of mouse, and potentially other mammals.

Beyond cell labeling and isolation, applications of this viral gene driving technology also include, without limitation, cell type-specific CRISPR Cas9 expression, gene overexpression, and optogenetic stimuli. The SNAIL platform will give users unprecedented control over circuit manipulation at the resolution of individual neuronal subtypes. In one example, the AAV-based cell type-specific label of parvalbumin positive neurons can be used to therapeutically manipulate this specific neuron population in humans. This particular cell type has been causally linked to schizophrenia, Parkinson's disease, and Alzheimer's disease. Additionally, AAV-based therapies have shown promise in human subjects.

The exon-only sequence encoding the N-truncated version of Sun1GFP (SEQ ID NO:3), the Rank72 candidate PV specific enhancer (SEQ ID NO:4), the full plasmid including Rank72 and Sun1GFP (SEQ ID NO:5), and the Sun1GFP amino acid sequence (SEQ ID NO:6) were as follows:

N-truncated exon-only version of Sun1GFP (SNAIL Sun1GFP DNA sequence)

```
ATGACCTCCAGGGTGTACTCCAGAGACAGGACTCTCAAACCACGGAAGGCAG
CCTCGGGAACCTTCTGGTGGCTAGGGAGCGGCTGGTACCAATTTGTTACTTTG
ATTTCTTGGCTGAATGTCTTTCTTCTTACCAGGTGCCTTCGAAATATTTGCAAG
GTTTTGTCTTGCTCCTCCACTCCTACTTTTACTAGGTGCTGGTGTCTCCCTG
TGGGGCCAGGGAACTTCTTCTCACTCCTACCAGTGCTGAACTGGACGGCCAT
GCAGCCAACACAGAGGGTGGACGATTCCAAGGGCATGCATAGACCTGGCCCT
CTTCCCCCGAGCCACCTCCAAAGGTTGATCACAAGGCTTCCAGTGGCCTCA
GGAGAGTGACATGGGGCAGAAGGTAGCTTCTTTGAGTGCGCAGTGCCACAAC
CATGATGAGAGACTTGCAGAGCTGACAGTCTCTGCTTCAGAACTACAGATAC
GGGTAGACCAAGTGGATGACGGCAGGGAAGGGCTGTCACTGTGGGTCAAGA
ATGTGGTTGGACAGCACCTGCAGGAGATGGGCACCATAGAACCACCTGATGC
TAAGACTGACTTCATGACTTTCCACCATGACCATGAAGTGCGTCTCTCCAAC
TGGAAGATGTTCTTAGAAAACTGACAGAAAAATCTGAGGCTATCCAGAAGGA
GCTGGAAGAAACCAAGCTGAAAGCAGGCAGCAGGGATGAAGAGCAGCCCT
CCTTGACCGTGTGCAGCACCTAGAAGTGAAGTGAACCTGTTGAAGTCACAG
CTGTCAGACTGGCAGCATCTGAAGACCAGCTGTGAGCAGGCTGGGGCCCGCA
TCCAGGAGACTGTGCAGCTCATGTTCTCTGAGGATCAGCAGGGCGGTTCCCTC
GAGTGGCTATTAGAGAAGCTTTCTTCTCGGTTCTGTGAGCAAGGATGAGCTGC
AGGTGCTCTTACATGACCTTGAGCTGAACTGCTGCAGAATATCACACACCA
CATCACCGTGACAGGACAGGCCCCGACATCCGAGGCTATTGTGTCTGCCGTG
```

AATCAGGCAGGGATTTTCAGGAATCACAGAAGCGCAAGCACATATCATTGTGA
 ACAATGCTCTGAAGCTGTACTCCCAAGACAAGACGGGGATGGTGGACTTTGC
 TCTGGAGTCTGGAGGTGGCAGCATCCTAAGCACTCGGTGCTCTGAGACCTAT
 GAGACCAAGACGGCACTGCTGAGCCTGTTTGGGGTCCCACTGTGGTACTTCTC
 ACAGTCACCTCGAGTGGTGATCCAGCCCGACATCTACCCAGGGAATTGCTGG
 GCGTTCAAAGGTTCCCAGGGGTACCTGGTGGTGCGGTTGTCCATGAAGATCT
 ACCCAACCACATTCACCATGGAACACATTCCAAAGACACTATCACCCACTGG
 TAACATCTCCAGTGCCCCCAAAGACTTTGCAGTCTATGGACTGGAAACGGAG
 TATCAAGAAGAGGGGGCAGCCTCTGGGACGGTTCACCTATGACCAGGAAGGAG
 ACTCACTCCAGATGTTCCACACACTGGAAAGACCTGACCAAGCCTTCCAGAT
 AGTAGAGCTCCGGGTCTGTCCAACCTGGGGCCACCCTGAGTACACTTGCCTCT
 ACCGGTTCCGAGTCCACGGAGAGCCCATCCAGATGAGCAAAGGAGAAGAAC
 TTTTCACTGGAGTTGTCCCAATTCTTGTGTAATTAGATGGTGATGTTAATGGG
 CACAAATTTTCTGTCCGTGGAGAGGGTGAAGGTGATGCTACAAACGGAAAAC
 TCACCCTTAAATTTATTTGCACTACTGGAAAACCTACCTGTTCCATGGCCAAAC
 CTTGTCACTACTCTGACCTATGGTGTTCAATGCTTTTCCCGTTATCCGGATCAC
 ATGAAACGGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGG
 AACGCACTATATCTTTCAAAGATGACGGGACCTACAAGACGCGTGCTGAAGT
 CAAGTTTGAAGGTGATACCCTTGTTAATCGTATCGAGTTAAAAGGTATTGATT
 TTAAAGAAGATGGAAACATTCTCGGACACAAACTCGAGTACAACCTTTAACTC
 ACACAATGTATACATCACGGCAGACAAACAAAAGAATGGAATCAAAGCTAA
 CTTCAAAATTCGCCACAACGTTGAAGATGGTTCGGTTCAACTAGCAGACCATT
 ATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCAT
 TACCTGTGACACAATCTGTCCTTTCGAAAGATCCCAACGAAAAGCGTGACC
 ACATGGTCTTCTTGAGTTTGTAAGTGTGCTGCTGGGATTACACATGGCATGGAT
 GAGCTCTACAAATAA (SEQ ID NO:3)

Rank72 candidate PV specific enhancer (SNAIL_Rank72_PromoterSequence)

attggtttgcttacacgctttgtgaactgcattgccatgactccatagagaaaaacacgcgaaaagcttctctaggtggtccc
 acatcttctgttgattccgtggactcgggtcaactagcagagtgatgtcagctgctatacatagactcacatggagtttgcctaaga
 caggctcactgctgaggcaagacacgttggaaggcaagacctgtagaggacacttggtgttttgaggeagtataaataggact
 caacggacgggtgatggaggctgggctgtgtgcataactagcttgcagtgcttcttggtctcgcgtcttactgaccttcgatttgt
 gagagcggcacagttgagaactccctggcatccctgggtggtcctaattctctctgtaactccagccaattcggctgaagcctg
 gcagttctgctaggtcgggccaccgctgctgatttctatttctatcccaactctaccaatctg (SEQ ID NO:4)

Full plasmid including Rank72 and Sun1GFP (SNAIL_Rank72.dna)

CCTGCAGGCAGCTGCGCGCTCGCTCGCTCACTGAGGCCGCCCGGGCAAAGCC
 CGGGCGTCCGGGCGACCTTTGGTCGCCCGGCCTCAGTGAGCGAGCGAGCGCGC
 AGAGAGGGAGTGGCCAACTCCATCACTAGGGGTTCTGCGGCCGACGCGTG
 CGATCGCCATATGAGGCCTCATTGGCCATTGTAATTGGTTTGCCTTACACGCT
 TTGTTGAACTGCATTTGCCATGACTCCATAGAGAAAAACACGCCGAAAAGCT

TCTCTAGGTGGTTCCACATCTTCTTGTGGATTCCGTGGACTCGGGTCAACTA
GCAGAGTGATGTCAGCTGCTATACATAGACTCACATGGAGTTTTGCTAAGAC
AGGCTCACGTGCTGAGGCAAGACACGTGGTAAGGCAAGACCTGTAGAGGAC
ACTTGGTGTTTTGAGGCAGTATAAATAGGACTCAACGGACGGTGATGGAGGC
TGGGCTTGCTTGCATAACTAGCTTTGCAGTGCTTCTTGGTCTCGCGTCTTCACT
GACCTTCGATTTGTTGAGAGCGGCACAGTTGAGAACTTCCCCTGGCATCCCTG
GTGGTCTTAATTCTCCTGCTAACTCCAGCCAATTTCGGCTGAAGCCTGGCAGT
TTCTGCTAGGTGCGGCCACCGCTGCTGATTTCGTATTTGCTATCCCAACTCTAC
CAATCTGGAATTAATTGCTGTCTGCGAGGGCCAGCTGTTGGGGTGAGTACTC
CCTCTCAAAGCGGGCATGACTTCTGCGCTAAGATTGTCAGTTTCCAAAACG
AGGAGGATTTGATATTCACCTGGCCCCGCGGTGATGCCTTTGAGGGTGGCCGC
GTCCATCTGGTCAGAAAAGACAATCTTTTTTGTGTCAAGCTTGAGGTGTGGCA
GGCTTGAGATCTGGCCATACACTTGAGTGACAATGACATCCACTTTGCCTTTC
TCTCCACAGGTGTCCACTCCCAGGTCCAAGTGCAGGTCGCGACTAGCGGATCC
AATGACCTCCAGGGTGTACTCCAGAGACAGGACTCTCAAACCACGGAAGGCA
GCCTCGGGAACCTTCTGGTGGCTAGGGAGCGGCTGGTACCAATTTGTTACTTT
GATTTCTTGGCTGAATGTCTTTCTTCTTACCAGGTGCCCTTCGAAATATTTGCAA
GGTTTTTGTCTTGCTCCTCCCACTCCTACTTTTACTAGGTGCTGGTGTCTCCCT
GTGGGGCCAGGGAACTTCTTCTCACTCCTACCAGTGCTGAACTGGACGGCC
ATGCAGCCAACACAGAGGGTGGACGATTCCAAGGGCATGCATAGACCTGGCC
CTCTTCCCCCGAGCCCACCTCCAAAGGTTGATCACAAGGCTTCCCAGTGGCCT
CAGGAGAGTGACATGGGGCAGAAGGTAGCTTCTTTGAGTGCGCAGTGCCACA
ACCATGATGAGAGACTTGCAGAGCTGACAGTCCTGCTTCAGAACTACAGAT
ACGGGTAGACCAAGTGGATGACGGCAGGGAAGGGCTGTCACTGTGGGTCAA
GAATGTGGTTGGACAGCACCTGCAGGAGATGGGCACCATAGAACCACTGAT
GCTAAGACTGACTTCATGACTTTCCACCATGACCATGAAGTGCGTCTCTCCAA
CTTGGAAGATGTTCTTAGAAAATGACAGAAAAATCTGAGGCTATCCAGAAG
GAGCTGGAAGAAACCAAGCTGAAAGCAGGCAGCAGGGATGAAGAGCAGCCC
CTCCTTGACCGTGTGCAGCACCTAGAACTGGAAGTGAACCTGTTGAAGTCAC
AGCTGTCAGACTGGCAGCATCTGAAGACCAGCTGTGAGCAGGCTGGGGCCCG
CATCCAGGAGACTGTGCAGCTCATGTTCTCTGAGGATCAGCAGGGCGGTTCC
CTCGAGTGGCTATTAGAGAAGCTTTCTTCTCGGTTTCGTGAGCAAGGATGAGCT
GCAGGTGCTCTTACATGACCTTGAGCTGAAACTGCTGCAGAATATCACACAC
CACATCACCGTGACAGGACAGGCCCCGACATCCGAGGCTATTGTGTCTGCCG
TGAATCAGGCAGGGATTTCAAGGAATCACAGAAGCGCAAGCACATATCATTTGT
GAACAATGCTCTGAAGCTGTACTCCCAAGACAAGACGGGGATGGTGGACTTT
GCTCTGGAGTCTGGAGGTGGCAGCATCCTAAGCACTCGGTGCTCTGAGACCT
ATGAGACCAAGACGGCACTGCTGAGCCTGTTTGGGGTCCCCTGTGGTACTTC
TCACAGTCACCTCGAGTGGTGATCCAGCCCGACATCTACCCAGGGAATTGCT
GGGCGTTCAAAGGTTCCCAGGGGTACCTGGTGGTGCAGGTTGTCCATGAAGAT
CTACCCAACCACATTACCATGGAACACATTCCAAAGACACTATCACCCACT
GGTAACATCTCCAGTGCCCCCAAAGACTTTGTCAGTCTATGGACTGGAAACGG

AGTATCAAGAAGAGGGGCAGCCTCTGGGACGGTTCACCTATGACCAGGAAGG
AGACTCACTCCAGATGTTCCACACACTGGAAAGACCTGACCAAGCCTTCCAG
ATAGTAGAGCTCCGGGTCCTGTCCAACCTGGGGCCACCCTGAGTACACTTGCCT
CTACCGGTTCCGAGTCCACGGAGAGCCCATCCAGATGAGCAAAGGAGAAGA
ACTTTTCACTGGAGTTGTCCCAATTCTTGTGTAATTAGATGGTGATGTTAATG
GGCACAAATTTTCTGTCCGTGGAGAGGGTGAAGGTGATGCTACAAACGGAAA
ACTCACCCCTTAAATTTATTTGCACTACTGGAAAACCTACCTGTTCCATGGCCAA
CACTTGTCACTACTCTGACCTATGGTGTTCATGCTTTTCCCGTTATCCGGATC
ACATGAAACGGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACA
GGAACGCACATATCTTTCAAAGATGACGGGACCTACAAGACGCGTGCTGAA
GTCAAGTTTGAAGGTGATACCCTTGTTAATCGTATCGAGTTAAAAGGTATTGA
TTTTAAAGAAGATGGAAACATTCTCGGACACAAACTCGAGTACAACTTTAAC
TCACACAATGTATACATCACGGCAGACAAACAAAAGAATGGAATCAAAGCTA
ACTTCAAAATTCGCCACAACGTTGAAGATGGTTCCGTTCAACTAGCAGACCAT
TATCAACAAAATACTCCAATTGGCGATGGCCCTGTCTTTTACCAGACAACCA
TTACCTGTGACACAATCTGTCTTTTCGAAAGATCCCAACGAAAAGCGTGACC
ACATGGTCTTCTTGAGTTTGTAACTGCTGCTGGGATTACACATGGCATGGAT
GAGCTCTACAAATAAGGCGCGGAATTCGATATCAAGCTTATCGATAATCAAC
CTCTGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTATGTTGCT
CCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTGCT
TCCCGTATGGCTTTTCATTTTCTCCTCCTTGATAAAATCCTGGTTGCTGTCTCTTT
ATGAGGAGTTGTGGCCCGTTGTCAAGCAACGTGGCGTGGTGTGCACTGTGTTT
GCTGACGCAACCCCCACTGGTTGGGGCATTGCCACCACCTGTCAGCTCCTTTC
CGGGACTTTTCGCTTTCCCCCTCCCTATTGCCACGGCGGAACCTCATCGCCGCCT
GCCTTGCCCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTG
GTGTTGTGCGGGAAATCATCGTCCTTTTCCTTGGCTGCTCGCCTGTGTTGCCAC
CTGGATTCTGCGCGGGACGTCTTCTGCTACGTCCCTTCGGCCCTCAATCCAG
CGGACCTTCTTCCCGCGGCCTGCTGCCGGCTCTGCGGCCTCTTCCGCGTCTTC
GCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCGCCTCCCCGCATCGA
TACCGAGCGCTGCTCGAGAGATCTACGGGTGGCATCCCTGTGACCCCTCCCCA
GTGCCTCTCCTGGCCCTGGAAGTTGCCACTCCAGTGCCCACCAGCCTTGTCCT
AATAAAATTAAGTTGCATCATTTTGTCTGACTAGGTGTCTTCTATAATATTAT
GGGGTGGAGGGGGGTGGTATGGAGCAAGGGGCAAGTTGGGAAGACAACCTG
TAGGGCCTGCGGGGTCTATTGGGAACCAAGCTGGAGTGCAGTGGCACAATCT
TGGCTCACTGCAATCTCCGCCTCCTGGGTTCAGCGATTCTCCTGCCTCAGCC
TCCCGAGTTGTTGGGATTCCAGGCATGCATGACCAGGCTCAGCTAATTTTTGT
TTTTTTGGTAGAGACGGGGTTTCACCATATTGGCCAGGCTGGTCTCCAACCTCC
TAATCTCAGGTGATCTACCCACCTTGGCCTCCCAAATTGCTGGGATTACAGGC
GTGAACCACTGCTCCCTTCCCTGTCCTTCTGATTTTGTAGGTAACCACGTGCG
GACCGAGCGGCCGAGGAACCCCTAGTGATGGAGTTGGCCACTCCCTCTCTG
CGCGCTCGCTCGCTCACTGAGGCCGGGGCGACCAAAGGTGCCCCGACGCCCGG
GCTTTGCCCGGGCGGCCTCAGTGAGCGAGCGAGCGCGCAGCTGCCTGCAGGG

GCGCCTGATGCGGTATTTCTCCTTACGCATCTGTGCGGTATTTACACCCGCA
TACGTCAAAGCAACCATAAGTACGCGCCCTGTAGCGGCGCATTAAAGCGCGGGC
GGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGC
CCGCTCCTTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCCGCCGGCTTTCCCCG
TCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGC
ACCTCGACCCCAAAAACTTGATTTGGGTGATGGTTCACGTAGTGGGCCATC
GCCCTGATAGACGGTTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATA
GTGGACTCTTGTTCCAACTGGAACAACACTCAACCCTATCTCGGGCTATTCT
TTTGATTTATAAGGGATTTTGCCGATTTTCGGCCTATTGGTTAAAAAATGAGCT
GATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTT
TATGGTGCACCTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGCCC
CGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGCTCTGCTCCCGG
CATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTGAGAG
GTTTTACCCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGC
CTATTTTTATAGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGG
CACTTTTCGGGGAAATGTGCGCGGAACCCCTATTGTTTTATTTTTCTAAATAC
ATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAA
TATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCC
CTTTTTTTCGGGCATTTTGCCTTCCTGTTTTTGTCTACCCAGAAACGCTGGTGAA
AGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTACATCGAACTG
GATCTCAACAGCGGTAAGATCCTTGAGAGTTTTTCGCCCCGAAGAAGCTTTTCC
AATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTG
ACGCCGGGCAAGAGCAACTCGGTGCGCCGATACACTATTCTCAGAATGACTT
GGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTA
AGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACT
TACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAA
CATGGGGGATCATGTAACCTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAA
GCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAA
CGTTGCGCAAACTATTAACCTGGCGAACTACTTACTCTAGCTTCCCGGCAACAA
TTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGG
CCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGG
TCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGT
AGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAG
ATCGCTGAGATAGGTGCCTCACTGATTAAGCATTTGGTAACTGTCAGACCAAG
TTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTTAATTTAAAAGGA
TCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAG
TTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTG
AGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGC
TACCAGCGGTGGTTTGTGTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAG
GTAACCTGGCTTCAGCAGAGCGCAGATACCAAAATACTGTCCCTTCTAGTGTAGCC
GTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTC
TGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACC

GGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCTGGGCTGAA
 CGGGGGGTTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACT
 GAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGA
 AAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCTGGAACAGGAGAGCGCACG
 AGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCTGGGTTTCG
 CCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCC
 TATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGG
 CCTTTTGCTCACATGT (SEQ ID NO:5)

Sun1GFP polypeptide

MTSRVYSRDRTLKPRKAASGTFWWLGSGWYQFVTLISWLNVLTRCLRNICKV
 FVLLLPLLLLLGAGVSLWGQGNFFSLPLVLTAMQPTQVRVDDSKGMHRPGPLP
 PSPPPKVDHKASQWPQESDMGQKVASLSAQCHNHDERLAELTVLLQKLQIRVDQ
 VDDGREGLSLWVKNVVGQHLQEMGTIEPPDAKTDFTFHHDEHVRLSNLEDVL
 RKLTEKSEAIQKELEETKLKAGSRDEEQPLLDRVQHLELELNLLKSQSLDWQHLK
 TSCEQAGARIQETVQLMFSEDQGGGSLEWLLEKLSSRFVSKDELQVLLHDLELKL
 LQNITHHITVTGQAPTSEAIVSAVNQAGISGITEAQAHIIIVNNALKLYSQDKTGMV
 DFALESGGGSILSTRCSETYETKTALLSLFGVPLWYFSQSPRVVIQPDYPGNCWA
 FKGSQGYLVVRLSMKIYPTTFTMEHIPKTLSPGNISSAPKDFAVYGLEYQEYG
 QPLGRFTYDQEGDSLQMFHTLERPDQAFQIVELRVLSNWGHPEYTCLYRFRVHG
 EPIQMSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTCLKFICTTGKLP
 VPWPTLVTTLTYGVCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAE
 VKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFSHNVYITADKQKNGIKANFKIRH
 NVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMLLEFVTA
 AGITHGMDELYK (SEQ ID NO:6) [residues 2-551 are amino acids 208-757 of the *Mus musculus* Sun1 transcript variant 5 (GENBANK® Accession No. NM_001256118.1), and residues 552-789 (in italics) are the superfolder GFP amino acid sequence]

Example 2 – Cell type-specific oxidative stress genomic signatures in the globus pallidus of dopamine depleted mice

Neuron subtype dysfunction is a key contributor to the motor deficits observed in dopamine depleted mouse models of Parkinson's disease. For example, cell type-specific optogenetic stimulation of external globus pallidus (GPe) parvalbumin-expressing neurons (PV+) – but not global GPe stimulation – rescues normal motor behavior in dopamine depleted mice. Despite the importance of these cells in the disease process, the molecular correlates of GPe PV+ neuron performance after DD remain unknown due to the difficulty of isolating specific neuron subtypes in the brain. To address this issue, the cSNAIL viral affinity purification strategy was developed to isolate Cre recombinase-expressing (Cre+) nuclei from the adult mouse brain. Specifically, cSNAIL was used to

isolate PV+ neurons from the cortex, striatum, and GPe of 6-hydroxydopamine (6-OHDA) lesioned mice and sham controls. In parallel, RNA-seq and ATAC-seq (Buenrostro et al. 2013, *supra*; and Buenrostro et al. 2015, *supra*) were conducted on cSNAIL-isolated PV+ populations and the remaining PV- cells in each brain region. With this strategy, GPe PV+ neuron-specific epigenetic and transcriptional responses to DD were identified that are related to cellular oxygen sensing and oxidative stress, providing new insight into the molecular pathology of PD.

Materials and Methods

AAV design. Several modifications were made to the INTACT system to allow it to be delivered using AAV. First, an N-truncated version of the human SUN1 gene (amino acids 208-913) was used. This version retained the ability to localize to the nuclear envelope in a manner that was resistant to detergent tissue homogenization as characterized by Haque et al. (*J Biol Chem* 285:3487-3498, 2010). Second, only the DNA sequence of exonic regions of SUN1 was incorporated into the construct. Third, the SUN1GFP fusion gene included one copy of superfolder GFP instead of two. Finally, the Cre-dependent mechanism was adjusted from a loxp-flanked stop codon to the double inverted orientation (DIO) system. In the initial cSNAIL genome, the SUN1GFP gene is in the reverse orientation with respect to the promoter such that the gene cannot be expressed. When the virus encounters Cre recombinase protein (Cre), it acts on the double flanking lox sites to excise the SUN1GFP gene and re-insert it in the correct orientation for transcription (FIG. 1A).

Vector construction. The modified SUN1GFP fusion sequence fragment was synthesized by Integrated DNA Technologies (IDT; Coralville, IA) and contained flanking restriction sites for *SgsI* and *BspolI*. The SUN1GFP sequence was inserted into an AAV DIO backbone vector with an Efla promoter using standard restriction cloning. pAAV-Efla-DIO-EGFP-WPRE-pA was obtained from Bernardo Sabatini (n2t.net/addgene:37084; RRID: Addgene_37084) (Saunders et al., *Front Neural Circuits* 6:47, 2012). The ligation was transformed into electrocompetent bacteria, and positive clones were selected and sequences were confirmed with Sanger sequencing. The cSNAIL genome vector (pAAV-Efla-DIO-SUN1GFP-WPRE-pA) is shown in FIG. 5A.

AAV production. The AAV was produced by triple co-transfection of pAAV-Efla-DIO-SUN1GFP-WPRE-pA, an AAV helper plasmid, and pUCmini-iCAP-PHP.eB

into AAVpro(R) 293T cells (Takara Bio USA, Inc., Mountain View, CA; cat #632273). pUCmini-iCAP-PHP.eB was obtained from Viviana Gradinaru (<http://n2t.net/addgene:103005>; RRID: Addgene 103005) (Chan et al., *supra*). After cell expansion, the AAV particles were precipitated with polyethylene glycol (PEG) and purified with ultracentrifugation on an iodixanol gradient. The virus was titrated using the AAVpro(R) Titration Kit (Takara, cat #6233), diluted in PBS to a concentration of 8.0×10^9 vg/ μ l, and stored at -80° until injection.

Animals. All molecular experiments were performed on Pvalb-2A-Cre mice (B6.Cg-Pvalb^{tm1.1(cre)}Aibs/J; The Jackson Laboratory, Bar Harbor, ME, stock no: 012358) (Madisen et al., *Nat Neurosci* 13:133-140, 2010). Imaging experiments for the validation of cSNAIL (**FIGS. 6A-6C**) were performed with double-transgenic Pvalb-2A-Cre and Ai14 mice (Ai14 strain; B6.Cg-Gt(ROSA)26Sor^{tm14(CAG-tdTomato)Hze}/J; Jackson Stock No: 007914) (Madisen et al., *supra*). Experiments related to Hif2a staining (**FIG. 9D**) were conducted on wild type mice (C57BL/6J; Jackson Stock No: 000664).

AAV delivery. Mice were anesthetized with isoflurane for 2-5 minutes until breathing slowed and the animal had no pedal reflex. AAV (4×10^{10} vg) was injected into the retro-orbital cavity. The mice received 0.5% Proparacaine Hydrochloride Ophthalmic Solution for comfort and were monitored for physical and behavioral abnormalities post-procedure, but no such abnormalities were observed. The virus incubated for 21 to 28 days to reach peak expression before downstream experiments.

Dopamine depletion surgery. The molecular experiments were designed to reproduce previous conditions and maximize cell yield. Therefore, bilateral acute 6-OHDA or saline injections were administered to the medial forebrain bundle as described elsewhere (Mastro et al., *Nat Neurosci* 20:815-823, 2017).

Immunofluorescence staining and imaging. Wherever possible, tissue for imaging was preserved by 4% paraformaldehyde (PFA) perfusion. The brains were dissected and incubated overnight at 4°C in 4% PFA. Shortly after, they were sectioned coronally with a vibratome at $100 \mu\text{m}$ for antibody staining and slide mounting. To visualize cSNAIL specificity (**FIG. 6B-6D**) tissues were stained with primary anti-NeuN (Cell Signaling Technology, Danvers, MA; #12943) and secondary Alexafluor 405 (Invitrogen, Waltham, MA; #A-31556) to label neurons. For images related to **FIG. 9D**, tissues were stained with primary anti-Pvalb (Swant, PV 27) paired with secondary Alexafluor 405 (Invitrogen, #A-31556) or Alexafluor 488 (Cell Signaling Technology, #4408S) and

primary anti-Hif2a (Novus Biologicals, Centennial, CO; #NB100-132) paired with secondary Alexafluor 594 (Cell Signaling Technology, #8890S). For tissues involved in both genomic assays and imaging, one fresh coronal slice including caudate putamen, but anterior to the GPe, was fixed in 4% PFA for 24 hours. Tissues were stained with primary anti-tyrosine hydroxylase (Pelfreez, Rogers, AR; #P40101-150) and secondary Alexafluor 647 (Invitrogen, #A-31573) to assess dopamine levels in the striatum. All imaging sections were mounted onto slides using ProLong Diamond Antifade mounting media (Life Technologies Corp., Rockville, MD; #P36961) and imaged with confocal microscopy.

Image analysis. Single-channel and co-labeled cells were manually counted for **FIGS. 6A-6C** and **FIG. 9D**. To quantify the levels of tyrosine hydroxylase as a proxy for dopamine depletion, mean pixel intensity was measured with Fiji (Schindelin et al., *Nat Methods* 9:676-682, 2019). Measurements were performed on one 500 x 500µm square from the striatum of each hemisphere.

Nuclei collection. Fresh mouse brain tissue was sliced on a vibratome in cold, oxygenated artificial cerebrospinal fluid. Brain regions of interest were then dissected under a light microscope and transferred into cold ATAC-seq lysis buffer (Buenrostro et al. 2015, *supra*). The nuclei were extracted from each tissue using 30 strokes of loose pestle dounce homogenization, followed by 70 µm filtration and one centrifugation wash for 10 minutes at 2000 x g.

Affinity Purification of SUN1GFP+ nuclei. The nuclei suspension was pre-cleared for 10-15 minutes with Protein G Dynabeads (Thermo Fisher Scientific, Waltham, MA; cat. 10004D) to remove nuclei or debris that had native affinity for the beads. Free nuclei were incubated with anti-GFP antibody (Invitrogen, #G10362) for 30 minutes before fresh beads were added to the reaction and incubated for another 20 minutes. All incubations took place at 4°C with 40 rpm end-to-end rotation. After this process, the bead-bound nuclei (SUN1GFP+ fraction) were purified and washed on a magnet, while the unbound nuclei were spun down and resuspended in water to preserve the SUN1GFP- fraction.

RNA-seq library construction. After setting aside 50,000 nuclei for ATAC-seq, total RNA was extracted from all remaining nuclei in each sample using the QIAGEN RNEASY® Micro Kit. (QIAGEN, Germantown, MD). Total RNA (10 pg to 10 ng per sample) was processed using the OVATION® SoLo RNA-seq System with mouse

AnyDeplete® (Tecan, Redwood City, CA; cat. 0501-32) to remove ribosomal RNA. Because many nuclear transcripts are immature, no polyA selection was performed. Libraries were paired-end sequenced at 3-9 million reads per sample on the NEXTSEQ™ 500 System (Illumina, Inc., San Diego, CA).

RNA-seq processing. The sequencing output files were mapped to the mm10 mouse genome assembly using Hisat2 and sorted. Duplicate UMIs were removed using NuDup (online at github.com/tecangenomics/nudup). The quality of the filtered data was assessed using Picard CollectRnaSeqMetrics (online at broadinstitute.github.io/picard). A counts per gene per sample matrix was constructed using Rsubread featureCounts (Liao and Smyth, *Nucleic Acids Res* 47:e47, 2019) for input into DESeq2 (Love et al., *supra*).

ATAC-seq library construction. A small subset of nuclei was stained with DAPI and fluorescent nuclei were manually counted using a hemocytometer. About 50,000 nuclei were transferred to a new tube in 22.5µl water and mixed with 2.5µl Tagment DNA Enzyme I and 25µl Tagment DNA Buffer (Illumina, cat. 20034198). The tn5 transposition reaction was carried out at 37°C, 300 rpm for 30 minutes. After tagmentation, the samples were taken through column purification, amplification, and double-sided size selection (100-1000 bp) with AmpureXP beads (Beckman Coulter, Indianapolis, IN; cat. A63881). To reduce amplification bias, a side qPCR reaction was evaluated for each sample to determine the optimal number of PCR cycles for each sample to reach 1/3 maximum intensity. The samples were pooled and paired-end sequenced on a Novaseq 6000 (Illumina) to about 25 million unique, non-mitochondrial reads per sample.

ATAC-seq processing. The fastq files from each sample were processed with the ENCODE ATAC-seq pipeline (online at github.com/kundajelab/atac_dnase_pipelines). Peaks were called on biological replicates of the same cell type, brain region, and treatment with an Irreproducible Discovery Rate (IDR) threshold 0.1 (github.com/kundajelab/idr; Li et al., *Annals Appl Stat* 5(3):1752-1779, 2011). The merged union of IDR peaks of the relevant cell populations were evaluated for differential accessibility using DESeq2 (Love et al., *supra*), where the reads per peak for each sample were quantified using Rsubread featureCounts (Liao and Smyth, *supra*) on the filtered bam alignments. For analysis of cellular identity (FIGS. 7A-7C and 8A-8F), the regions tested were the union of IDR peaks from all sham animal samples. For FIG. 8E, instead of ATAC-seq peaks, counts were computed over 10kb windows around the

transcription start sites for all genes assessed in the RNA-seq analysis. For assessment of DD-affected regions (**FIG. 10**), six separate count tables were constructed for the six cell populations and represented the union of IDR peaks from sham and DD animals for the given cell type and brain region.

snATAC-seq and snRNA-seq processing. To compare cSNAIL-isolated cell types with single nucleus sequencing data, two publicly available datasets from adult mouse cortex were reprocessed. 10x snATAC-seq data was downloaded and processed following the Signac vignette for the purposes of **FIG. 7B** (online at github.com/timoast/signac). For the snRNA-seq comparison, raw count matrices derived from 10x chromium v3 experiments conducted in primary motor cortex were downloaded from NEMO (portal.nemoarchive.org). The sample meta data is provided online at drive.google.com/open?id=1uDeGI-IPZoiX4aYGNZ1WFcGWjaNqTHCd. Empty droplets were removed using the DropletUtils package (Griffiths et al., *Nat Commun* 9:2667, 2018) *defaultDrops* method. Doublets were predicted and removed using the SCDS package (Bais et al., *Bioinformatics* 36:1150-1158, 2019) *cxds_beds_hybrid* method with score cutoff of 1.0. The UMI count values were normalized using pooled size factor normalization (Lun et al., *Genome Biol* 17:75, 2016) implemented in the Scanpy package (Lun et al., *Fl1000 Res* 5:2122, 2016). Size factor normalization requires a preclustering step which was computed using the Scanpy (Wolf et al., *Genome Biol* 19:15, 2018) package's Leiden community detection algorithm (Traag et al., *Sci Rep* 9:5233, 2018). To create the UMAP (McInnes et al., *arXiv* 1802.03426, 2018) for visualization the Scanpy package was also used. Principle components were computed on the top 7500 most variable genes. The UMAP was then computed based on the PCA reduced single nucleus gene expression matrix.

Experimental Design and Statistical Analysis. The statistical methods used for the quantification of cSNAIL specificity and efficiency by imaging co-fluorescence are described in the description of **FIG. 6A-6D**. For molecular assessment of cSNAIL isolated cell fractions (**FIGS. 7A-7C and 8A-8F**), data were collected from two sham animals, one male and one female. Differentially expressed genes and differentially accessible ATAC-seq regions between SUN1GFP+ and SUN1GFP- samples in each brain region were assessed using the negative binomial model in DESeq2: $\text{read count} \sim \text{sex} + \text{SUN1GFPstatus}$. To restrict DESeq2 dispersion estimates to the relevant samples, each model was built separately and only included the samples from one brain region.

For comparing cSNAIL ATAC-seq markers to snATAC-seq cluster markers (FIG. 7B), the hypergeometric enrichment test was used, sampling cell type-specific cSNAIL peaks in each brain region from an expected frequency of cluster-specific snATAC-seq peaks. Only the set of peaks present in both data (112,082) were evaluated. For simplicity, cluster-specific peaks from all excitatory neuron clusters were combined into one category. For visualization, the hypergeometric test enrichments were converted into a z-score for each cSNAIL cell population and plotted using gplots heatmap.2 in R (online at github.com/talgalili/gplots). Annotation enrichments were assessed in differential gene sets using g:Profiler (FIG. 8C) (Peterson et al., *supra*).

Differential motif enrichment analyses was performed with AME (McLeay and Bailey, *supra*), with peak sets of interest summit-centered and unified to 500 bp. Where more than one summit was called within a peak and these summits were more than 100 bp apart, both entries were included in the analysis. Enrichments were assessed relative to a background of all ATAC-seq peaks in that cell type or population.

Unless otherwise noted, molecular comparisons between DD and sham animals were conducted with data from two sham animals and four DD animals. Both treatment groups consisted of equal numbers of males and females. The three female animals were littermates at 12 weeks of age at endpoint and the three male animals were littermates 9 weeks of age, all from the same colony. Animals were euthanized consistently in the morning, 2-3 hours after the light cycle began. ATAC-seq and RNA-seq were performed on partitions of the same source pools of nuclei.

Sample similarities for RNA-seq and ATAC-seq were assessed by principal component analysis (PCA) and hierarchical clustering in DESeq2 (FIGS. 11A-11B, and 12A-12D). Because transcript levels are quantitative, the RNA-seq counts were modeled continuously according to the anti-TH imaging intensities of each animal, which reflect the extent of DD: DESeq2 read count = $\sim$ sex + THintensity. To follow up on RNA-seq results related to Hif2a, image quantification was performed as noted in the description of FIGS. 10A-10C and related results. For DD-affected open chromatin analyses (Fig. 10A-10C), a binary model of treatment state (DD vs. saline) was used, controlling for sex differences: DESeq2 read count = $\sim$ sex + treatment. Again, models for each cell population were built separately and only included samples for one comparison. DD-affected ATAC-seq peak sets were subjected to pathway enrichment analysis using

GREAT version 3.0.0 (McLean et al., *Nat Biotechnol* 28:495-501, 2010) and motif enrichment using AME as described above.

Results

cSNAIL isolates Cre+ nuclei with high precision. cSNAIL technology is a viral strategy for labeling and isolating the nuclei of Cre-expressing cells. The cSNAIL viral genome expresses a nuclear envelope anchored GFP fusion protein in a Cre-dependent manner. To achieve nuclei isolation compatible epigenomic profiling, a similar strategy to the SUN1GFP INTACT transgenic mouse strain (Mo et al., *supra*) was used. SUN1 is a highly conserved nuclear envelope protein that helps connect the nucleoskeleton and the cytoskeleton (Haque et al., *supra*). Because of its tight association with the inner nuclear envelope, the SUN1GFP fusion protein localizes the GFP protein on the nuclear surface. This positioning allows for affinity purification of nuclei, whereby after tissue is homogenized into a single-nuclei suspension, magnetic beads coated with anti-GFP antibody specifically bind SUN1GFP+ nuclei, thus separating them from SUN1GFP- nuclei. Several modifications to the INTACT mouse transgene design (described above) to package the cSNAIL SUN1GFP vector into AAV variant PHP.eB, which is capable of crossing the blood brain barrier to broadly transduce many cells in the central nervous system (Chan et al., *supra*).

The ability of cSNAIL to drive Cre-specific nuclear anchored independent labeling in Pvalb-2A-Cre mouse lines was evaluated using imaging and molecular data (FIGS. 6A-6C and 7A-7C). Fluorescent imaging confirmed that cSNAIL promoted SUN1GFP expression that was properly localized to the nuclear envelope in tissue (FIG. 6A). The expression was sufficient to bind anti-GFP coated magnetic beads after tissue dissociation (FIG. 6B). To quantify the ability of cSNAIL to correctly target Cre+ cells, the expression of cSNAIL was compared to the expression of Cre reporter tdTomato of the Ai14 mouse strain (FIG. 6C). For the purposes of this analysis, virus *specificity* was defined as the percent of SUN1GFP+ cells that also expressed tdTomato, and virus *efficiency* was defined as the percent of tdTomato+ cells that also expressed SUN1GFP. cSNAIL exhibited very strong (>94%) specificity in the cortex, striatum, and GPe. The efficiency ranged from 53% to 81%, reflecting the expected transduction of the viral capsid in these brain regions (Chan et al., *supra*).

After affinity purification, bead-bound nuclei and free nuclei were counted using a manual hemocytometer on a fluorescent microscope to determine the purity and yield of cSNAIL nuclei isolation. Of the nuclei bound to beads, 97.18% had observable SUN1GFP expression. Based on the numbers of SUN1GFP+ nuclei in the bound and unbound fractions, it was estimated that the bound fraction captured up to 65% of the total SUN1GFP+ nuclei. These results demonstrated that nuclei isolation with cSNAIL was as good as SUN1GFP nuclei affinity purification using transgenic expression.

To confirm the molecular identity of the cSNAIL-isolated nuclei from Pvalb-2A-Cre mice, ATAC-seq and RNA-seq from sham animal SUN1GFP+ and SUN1GFP- fractions were compared to publicly available PV+ and PV- data. cSNAIL ATAC-seq and RNA-seq exhibited the expected cell type-specific patterns at the Pvalb locus compared with signals from INTACT-isolated populations (Mo et al., *supra*) (**FIG. 7A**). When comparing the identified cell type-specific ATAC-seq markers to single nucleus (sn)ATAC-seq clusters from the adult mouse cortex (github.com/timoast/signac), the SUN1GFP+ peaks from each brain region were enriched for PV+ cluster markers (**FIG. 7B**). Aligning with the expected cell type proportions in each brain region, SUN1GFP- fractions were most enriched for snATAC-seq markers of excitatory neuron clusters (cortex only) and glial clusters.

Comparison of cSNAIL RNA-seq in Pvalb-2A-Cre mice with mouse cortex snRNA-seq further corroborated the desired enrichment. For this analysis, cSNAIL SUN1GFP+ and cSNAIL SUN1GFP- gene expression signatures (i.e., strongly differentially expressed genes) were defined in the cortex. snRNA-seq cells were then assessed for similar expression patterns by scoring their expression of the cSNAIL SUN1GFP+ gene signature (**FIGS. 7C, 13A, and 13B**). The SUN1GFP+ scores were computed by averaging across the scaled (0 mean centered, unit variance) expression of signature genes across all cells. As expected, cells in the PV+ interneuron cluster (circled in **FIG. 13B**) had the strongest expression of the cSNAIL SUN1GFP+ gene signature, followed by other interneuron clusters. These comparisons provided evidence that cSNAIL SUN1GFP+ samples are indeed strongly enriched for PV+ neurons. Therefore, throughout the remainder of this Example, cSNAIL SUN1GFP+ and SUN1GFP- cells are referred to as PV+ and PV- cells, respectively.

Molecular signatures of cSNAIL-isolated PV+ and PV- cells. In each brain region, hundreds of differentially expressed genes and thousands of differentially accessible

ATAC-seq regions were recovered between PV+ and PV- cell populations (**FIGS. 8A-8F**). 35 PV+ and 29 PV- marker genes with high specificity (DESeq2 $p_{adj} < 0.01$ and $|LFC| > 1$) across all three tissues were defined. Of these genes, a small number also had corresponding cell type-specific chromatin accessibility within a 10kb window around the gene transcription start site, including those highlighted in **FIGS. 8B and 8E**. The PV+ marker genes *Coro6*, *Flt3*, and *Cacna2d2* also tended to be restricted to PV+ clusters in DropViz snRNA-seq from the mouse cortex, striatum, and globus pallidus, although *Cacna2d2* also was prevalent in Chat neurons (**FIG. 14A**) (Saunders et al., *Cell* 174:1015-1030.e16, 2018). This further demonstrated that cSNAIL enrichment for PV+ neurons was generalized to brain regions outside the cortex. PV- gene expression signatures that were consistent across all three brain regions appeared to be dominated by glial cells, especially oligodendrocytes (e.g., *Hepacam* and *Kcnj10*) (**FIG. 14B**). This was not surprising given that neuron composition varies greatly between the cortex, striatum, and GPe. Functional gene enrichment on pan-PV+ and pan-PV- differential genes was consistent with these observations, implicating neuron-related pathways among PV+ genes and glial pathways among PV- genes (**FIG. 8C**).

Next, studies were conducted to identify transcription factor binding motifs that might play a role in epigenetically defining PV+ neurons, by testing for enriched motifs in sequences underlying PV+ and PV- specific ATAC-seq peaks in each brain region (**FIG. 8F**). Among the most significantly enriched PV+ motifs in all three brain regions was *Tcf3*, which was previously implicated in PV+ neuron-specific gene regulation (Mo et al., *supra*). Additionally, consistent PV+ enrichment for the *Esrrg* and *Hoxa2* motifs was observed. The *Esrrg* transcript itself also was upregulated in PV+ neurons in all three brain regions ($p_{adj} < 0.001$), but *Tcf3* or *Hoxa2* were not. Consistent PV- motif enrichments across all brain regions included the motif for *Sox10*, which is necessary for the differentiation and survival of oligodendrocytes (Pozniak et al., *Proc Natl Acad Sci USA* 107:21795-21800, 2010; and Takada et al., *Glia* 58:996-1006, 2010). There also were significant enrichments for many other Sox family motifs and the *Meis1* motif among PV- ATAC-seq sequences. The representation of glial signatures among PV- cell ATAC-seq and RNA-seq indicates that the nuclei dissociation during cSNAIL affinity purification was not restricted to neurons.

Gene expression differences in dopamine depleted animals. Striatal tissue from 6-OHDA lesioned animals had significantly lower levels of dopamine production enzyme

tyrosine hydroxylase than sham animals, indicating successful depletions (FIGS. 11A-11B). All depletions were bilateral except in one animal, DD2, where one hemisphere was only weakly depleted. The RNA-seq data was high quality for all samples except one striatum PV- sample from a DD animal, which was excluded from the analysis. Samples were separated by tissue and cell type in PCA and hierarchical clustering, indicating higher variability between cell types than between biological replicates (FIGS. 15A and 15B).

In the RNA-seq data from DD and sham animals, very few significant DD-affected genes were recovered in any cortical cell types, striatal cell types, or the GPe PV- population. In contrast, GPe PV+ neurons contained 29 differentially expressed genes in DD ($p_{\text{adj}} < 0.05$) (FIGS. 9A and 9B). It was readily apparent that there were many genes involved in cellular oxygen homeostasis and neuroprotection among the fourteen genes that were upregulated in the GPe PV+ neurons of DD animals. These included *Epas1*, which encodes for Hypoxia-inducible factor 2 alpha (transcription factor Hif2a), and three of its targets: *Cp*, *Slc2a1*, and *Flt1* (Schofield and Ratcliffe, *Nat Rev Mol* 5:343-354, 2004; Dengler et al., *Crit Rev Biochem Mol Biol* 49:1-15, 2015; and Smeyne et al., *Neurosci* 295:23-38, 2015) (FIG. 9C). The downregulated set also included *Elob*, a moderator of free oxygen sensing that inhibits Hif2a (Ohh et al., *Nat Cell Biol* 2:423-427, 2000). These results suggested that GPe PV+ neurons may react to DD through hypoxia-inducible factor (HIF) signaling and associated transcriptional response.

To see whether increased *Epas1* transcription in GPe PV+ neurons was accompanied by increased Hif2a protein levels, double immuno-fluorescent staining for Pvalb and Hif2a was performed in healthy and DD mouse tissue. Indeed, a higher proportion of GPe PV+ neurons expressed Hif2a in images from DD animals compared with healthy animals (FIGS. 9D and 16). The change in mean proportion was 11.3% and this difference was significant (t-test, $p = 0.007$).

Other differentially expressed genes in the GPe PV+ neurons of DD mice have also been implicated in processes of neurodegeneration and/or neuroprotection. For example, overexpression of *Timp3*, a key inhibitor of matrix metalloproteinases that contribute to dopaminergic neuron apoptosis in PD (Kim et al., *FASEB J* 21:179-187, 2007), was observed. *Timp3* expression in neurons is protective against blood-brain barrier damage, but high levels of *Timp3* can lead to cell death (Rosenberg, *Lancet Neurol* 8:205-216, 2009). *Lpcat3*, a gene found to have reduced expression in GPe PV+

neurons upon DD, is necessary for ferroptosis, which has been linked to neurodegeneration in PD and Alzheimer's disease (Angeli et al., *Cell* 171:273-285, 2017). Due to the small number of regulated genes, functional enrichment for specific pathways was low.

Open chromatin differences in dopamine depleted animals. Based on standard QC metrics (see above), the ATAC-seq data was high quality for all samples except one cortex PV+ sample from a DD animal, which was excluded from the analysis (FIGS. 12A-12D). In DD animals, a small number of confident open chromatin changes were observed in PV+ cell types, and many changes were observed within PV- cell types that met significance at a cutoff of $p_{\text{adj}} < 0.05$ (FIG. 10A). The DD-affected open chromatin regions in different cell types had some redundancy, but the majority were cell type-specific at the level of the individual regulatory element (FIG. 17). However, DD-affected regions in multiple cell types tended to converge around areas of the genome enriched for certain functions including neurotransmission, immune response, and the methionine cycle (FIG. 18).

Further studies were conducted to assess whether the observed differences in open chromatin could be responsible for the transcriptional differences observed in the GPe PV+ population. Genes with increased expression in DD had higher levels of open chromatin nearby (hypergeometric $p=0.0003$), while genes that decreased in expression had lower levels (hypergeometric $p=0.002$) (FIG. 19). However, there were no overlaps in the most significantly differential genes and peaks.

Due to the increased expression of HIF-related transcripts in GPe PV+ neurons in DD, it was hypothesized that changing open chromatin regions in this cell type would contain HIF transcription factor binding motif sites. Upon activation, Hif1a subunits heterodimerize with Hif1b and the complex binds regulatory elements in the genome to affect transcription. Therefore, a motif enrichment analysis was conducted for four HIF family motifs with Fisher's exact test statistics. These included the JASPAR motifs for (i) Hif1a (MA1106.1), which has core motif redundancy with Hif2a (Schodel et al., *Nat Rev Mol* 5:343-354, 2011), (ii) Arnt aka Hif1b (MA0004.1), (iii) the Hif1a-Hif1b complex (MA0259.1), and a related Hif1a binding partner Arnt1 (M0603.1) (FIG. 10B). Because there are few examples of DD-affected PV+ peaks at this threshold, PV+ enrichments were assessed within peaks that met a more lenient threshold of adjusted p-value < 0.2 in DESeq2. As predicted, all four motifs were highly enriched within GPe PV+ DD-

increasing peaks (the set of ATAC-seq peaks with increased accessibility upon DD) (FIG. 10C). Moreover, the motifs were not enriched in DD-increasing peak sets from other cell types with the exception of some weaker enrichment in Striatal PV+ neurons. Finally, sets of ATAC-seq peaks with decreased accessibility in DD (DD-decreasing) were not enriched for HIF motifs in any cell type including GPe PV+. These results provided further support for a cell type-specific induction of HIF signaling in GPe PV+ neurons upon DD.

To assess which additional transcription factors may be involved in the epigenetic response to DD, the analysis was extended to all motifs within the non-redundant JASPAR 2018 core vertebrates database (Khan et al., *supra*). Specific to DD-increasing GPe PV+ peaks and to a lesser extent also DD-increasing Striatal PV+ peaks, an enrichment for Ets-related transcription factor motifs including Gabpa was observed (GPe $p_{\text{adj}} = 1.25\text{e-}30$; Striatum $p_{\text{adj}} = 1.41\text{e-}10$). Gabpa is an essential regulator of many cellular respiration genes, including multiple cytochrome c oxidase subunits (Ongwijitwat et al., *Gene* 374:39-49, 2006).

DD-affected open chromatin regions in other cell types contained enrichments for several additional transcription factor binding motifs. Notably, DD-increasing peak sets in every cell population except for GPe PV+ were all enriched for the binding motifs of glucocorticoid receptor transcription factors Nr3c1, Nr3c2, and Ar ($p_{\text{adj}} < 0.01$). Glucocorticoid receptors are emerging as key regulators of neuroinflammation and disruption of their proper regulation is thought to play a role in the onset of PD by allowing the infiltration of cytotoxins in the SNpc (Herrero et al., *Front Neuroanat* 9:1-12, 2015). These findings were consistent with similar stressors in the cortex, striatum, and GPe, which may contribute to the progression of PD.

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

WHAT IS CLAIMED IS:

1. A method for generating, by a data processing system, data representing a synthetic genetic sequence configured for labeling at least one cell type by causing expression of a marker in the at least one cell type, the method comprising:
 - receiving training data associating at least one feature of a genetic sequence of the at least one cell type with expression of a hallmark of the at least one cell type;
 - training, based on the training data, a model configured to generate data representing a synthetic genetic sequence;
 - receiving input data comprising the at least one feature; and
 - generating the synthetic genetic sequence in response to receiving the input data comprising the at least one feature.
2. The method of claim 1, wherein the model comprises a generative adversarial network (GAN).
3. The method of claim 2, wherein the GAN is a deep convolutional GAN (DCGAN).
4. The method of claim 2, wherein the GAN comprises:
 - a generator network configured to receive latent random noise data as the input data and generate the synthetic genetic sequence; and
 - a discriminator network configured to generate a probability value representing whether the input sequence is drawn from the synthetic genetic sequence from the generator network or from a distribution of natural genetic sequences.
5. The method of claim 4, comprising:
 - training the generator network and the discriminator network of the GAN by alternating the input data between synthetic sequence derived from the latent noise data input to the generator network and natural genetic sequence data.
6. The method of claim 5, comprising:
 - receiving input data representing endogenous genomic sequences underlying previously profiled open chromatin regions of at least two cell types; and

updating the discriminator network for distinguishing between the at least two cell types.

7. The method of claim 6, wherein the at least two cell types include parvalbumin positive (PV+) and parvalbumin negative (PV-) neurons; and wherein the discriminator network is configured to distinguish between the PV+ and PV- neurons based on the input data.

8. The method of claim 5, comprising optimizing a feature of the synthetic genetic sequence by applying a class label to both the generator network and the discriminator network.

9. The method of claim 8, wherein the class label is configured to force a first probability value for a first input type and a second probability value for a second input type that is different from the first input type.

10. The method of claim 8, wherein the feature represents an enhancer of an activity in a cell type.

11. The method of claim 1, wherein the method further comprises: generating a nucleic acid sequence comprising the synthetic genetic sequence.

12. The method of claim 11, wherein the nucleic acid sequence comprises the synthetic genetic sequence operably linked to a nucleotide sequence encoding a marker.

13. The method of claim 12, wherein the marker is a tagged Sun1 fusion polypeptide.

14. The method of claim 11, wherein the nucleic acid sequence further comprises a virus nucleic acid sequence.

15. The method of claim 14, wherein the virus is adeno-associated virus (AAV) or lentivirus.

16. The method of claim 1, comprising:
receiving results data representing a delivery of a nucleic acid including the synthetic genetic sequence to an organism, the results data representing a successful labeling of the at least one cell type or an unsuccessful labeling of the at least one cell type;
updating the model using the results data; and
generating an updated synthetic genetic sequence based on the updated model.
17. The method of claim 1, further comprising
receiving training data including the hallmark representing marker positive results marker negative results, or both;
extracting at least one feature corresponding to the marker positive result or to the marker negative result; and
adding the at least one feature to the model.
18. The method of claim 17, wherein the at least one feature is a set of k-mer or gapped k-mer counts, and wherein extracting the at least one feature comprises scanning the sequence to determine the set of k-mer counts that form the sequence.
19. The method of claim 1, wherein the model comprises a support vector machine comprising:
a feature space; and
a support vector representing a classification border in the feature space,
wherein the method comprises adding, to the support vector of the support vector machine, a given set of k-mer counts within a predefined distance of the classification border in the feature space of the support vector machine.
20. The method of claim 1, wherein the model comprises a neural network.
21. The method of claim 20, wherein the neural network is a convolutional neural network.

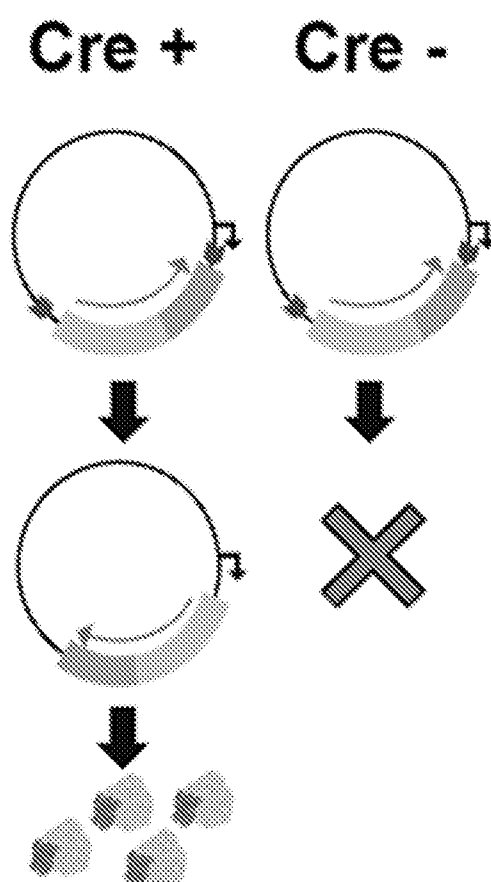
22. The method of claim 20, wherein the neural network comprises one or more weight values each associated with a feature of the synthetic genetic sequence.

23. The method of claim 20, wherein the feature comprises a set of k-mer counts of the genetic sequence.

24. The method of claim 20, wherein the feature represents a transcription factor binding motif.

25. The method of claim 1, wherein the synthetic genetic sequence is configured to distinguish between either:

- parvalbumin positive (PV+) and parvalbumin negative (PV-) neurons;
- PV+ and excitatory (EXC) neurons; or
- PV+ and vasoactive intestinal peptide-expressing (VIP+) neurons.

**FIG. 1A**

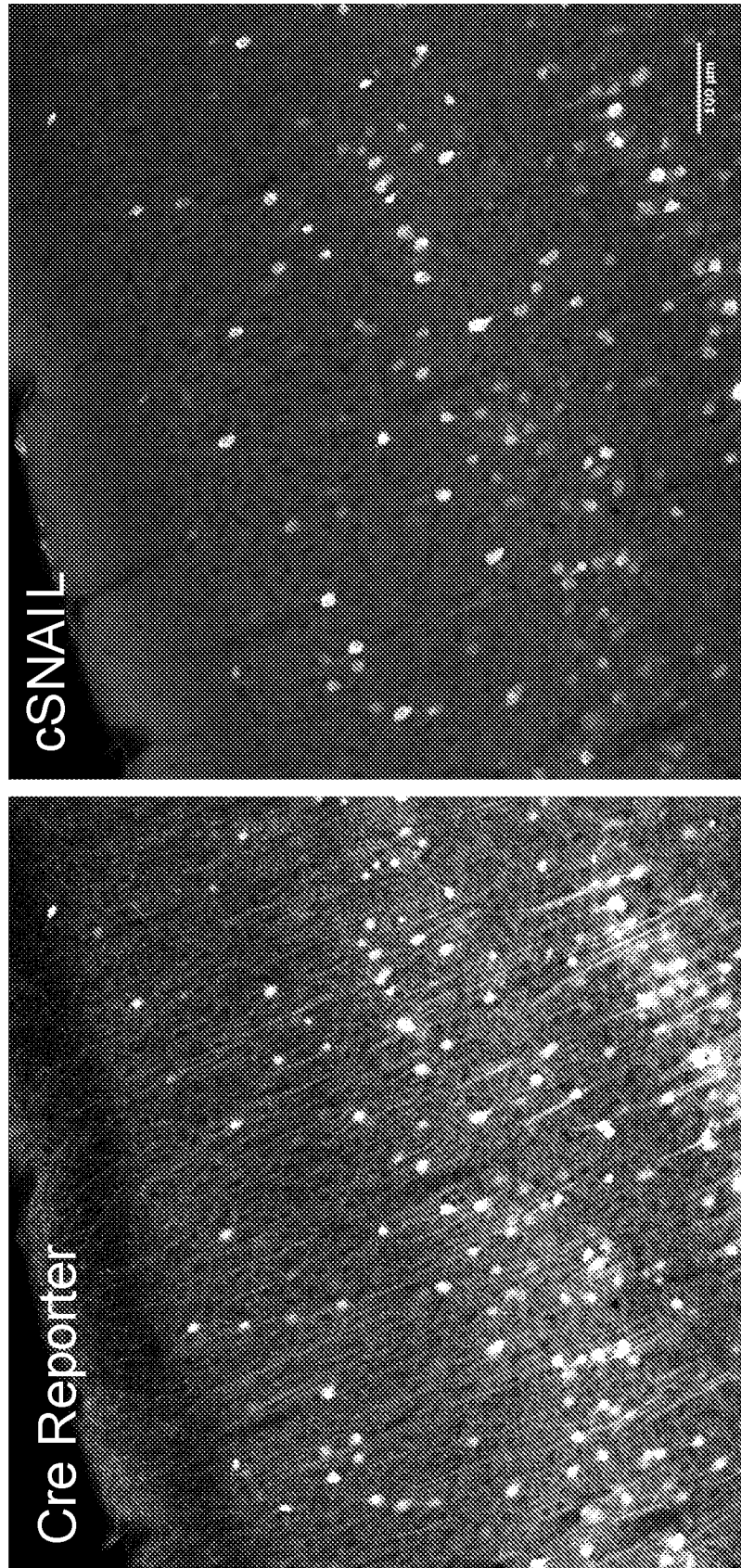
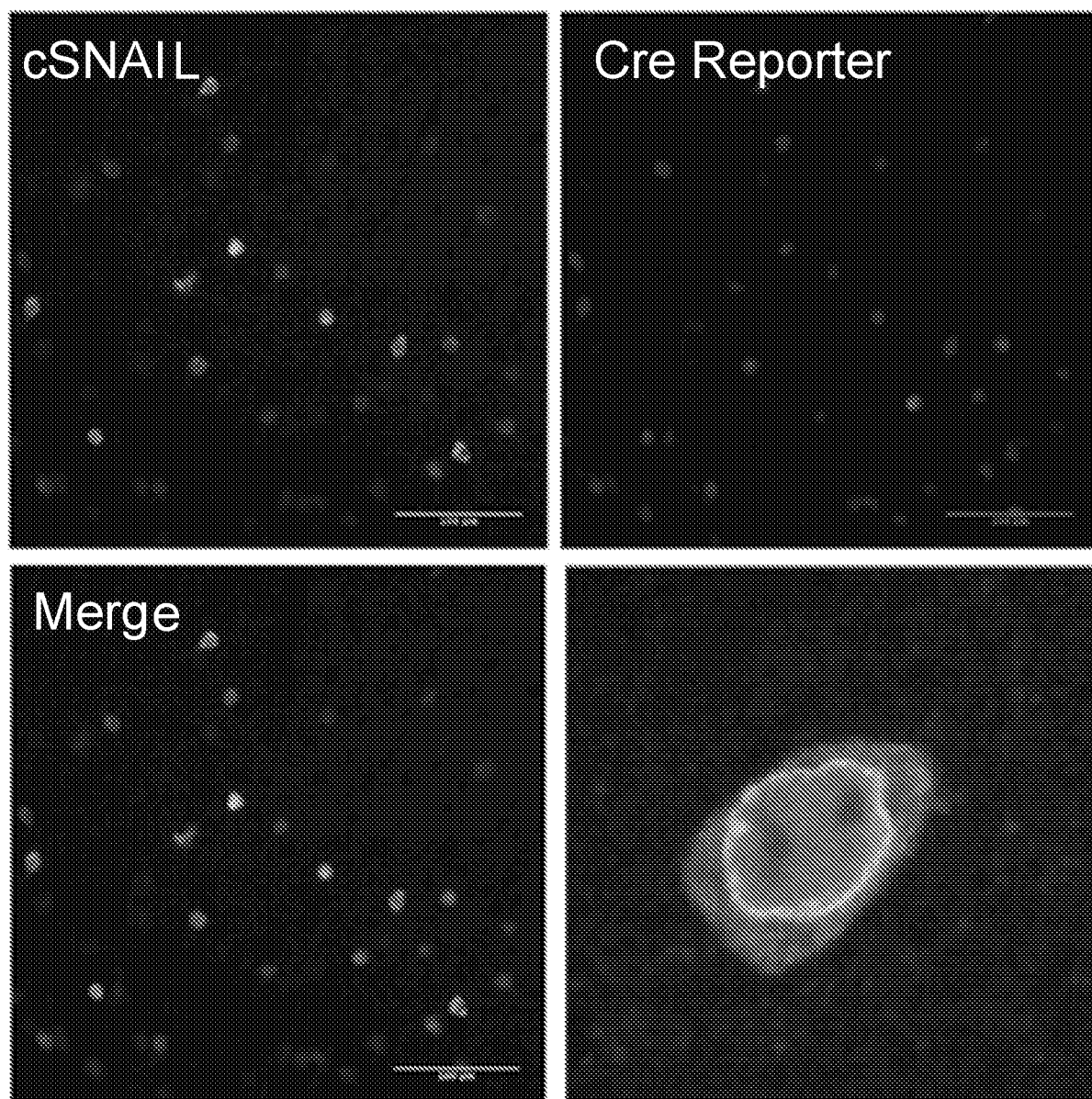
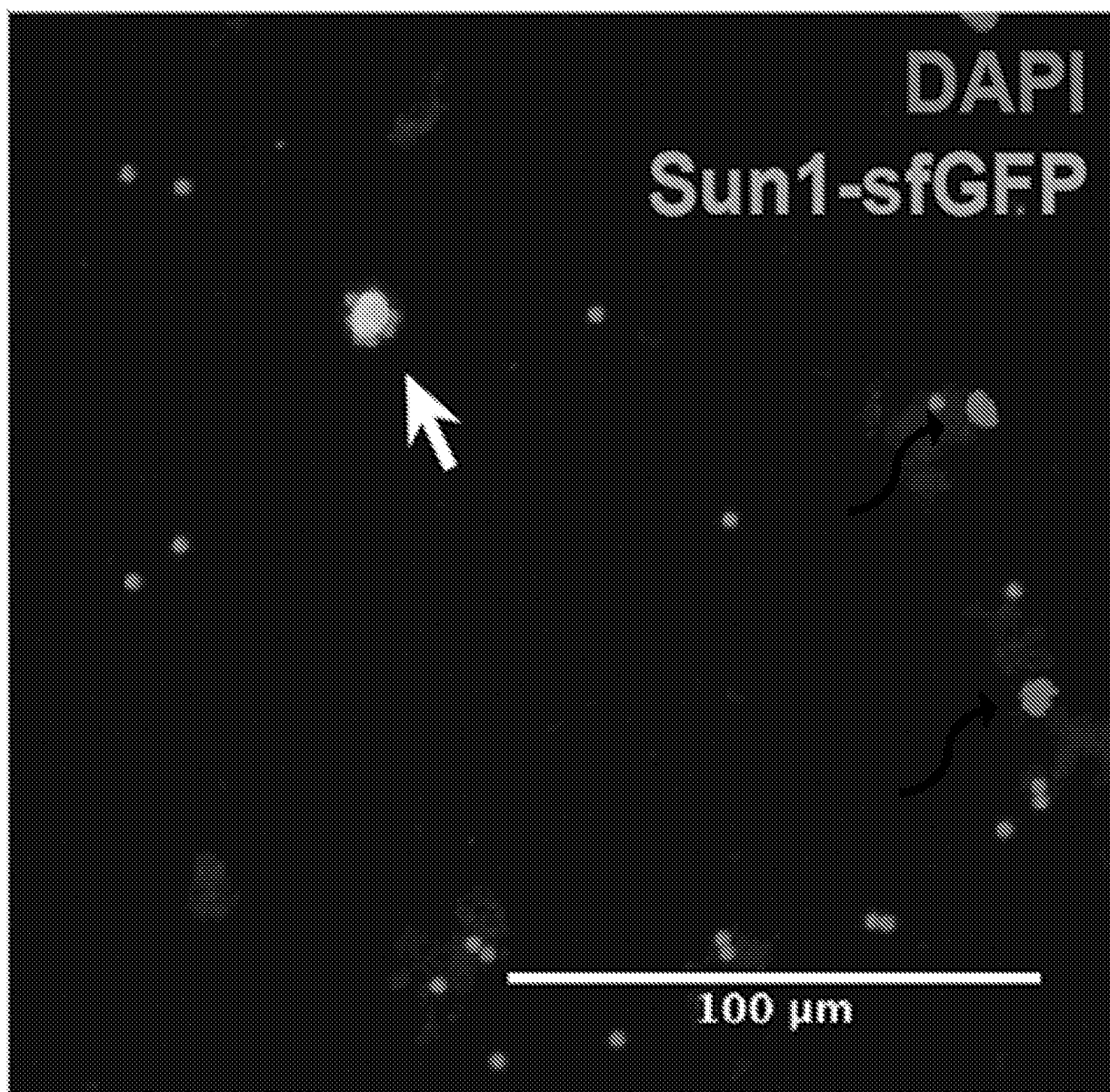


FIG. 1B

**FIG. 1C**

**FIG. 1D**

5/32

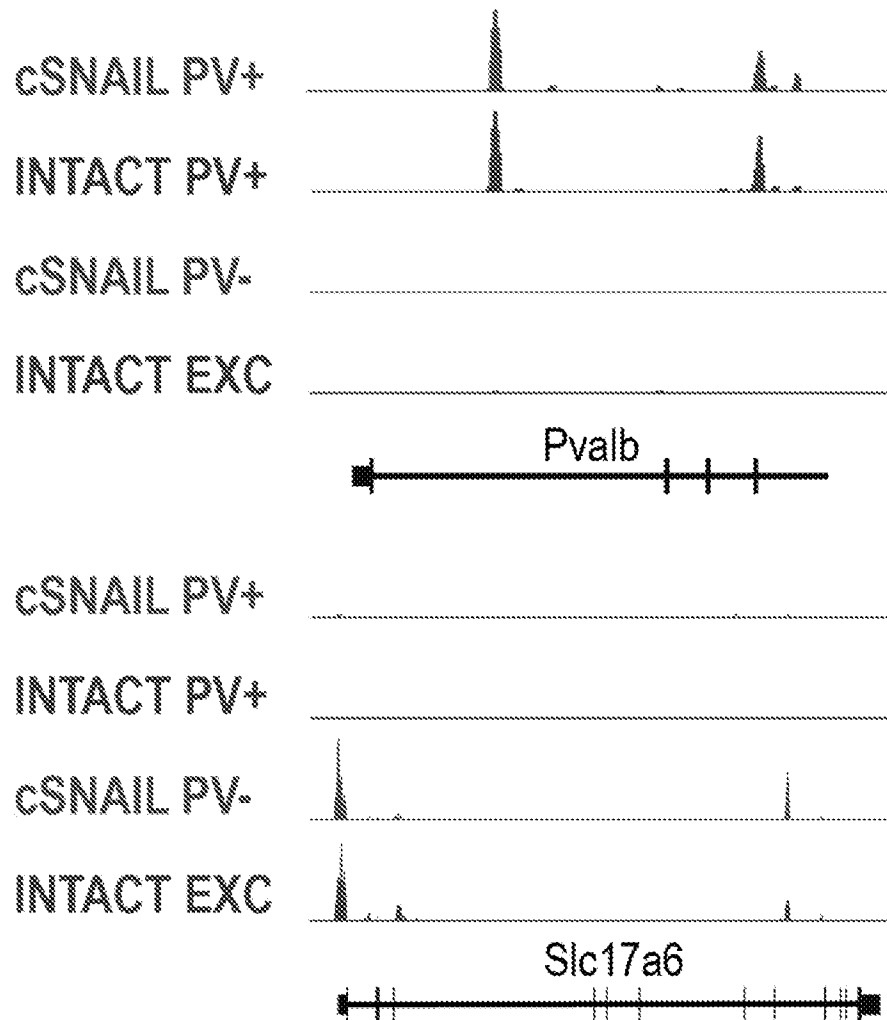
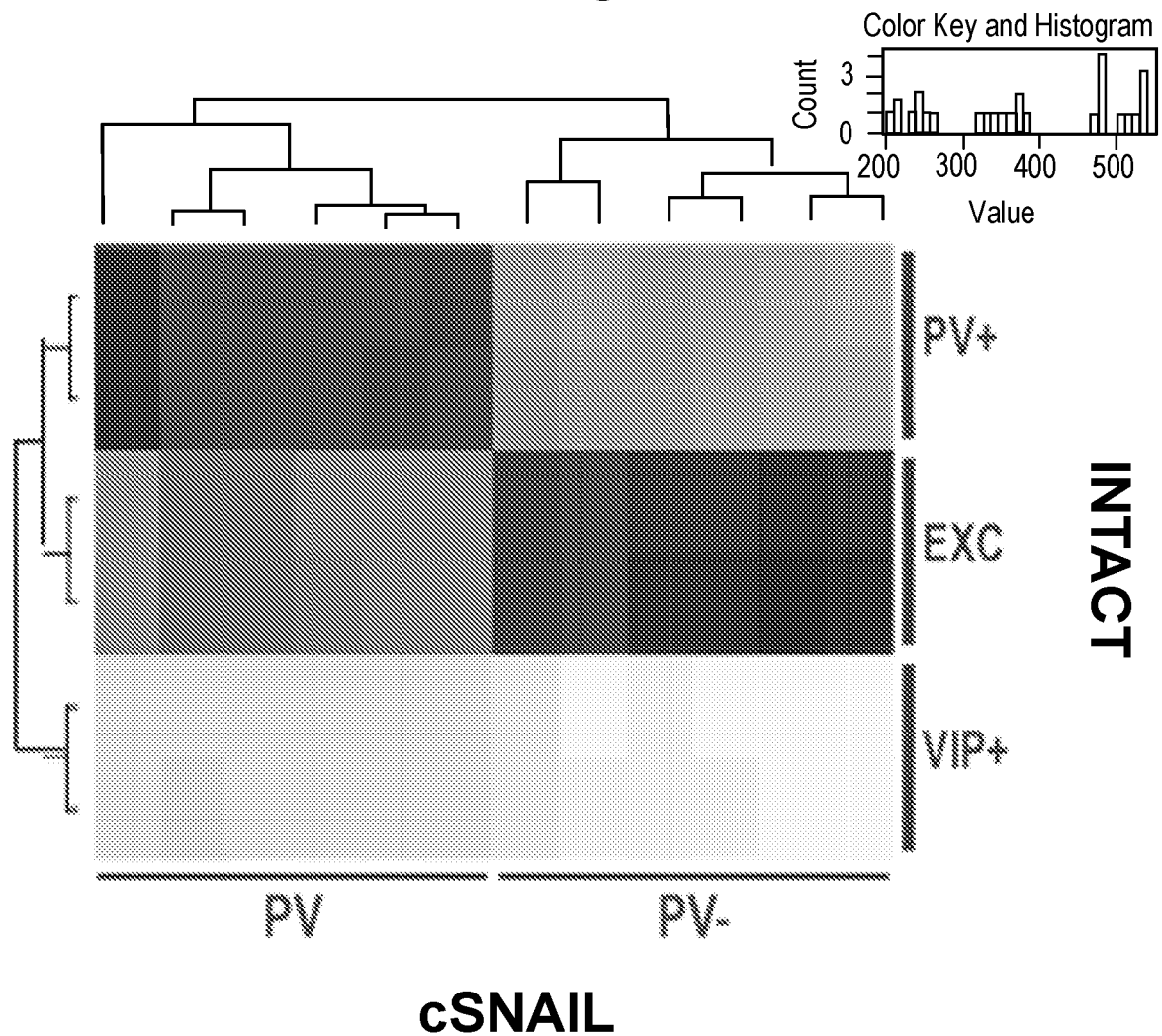
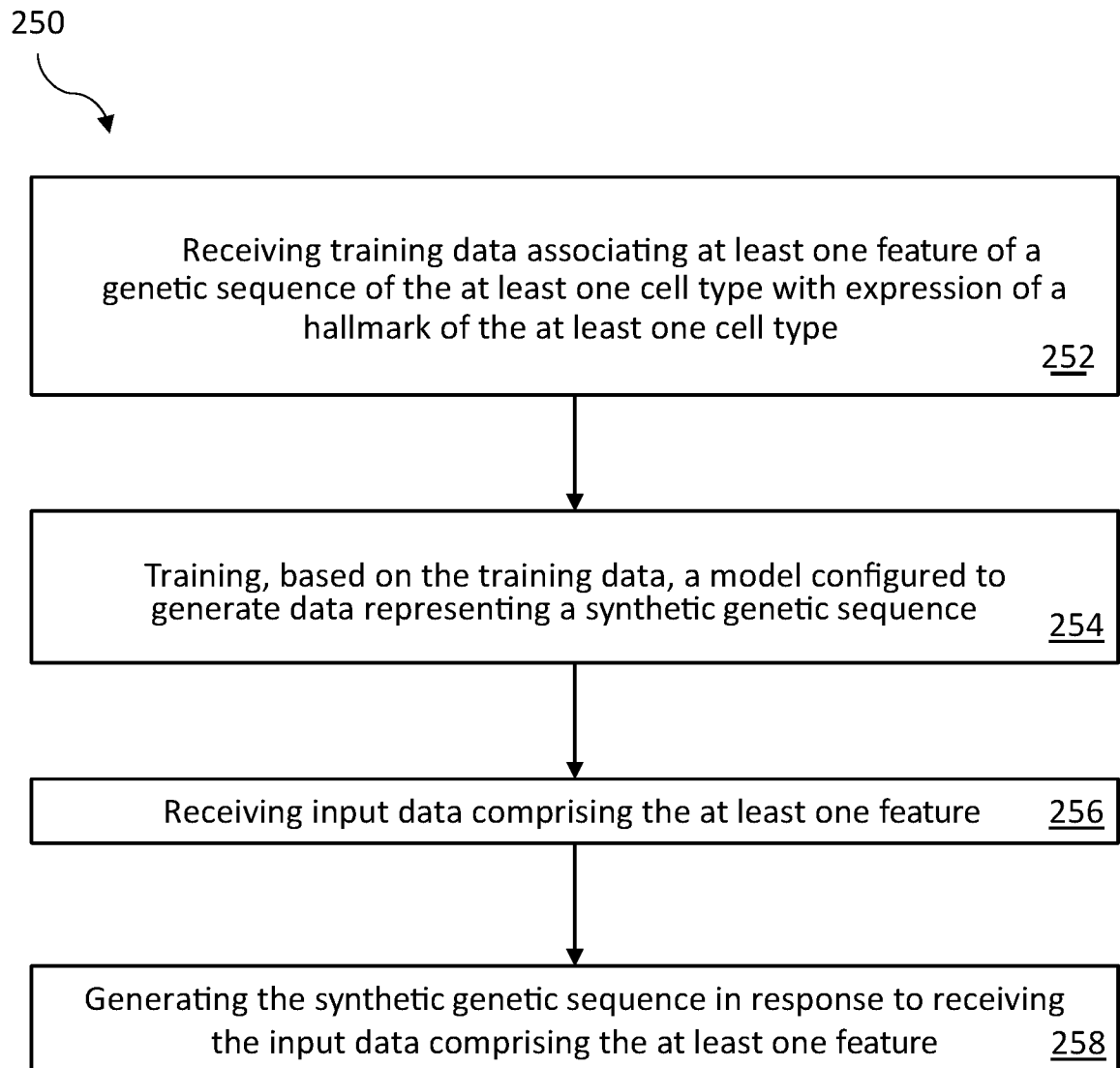
FIG. 1E

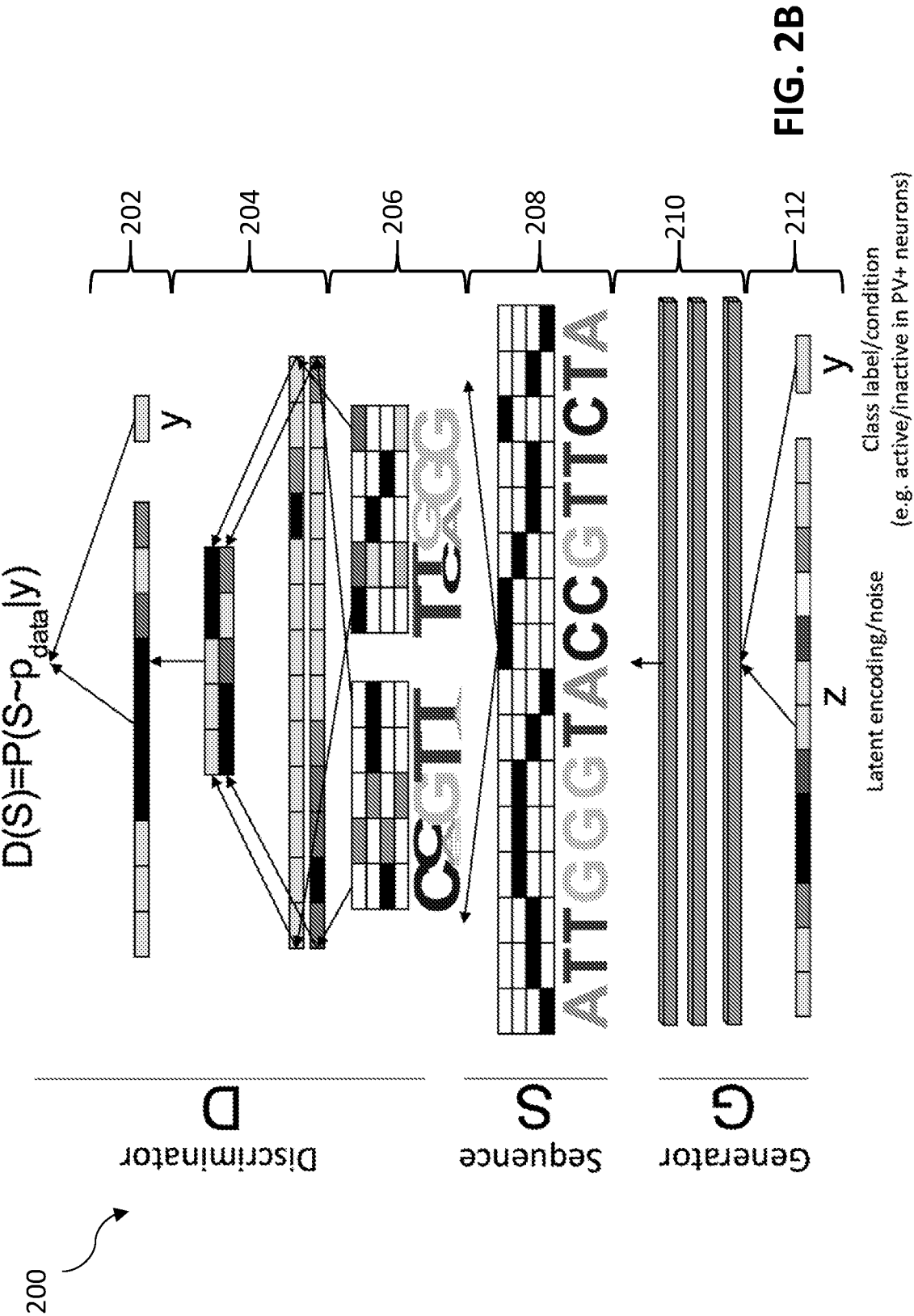
FIG. 1F

Sample Distances Between Viral Probe and Transgenic Methods

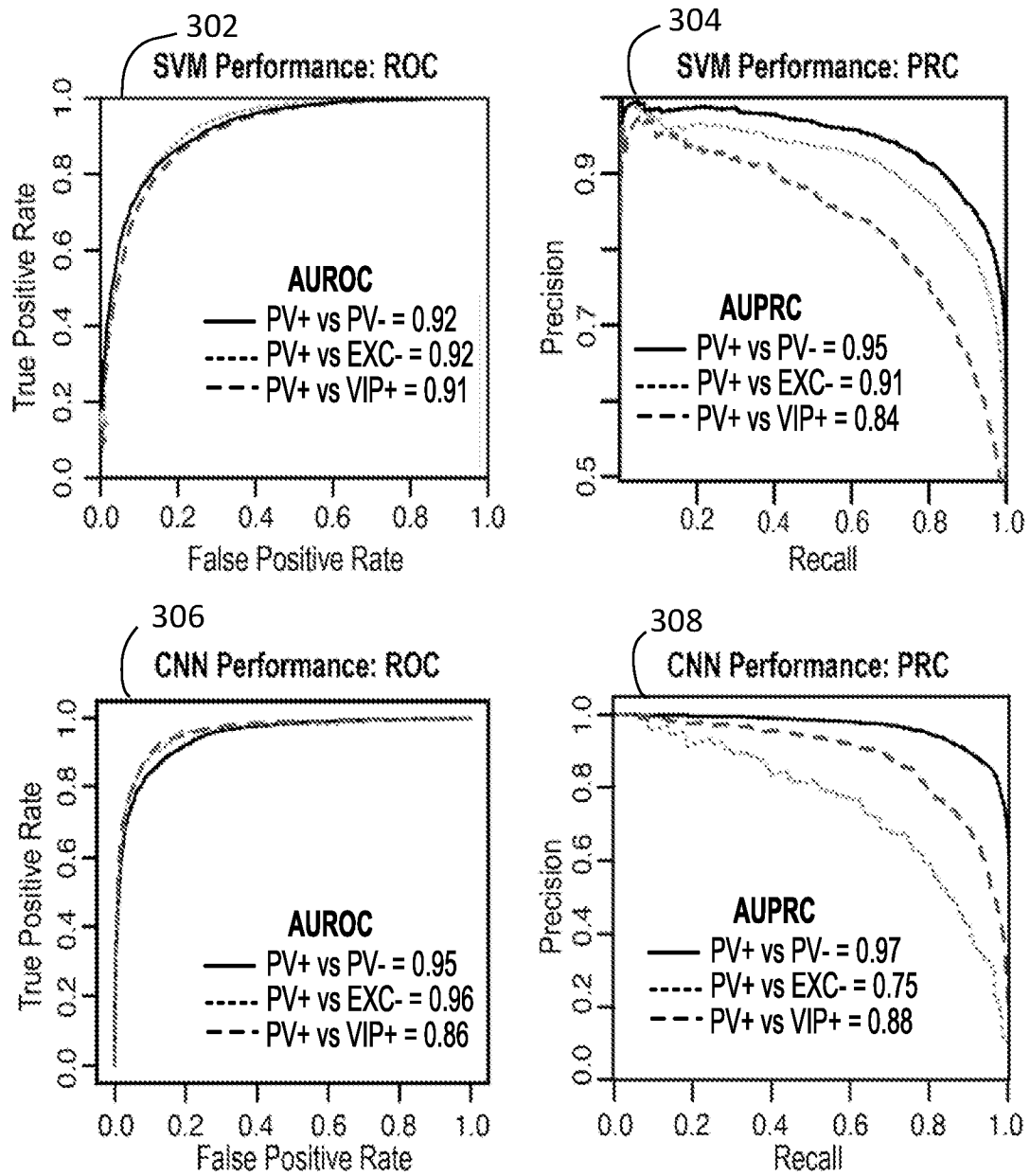


7/32

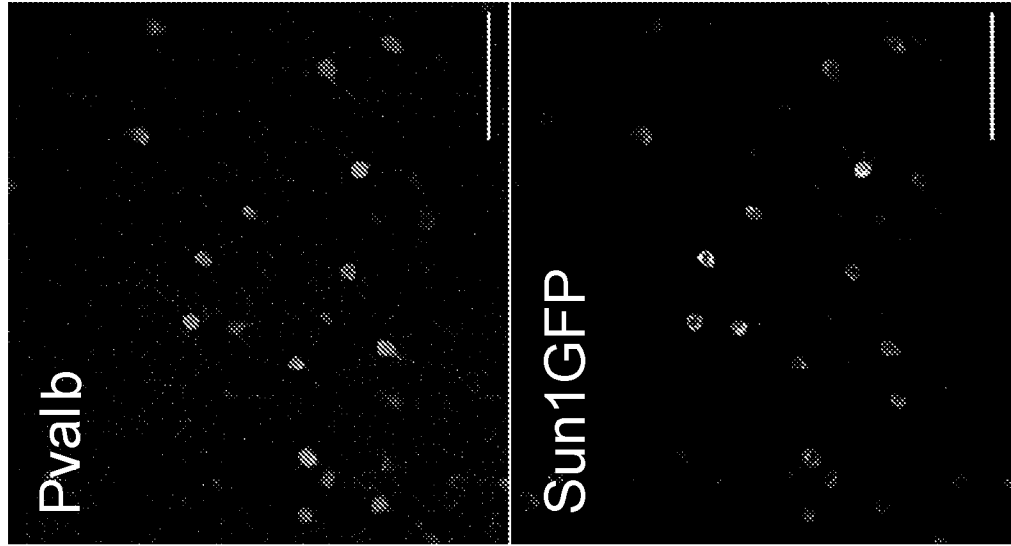
**FIG. 2A**



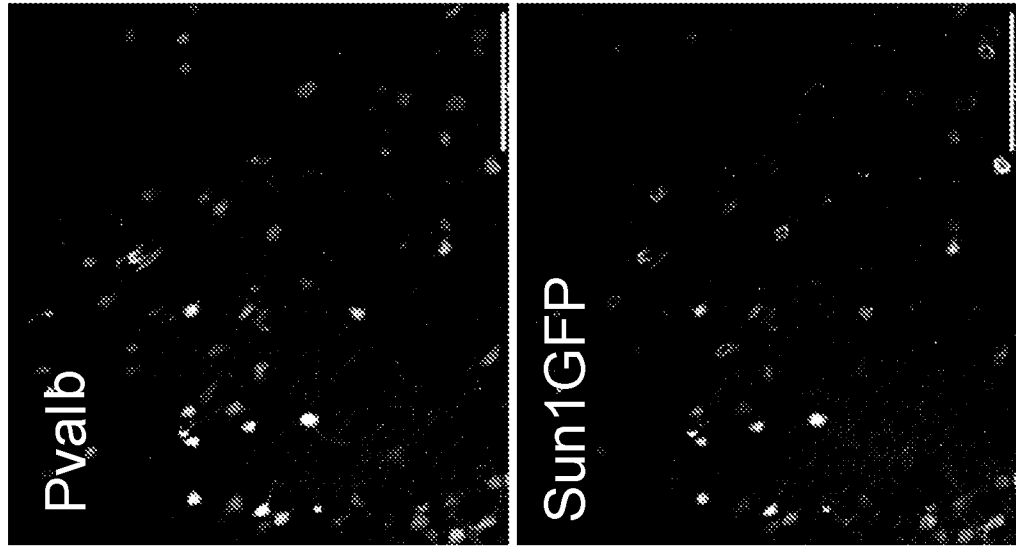
9/32

**FIG. 3**

Rank72



Rank58



Non-specific control

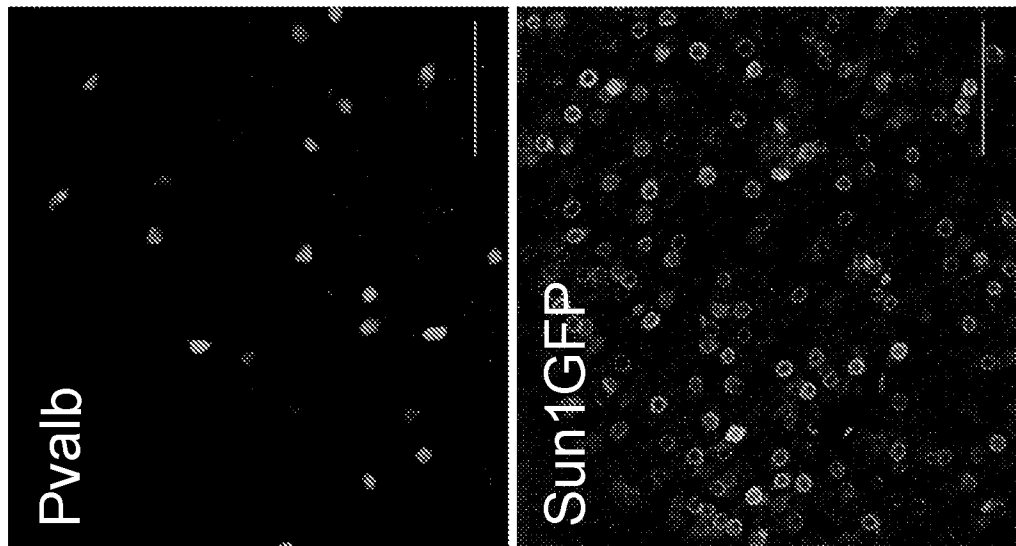
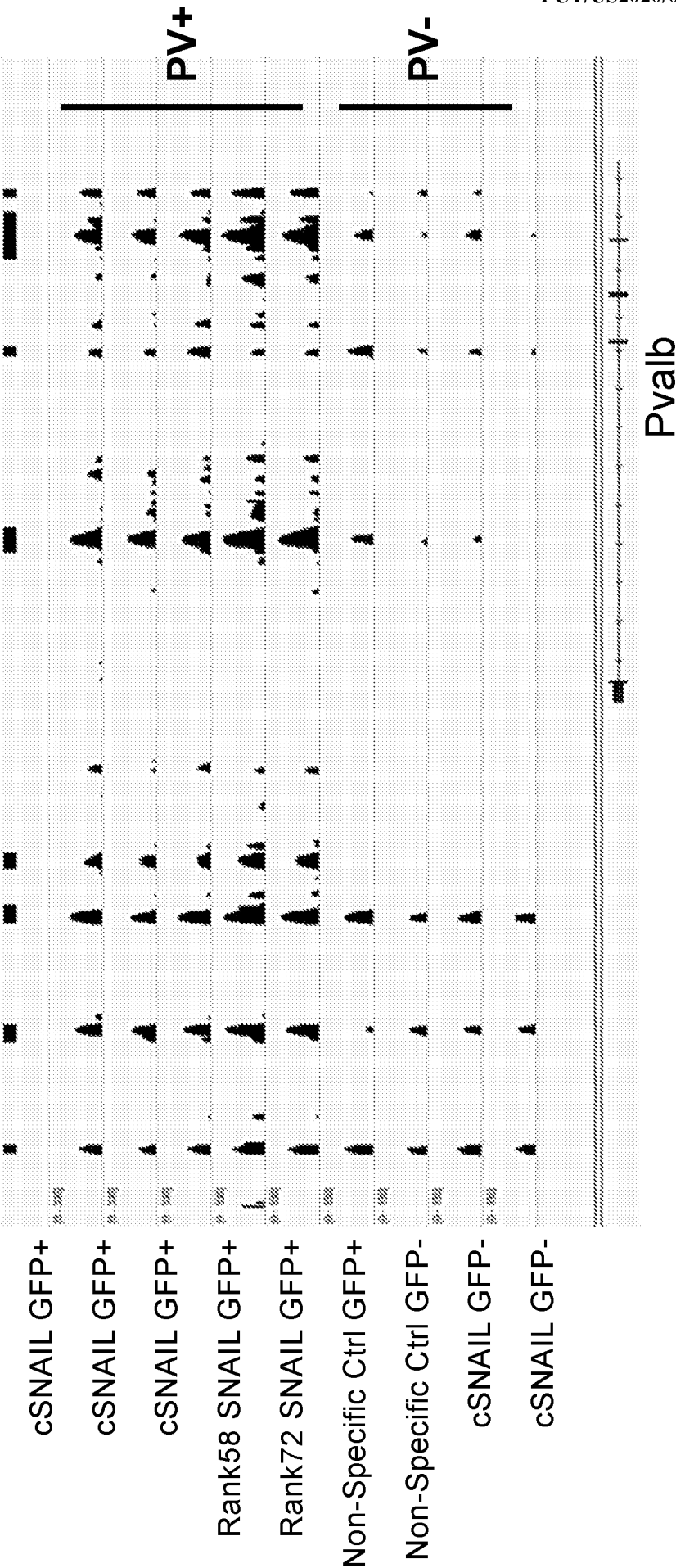
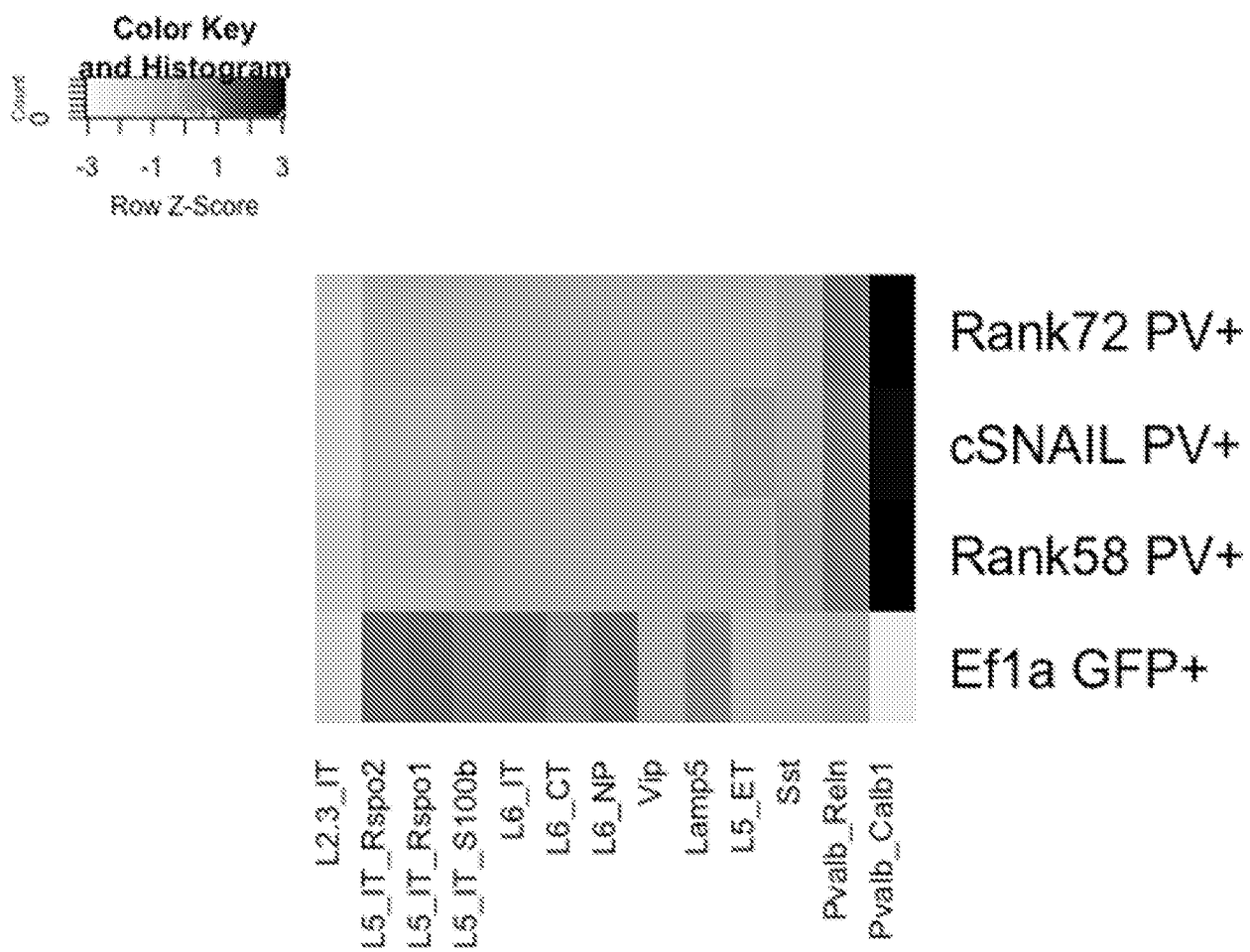


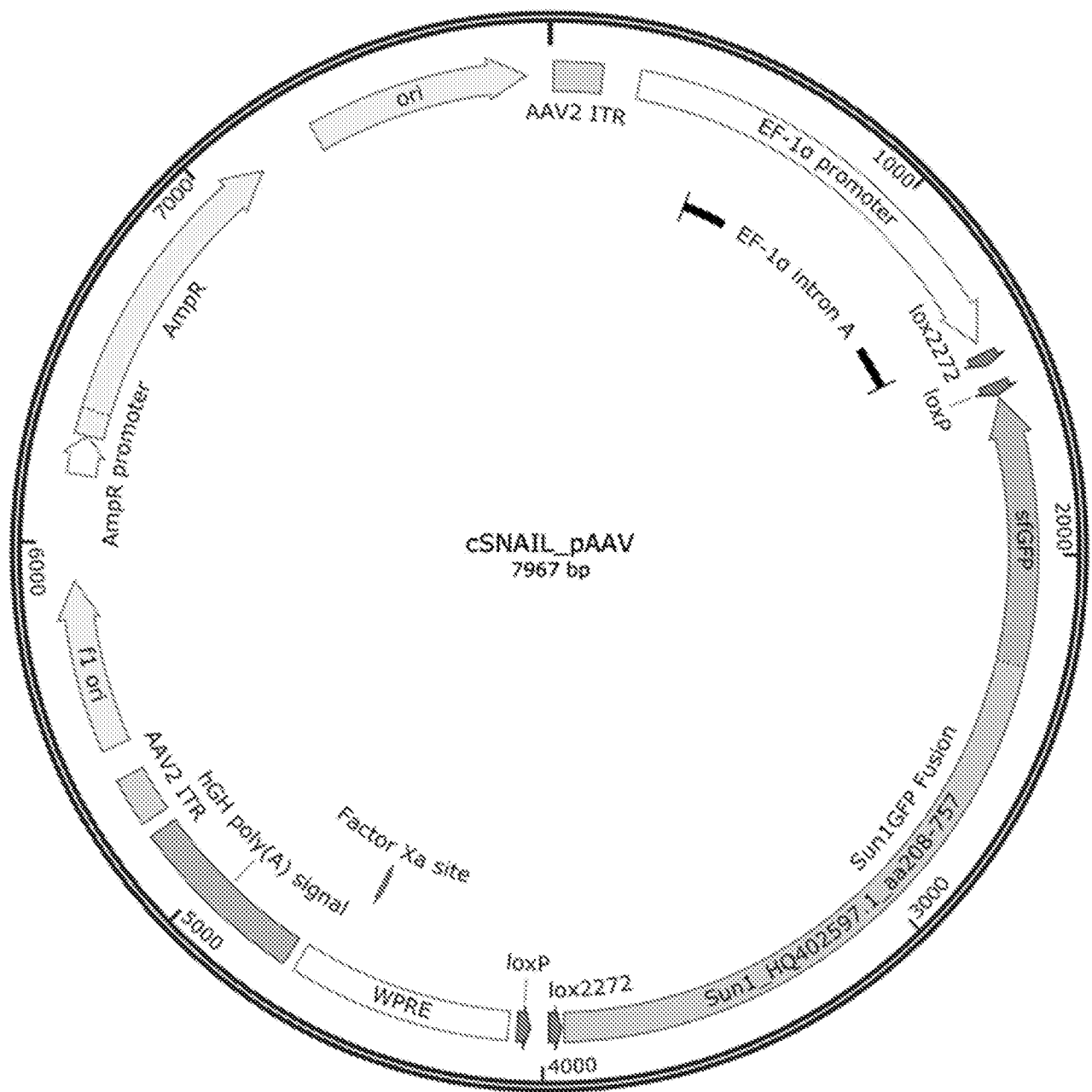
FIG. 4A

FIG. 4B



12/32

**FIG. 4C**

**FIG. 5A**

14/32

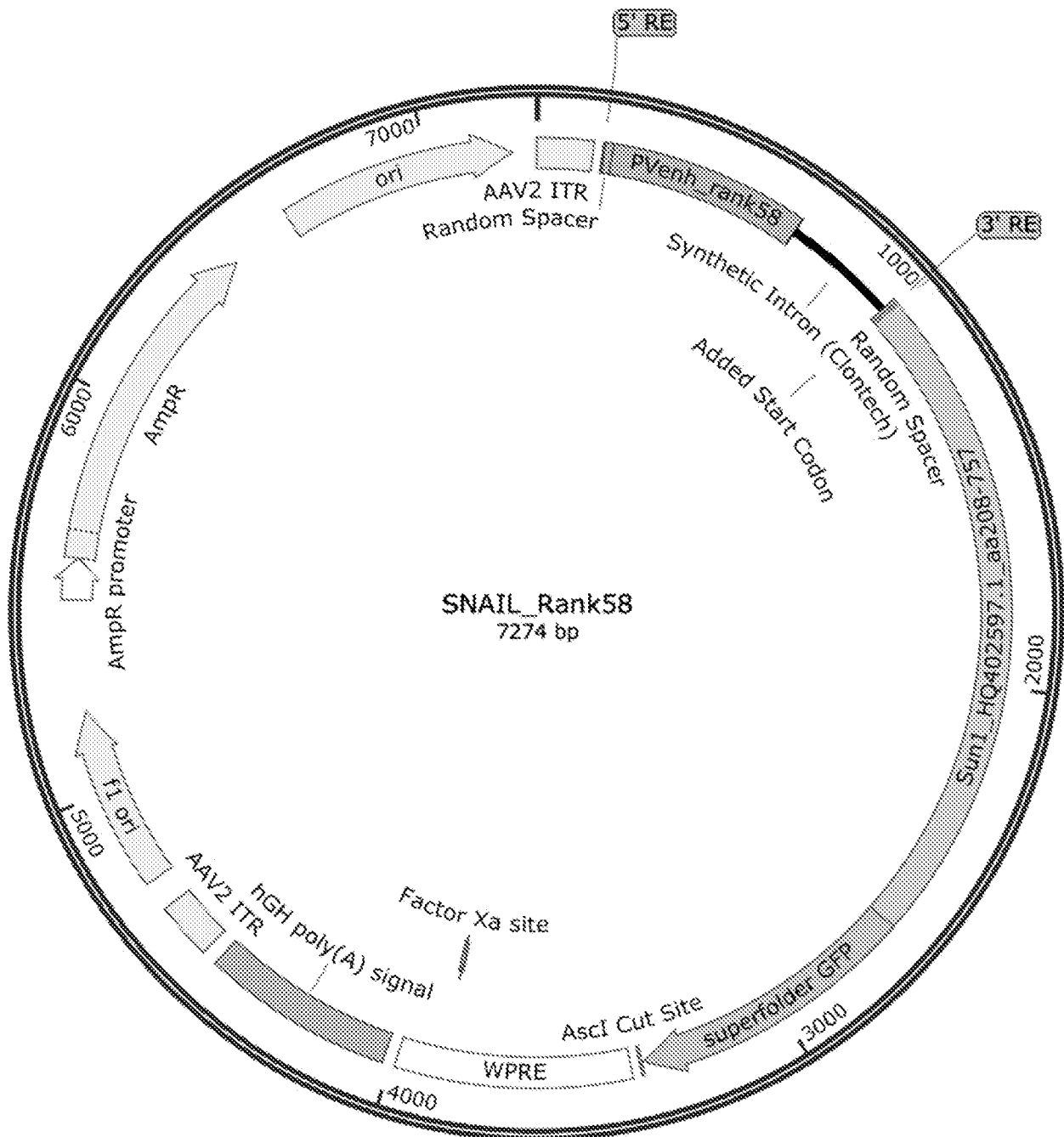


FIG. 5B

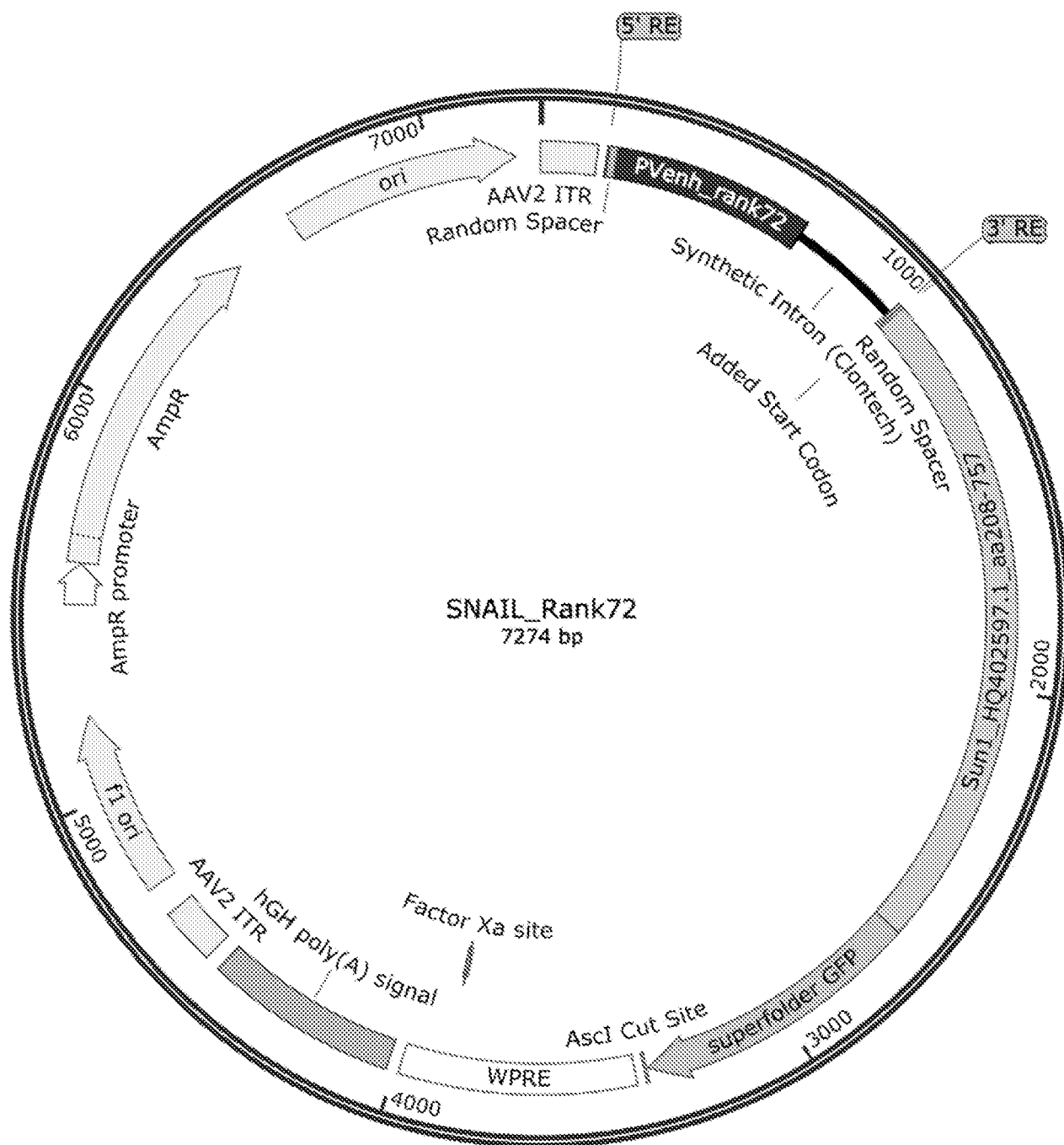


FIG. 5C

16/32

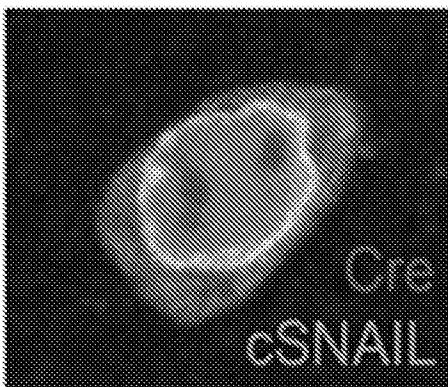


FIG. 6A

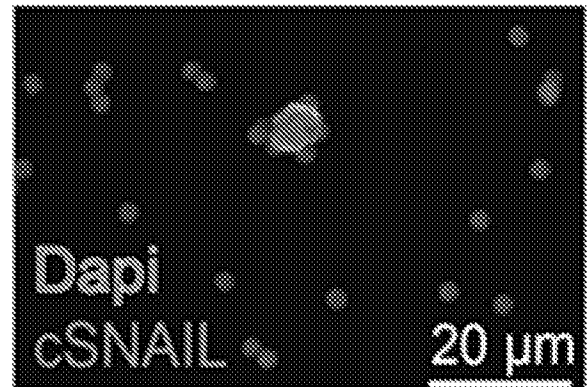
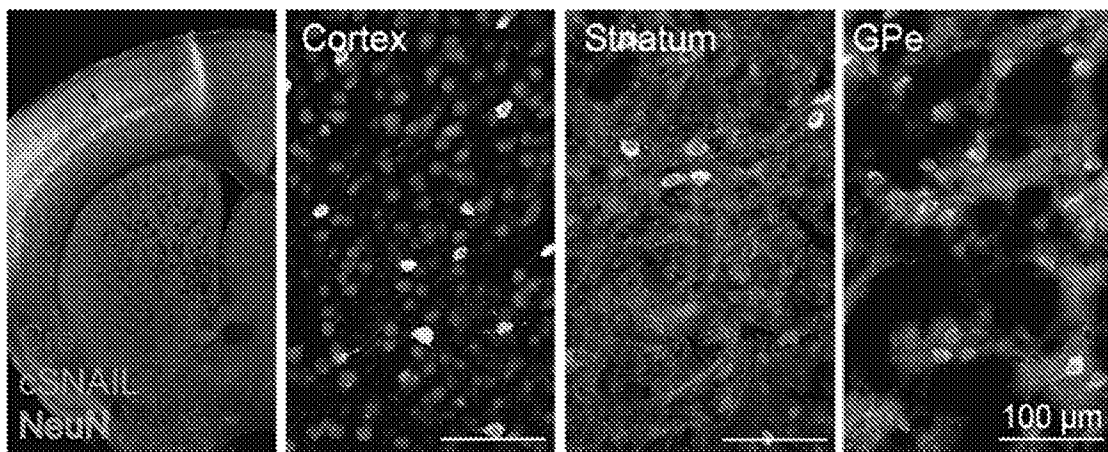


FIG. 6B



Specificity ●/(●+○)	94.0% ± 0.7	94.1% ± 2.8	94.5% ± 0.6
Efficiency ●/(●+●)	80.7% ± 2.3	55.2% ± 4.8	53.3% ± 1.3
Expected transduction	69%	55%	ND

FIG. 6C

17/32

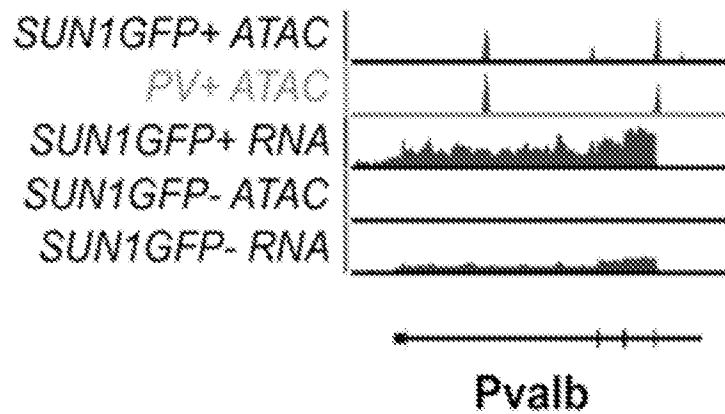


FIG. 7A

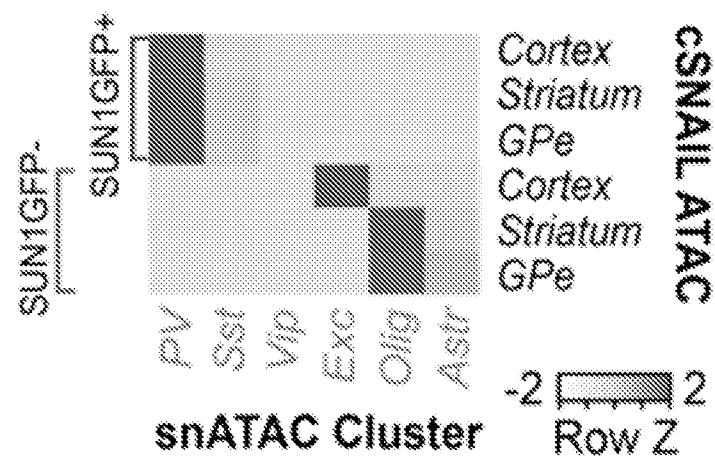


FIG. 7B

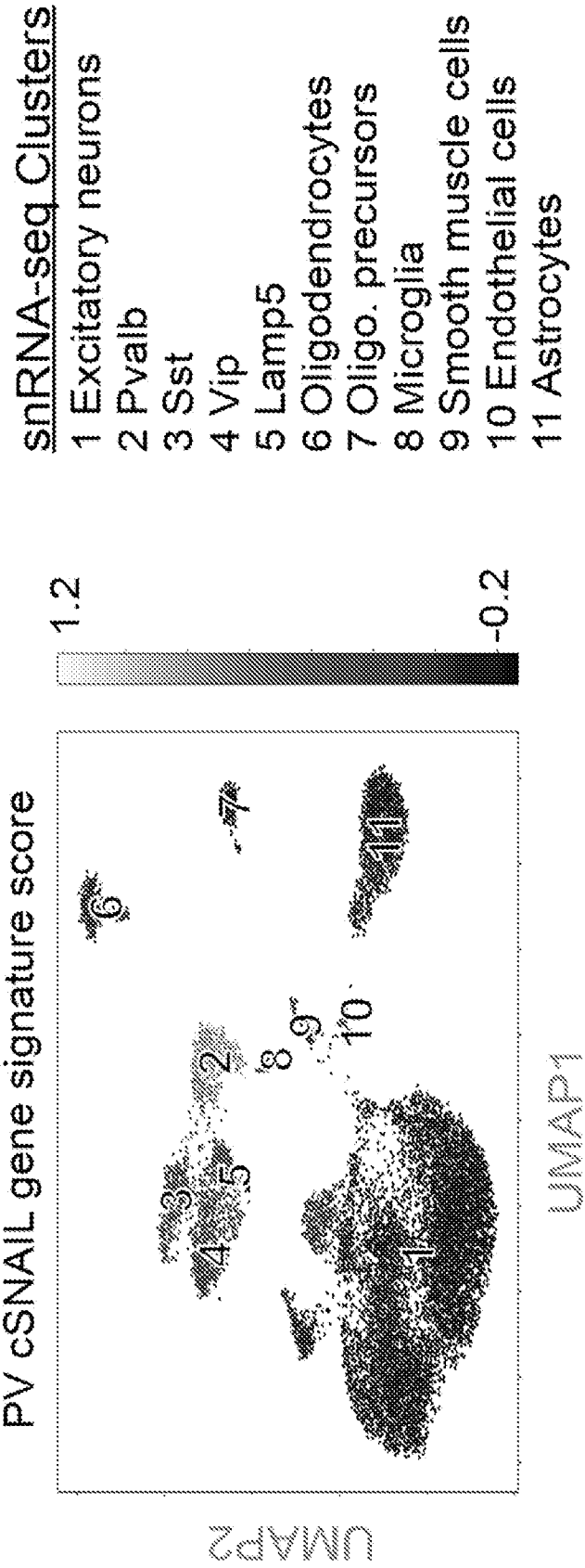


FIG. 7C

FIG. 8A

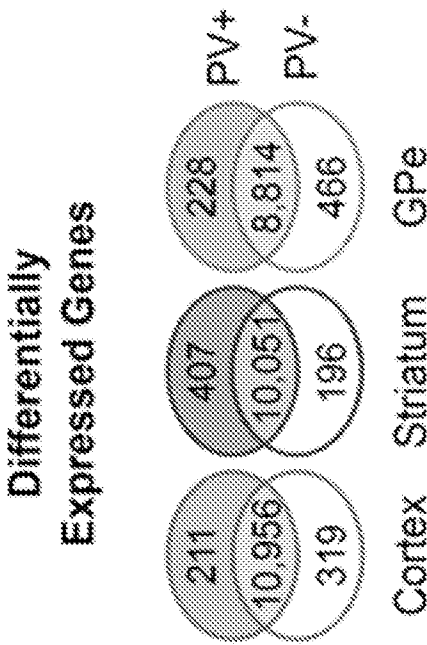


FIG. 8B

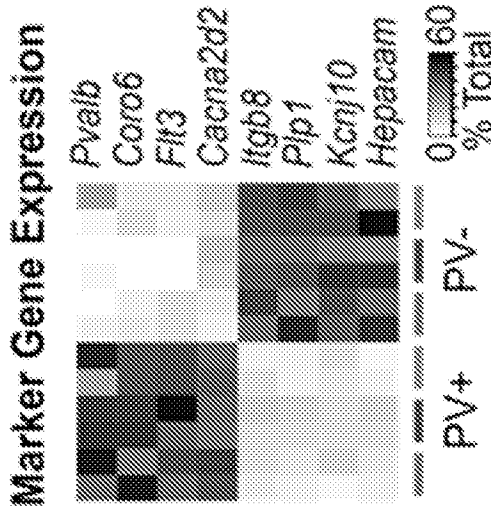


FIG. 8C

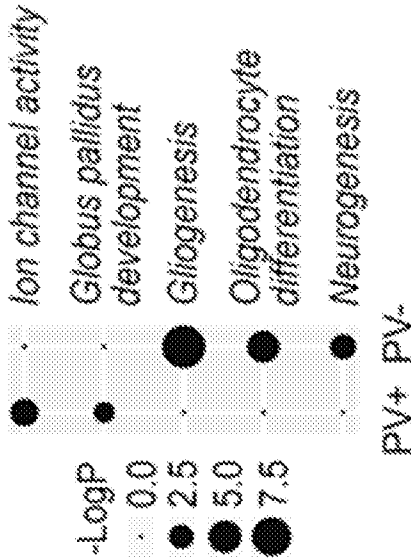


FIG. 8D

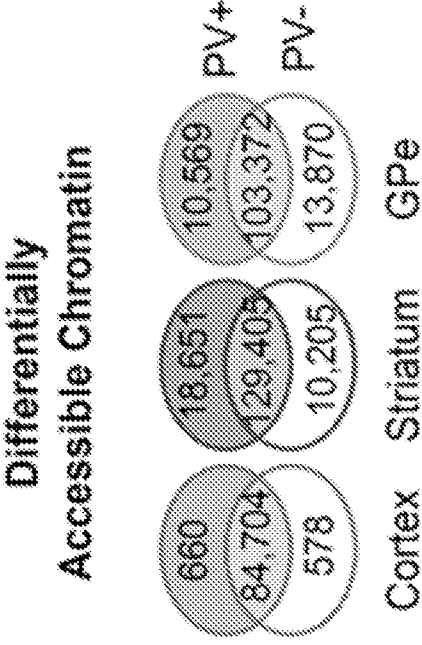


FIG. 8E

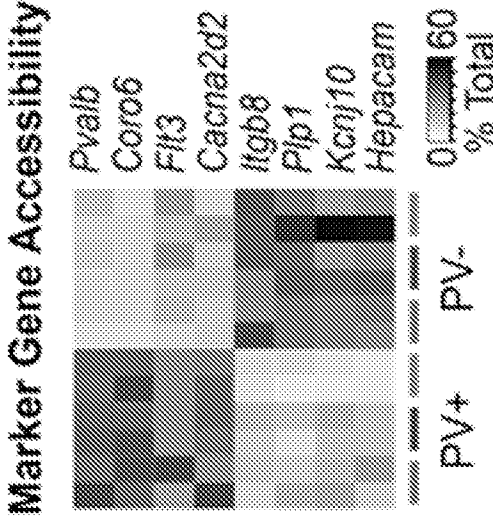
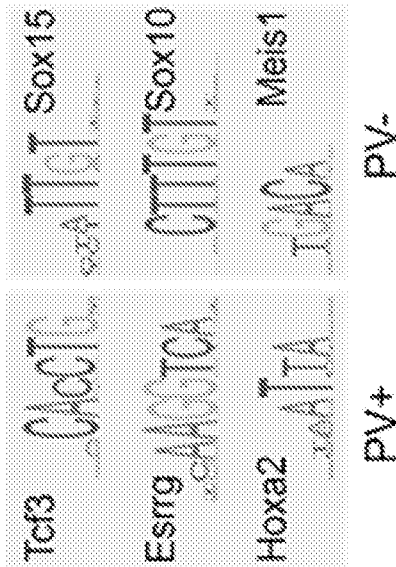


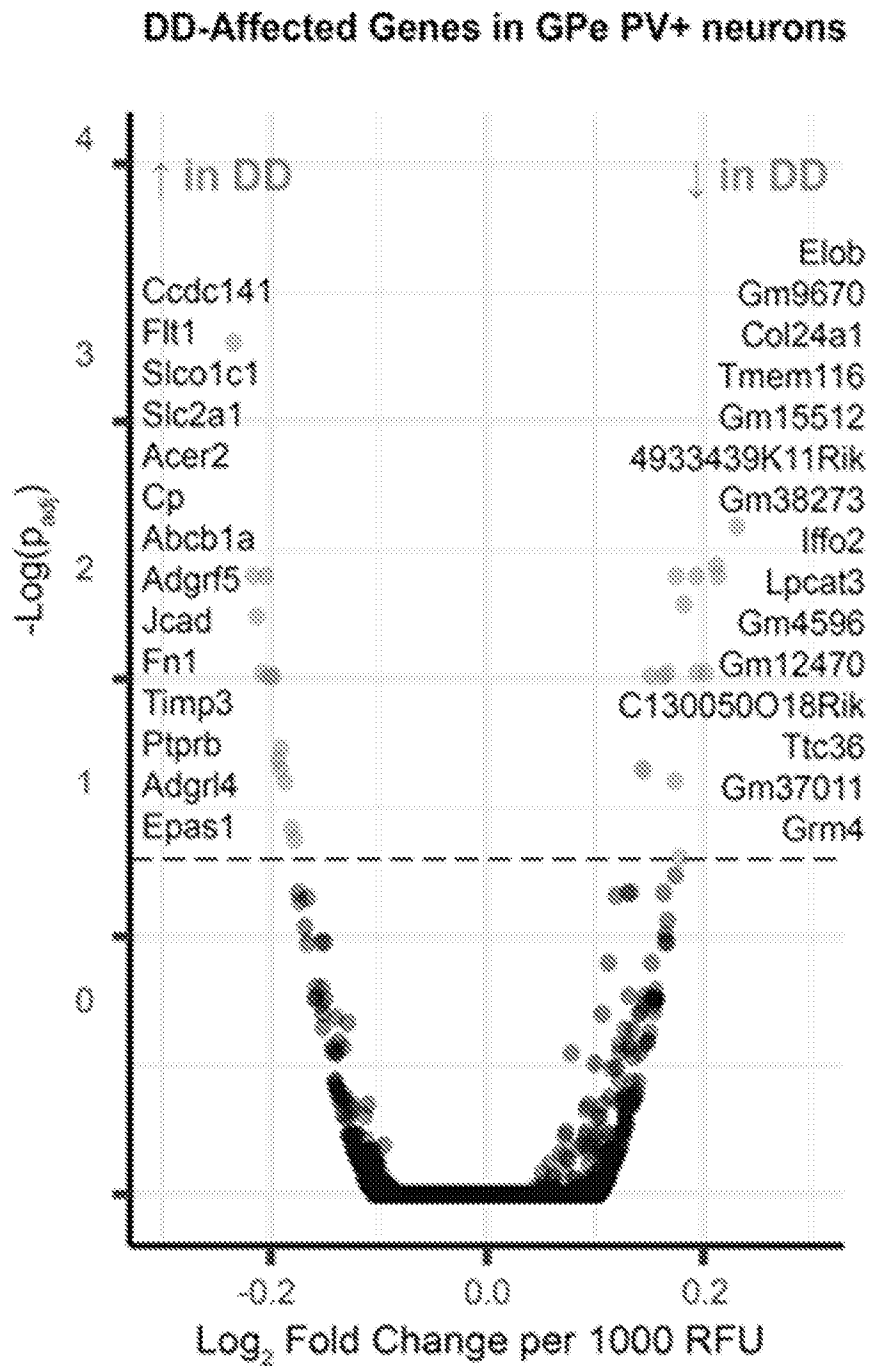
FIG. 8F



RNA-seq	Cortex PV+	Cortex PV-	Striatum PV+	Striatum PV-	GPe PV+	GPe PV-
Genes Tested	26245	26860	25149	24737	17430	26325
↑ in DD	0	0	0	0	14	0
↓ in DD	0	3	0	0	15	3

$p_{adj} < 0.05$

FIG. 9A

**FIG. 9B**

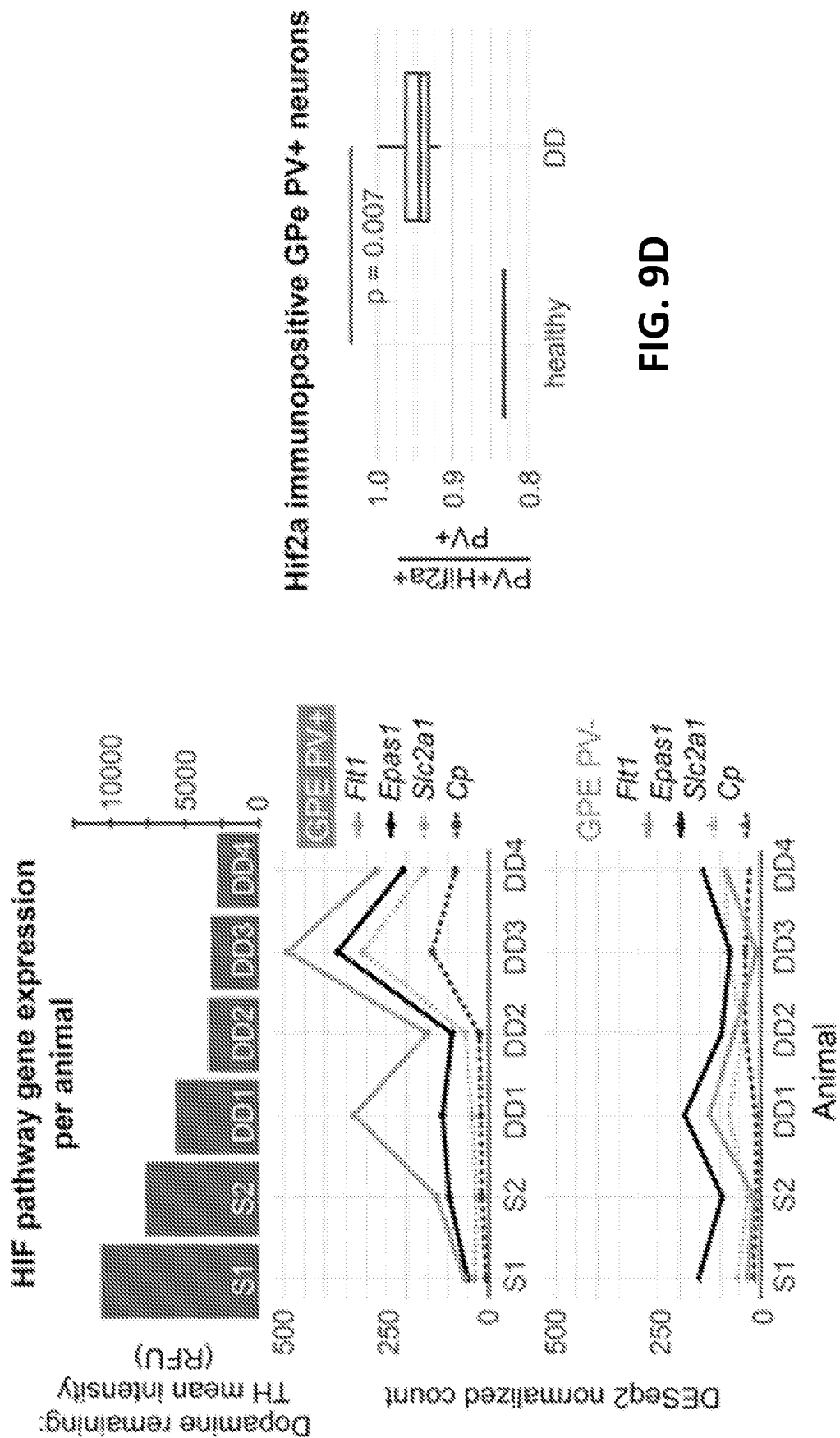


FIG. 9C

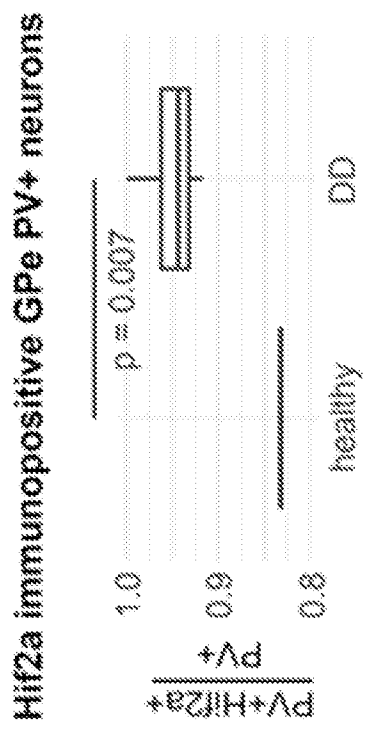


FIG. 9D

ATAC-seq	Cortex PV+	Cortex PV-	Striatum PV+	Striatum PV-	GPe PV+	GPe PV-
Regions Tested	118733	127730	108126	125257	95235	112529
↑ in DD	0	2361	26	3075	16	403
↓ in DD	1	2483	15	3622	18	431

$p_{adj} < 0.05$ & $|LFC| > 0.5$

FIG. 10A

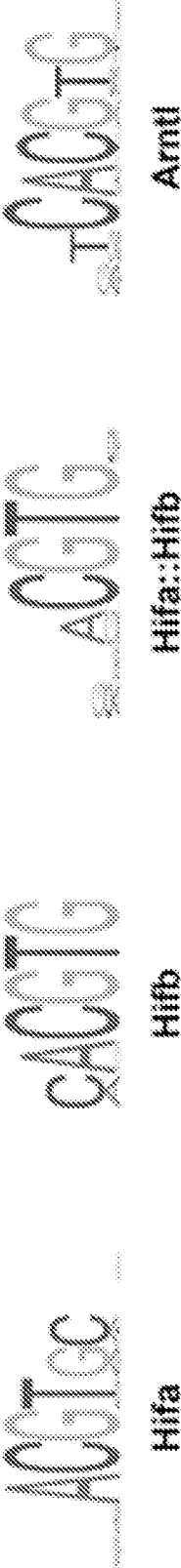


FIG. 10B

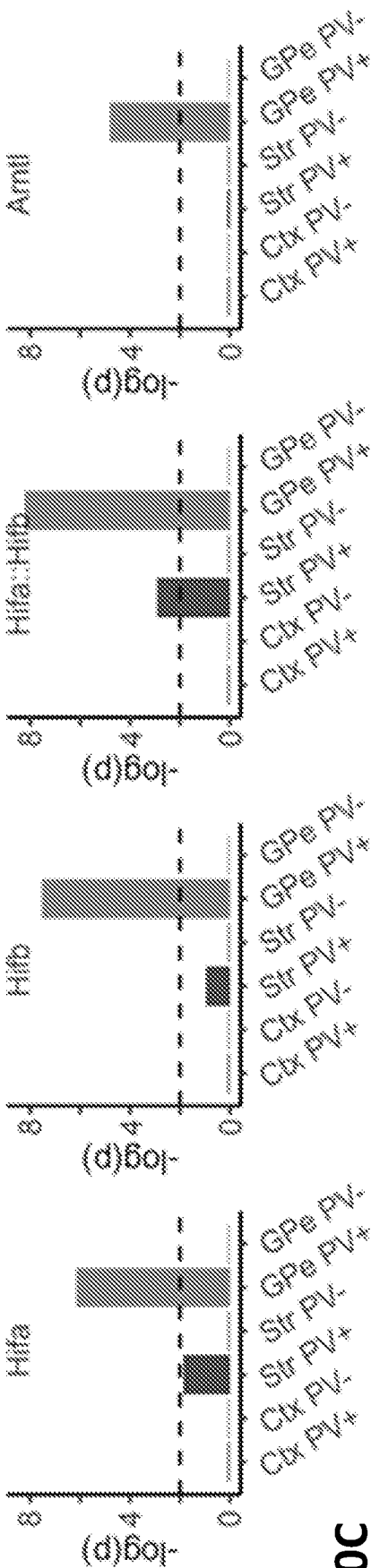


FIG. 10C

Differential Peak Sets: ↑ in DD

FIG. 11A

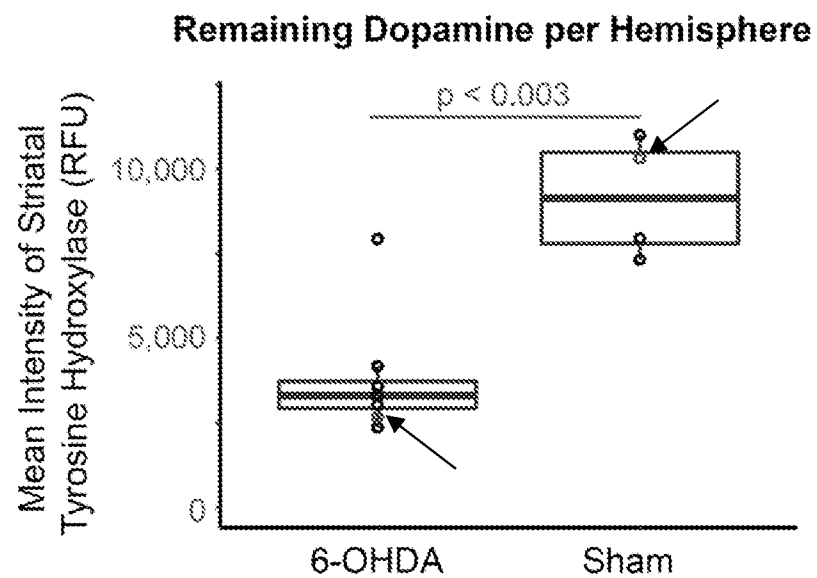


FIG. 11B

Example Anti-Tyrosine Hydroxylase Images

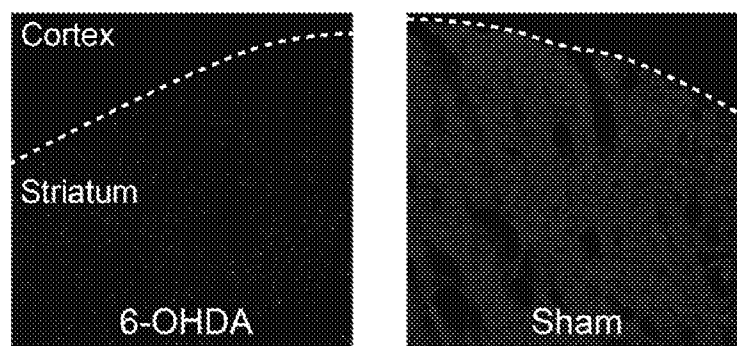


FIG. 12A

PCA for ATAC-seq samples

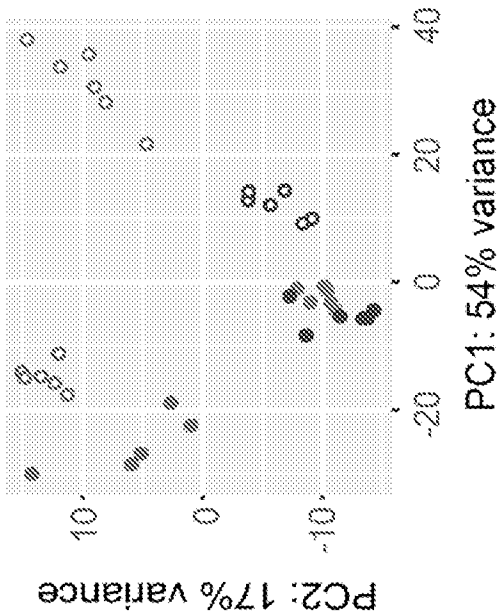
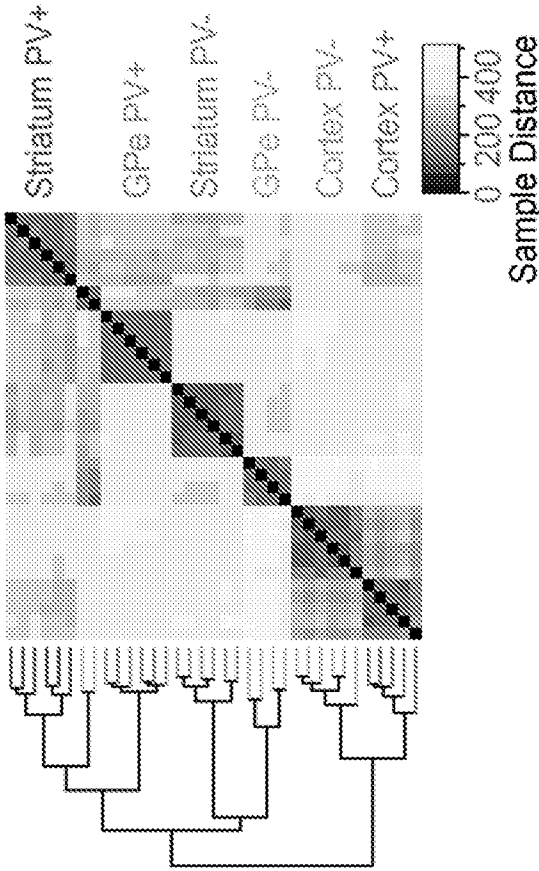
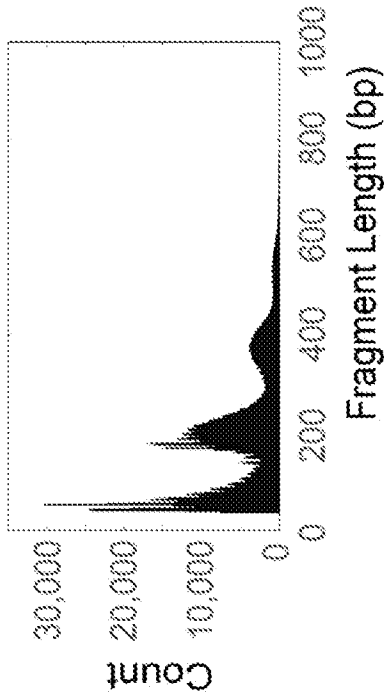


FIG. 12B

Hierarchical clustering of ATAC-seq samples



Fragment Length Distribution



TSS Enrichment of ATAC-seq signal

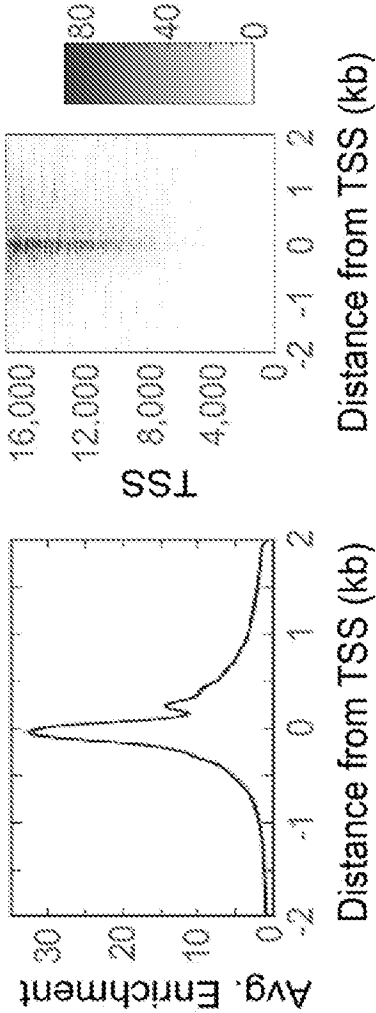


FIG. 12C

FIG. 12D

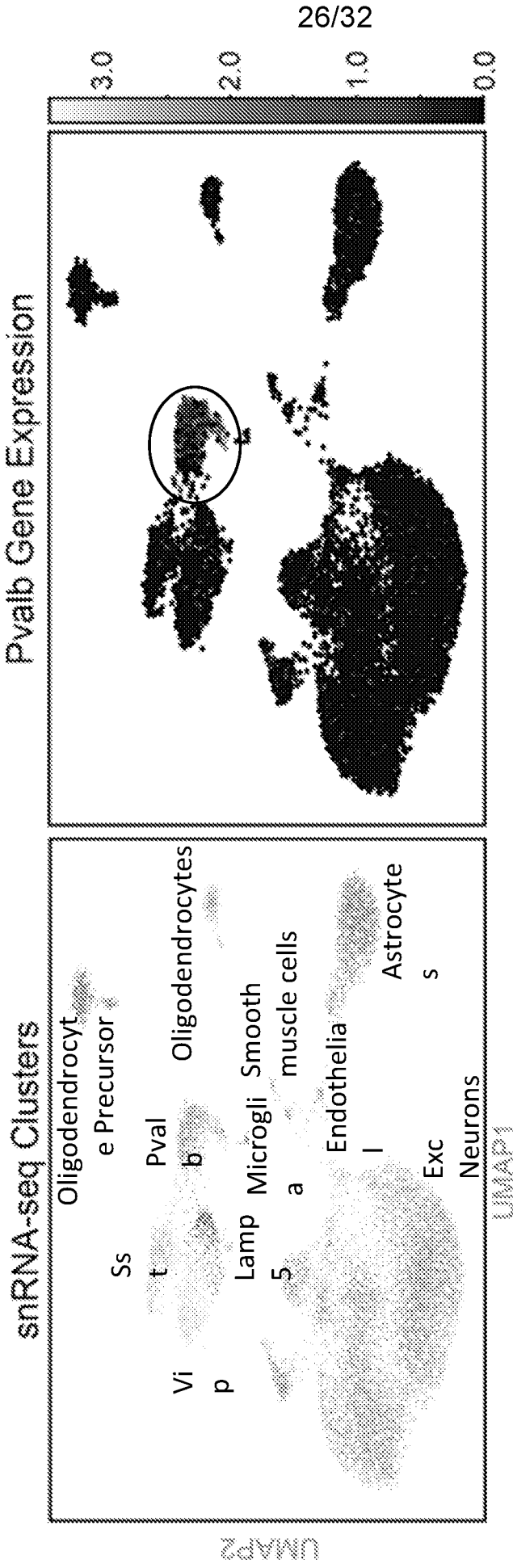


FIG. 13B

FIG. 13A

FIG. 14B

cSNAIL PV- marker gene expression in scRNA-seq

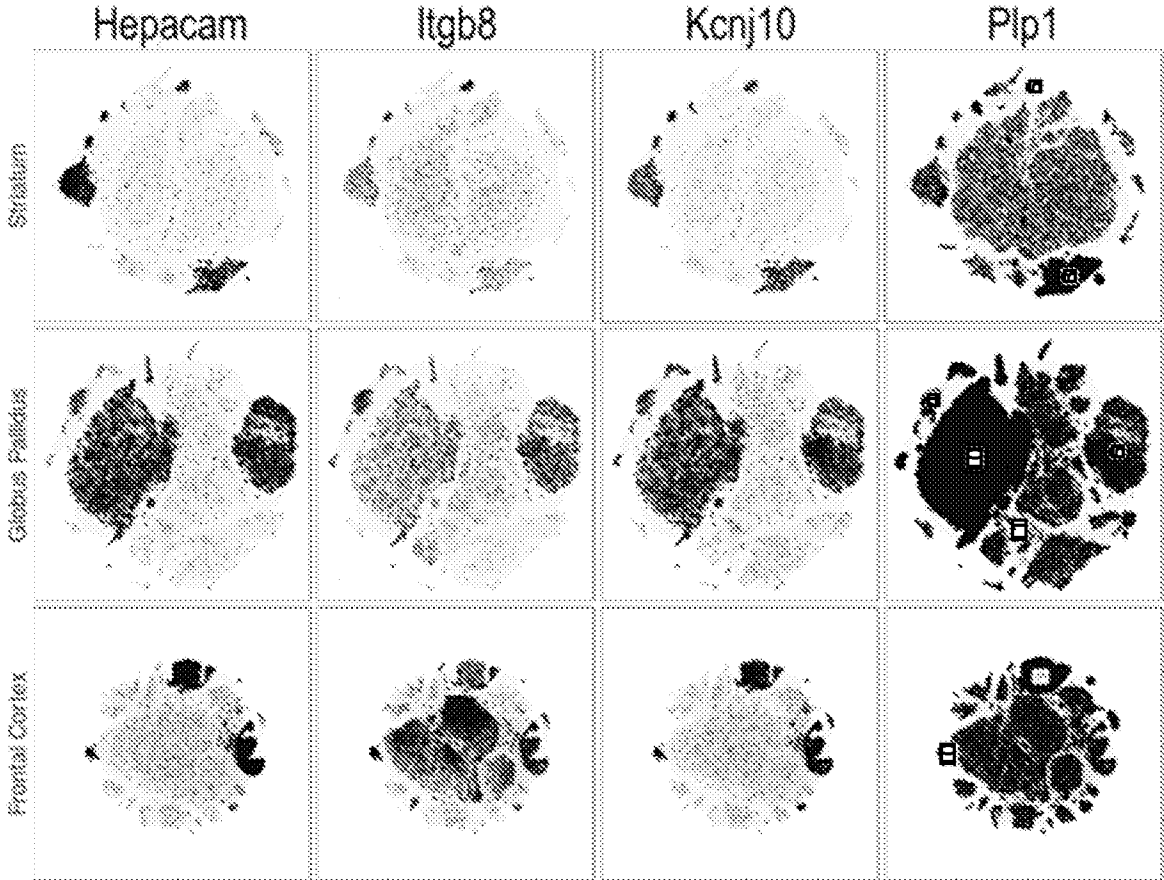
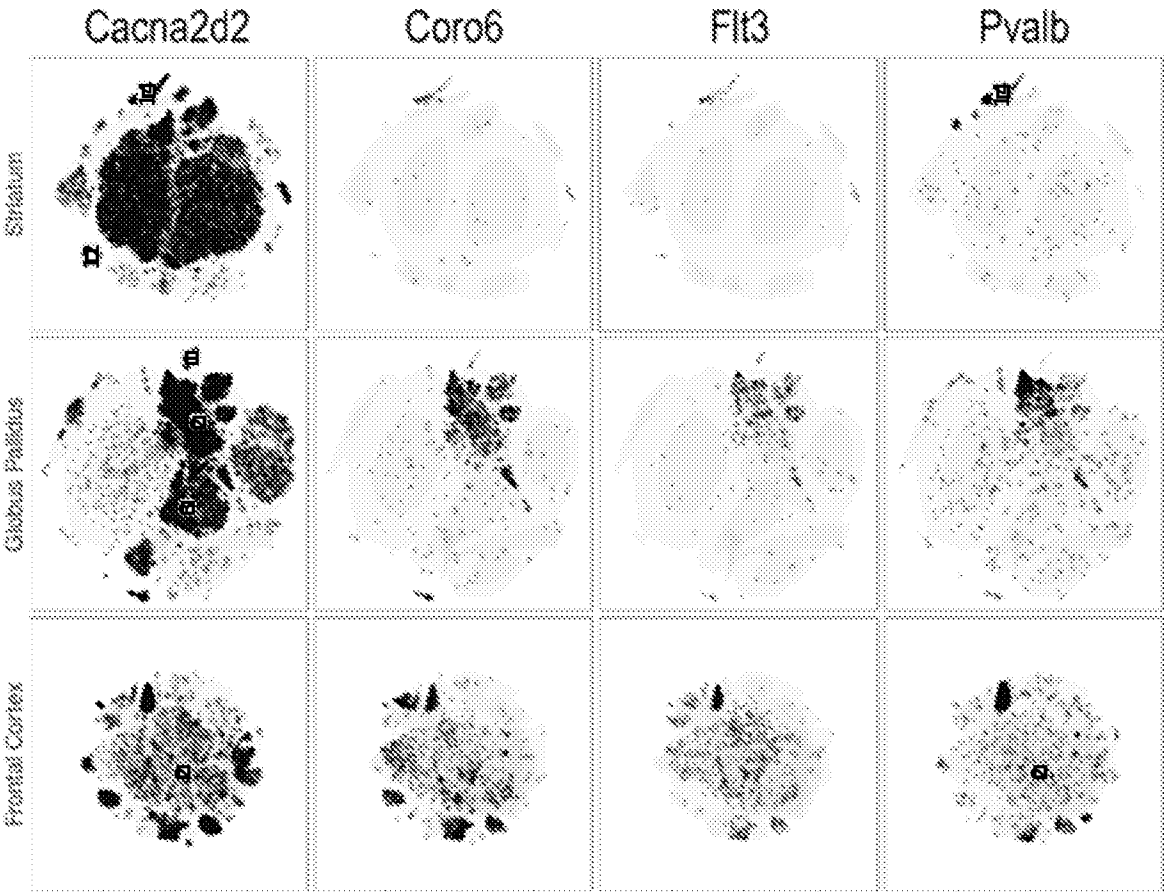


FIG. 14A

cSNAIL PV+ marker gene expression in scRNA-seq



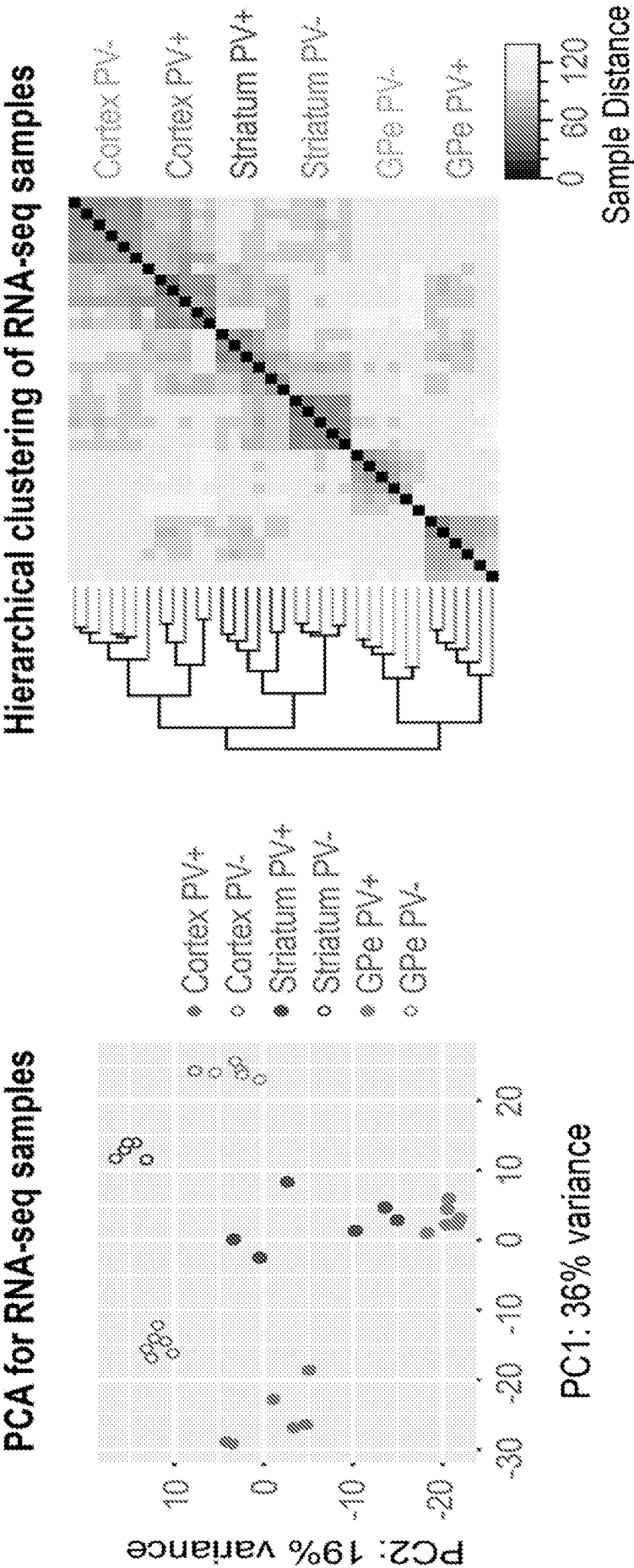


FIG. 15B

FIG. 15A

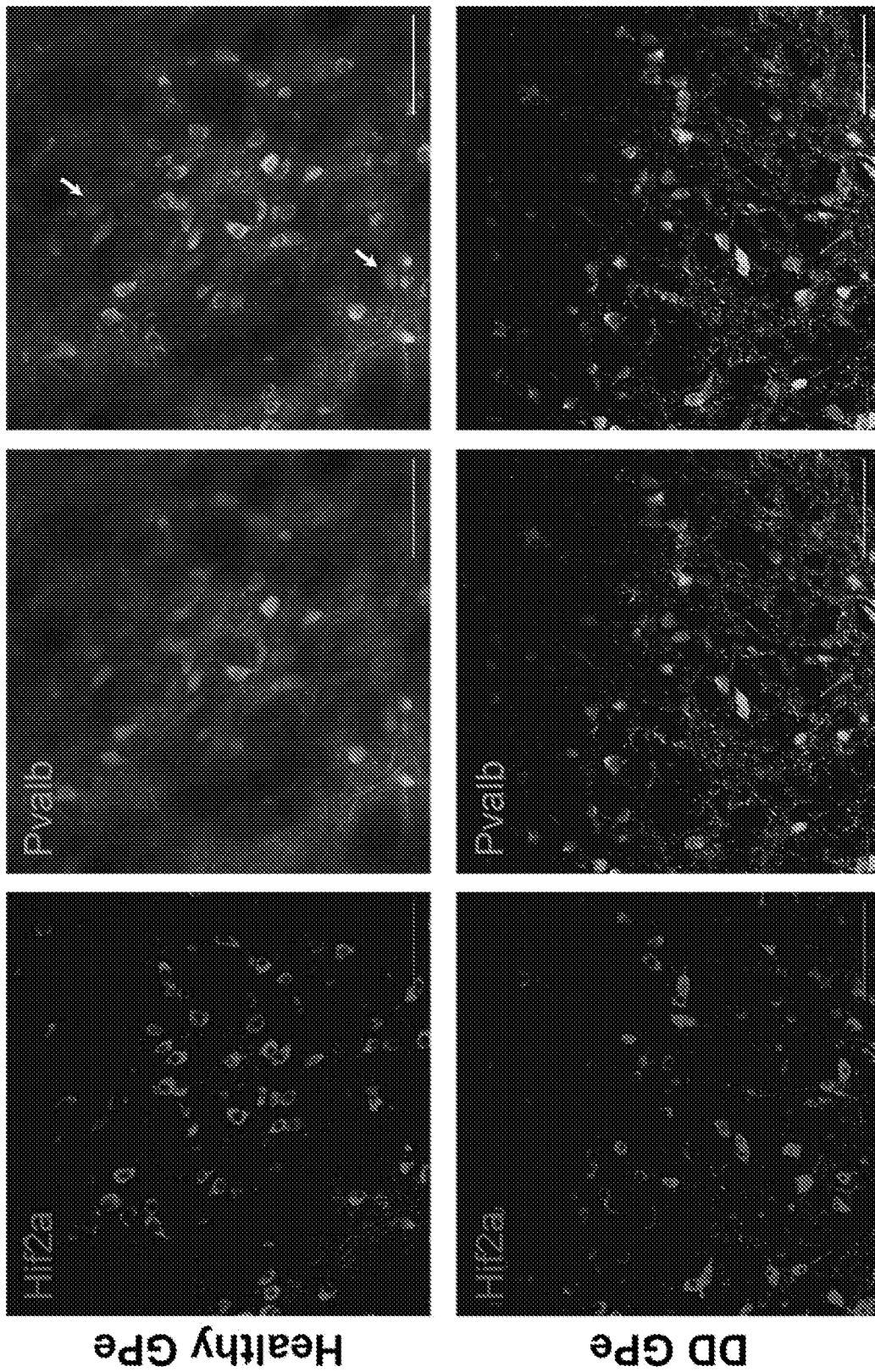


FIG. 16

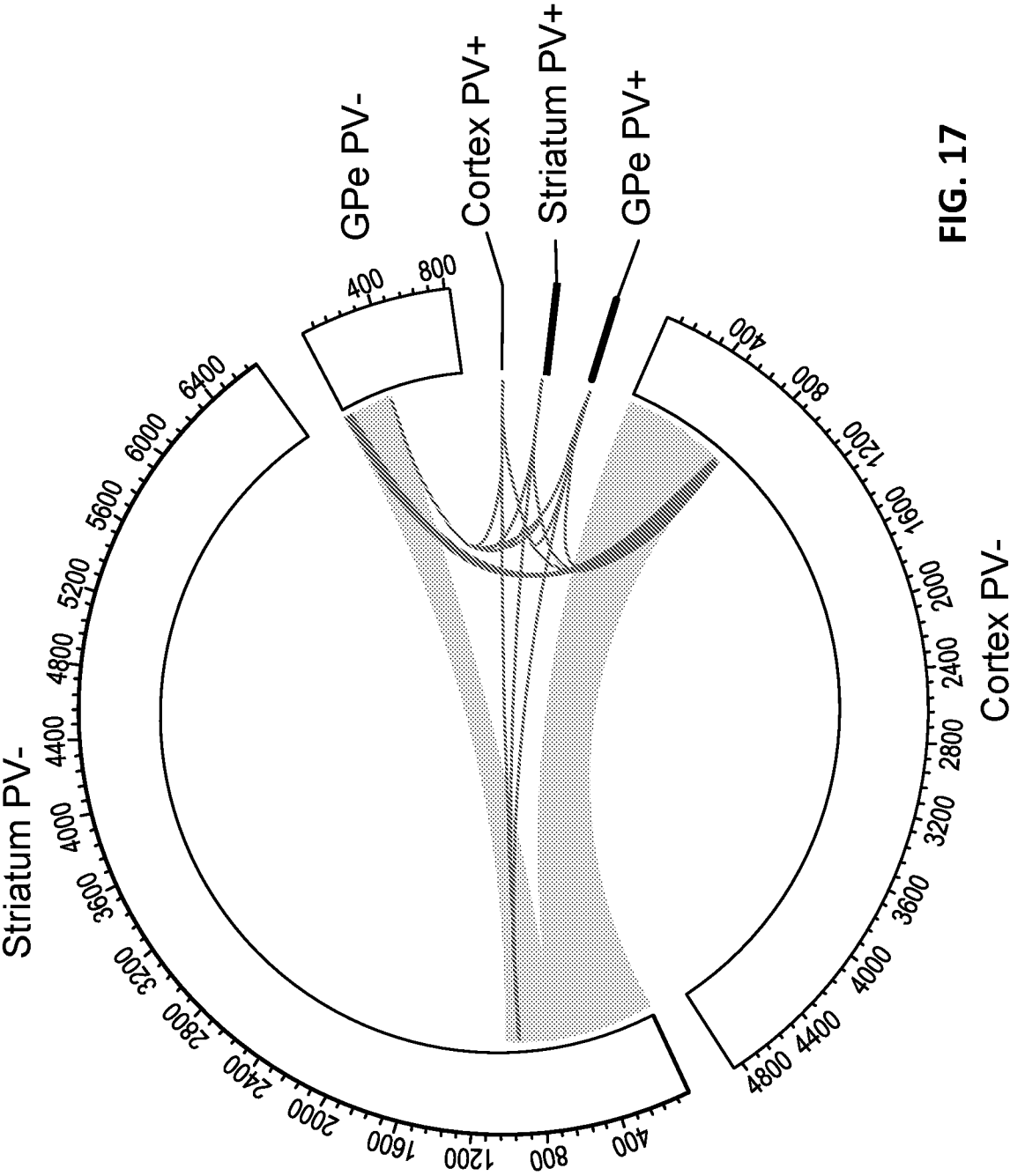


FIG. 17

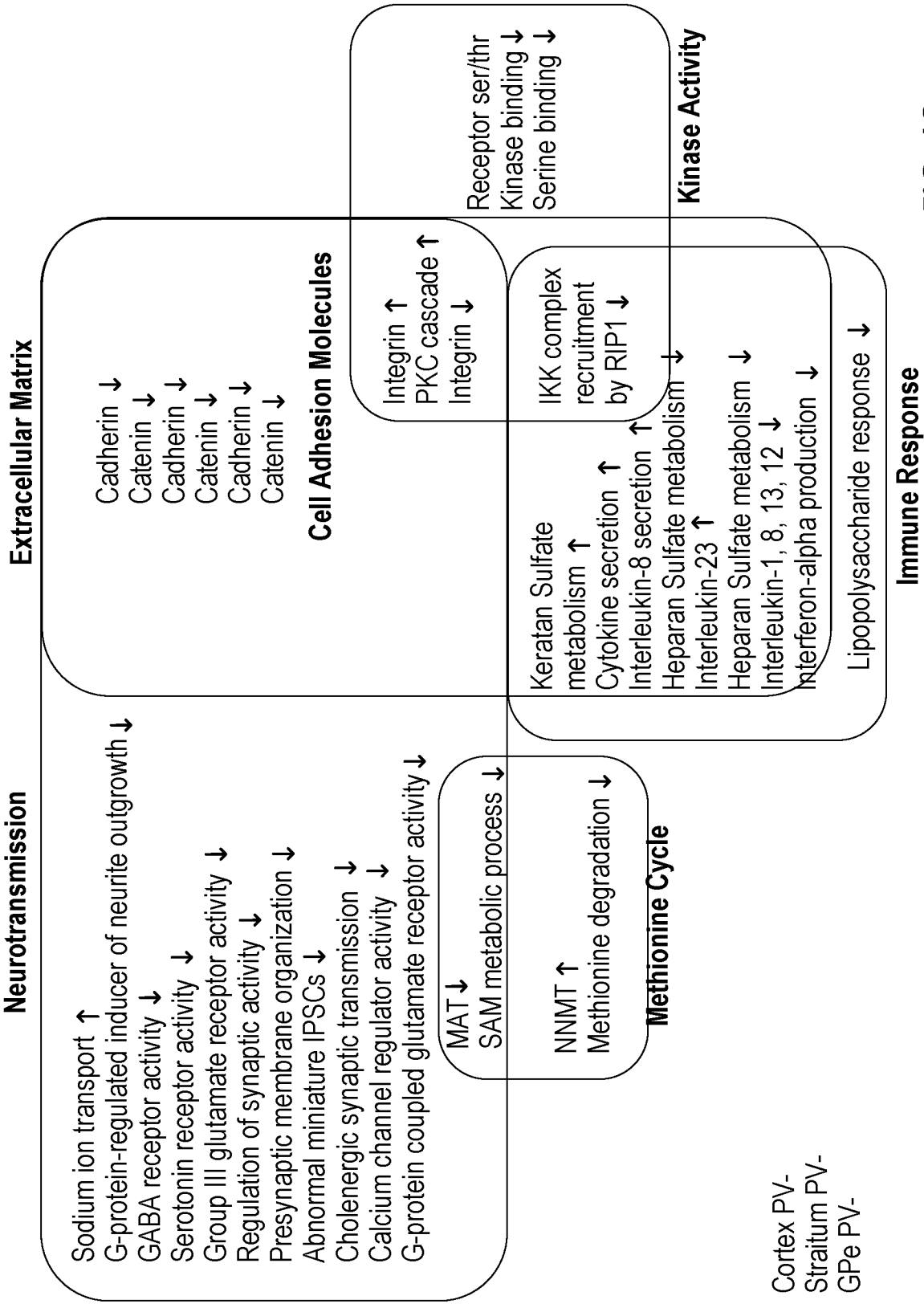
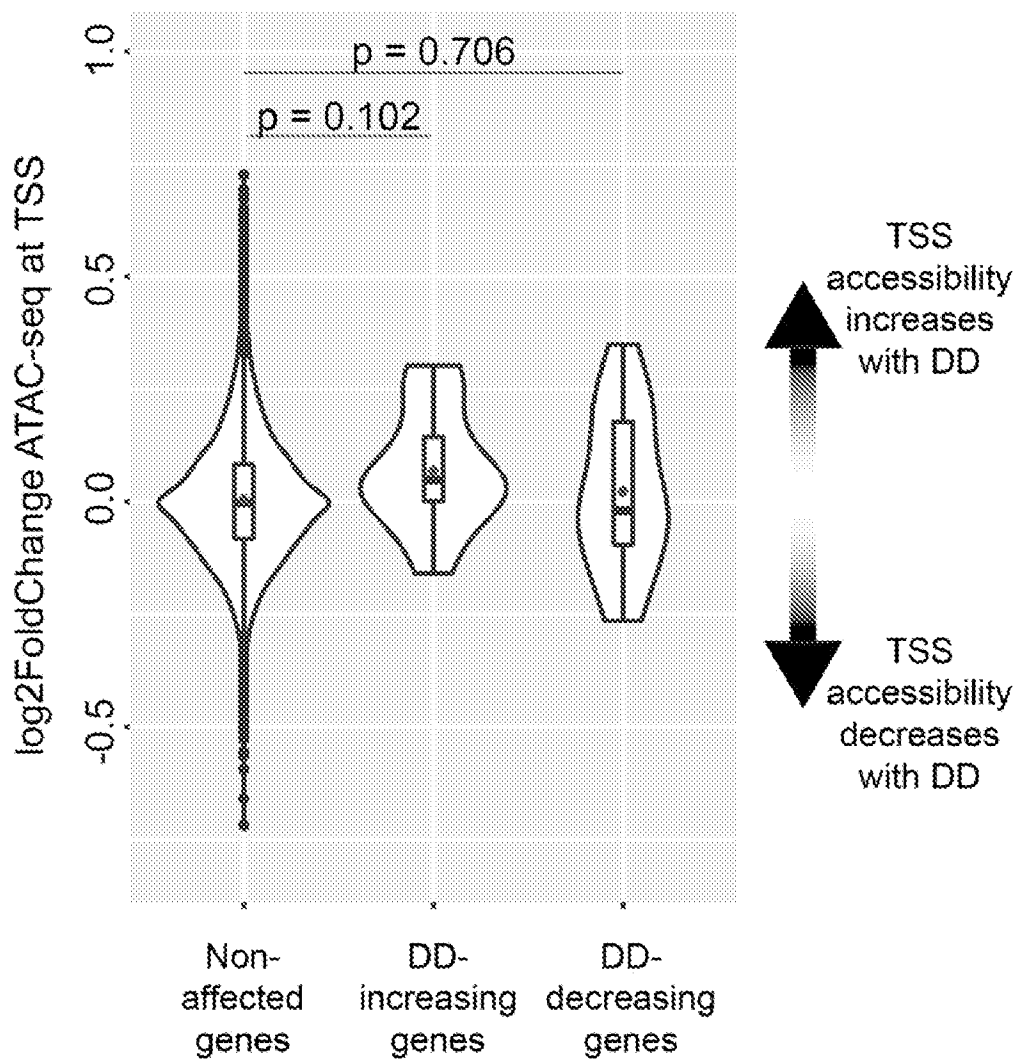


FIG. 18

32/32

**FIG. 19**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/38528

A. CLASSIFICATION OF SUBJECT MATTER

IPC - G16C 20/70, G16B 40/20 (2020.01)

CPC - G16C 20/80, G06N 7/046, G06N 7/005, G16B 30/00, G06N 3/0635

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GJONESKA et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature, 18 February 2015, vol. 518, No. 7539, pp 365-369 and methods pp 1-3; abstract, pg 365, col 2, para 3, pg 366, col 2, para 6, pg 368, col 2, para 1, methods pg 1, col 1, para 6 - col 2, para 1, methods pg 2, col 1, para 1-5, col 2 para 3-4	1, 11, 16 ----- 2-10, 12-15, 17-25
Y	WANG et al. Synthetic Promoter Design in Escherichia coli based on Generative Adversarial Network. bioRxiv 563775, 25 April 2019 [online]. [Retrieved on 18 September 2020]. Retrieved from the internet <URL: https://doi.org/10.1101/563775 >, pg 2, col 1, para 2, pg 2, col 1, para 1, para 2, para 3 - col 2 para 1, pg 3, col 1, para 1, para 2, col 2, para 3, pg 5 col 2, para 2, pg 7, col 2, para 3, Fig. 1	2-10, 17-25
Y	US 2012/0124685 A1 (HENIKOFF et al.) 17 May 2012 (17.05.2012) para [0016], [0079], [0156], [0167], [0195]	7, 12-15, 25
Y	WOODRUFF et al. Networks of Parvalbumin-Positive Interneurons in the Basolateral Amygdala. Journal of Neuroscience 17 January 2007, vol. 27 no. 3, pp 553-563, abstract	7, 25
Y	MENG et al. Construction of precise support vector machine based models for predicting promoter strength. Quantitative Biology, 30 March 2017, vol. 5, pp 90-98, abstract	19
Y	US 2012/0254077 A1 (PURIKL et al.) 04 October 2012 (04.10.2012) abstract, para [0058], [0059]	19

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

18 September 2020

Date of mailing of the international search report

30 SEP 2020

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300