Engineering tools to probe and manipulate the immune system at single-cell resolution

Thesis by

Tatyana Dobreva

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2022

(Defense August 30th, 2021)

Tatyana Dobreva
ORCID: 0000-0002-2625-8873

ACKNOWLEDGEMENTS

My parents for sacrificing their jobs, friends, and family to come to United States to give my brothers and I an opportunity to achieve our full potential.

My old friends who have stuck with me through the years and made time to keep up with me! Especially Chris Smith for reminding how lucky and privileged I am when I fail to see that.

I would like to acknowledge and thank many friends whom I made at Caltech.

There is never enough appreciation for the arts that inspire us and make life worth living. To writers, musicians, video game designers, movie directors, and actors – thank you!

Elisha Mackey and Keith Beadle for teaching me lab work and persevering through my slew of curiosities when I came in with zero experience!

Michael Elowitz, Ellen Rothenberg, Lea Goentoro, and Matt Thomson for always being excited to chat science and get into details, what-ifs, and never discouraging me from pursuing my ideas.
Thank you also, Helen McBride and Matt Thomson, for unwavering support in my pursuit of entrepreneurship!

Viviana Gradinaru for taking me on as a volunteer while I worked as a full-time engineer at JPL so I can learn the shortcomings and opportunities of engineering biological systems. Thank you also for supporting my pursuit of unconventional projects!

My pet turtle, Spunky, who always reminds me to never give up despite there being a lack of free will.

And of course, David Brown, who has been my partner, best friend, and fellow adventurer. Pursuits to one day have synthetic organs and cybernetic brains would be very lonely and challenging without you!

I would like to dedicate my PhD thesis in the memory of my grandparents. They have all been instrumental in nurturing my imagination, playfulness, and teaching me importance of philosophical thinking at an early age. I miss you all very much.

ABSTRACT

My thesis focuses on developing experimental and computational tools to probe and manipulate cellular transcriptomes in the context of human health and disease. Chapters 1 and 2 focus on published work where we leverage single-cell RNA sequencing (scRNA-seq) to understand human immune variability, characterize cell-type specific biases of multiple viral variants within an animal, and assess temporal immune response in the brain to delivery of genetic cargo via an adeno-associated virus (AAV). Chapters 3 and 4 present progress I have made on tools for exporting RNA extracellularly and engineering of a transcription factor for modulating macrophage state.

For probing cellular transcriptome states, we have developed a platform using multiplexed single-cell sequencing and out-of-clinic capillary blood extraction to understand temporal and inter-individual variability of gene expression within immune cell types. Our platform enables simplified, cost-effective profiling of the human immune system across subjects and time at single-cell resolution. To demonstrate the power of our platform, we performed a three day time-of-day study of four healthy individuals, generating gene expression data for 24,087 cells across 22 samples. We detected genes with cell type-specific time-of-day expression and identified robust genes and pathways particular to each individual, all of which could have been missed if analyzed with bulk RNA-sequencing. Also, using scRNA-seq, we have developed a method to screen and characterize cellular tropism of multiple AAV variants. Additionally, I have looked at AAV-mediated transcriptomic changes in animals injected with AAV-PHP.eB three days and twenty-five days post-injection. I have found that there is an upregulation of genes involved in p53 signaling in endothelial cells three days post-injection.

In the context of manipulating cellular transcriptomic states, I demonstrate that a fusion between RNA targeting enzyme, dCas13, and capsid-forming neuronal protein, Arc, is able to form a capsid-like structure capable of encapsulating RNA. I also present methods and preliminary data for tuning macrophage states through mutations in transcription factor EB (TFEB) using scRNA-seq as a readout.

PUBLISHED CONTENT AND CONTRIBUTIONS

(1) Dobreva, T., Brown, D., Park, J. H., & Thomson, M. (2020). Single cell profiling of capillary blood enables out of clinic human immunity studies. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-77073-3

T.D. conceived the project, designed the experiments, performed the experiments, performed the analysis, and wrote the manuscript.

(2) Brown, D., Altermatt, M., **Dobreva, T**., Chen, S., Wang, A., Thomson, M., & Gradinaru, V. (2021). Deep parallel characterization of AAV tropism and AAV-mediated transcriptional changes via single-cell RNA sequencing. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2021.06.25.449955

T.D. conceived the project, designed the experiments, performed the experiments, performed the analysis, and wrote the manuscript which aligned with her contributions.

(3) Ravindra Kumar, S., Miles, T. F., Chen, X., Brown, D., **Dobreva, T**., Huang, Q., Ding, X., Luo, Y., Einarsson, P. H., Greenbaum, A., Jang, M. J., Deverman, B. E., & Gradinaru, V. (2020). Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. Nature Methods, 17(5), 541–550.

https://doi.org/10.1038/s41592-020-0799-7

T.D. contributed to data analysis and writing of the manuscript

(4) Seo, J. W., Ingham, E. S., Mahakian, L., Tumbale, S., Wu, B., Aghevlian, S., Shams, S., Baikoghli, M., Jain, P., Ding, X., Goeden, N., **Dobreva, T**., Flytzanis, N. C., Chavez, M., Singhal, K., Leib, R., James, M. L., Segal, D. J., Cheng, R. H., … Ferrara, K. W. (2020). Positron emission tomography imaging of novel AAV capsids maps rapid brain accumulation. Nature Communications, 11(1). https://doi.org/10.1038/s41467-020-15818-4 T.D. contributed to experiments and writing of the manuscript

TABLE OF CONTENTS

Chapter 1

SINGLE CELL PROFILING OF CAPILLARY BLOOD ENABLES OUT-OF-CLINIC HUMAN IMMUNITY STUDIES

Adopted from (1)

1.1 Introduction

Increasing evidence implicates the immune system in an overwhelming number of diseases, and distinct cell types play specific roles in their pathogenesis.[1,2] Studies of peripheral blood have uncovered a wealth of associations between gene expression, environmental factors, disease risk, and therapeutic efficacy.[3,4,49] For example, in rheumatoid arthritis, multiple mechanistic paths have been found that lead to disease, and gene expression of specific immune cell types can be used as a predictor of therapeutic non-response.[4] Furthermore, vaccines, drugs, and chemotherapy have been shown to yield different efficacy based on time of administration, and such findings have been linked to the time-dependence of gene expression in downstream pathways.[5,6,7] However, human immune studies of gene expression between individuals and across time remain limited to a few cell types or time points per subject, constraining our understanding of how networks of heterogeneous cells making up each individual's immune system respond to adverse events and change over time.

The advent of single-cell RNA sequencing (scRNA-seq) has enabled the interrogation of heterogeneous cell populations in blood without cell type isolation and has already been employed in the study of myriad immune-related diseases.[2,8,9,48] Recent studies employing scRNA-seq to study the role of immune cell subpopulations between healthy and ill patients, such as those for Crohn's disease [10], Tuberculosis [11], and COVID-19 [12], have identified cell type-specific disease relevant signatures in peripheral blood immune cells; however, these types of studies have been limited to large volume venous blood draws which can tax already ill patients, reduce the scope of studies to populations amenable to blood draws, and often

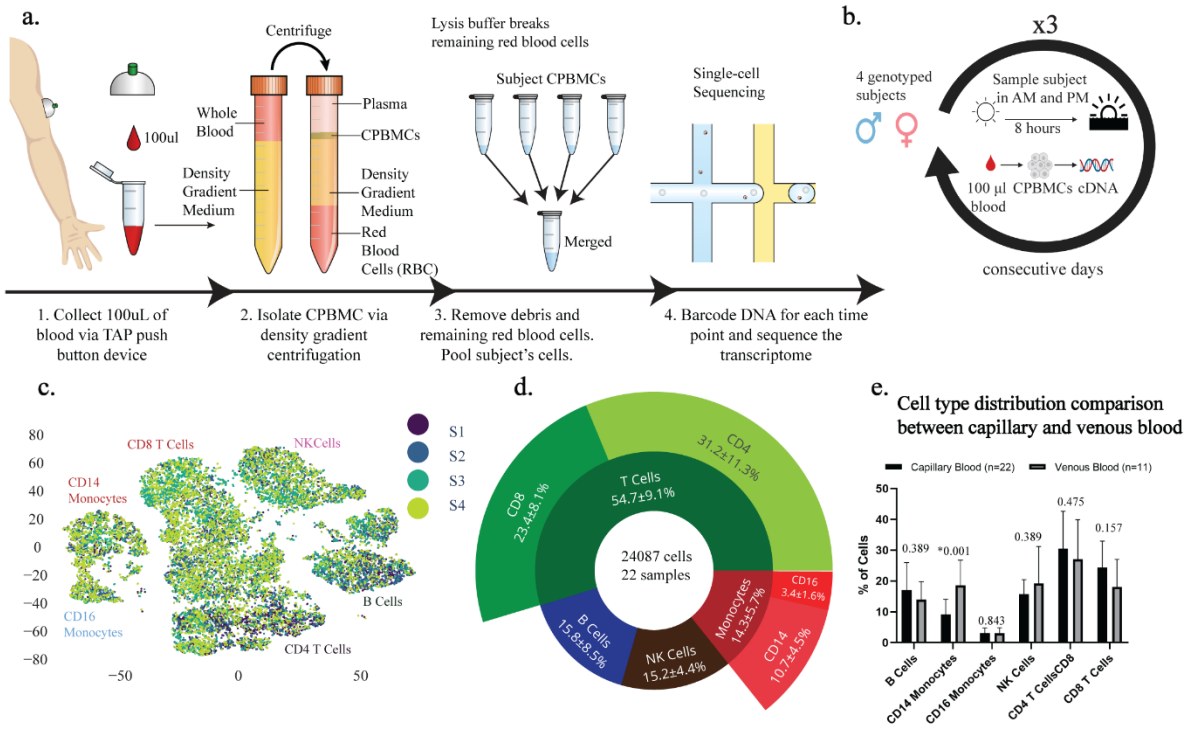


Fig.1 | Experimental workflow and consistency of capillary blood sampling (a) Experimental workflow for capillary blood immune profiling 1. Blood is collected using the TAP device from the deltoid. 2. Capillary peripheral blood mononuclear cells (CPBMCs) are separated via centrifugation. 3. Red blood cells are lysed and removed, and samples from different subjects are pooled together. 4. Cell transcriptomes are sequenced using single-cell sequencing. **(b)** Time-course study design. CPBMCs are collected and profiled from 4 subjects (2 male, 2 female) each morning (AM) and afternoon (PM) for 3 consecutive days. **(c)** 2-dimensional t-SNE projection of the transcriptomes of all cells in all samples. Cells appear to cluster by major cell type (Fig. S6) **(d)** Immune cell type percentages across all samples shows stable cell type abundances (includes cells without subject labels). **(e)** Cell type ratios between capillary blood from this study, and venous blood from 3 other studies were the same, with the exception of $CD14^{+}$ Monocytes, which are more abundant in venous blood (FDR < 0.05, 2-sided student t-test, multiple comparison corrected) The q-values are displayed for each cell type comparison.

require larger research teams to handle the patient logistics and sample processing costs and labor. In particular, getting repeated venous blood draws within a single day and/or multiple days at the subject's home has been a challenge for older people with frail skin and those on low dosage Acetylsalicylic acid.[13] This dependence on venous blood dramatically impacts our ability to understand the high temporal dynamics of health and disease.

Capillary blood sampling is being increasingly used in point-of-care testing and has been advised for obese, elderly, and other patients with fragile or inaccessible veins.[14,15,16,17] The reduction of patient burden via capillary blood sampling could enable researchers to perform

studies on otherwise difficult or inaccessible populations, and at greater temporal resolution. Additionally, capillary blood is being shown to be comparable to traditional venous blood draws for a variety of applications. For example, Catala, et al (2018) have shown that 39 out of 45 clinically relevant metabolites had overlapping ranges between capillary blood vs traditional venous blood draws[18], and Toma et al. have shown strong correlation (Spearman correlation coefficient >= 0.95) between bulk RNA sequencing data between capillary and venous blood from the same donor.[19] However, to date, scRNA-seq of human capillary blood has not yet been validated nor applied to study the immune system. In order to make small volumes of capillary blood (100 ul) amenable to scRNA-seq we have developed a platform which consists of a painless vacuum-based blood collection device, sample de-multiplexing leveraging commercial genotype data, and an analysis pipeline used to identify time-of-day and subject specific genes. The potential of our platform is rooted in enabling large scale studies of immune state variation in health and disease across people. The high-dimensional temporal transcriptome data could be paired with computational approaches to predict and understand emergence of pathological immune states. Most importantly, our platform makes collection and profiling of human immune cells less invasive, less expensive, and as such more scalable than traditional methods rooted in large venous blood draws.

1.2 Platform for low-cost interrogation of single-cell immune gene expression profiles

Our platform is composed of a protocol for isolating capillary peripheral blood mononuclear cells (CPBMCs) using a touch activated phlebotomy device (TAP)[14], pooling samples to reduce per-sample cost using genome-based demultiplexing[20], and a computational package that leverages repeated sampling to identify genes that are differentially expressed in individuals or between time points, within subpopulations of cells (Fig. 1a). Using a painless vacuum-based blood collection device such as the commercial FDA-approved TAP to collect capillary blood makes it convenient to perform at-home self-collected sampling and removes the need for a trained phlebotomist, increasing the ease of acquiring more samples. The isolation of CPBMCs is done using gradient centrifugation and red blood cells are further removed via a red blood cell lysis buffer. The cells from the different subjects are pooled,

sequenced via scRNA-seq using a single reagent kit, and demultiplexed [20] via each subject's single-nucleotide polymorphisms (SNPs), reducing the per-sample processing cost. By pooling the data across all 6 time points, and using a genotype-free demultiplexing software (popscle), we were able to identify which cells belonged to which subject across time points, removing the need for a separate genotyping assay to link subjects together across batches.

1.3 Single-cell RNA sequencing (scRNA-seq) of low volume capillary blood recovers distinct immune cell populations stably across time

As a proof-of-concept, we leveraged our scRNA-seq of capillary blood platform to identify genes that exhibit diurnal behavior in subpopulations of cells and find subject-specific immune relevant gene signatures. We performed a three-day study in which we processed capillary blood from four subjects in the morning and afternoon, totaling 24,087 cells across 22 samples (Fig. 1b). Major immune cell types such as T cells ($CD4^+$, $CD8^+$), Natural Killer cells, Monocytes ($CD14^+$, $CD16^+$), and B cells are present in all subjects and time points with stable expression of key marker genes (Fig. 1d, Fig. S1), demonstrating that these signals are robust to technical and biological variability of CPBMC sampling (Fig. 1c). In order to compare cell type distributions derived from our method with venous blood draws, we used data from 11 healthy subjects provided by three independent studies[21,22,23] (Table S4). $CD14^+$ Monocytes make up a higher percentage of PBMCs in venous blood (n=11) versus capillary blood (n=22) (FDR < 0.05, 2-sided student t-test, multiple comparison corrected), while other cell types do not have a significant difference in distributions (Fig. 1e).

1.4 High frequency scRNA-seq unveils new diurnal cell type-specific genes

Genes driven by time-of-day expression, such as those involved in leukocyte recruitment[24] and regulation of oxidative stress[25], have been determined to play an important role in both innate and adaptive immune cells[26]. Medical conditions such as atherosclerosis, parasite infection, sepsis, and allergies display distinct time-of-day immune responses in leukocytes[27], suggesting the presence of diurnally expressing genes that could be candidates for optimizing therapeutic efficacy via time-of-day dependent administration. However,

studies examining diurnal gene expression in human blood have been limited to whole blood gene panels via qPCR, or bulk RNA-seq.[28,29,30]

Leveraging our platform, which enables single-cell studies of temporal human immune gene expression, we detected 395 genes (FDR < 0.05, multiple comparison corrected) exhibiting diurnal activity within at least one cell subpopulation (Fig. 2a). Among the 20 top diurnally classified genes, we found that 40% of those genes were previously correlated with circadian behavior (Table S1), such as DDIT4[28] (Fig. 2b), SMAP2[31], and PCPB1[32]. However, only

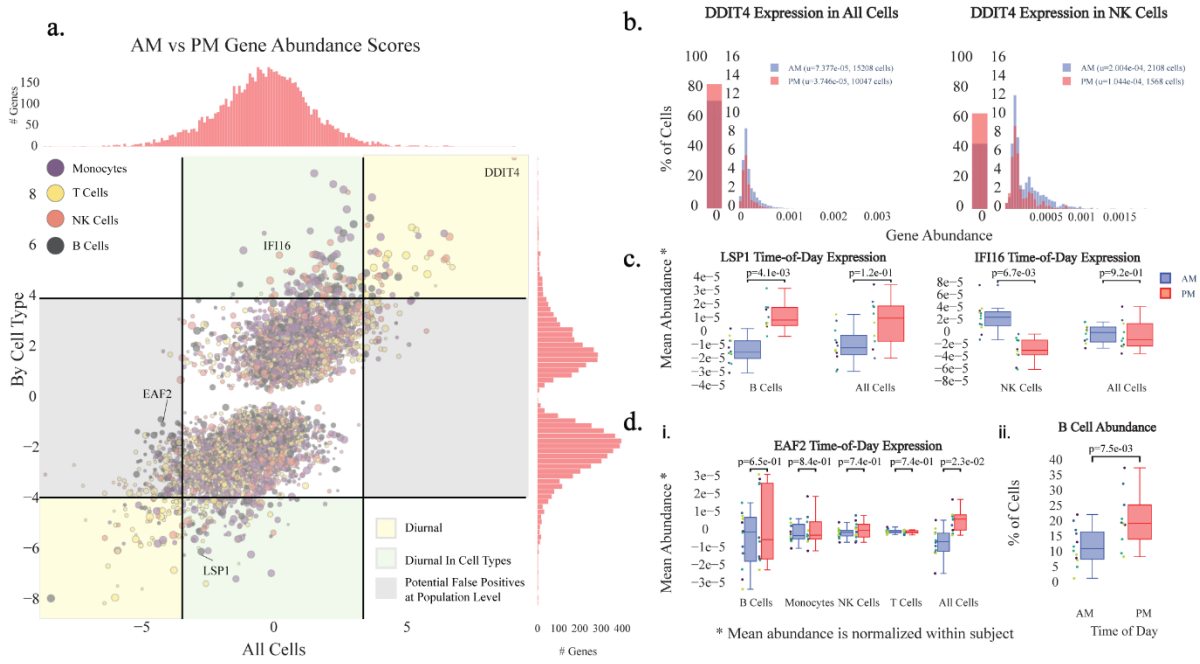


Fig. 2 | Diurnal variability in subpopulations of capillary blood (a) Magnitude (Z-score) of the difference in AM vs PM gene expression across the whole population of cells (x) vs the cell type with the largest magnitude Z-score (y). Points above or below the significance lines (FDR < 0.05, multiple comparison correction) display different degrees of diurnality. The size of each marker indicates the abundance of the gene (the largest percent of cells in a subpopulation that express this gene). **(b)** Distribution of expression of DDIT4, a previously identified circadian rhythm gene[9], shows diurnal signal across all cells, as well as individual cell types, such as natural killer (NK) cells. u indicates the mean fraction of transcripts per cell (gene abundance). **(c)** Example of newly identified diurnal genes, LSP1 and IFI16 that could be missed if analyzed at the population level **(d)** Example of a gene, EAF2, that could be falsely classified as diurnal **(i)** without considering cell type subpopulations due to a diurnal B cell abundance shift **(ii).**

119/395 (30.1%) of these genes are detected as diurnal at the whole population level (FDR < 0.05, multiple comparison corrected), suggesting there may be many more diurnally-varying genes than previously discovered. For example, IFI16 and LSP1 (Fig. 2c) have diurnal expression only in NK cells and B cells, respectively, and display previously unreported transcriptional diurnal patterns. In particular, LSP1 has been implicated in numerous leukemias and lymphomas of B cell origin.[33] Given previous evidence of increased efficacy of time-dependent chemotherapy administration[7,34] and tumor cells exhibiting out-of-sync behavior compared to normal cells[35], understanding LSP1's diurnal expression pattern can potentially guide timely administration of candidate therapeutics. Out of the identified 395 diurnally-varying genes, 114 (29%) are considered druggable under the drug gene interaction database (http://www.dgidb.org/).

1.5 scRNA-seq profiling distinguishes diurnal gene expression from cell type abundance changes

We also detected 406 genes (FDR < 0.05, multiple comparison corrected) exhibiting diurnal behavior when analyzed at the population level, such as EAF2, that do not display diurnal variation in any of our major cell types (Fig. 2d.i). Such false positives may come from diurnal shifts in cell type abundance rather than up- or down-regulation of genes. In the case of EAF2, which is most abundant in B cells, we hypothesized that the diurnality detected at the population level was a result of an increase of B cell abundance in the afternoon, and verified this in our data ($p=7.5x10^{-3}$, one-sided student-t test) (Fig. 2d.ii). This finding highlights the importance of looking at expression within multiple cell types to avoid potentially misleading mechanistic hypotheses.

1.6 Individuals exhibit robust cell type-specific differences in genes and pathways relevant to immune function

Gene expression studies of isolated cell subpopulations across large cohorts of people have revealed a high degree of variability between individuals that cannot be accounted for by genetics alone, with environmental effects that vary over time likely playing a critical

role.[36,37] Furthermore, these transcriptomic differences have been linked to a wide range of therapeutic responses, such as drug-induced cardiotoxicity.[38] However, while immune system composition and expression have been shown to be stable over long time periods

within an individual, acute immune responses generate dramatic immune system changes, meaning that large single time point population studies are unable to establish whether variability between individuals is stable or the result of dynamic response to stimuli.[39]

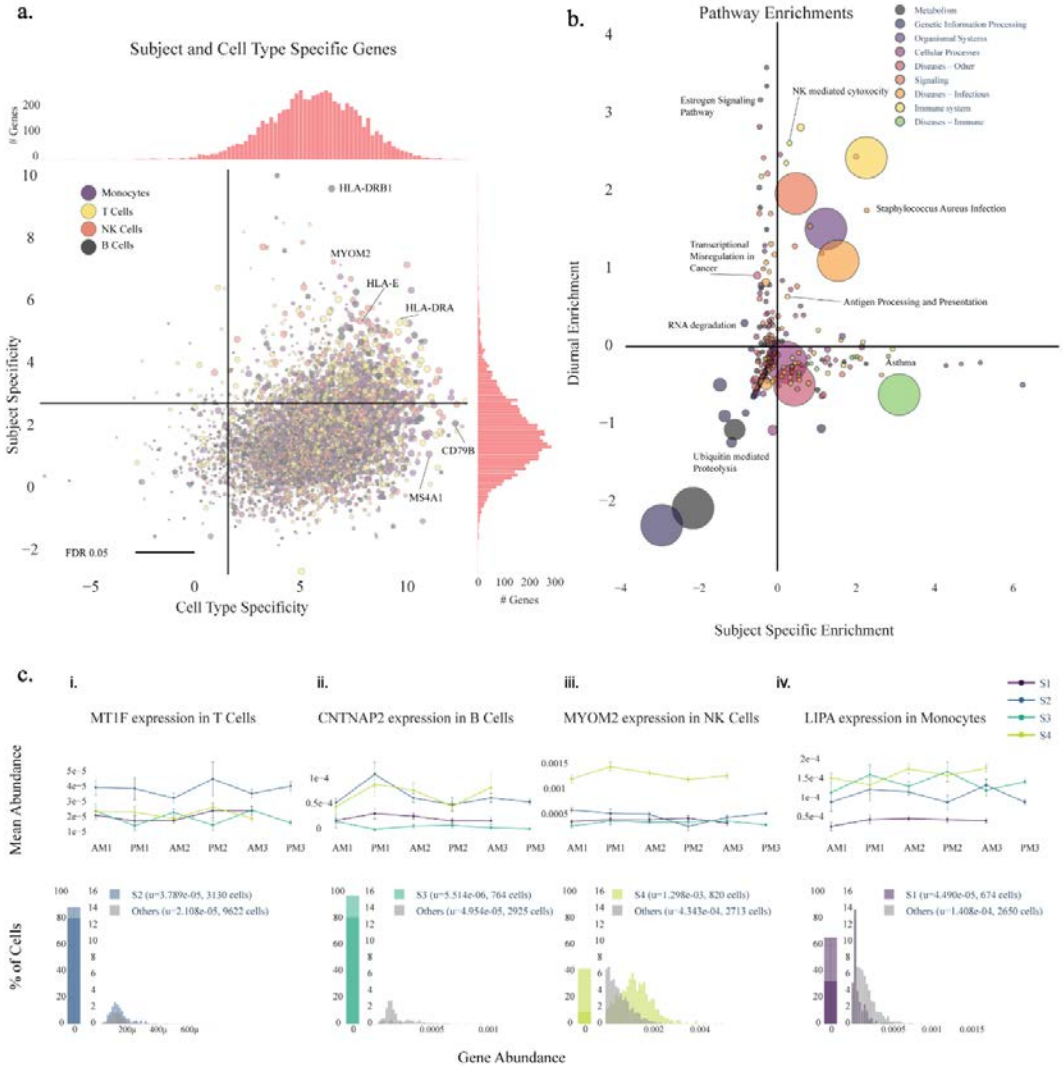


Fig.3 | Subject variability in immune and disease-relevant genes and pathways (a) Magnitude ($\log_2$ F statistic) of the variability in expression of genes between different cell types (x) and between subjects (y). 1284/7034 (18.3%) of genes are above the subject specificity significance line (FDR < 0.05, multiple comparison correction) and are classified as subject-specific. Several MHC class II genes (HLA-X) are strongly subject-specific, consistent with previous findings[5]. **(b)** KEGG pathways grouped into categories and their enrichment (Z-score from 2-proportion Z-test) among the top 250 diurnally and subject-varying genes vs all genes. Immune system and disease pathways are significantly enriched (p=0.029), supportive of the conclusion that immune and disease-related genes are highly subject dependent. The large circles indicate the enrichment of the category overall, and the sizes of the smaller pathway points indicate the number of genes associated with the pathway. **(c)** Subject and cell type specific gene examples for each subject and cell type with the upper row displaying the trace of mean gene expression across time-points and the bottom row showing gene abundance shifts for the subjects of interest.

To probe the stability of individual gene expression signatures at the single-cell level, we used our pipeline to identify genes whose variation in gene expression is most likely caused by intrinsic intersubject differences rather than high frequency immune system variability. We compared the mean gene expressions of all time points between subjects in all cell types and identified 1284 genes (FDR < 0.05, multiple comparison corrected) that are differentially expressed in at least one subpopulation of cells. Like Whitney, et al., we found MHC class II genes, such as HLA-DRB1, HLA-E, and HLA-DRA (Fig. 3a) to be among the largest sources of variation between subjects.[40] Additionally, we found that DDX17, which was classified by Whitney et al. as a gene with high intersubject variability, but low intrasubject variability via repeat sampling over longer time scales, may be a new class of temporally varying gene that varies by day of week, having consistently increasing expression each subsequent sampling day. This stresses the importance of high frequency sampling for identifying genes with the most intrinsic interindividual variability.

1.7 Numerous subject-specific genes are revealed in specific immune cell types

Within the 1284 genes with intrinsic interindividual variability, we found myriad disease-relevant genes for all subjects and cell types, which can be explored at our interactive online portal (http://capblood- seq.caltech.edu). As just one example, subject S1's monocytes have a consistent downregulation ($p=9.1x10^{-7}$, two-sided student t-test) of LIPA, a gene that is implicated in Lysosomal Acid Lipase Deficiency (Fig. 3c). Given the low abundance of monocytes in blood samples, such findings would typically only be discovered from a targeted blood test or RNA sequencing of isolated monocytes, either of which would only be performed if the disease was already suspected; this showcases how automated discovery in heterogeneous cell populations can be leveraged for personalized, preventative care.

1.8 Immune function and disease pathways are enriched in subject-specific genes

Given that genes do not act alone, we also found cell type-specific pathway differences among subjects. In particular, Subject 2's S100A8, S100A9, and S100A12 genes, calcium-binding proteins that play an important role in macrophage inflammation, are significantly downregulated in monocytes ($p_{S100A8}=1.3x10^{-5}$, $p_{S100A9}=9.0x10^{-5}$, $p_{S100A12}=3.0x10^{-4}$, two-sided student t-test) compared to other subjects (Fig. S2). We further explored our findings by inspecting the pathways that are most enriched in individual and time-varying genes, and found that genes that are implicated in immune system function (p=0.085) and immune diseases (p=0.029) are more present in subject-specific genes (Fig. 3b). This stands in contrast to pathways of core cellular functions such as genetic information processing (p=0.029) and metabolism (p=0.095), which are less present in subject-specific genes.

1.9 Discussion

Genome and transcriptome sequencing projects have unveiled millions of genetic variants and associated gene expression traits in humans.[41,47] However, large-scale studies of their functional effects performed through venous blood draws require tremendous effort to undertake, and this is exacerbated by the cost and complexity of single-cell transcriptome sequencing. Efforts such as the Immune Cell Census[42] are already underway to perform single-cell profiling of large cohorts, but reliance on venous blood draws of PBMCs will likely limit the diversity and temporal resolution of their sample pool. Our platform gives researchers direct, scalable access to high resolution immune system transcriptome information of human subjects, lowering the barrier of entry for myriad new research avenues. Examples of such studies include: 1.) tracking vulnerable populations over time, such as monitoring clonal expansion of $CD8^+$ T cells in Alzheimer's disease progression[2]; 2.) profiling of individuals who are under home care to track disease progression and therapeutic response, such as transplant patients and people under quarantine; and 3.) tracking how stress, diet, and environmental conditions impact the immune system at short and long time scales, particularly in underrepresented populations who do not have easy access to hospitals or research institutions, such as people in rural or underdeveloped areas. Larger, more diverse subject pools coupled with time course studies of cell type gene

expression in health and disease will have a dramatic impact on our ability to understand the baseline and variability of immune function.

Online Content

Online web portal is available to explore data presented in the main figures for study summary, with diurnal and subject specific genes via http://capblood-seq.caltech.edu.

Data Availability

Gene expression matrix and relevant metadata are available on https://data.caltech.edu/records/1407. FASTQ files are not being released to protect the identity of the subjects.

Code Availability

Custom code made for diurnal and subject specific gene detection is available on https://github.com/thomsonlab/capblood-seq

1.10 Methods

Human Study Cohort

This study was conducted at Caltech. Four healthy adults (2 male, 2 female) were recruited (Table S3). All participants provided written informed consent. The study was approved by the Institutional Review Board (IRB) at Caltech and all methods were performed in compliance with relevant guidelines and regulations. The blood collection took place in a non-BSL room to make sure the subjects were not exposed to pathogens. Subject blood was collected roughly 8 hours apart over three consecutive days.

CPBMC isolation

100 µl of capillary blood was collected via push-button collection device (TAP from Seventh Sense Biosystems). For each blood draw, the site of collection was disinfected with an alcohol wipe and the TAP device was placed on the deltoid of the subject per device usage instructions. The button was pushed, and then blood was collected for 2-7 minutes until the indicator turned red. Blood was extracted from the TAP device by gently breaking the seal foil, and mixed with PBS + 2% FBS to 1 ml. The mixture was slowly added to the side of a SepMate tube (SepMate-15 IVD, Stem Cell Technologies) containing 4.5 ml of Lymphoprep (#07811, Stem Cell Technologies) and centrifuged for 20 minutes at 800 RPM. Approximately 900 µl of CPBMC layer was extracted below the plasma layer. To further remove red blood cells, 100 µl of red blood cell lysis buffer (eBioscience 10X RBC Lysis Buffer, #00-4300-54) was added to the CPBMCs and incubated at RT for 15 minutes. The CPBMC pellet was washed twice with PBS and centrifuged at 400 rpm for 5 minutes. Cells were counted using trypan blue via an automated detector (Countess II Automated Cell Counter) and subjects' cells were pooled together for subsequent single-cell RNA sequencing.

Single-cell RNA sequencing

Subject pooled single-cell suspensions were loaded onto a Chromium Single Cell Chip (10X Genomics) based on manufacturer's instructions (targeted 10,000 cells per sample, 2,500 cells per person per time point). Captured mRNA was barcoded during cDNA synthesis and pooled for Illumina sequencing (Chromium Single Cell 3' solution - 10X Genomics). Each time point was barcoded with a unique Illumina sample index, and then pooled together for sequencing in a single Illumina flow cell. The libraries were sequenced with an 8-base index read, 26-base read 1 containing cell-identifying barcodes and unique molecular identifiers (UMIs), and a 91-base read 2 containing transcript sequences on a NovaSeq 6000.

Single-cell Dataset Generation

FASTQ files from Illumina were demultiplexed and aligned using Cell Ranger v3.0 (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger) and the hg19 reference genome with all options set to their defaults.

Sample Demultiplexing

FASTQ files from the single-cell sequencing Illumina libraries were aligned against the hg19 (human) reference genome using Cellranger v3.0 count function. SNPs were detected in the aligned data using freebayes (https://github.com/ekg/freebayes), which creates a combined variant call format (VCF) file, one per sample. SNPs were then grouped by cell barcode using popscle dsc-pileup (https://github.com/statgen/popscle). The SNP files for all samples were then merged into a single dsc-pileup file, and cell barcodes were disambiguated by providing a unique identifier per sample. Freemuxlet (popscle freemuxlet) was then run with default parameters to group cells into 4 subjects. This generates a probability of whether each cell barcode belongs to each subject, given the detection of single nucleotide polymorphism (SNPs) in reads associated with that cell barcode. Each cell was then assigned to the subject with the highest probability. Cells with low confidence (ambiguous cells) and high confidence in more than one subject (multiplets) were discarded, using popscle's default confidence thresholds. See the README at https://github.com/thomsonlab/capblood-seq for detailed instructions.

Debris Removal

The raw cell gene matrix provided by Cell Ranger contains gene counts for all barcodes present in the data. To remove barcodes representing empty or debris-containing droplets, a debris removal step was performed. First, a UMI count threshold was determined that yielded more than the expected number of cells based on original cell counts (15,000). All barcodes below this threshold were discarded. For the remaining barcodes, principal component analysis (PCA) was performed on the log-transformed cell gene matrix, and agglomerative clustering was used to cluster the cells. The number of clusters was automatically determined

by minimizing the silhouette score among a range of numbers of clusters (6 to 15). For each cluster, a barcode dropoff trace was calculated by determining the number of barcodes remaining in the cluster for all thresholds in increments of 50. These cluster traces were then clustered into two clusters using agglomerative clustering — the two clusters representing "debris" with high barcode dropoff rates and "cells" with low barcode drop-off rates. All clusters categorized as "debris" were then removed from the data.

Gene Filtering

Before cell typing, genes that have a maximum count less than 3 are discarded. Furthermore, after cell typing, any genes that are not present in at least 10% of one or more cell types are discarded.

Data Normalization

Gene counts were normalized by dividing the number of times a particular gene appears in a cell (gene cell count) by the total gene counts in that cell. Furthermore, for visualization only, the gene counts were multiplied by a constant factor (5000), and a constant value of 1 was added to avoid zeros and then log transformed.

Cell Typing

We used single cell Variational Inference (scVI) to transform the raw cell gene expression data into a 10-dimensional variational autoencoder latent space.[43] The variational autoencoder is conditioned on sample batch, creating a latent space which is independent of any batch-specific effects. The variational auto-encoder parameters: learning rate = 1e-3, number of epochs = 50

Agglomerative clustering (sci-kit learn) was used to generate clusters from the latent cell gene expression data. These clusters were then annotated based on known cell type marker genes (Fig. S1).

In order to resolve specific cell subtypes, such as those of T cells and Monocytes, we specified 13-15 clusters as an input for agglomerative clustering. For each study, we started at 13 clusters and incremented until all 4 major cell types and 2 subtypes were separable. In cases where agglomerative clustering yielded multiple clusters of the same cell type, these clusters were merged into a single cell type for analysis.

Venous and Capillary Blood Comparison

In order to compare venous blood cell type distributions to capillary blood, raw gene count data was downloaded from each of the respective studies, and we performed the same cell typing pipeline as for our capillary data, first projecting the data into a latent space via scVI, followed by agglomerative clustering and manual annotation based on known cell type marker genes.

Diurnal Gene Detection

To identify genes that exhibit diurnal variation in distinct cell types, we developed a statistical procedure that detects robust gene expression differences between morning (AM) and evening (PM) samples. Given that gene expression is different between subjects, we first normalize the mean gene expression within each subject for each cell type.

$$\mu'_{g_i,s_j,c_n,k} = \mu_{g_i,s_j,c_n,k} - \left(\frac{\sum_{k=1}^{N_{s_j}} 1_{k\in AM}\mu_{g_i,s_j,c_n,k}}{2\sum_{k=1}^{N_{s_j}} 1_{k\in AM}} + \frac{\sum_{k=1}^{N_{s_j}} 1_{k\in PM}\mu_{g_i,s_j,c_n,k}}{2\sum_{k=1}^{N_{s_j}} 1_{k\in PM}}\right) \qquad \text{(Eq. 1)}$$

We take the mean gene expression μ for each gene g_i in all samples k for cell type c_n and subject s_j and renormalize it into μ' by subtracting the equally weighted mean of AM and PM samples (Eq. 1). We then split the mean gene values into an AM group and a PM group and perform a statistical test (two-tailed student-t test) to determine whether to reject the null hypothesis that gene expression in AM and PM samples come from the same distribution. We then perform Benjamini-Hochberg multiple comparison correction at an FDR of 0.05 on

all gene and cell type p-values to determine where to plot the significance threshold. For plotting the genes, we choose the Z-statistic corresponding to the minimum p-value among cell types for that gene. To determine diurnality at the population level, we repeated the procedure above with all cells pooled into a single cell type.

Subject and Cell Type Specific Gene Detection

To classify genes as subject specific, we detect genes with mean gene expression levels that are robustly different between subjects in at least one cell type. For each cell type c_n and gene g_i, we create subject groups containing the mean gene expression values from each sample. To determine whether the gene expression means from the different subjects do not originate from the same distribution, we perform an ANOVA one-way test to get an F-statistic and p-value for each gene. We then perform Benjamini-Hochberg multiple comparison correction at an FDR of 0.05 on all gene and cell type p-values. For plotting the genes, we chose the F-statistic corresponding to the minimum p-value among cell types for that gene.

For determining gene cell type specificity, we performed a similar procedure. In particular, for each gene g_i, we create cell type groups containing the mean gene expression values for that cell type from each sample. We then perform a one-way ANOVA, and Benjamini-Hochberg multiple comparison correction at an FDR of 0.05.

Pathway Enrichment Analysis

Pathways from the KEGG database (python bioservices package) were used to calculate pathway enrichment for genes that were among the top 250 most diurnal and individual specific. All remaining genes present in the data were considered background. In order to normalize for gene presence across pathways, each gene was weighted by dividing the number of pathways in which that gene appears. For each KEGG pathway[44,45,46], the test statistic for a two-proportion z-test (python statsmodel v0.11.1) is used to determine pathway enrichment. From the top level pathway classes, we broke out “Diseases” into “Other”, “Immune Diseases”, and “Infectious Diseases” and separated “Immune System” from

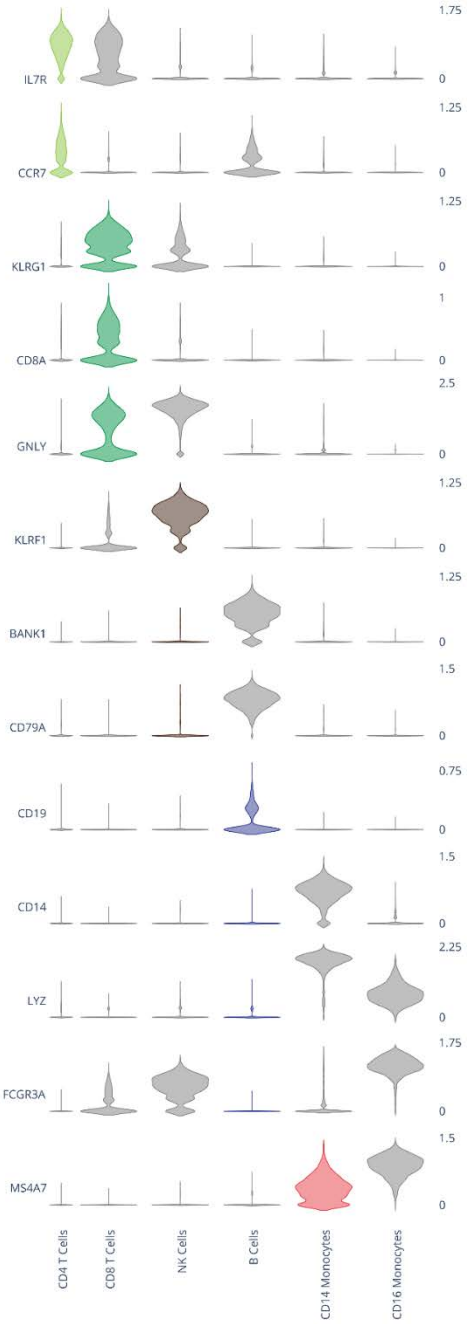
"Organismal System" to understand diurnal and subject-specific genes in an immune relevant context.

Figure Art

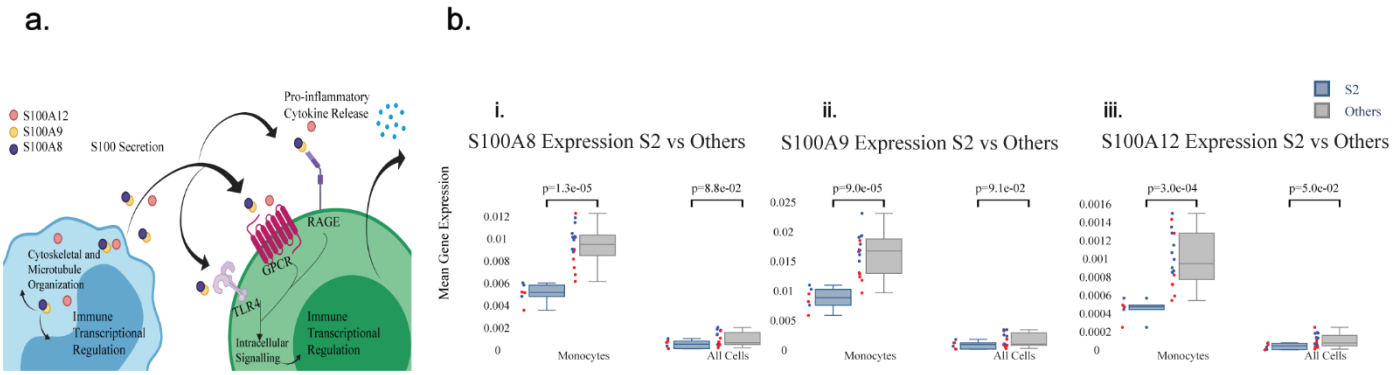
All drawings (Fig. 1a,b and Fig. S2a) are generated using BioRender.com

Fig. 1e was generated using GraphPad Prism 8.3.1

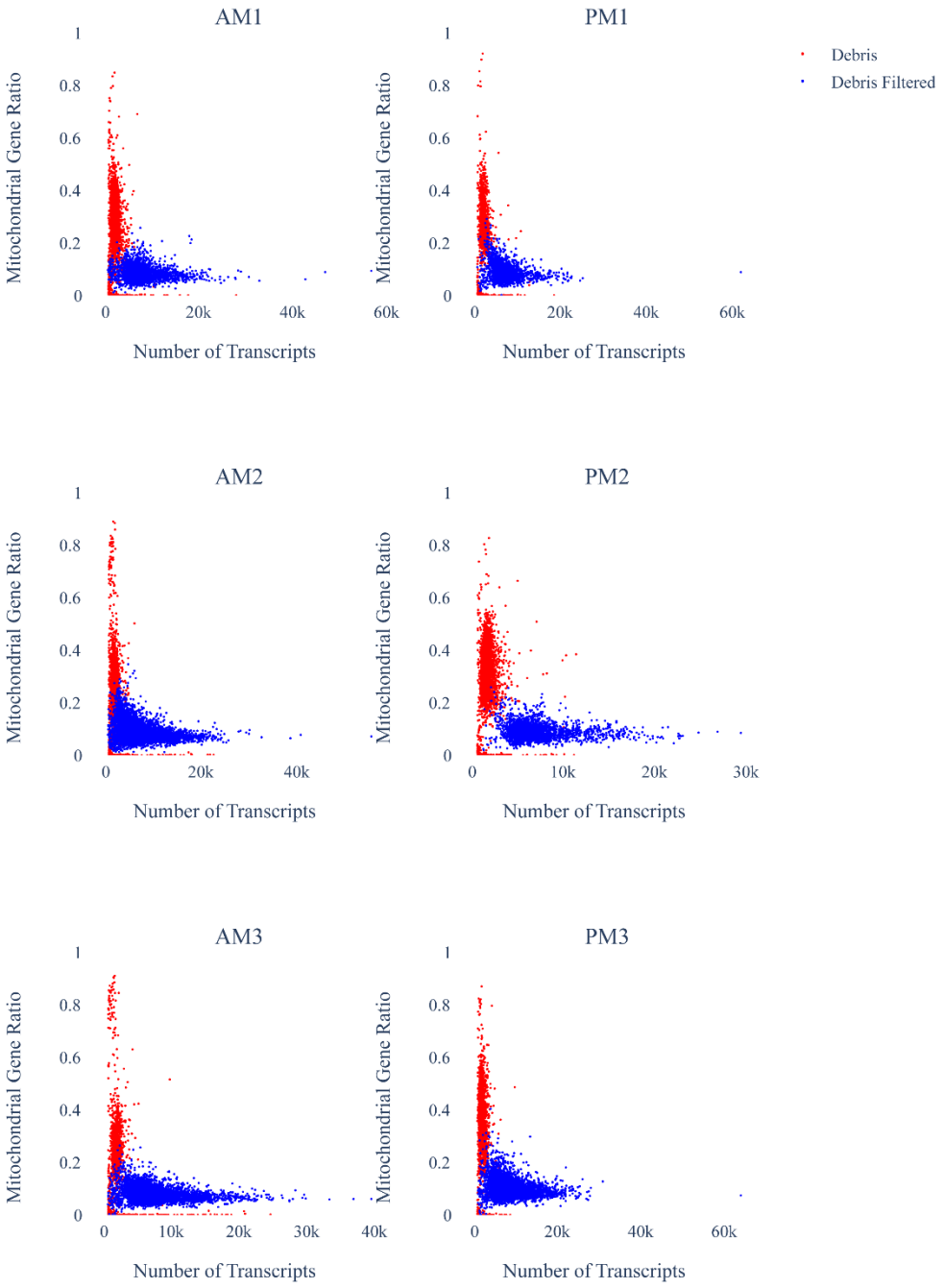
1.11 Supplementary Material



Supplementary Fig. 1 | Cell type marker gene expression in cell clusters Violin plots of log-normalized gene expression (y-axis, right hand side) for cell type markers (y-axis, left hand side) used to annotate cell clusters (x-axis) for known cell types. The colors correlate to clusters from Figure 1.d.

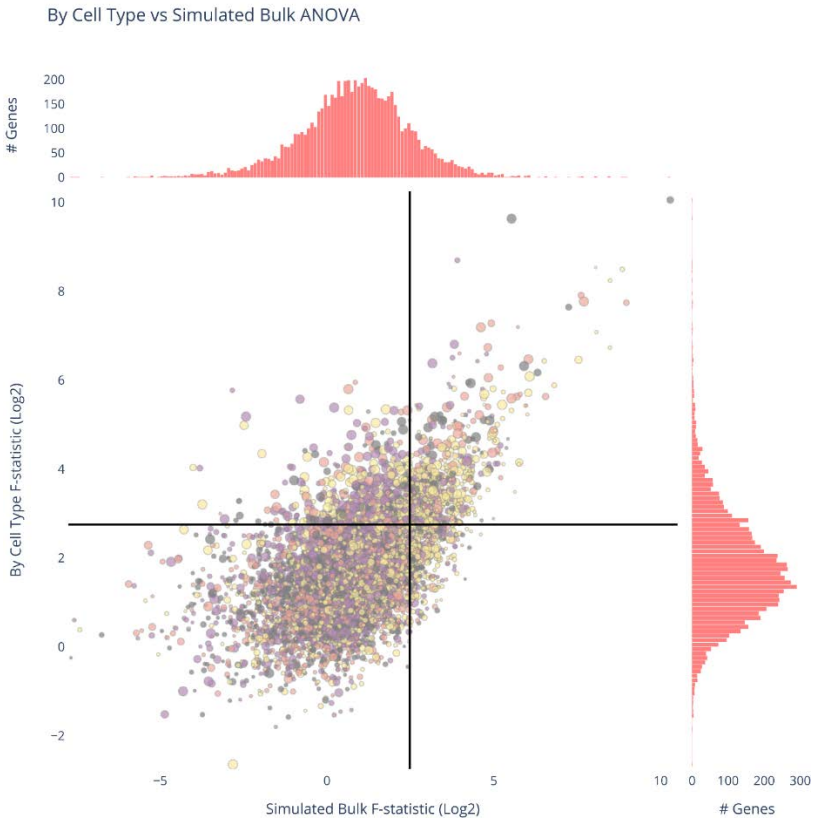


Supplementary Fig. 2 | S100 pathway exhibits individual-specific regulation (a) Simple schematic illustrating the role of S100A8, S100A9, and S100A12 genes in immune regulation. **(b)** Normalized mean gene expression of S100A8, S100A9, and S100A12 genes for S2 showing significant downregulation in monocytes as compared to all cells.



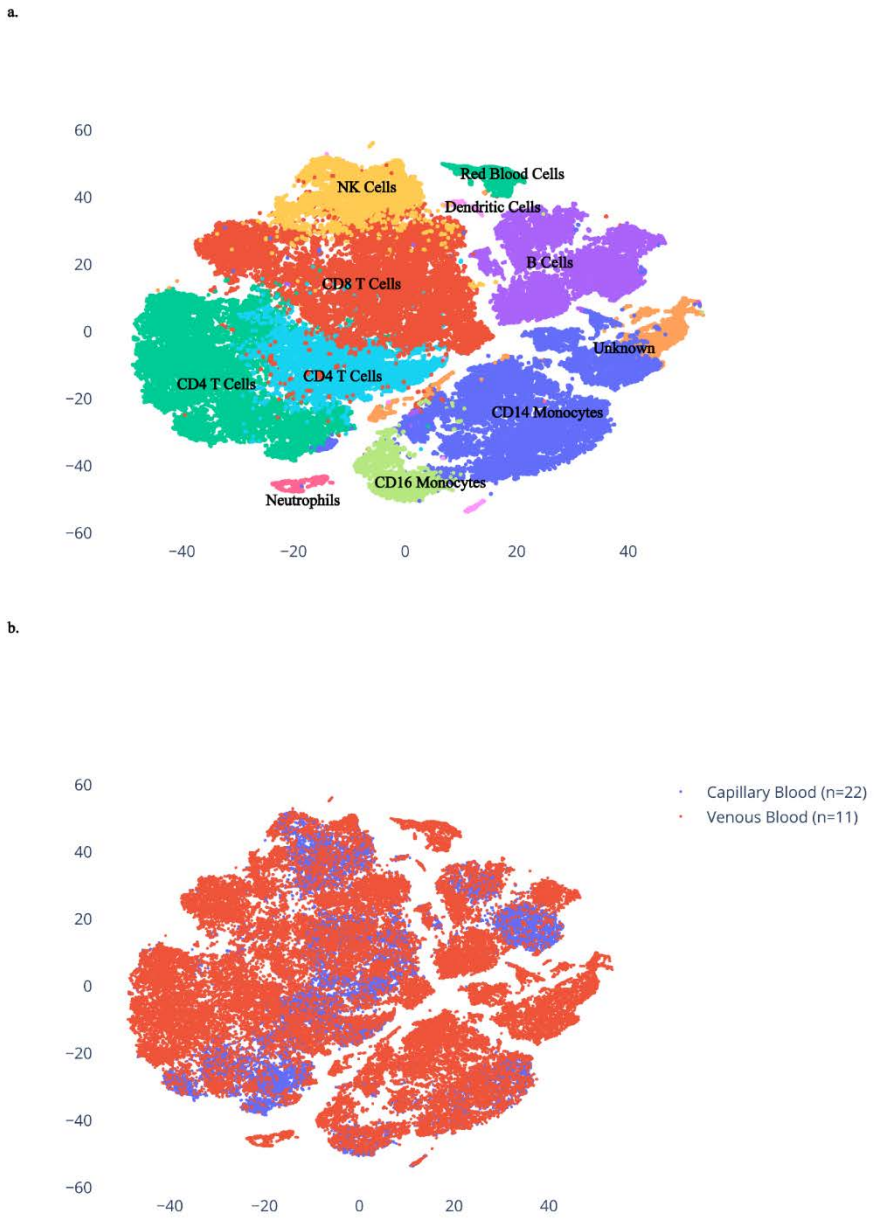
Supplementary Fig. 3 | Characterization of debris removal pipeline across each time sample. Scatter plots of the total number of transcripts (UMIs) detected for each barcode (x-

axis), and the ratio of transcripts that are mitochondrial (y-axis). These barcodes are the union of barcodes called by 10X Cellranger and our debris filtering pipeline. Barcodes colored red were flagged as debris and removed. The debris filtering pipeline appears to detect barcodes that have both a low transcript count, and a high mitochondrial gene ratio, or a rare number of cells that appear to have 0 mitochondrial genes. The counts of barcodes removed for each sample are in Table S6.



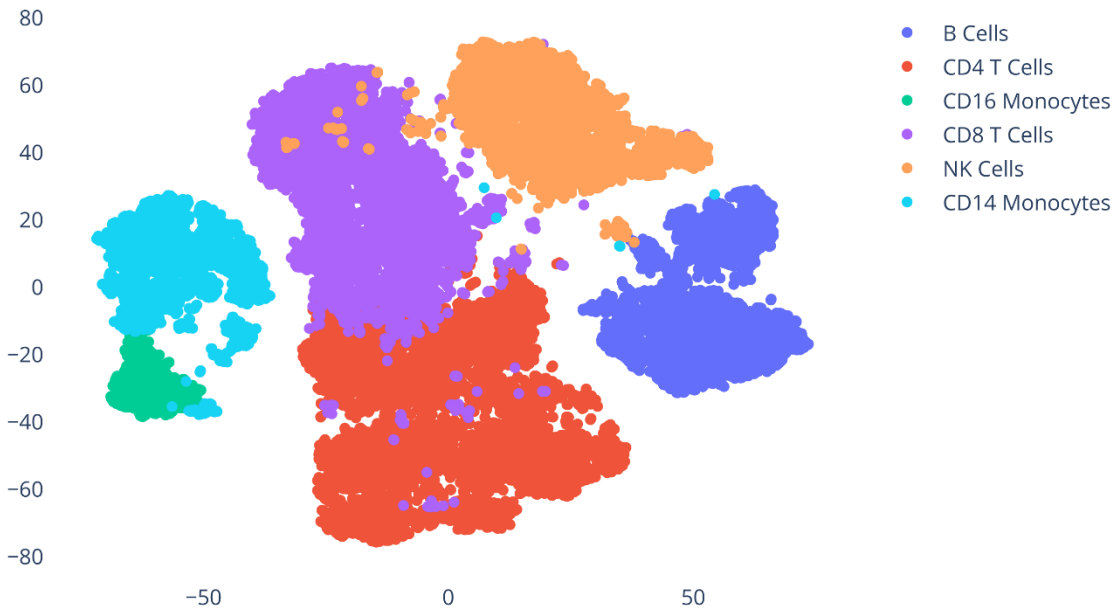
Supplementary Fig. 4 | Comparison of individual specificity by cell type vs in simulated bulk data. Magnitude ($\log_2$ F statistic) of the variability in expression of genes between subjects, accounting for each cell type separately (y) and in simulated bulk (x). 1284/7034 (18.3%) of genes are above the subject specificity significance line (FDR < 0.05, multiple comparison corrected) and are classified as subject-specific. Of these, only 637/1284 (49.6%) are also detected as subject-specific when simulating bulk RNA reads, despite the

significantly lower multiple comparison correction burden (7034 tests as compared to 28,136 tests in the cell type case).



Supplemental Fig. 5 | Merged projection of capillary and venous blood cells. Capillary blood cells from this study (n=22) and venous blood cells from 3 other studies (n=11) were

projected into a joint latent space using scVI. (a) Agglomerative clustering with n=13 clusters was performed to identify cell types, and annotated using known cell type markers (b) Capillary blood cells cluster together with venous blood cells, with the exception of one cluster of B cells unique to capillary cells, as well as 3 cell types unique to the venous blood sample: red blood cells, dendritic cells, and neutrophils, which are likely filtered out via laboratory procedures and the computational debris filtering pipeline.



Supplemental Fig. 6 | Immune cell type clusters detected in capillary blood. 2-dimensional t-SNE projection of the transcriptomes of all cells in all samples obtained from agglomerative clustering of latent gene expression. Cell clusters were annotated and grouped

based on the markers presented in Table S2. Small unidentifiable clusters were are not included in the figure.

Table S1: Genes that ranked in top 20 that had pre-existing literature tying to circadian/diurnal expression

Gene	**DOI Reference**
DDIT4	10.7554/eLife.20214.001, 10.1073/pnas.1800314115
SMAP2	10.1038/s41398-019-0671-7
RPL19	10.1128/MCB.00701-15
RPS9	10.1073/pnas.1515308112
PCPB1	10.1038/s41556-019-0441-z
RPS2	10.1073/pnas.1601895113
COX5B	10.1152/physiolgenomics.00066.2007

Table S2: Marker genes used to annotate clusters with specified cell population identity.

Cells	**Marker Genes**
CD14 Monocytes	CD14, LYZ
CD16 Monocytes	FCGR3A, MS4A7
CD4 T Cells	IL7R,CCR7
CD8 T Cells	KLRG1, CD8A, CD8B
Natural Killer (NK) Cells	GNLY, KLRF1, KLRD1
B Cells	BANK1, CD79A, CD79B, CD19

Table S3: Subject age and demographics. All subjects indicated to be healthy during the study.

Subject	**Age**	**Gender**
S1	32	M
S2	41	M
S3	34	F
S4	26	F

Table S4: Details of studies used to get healthy venous blood single-cell RNA sequencing dataset for comparison with capillary blood.

Subject	Age	Gender	Corresponding DOI	Corresponding Study Identification
S1	21	M	https://doi.org/10.1038/s41598-020-59827-1	Pre-THC-S1
S2	21	M	https://doi.org/10.1038/s41598-020-59827-1	Pre-THC-S2
S3	63	F	https://doi.org/10.1126/sciimmunol.abd1554	Sample 5_Normal 1 scRNA-seq [SW107]
S4	54	F	https://doi.org/10.1126/sciimmunol.abd1554	Sample 13_Normal 2 scRNA-seq [SW115]
S5	67	F	https://doi.org/10.1126/sciimmunol.abd1554	Sample 14_Normal 3 scRNA-seq [SW116]
S6	63	M	https://doi.org/10.1126/sciimmunol.abd1554	Sample 19_Normal 4 scRNA-seq [SW121]
S7	50	M	https://doi.org/10.1073/pnas.1907883116	CT1
S8	70	F	https://doi.org/10.1073/pnas.1907883116	CT2
S9	60	F	https://doi.org/10.1073/pnas.1907883116	CT3
S10	70	F	https://doi.org/10.1073/pnas.1907883116	CT4
S11	80	M	https://doi.org/10.1073/pnas.1907883116	CT5

Table S5: Number of genes in different cell types that is specific to each subject.

	B Cells	Monocytes	NK Cells	T Cells	Any
S1	55	67	58	269	400
S2	24	94	49	58	190
S3	55	149	70	150	353
S4	49	36	34	44	131

Table S6: Statistics for debris removal pipeline.

	Cellranger Called	Removed	Added	Final # Cells	% Removed
AM1	5808	2662	21	3167	45.83
PM1	3144	1302	12	1854	41.41
AM2	8772	2037	20	6755	23.22
PM2	6172	3587	0	2585	58.12

AM3	6684	1408	10	5286	21.07
PM3	7974	2370	4	5608	29.72

	Description	**File Name**
Table S7	Differential expression analysis for each cluster and cell type of the combined capillary blood (n=22) dataset	cluster_differential_expression.xlsx
Table S8	Differential expression analysis for all clusters between capillary blood (n=22, this study), and venous blood (n=11, external studies)	capillary_vs_venous_differential_expression.xlsx

1.12 References

1. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2014).

2. Gate, D. et al. Clonally expanded CD8 T cells patrol the cerebrospinal fluid in Alzheimer's disease. Nature 577, 399–404 (2020).

3. De Jager, P. L. et al. ImmVar project: Insights and design considerations for future studies of "healthy" immune variation. Seminars in Immunology 27, 51–57 (2015).

4. Sumitomo, S. et al. Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review. Inflamm Regener 38, (2018).

5. Long, J. E. et al. Morning vaccination enhances antibody response over afternoon vaccination: A cluster-randomised trial. Vaccine 34, 2679–2685 (2016).

6. Kobayashi, M., Wood, P. A. & Hrushesky, W. J. M. Circadian chemotherapy for gynecological and genitourinary cancers. Chronobiology International 19, 237–251 (2002).

7. Lévi, F. et al. Implications of circadian clocks for the rhythmic delivery of cancer therapeutics. Advanced Drug Delivery Reviews 59, 1015–1035 (2007).

8. Kazer, S. W. et al. Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. Nat Med 26, 511–518 (2020).

9. Uniken Venema, W. T. et al. Single-Cell RNA Sequencing of Blood and Ileal T Cells From Patients With Crohn's Disease Reveals Tissue-Specific Characteristics and Drug Targets. Gastroenterology 156, 812-815.e22 (2019).

10. Martin, J. C. et al. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. Cell 178, 1493-1508.e20 (2019).

11. Cai, Y. et al. Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. EBioMedicine 53, 102686 (2020).

12. Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. Sci. Immunol. 5, eabd1554 (2020).

13. David Bennett, Head of Religious Orders Study and Rush Memory and Aging Project, Personal Communicate, 07/16/2020

14. Blicharz, T. M. et al. Microneedle-based device for the one-step painless collection of capillary blood samples. Nat Biomed Eng 2, 151–157 (2018).

15. Lenicek Krleza, J., Dorotic, A., Grzunov, A. & Maradin, M. Capillary blood sampling: national recommendations on behalf of the Croatian Society of Medical Biochemistry and Laboratory Medicine. Biochem Med 335–358 (2015).

16. Tang, R. et al. Capillary blood for point-of-care testing. Critical Reviews in Clinical Laboratory Sciences 54, 294–308 (2017).

17. Robison, E. H. et al. Whole genome transcript profiling from fingerstick blood samples: a comparison and feasibility study. BMC Genomics 10, 617 (2009).

18. Catala, A., Culp-Hill, R., Nemkov, T. & D'Alessandro, A. Quantitative metabolomics comparison of traditional blood draws and TAP capillary blood collection. Metabolomics 14, (2018).

19. Toma, R. et al. A clinically validated human capillary blood transcriptome test for global systems biology studies. BioTechniques (2020) doi:10.2144/btn-2020-0088

20. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 36, 89–94 (2017).

21. Hashimoto, K. et al. Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. Proc Natl Acad Sci USA 116, 24242–24251 (2019).

22. Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. Sci. Immunol. 5, eabd1554 (2020).

23. Hu, Y. et al. Single-cell Transcriptome Mapping Identifies Common and Cell-type Specific Genes Affected by Acute Delta9-tetrahydrocannabinol in Humans. Sci Rep 10, (2020).

24. He, W. et al. Circadian Expression of Migratory Factors Establishes Lineage-Specific Signatures that Guide the Homing of Leukocyte Subsets to Tissues. Immunity 49, 1175-1190.e7 (2018).

25. Zhao, Y. et al. Uncovering the mystery of opposite circadian rhythms between mouse and human leukocytes in humanized mice. Blood 130, 1995–2005 (2017).

26. Keller, M. et al. A circadian clock in macrophages controls inflammatory immune responses. PNAS 106, 21407–21412 (2009).

27. Pick, R., He, W., Chen, C.-S. & Scheiermann, C. Time-of-Day-Dependent Trafficking and Function of Leukocyte Subsets. Trends in Immunology 40, 524–537 (2019).

28. Braun, R. et al. Universal method for robust detection of circadian state from gene expression. Proc Natl Acad Sci USA 115, E9247–E9256 (2018).

29. Lech, K. et al. Dissecting Daily and Circadian Expression Rhythms of Clock-Controlled Genes in Human Blood. J Biol Rhythms 31, 68–81 (2015).

30. Kusanagi, H. et al. Expression profiles of 10 circadian clock genes in human peripheral blood mononuclear cells. Neuroscience Research 61, 136–142 (2008).

31. Foo, J. C. et al. Longitudinal transcriptome-wide gene expression analysis of sleep deprivation treatment shows involvement of circadian genes and immune pathways. Transl Psychiatry 9, (2019).

32. Chang, J. et al. Circadian control of the secretory pathway maintains collagen homeostasis. Nat Cell Biol 22, 74–86 (2020).

33. PULFORD, JONES, BANHAM, HARALAMBIEVA & MASON. Lymphocyte-specific protein 1: a specific marker of human leucocytes. Immunology 96, 262–271 (1999).

34. Hermida, R. C., Ayala, D. E., Chayán, L., Mojón, A. & Fernández, J. R. Administration-Time-Dependent Effects of Olmesartan on the Ambulatory Blood Pressure of Essential Hypertension Patients. Chronobiology International 26, 61–79 (2009).

35. Ramsey, M. R. & Ellisen, L. W. Circadian function in cancer: Regulating the DNA damage response. Proceedings of the National Academy of Sciences 108, 10379–10380 (2011).

36. Ye, C. J. et al. Intersection of population variation and autoimmunity genetics in human T cell activation. Science 345, 1254665–1254665 (2014).

37. Thomas, D. Gene–environment-wide association studies: emerging approaches. Nat Rev Genet 11, 259–272 (2010).

38. Matsa, E. et al. Transcriptome Profiling of Patient-Specific Human iPSC-Cardiomyocytes Predicts Individual Drug Safety and Efficacy Responses In Vitro. Cell Stem Cell 19, 311–325 (2016).

39. Brodin, P. & Davis, M. M. Human immune system variation. Nat Rev Immunol 17, 21–29 (2016).

40. Whitney, A. R. et al. Individuality and variation in gene expression patterns in human blood. Proceedings of the National Academy of Sciences 100, 1896–1901 (2003).

41. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013).

42. The Immune Cell Census (2020, March). Retrieved from https://www.immunecensus.org/

43. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053–1058 (2018).

44. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27–30 (2000).

45. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. Nucleic Acids Research 47, D590–D595 (2018).

46. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. Protein Science 28, 1947–1951 (2019).

47. Chen, L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 167, 1398-1414.e24 (2016).

48. Der, E. et al. Tubular cell and keratinocyte single-cell transcriptomics applied to lupus nephritis reveal type I IFN and fibrosis relevant pathways. Nat Immunol 20, 915–927 (2019).

49. Fairfax, B. P. & Knight, J. C. Genetics of gene expression in immunity to infection. Current Opinion in Immunology 30, 63–71 (2014).

Chapter 2

DEEP PARALLEL CHARACTERIZATION OF AAV TROPISM AND AAV-MEDIATED TRANSCRIPTIONAL CHANGES VIA SINGLE-CELL RNA SEQUENCING

Adapted from (2).

2.1 Introduction

Recombinant AAVs (rAAVs) have become the preferred gene delivery vehicles for many clinical and research applications (Bedbrook et al., 2018; Samulski and Muzyczka, 2014) owing to their broad viral tropism, ability to transduce dividing and non-dividing cells, low immunogenicity, and stable persistence as episomal DNA ensuring long-term transgene expression (Daya and Berns, 2008; Deverman et al., 2018; Gaj et al., 2016; Hirsch and Samulski, 2014; Naso et al., 2017; Wu et al., 2006). However, current systemic gene therapies using AAVs have a relatively low therapeutic index (Mével et al., 2020). High doses are necessary to achieve sufficient transgene expression in target cell populations, which can lead to severe adverse effects from off-target expression (Hinderer et al., 2018; Srivastava, 2020; Wilson and Flotte, 2020). Increased target specificity of rAAVs would reduce both the necessary viral dose and off-target effects: thus, there is an urgent need for AAV gene delivery vectors that are optimized for cell-type-specific delivery (Paulk, 2020). Lower viral doses would also alleviate demands on vector manufacturing and minimize the chances of undesirable immunological responses (Calcedo et al., 2018; Gao et al., 2009; Mingozzi and High, 2013). Capsid-specific T-cell activation was reported to be dose-dependent in vitro (Finn et al., 2010; Pien et al., 2009) and in humans (Mingozzi et al., 2009; Nathwani et al., 2011). Shaping the tropism of existing AAVs to the needs of a specific disease has the potential to reduce activation of the immune system by detargeting cell types, such as dendritic cells, that have an increased ability to activate T-cells (Herzog et al., 2019;

Rogers et al., 2017; Rossi et al., 2019; Somanathan et al., 2010; Vandenberghe et al., 2006; Zhu et al., 2009).

Several studies have demonstrated that the transduction efficiency and specificity of natural AAVs can be improved by engineering their capsids using rational design (Bartlett et al., 1999; Davidsson et al., 2019; Davis et al., 2015; Lee et al., 2018; Sen, 2014) or directed evolution (Chan et al., 2017; Dalkara et al., 2013; Deverman et al., 2016; Excoffon et al., 2009; Grimm et al., 2008; Körbelin et al., 2016a; Kotterman and Schaffer, 2014; Maheshri et al., 2006; Müller et al., 2003; Ogden et al., 2019; Ojala et al., 2018; Pekrun et al., 2019; Pulicherla et al., 2011; Ravindra Kumar et al., 2020; Tervo et al., 2016; Ying et al., 2010). These engineering methods yield diverse candidates that require thorough, preferably high-throughput, in vivo vector characterization to identify optimal candidates for a particular clinical or research application. Toward this end, conventional immunohistochemistry (IHC) and various in situ hybridization (ISH) techniques are commonly employed to profile viral tropism by labeling proteins expressed by the viral transgene or viral nucleic acids, respectively (Arruda et al., 2001; Chan et al., 2017; Deleage et al., 2016, 2018; Deverman et al., 2016; Grabinski et al., 2015; Hinderer et al., 2018; Hunter et al., 2019; Miao et al., 2000; Polinski et al., 2015, 2016; Puray-Chavez et al., 2017; Ravindra Kumar et al., 2020; Wang et al., 2020; Zhang et al., 2016; Zhao et al., 2020).

Although these histological approaches preserve spatial information, current technical challenges limit their application to profiling the viral tropism of just one or two AAV variants across a few gene markers, thus falling short of efficiently characterizing multiple AAVs across many complex cell types characteristic of tissues in the central nervous system (CNS). The reliance on known marker genes also prevents the unbiased discovery of tropisms since such marker genes need to be chosen *a priori*. Choosing marker genes is particularly challenging for supporting cell types, such as pericytes in the CNS microvasculature and oligodendrocytes, which often have less established cell type identification strategies (Liu et al., 2020; Marques et al., 2016). The advent of single-cell RNA sequencing (scRNA-seq) has enabled comprehensive transcriptomic analysis of entire

cell-type hierarchies, and brought new appreciation to the role of cell subtypes in disease (Berto et al., 2020; Gokce et al., 2016; Tasic et al., 2016, 2018; Zeisel et al., 2018). However, experimental and computational challenges, such as the sparsity of RNA capture and detection, strong batch effects between samples, and the presence of ambient RNA in droplets, reduce the statistical confidence of claims about individual gene expression (Lähnemann et al., 2020; Yang et al., 2020; Zheng et al., 2017). Computational methods have been developed to address some of these challenges, such as identifying contaminating RNA (Yang et al., 2020), accounting for or removing batch effects (Korsunsky et al., 2019; Lin et al., 2019; Lopez et al., 2018), and distinguishing intact cells from empty droplets (Lun et al., 2019; Macosko et al., 2015; Zheng et al., 2017). However, strategies for simultaneously processing transcripts from multiple delivery vehicles and overcoming the computational challenges of confidently detecting individual transcripts have not yet been developed for probing the tropism of AAVs in complex, heterogeneous cell populations.

Collecting the entire transcriptome of injected and non-injected animals offers an opportunity to study the effects of AAV transduction on the host cell transcriptome. A similar investigation has been conducted with G-deleted rabies virus (Huang and Sabatini, 2020). This study demonstrated that virus infection led to the downregulation of genes involved in metabolic processes and neurotransmission in host cells, whereas genes related to cytokine signaling and the adaptive immune system were upregulated. At present, no such detailed examination of transcriptome changes upon systemic AAV injection has been conducted. High-throughput single-cell transcriptomic analysis could provide further insight into the ramifications of AAV capsid and transgene modifications with regard to innate (Duan, 2018; Hösel et al., 2012; Martino et al., 2011; Shao et al., 2018; Zaiss et al., 2008) and adaptive immune recognition (George et al., 2017; Manno et al., 2006; Mingozzi et al., 2007; Nathwani et al., 2011, 2014). Innate and adaptive immune responses to AAV gene delivery vectors and transgene products constitute substantial hurdles to their clinical development (Colella et al., 2018; Shirley et al., 2020). The study of brain immune response to viral gene therapy has been limited to antibody staining and observation of brain tissue slices post direct injection. In particular, prior studies have shown that intracerebral injection of rAAV vectors

in rat brains does not induce leukocytic infiltration or gliosis (Chamberlin et al., 1998; McCown et al., 1996); however, innate inflammatory responses were observed (Lowenstein et al., 2007). Results reported by these methods are rooted in single-marker staining and thus prevent the discovery of unexpected cell-type-specific responses. A comprehensive understanding of the processes underlying viral vector or transgene-mediated responses is critical for further optimizing AAV gene delivery vectors and treatment modalities that mitigate such immune responses.

Here, we introduce an experimental and bioinformatics workflow capable of profiling the viral tropism and response of multiple barcoded AAV variants in a single animal across numerous complex cell types by taking advantage of the transcriptomic resolution of scRNA-seq techniques (Figure 1 A). For this proof-of-concept study, we profile the tropism of previously-characterized AAV variants that emerged from directed evolution with the CREATE (AAV-PHP.B, AAV-PHP.eB) (Chan et al., 2017; Deverman et al., 2016) or M-CREATE (AAV-PHP.C1, AAV-PHP.C2, AAV-PHP.V1, AAV.CAP-B10) (Flytzanis et al., 2020; Ravindra Kumar et al., 2020) platforms. We selected the AAV variants based on their unique CNS tropism following intravenous injection. AAV-PHP.B and AAV-PHP.eB are known to exhibit overall increased targeting of the CNS compared with AAV9 and preferential targeting of neurons and astrocytes. Despite its sequence similarity to AAV-PHP.B, the tropism of AAV-PHP.V1 is known to be biased toward transducing brain vascular cells. AAV-PHP.C1 and AAV-PHP.C2 have both demonstrated enhanced blood–brain barrier (BBB) crossing relative to AAV9 across two mouse strains (C57BL/6J and BALB/cJ). Finally, AAV.CAP-B10 is a recently-developed variant with a bias toward neurons compared to AAV-PHP.eB (Flytzanis et al., 2020).

In our initial validation experiment, we quantify the transduction biases of AAV-PHP.eB and AAV-CAP-B10 across major cell types using scRNA-seq, and our results correlate well with both published results and our own conventional IHC-based quantification. We then demonstrate the power of our transcriptomic approach by going beyond the major cell types to reveal significant differences in sub-cell-type transduction specificity. Compared with

AAV-CAP-B10, AAV-PHP.eB displays biased targeting of inhibitory neurons, and both variants transduce Sst+ or Pvalb+ inhibitory neurons more efficiently than Vip+ inhibitory neurons. We validate these results with fluorescent in situ hybridization – hybridization chain reaction (FISH-HCR). We then develop and validate a barcoding strategy to investigate the tropism of AAV-PHP.V1 relative to AAV-PHP.eB in non-neuronal cells and reveal that pericytes, a subclass of vascular cells, evade transduction by this and other variants. We further use scRNA-seq to profile cell-type-specific responses to AAV.PHP-eB at 3 and 25 days post-injection (DPI), finding, for example, numerous genes implicated in the p53 pathway in endothelial cells to be upregulated at 3 DPI. While most upregulated genes across cell types return to control levels by day twenty-five, excitatory neurons show a persistent upregulation of genes involved in MAPK signaling extending to 25 days. Finally, we showcase the capabilities of parallel characterization by verifying the preceding findings in a single animal with seven co-injected AAV variants and reveal the unique non-neuronal tropism bias of AAV-PHP.C2.

2.2 Multiplexed single-cell RNA sequencing-based AAV profiling pipeline

To address the current bottleneck in AAV tropism profiling, we devised an experimental and computational workflow (Figure 1 A) that exploits the transcriptomic resolution of scRNA-seq to profile the tropism of multiple AAV variants across complex cell-type hierarchies. In this workflow, single or multiple barcoded rAAVs are injected into the retro-orbital sinus of mice followed by tissue dissociation, single-cell library construction using the 10X Genomics Chromium system, and sequencing with multiplexed Illumina next-generation sequencing (NGS) (Zheng et al., 2017). The standard mRNA library construction procedure includes an enzymatic fragmentation step that truncates the cDNA amplicon such that its final size falls within the bounds of NGS platforms (Figure 1 B). These cDNA fragments are only approximately 450 bp in length and, due to the stochastic nature of the fragmentation, sequencing from their 5’ end does not consistently capture any particular region. The fragment length limit and heterogeneity pose a problem for parallelizing AAV tropism profiling, which requires reliable recovery of regions of the transgene that identify the

originating AAV capsid. For example, posttranscriptional regulatory elements, such as the 600 bp Woodchuck hepatitis virus posttranscriptional regulatory element (WPRE), are commonly placed at the 3' end of viral transgenes to modulate transgene expression. The insertion of such elements pushes any uniquely identifying cargo outside the 450 bp capture range, making them indistinguishable based on the cDNA library alone (Supplemental Figure 1 A). An alternative strategy of adding barcodes in the 3' polyadenylation site also places the barcode too distant for a 5' sequencing read, and reading from the 3' end would require sequencing through the homopolymeric polyA tail, which is believed to be unreliable in NGS platforms (Chang et al., 2014; Shin and Park, 2016).

We circumvented these limitations in viral cargo identification by taking an aliquot of the intact cDNA library and adding standard Illumina sequencing primer recognition sites to the viral transcripts using PCR amplification such that the identifying region is within the two Illumina primer target sequences (e.g. Figure 2 B). The cell transcriptome aliquots undergoing the standard library construction protocol and the amplified viral transcripts are then sequenced as separate NGS libraries. We sequence shorter viral transcripts in the same flow cell as the cell transcriptomes and longer viral transcripts on the Illumina MiSeq, which we found to be successful at sequencing cDNAs up to 890 bp long. The sequencing data undergoes a comprehensive data processing pipeline (see Methods). Using a custom genome reference, reads from the cell transcriptome that align to the viral cargo plasmid sequences are counted as part of the standard 10X Cell Ranger count pipeline (see Methods and Supplemental Figure 1 C). In parallel, reads from the amplified viral transcripts are used to count the abundance of each viral barcode associated with each cell barcode and unique molecular identifier (UMI). The most abundant viral barcode for each cell barcode and UMI is assumed to be the correct viral barcode, and is used to construct a variant lookup table. This lookup table approach identifies an originating capsid in 69.1 ± 1.9% of viral reads detected in the cell transcriptome aliquots (Supplemental Table 4).

For determining viral cell-type tropism, we developed a method to estimate the fraction of cells within a cell type that express viral transcripts. Viral RNA expression levels depend on

both the multiplicity of infection and the transcription rate of the delivered cargo. Thus, directly using viral RNA counts to determine tropism is confounded by differences in transcription rate between cell types, limiting comparison with imaging-based tropism quantification methods. As evidence of this, we detected that viral RNA expression levels can vary by cell type but are not perfectly rank correlated with the percent of cells detected as expressing that transcript (Supplemental Figure 2 B). An additional confound arises from the ambient RNA from cellular debris co-encapsulated with cell-containing droplets, which can lead to false positives, i.e., detecting viral RNA in droplets containing a cell that was not expressing viral RNA. For example, we detected low levels of viral transcripts in large percentages of cells, even in cell types suspected to evade transduction, such as immune cells (Supplemental Figure 2 A). To reduce the effect of both variability in expression and ambient RNA, we developed an empirical method to estimate the percentage of cells expressing transcripts above the noise, wherein the distribution of viral transcript counts in a set of cells of interest is compared to a background distribution of cell-free (empty) droplets (see Methods, Supplemental Figure 2 C). In simulation, this method accurately recovers the estimated number of cells expressing transcripts above background across a wide range of parameterizations of negative binomial distributions (see Methods, Supplemental Figure 2 D).

To address several additional technical problems in default single-cell pipelines, we developed a simultaneous quality control (QC) and droplet identification pipeline. Our viral transduction rate estimation method described above relies on having an empirical background distribution of viral transcript counts in empty droplets to compare against the cell type of interest. However, the default cell vs. empty droplet identification method provided by the 10X Cell Ranger software, which is based on the EmptyDrops method (Lun et al., 2019), yielded unexpectedly high numbers of cells and clusters with no recognizable marker genes, suggesting they may consist of empty droplets of ambient RNA or cellular debris (Supplemental Figure 3 A, B). Additionally, we sought to remove droplets containing multiple cells (multiplets) from our data due to the risk of falsely attributing viral tropism of one cell type to another. However, using Scrublet (Wolock et al., 2019), an established

method for identifying droplets containing multiplets, failed to identify multiplets in some of our samples and only identified small proportions of clusters positive for known non-overlapping marker genes, such as Cldn5 and Cx3cr1 (Supplemental Figure 3 C). To address both the empty droplet and multiplet detection issues, we built a droplet classification pipeline based on scANVI, a framework for classifying single-cell data via neural-network-based generative models (Xu et al., 2021). Using clusters with a high percentage of predicted multiplets from Scrublet as training examples of multiplets, and clusters positive for known neuronal and non-neuronal marker genes as training examples of neurons and non-neuronal cells, we trained a predictive model to classify each droplet as a neuron, non-neuron, multiplet, or empty droplet (see Methods, Supplemental Figure 4 A). This model performed with 97.6% accuracy on 10% of cells held out for testing, and yielded a database of 270,982 cortical cells (Supplemental Figure 4 B). Inspection of the cells classified as empty droplets reveals that these droplets have lower transcript counts and higher mitochondrial gene ratios, consistent with other single-cell quality control pipelines (Supplemental Figure 4 D). Critically, we discovered that non-neuronal clusters contained significantly more cells that had been previously removed by the Cell Ranger filtering method as compared to neuronal clusters ($P = 0.02$, 2-sided student t-test). In some clusters, such as $Gpr17^{+}$ $C1ql1^{+}$ oligodendrocytes and $Gper^{+}$ $Myl9^{+}$ vascular cells, we identified up to 85% more cells than what were recovered via Cell Ranger in some samples.

Using our combined experimental and computational pipeline for viral transcript recovery and droplet identification, we can recover a lower bound on the expected number of cells expressing each unique viral cargo within groups of cells in heterogeneous samples.

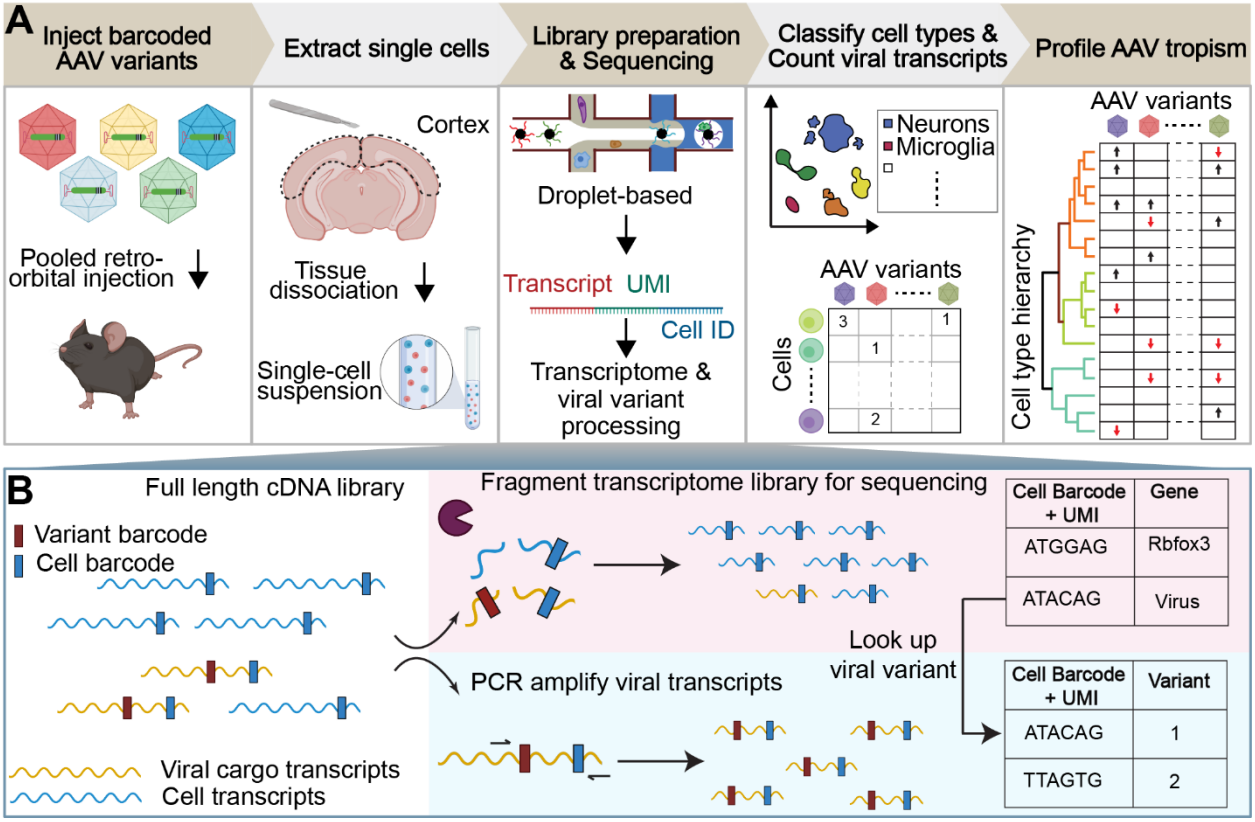


Figure 1. Workflow of AAV tropism characterization by scRNA-seq. *(A) (I) Injection of a single AAV variant or multiple barcoded AAV variants into the retro-orbital sinus. (II) After 3–4 weeks post-injection, the brain region of interest is extracted and the tissue is dissociated into a single-cell suspension. (III) The droplet-based 10x Genomics Chromium system is used to isolate cells and build transcriptomic libraries (see B). (IV) Cells are assigned a cell-type annotation and a viral transcript count. (V) AAV tropism profiling across numerous cell types. (B) The full length cDNA library is fragmented for sequencing as part of the single-cell sequencing protocol (top). To enable viral tropism characterization of multiple rAAVs in parallel, an aliquot of the intact cDNA library undergoes further PCR amplification of viral transcripts (bottom). During cDNA amplification, Illumina sequencing primer targets are added to the viral transcripts such that the sequence in between the Illumina primer targets contains the AAV capsid barcode sequence. Viral cargo in the cell transcriptome is converted to variant barcodes by matching the corresponding cell barcode + UMI in the amplified viral transcript library (right).*

2.3 Single-cell RNA sequencing recapitulates AAV capsid cell-type-specific tropisms

As a first step, we validated our method by comparing the quantification of AAV transduction of major cell types via scRNA-seq to conventional IHC. For this purpose, we characterized the tropism of two previously reported AAV variants, AAV-PHP.eB (Chan et al., 2017) and AAV-CAP-B10 (Flytzanis et al., 2020) (Figure 2 A). In total, four animals received single or dual retro-orbital injections of AAV-PHP.eB and/or AAV-CAP-B10 with 1.5×10^{11} viral genomes (vg) per variant. Co-injection of both variants served to test the ability of our approach to parallelize tropism profiling. By having each variant package a distinct fluorophore, tropism could be simultaneously assessed via multi-channel fluorescence and mRNA expression of the distinct transgene. After 3–4 weeks of expression, we harvested the brains and used one hemisphere for IHC and one hemisphere for scRNA-seq. To recover viral transcripts, we chose primers such that enough of the XFP sequence was contained within the Illumina primer target sequences to differentiate the two variants (Supplemental Table 1). For this comparison, we focused on the transduction rate for neurons (Rbfox3), astrocytes (S100b), and oligodendrocytes (Olig2). For IHC, a cell was classified as positive for the marker gene on the basis of antibody staining, and was classified as transduced on the basis of expression of the delivered fluorophore. For scRNA-seq, all cells that passed our QC pipeline were projected into a joint scVI latent space and clustered. To most closely match our imaging quantification, we considered all clusters that were determined to be positive for the respective marker gene as belonging to the corresponding cell type (see Methods). All clusters of the same marker gene were grouped together, and the transduction rate of the combined group of cells was determined using our viral transduction rate estimation method.

Our analysis of the scRNA-seq data demonstrates that the viral tropism biases across the three canonical marker genes are consistent with previous reports (Figure 2 C) (Chan et al., 2017; Flytzanis et al., 2020). In contrast to AAV-PHP.eB, AAV-CAP-B10 preferentially targets neurons over astrocytes and oligodendrocytes. No marked discrepancies in viral tropism characterization were observed with single versus dual injections.

To quantify the similarity of the AAV tropism characterizations obtained with IHC and scRNA-seq, we directly compared the transduction rate of each AAV variant for every cell type and its corresponding marker gene (i.e., Rbfox3, S100b, or Olig2) as determined by each technique and noticed a good correlation (Figure 2 D). Despite the different underlying biological readouts–protein expression in IHC and RNA molecules in labeled cell types for scRNA-seq–the two techniques reveal similar viral tropisms.

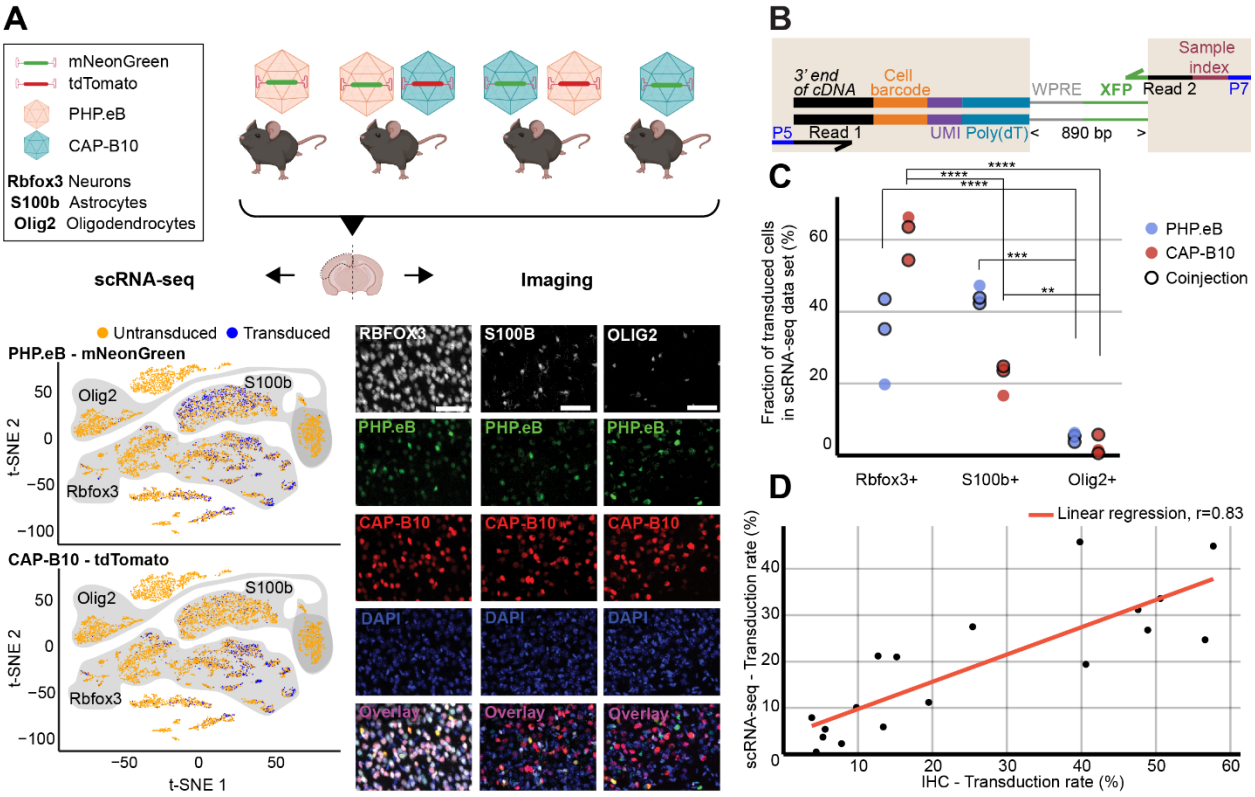


Figure 2. Comparison of viral tropism profiling with traditional IHC and scRNA-seq. *(A) Overview of the experiment. Four animals were injected with 1.5 × 10^{11} viral genomes (vg) packaged in AAV-PHP.eB and/or AAV-CAP-B10. The bottom panels show a representative dataset collected from an animal that was co-injected with AAV-PHP.eB and AAV-CAP-B10. The left side displays the scRNA-seq data set in the lower dimensional t-SNE space, with cells colored according to transduction status. The shaded areas indicate clusters with high expression of the corresponding gene marker. The right side shows representative confocal images of cortical tissue labeled with IHC. Scale bar, 50 µm. (B) Viral transcript*

*recovery strategy. The shaded areas highlight sequences added during library construction. (C) The fraction of the total number of transduced cells labeled as expressing the corresponding marker gene. For each AAV variant, the results of a two-way ANOVA with correction for multiple comparisons using Sidak's test are reported with adjusted P-values (****P ≤ 0.0001, ***P ≤ 0.001, and **P ≤ 0.01 are shown; P > 0.05 is not shown). (D) Comparison of transduction rates based on quantification via scRNA-seq or IHC. Transduction rate was calculated as (number of transduced cells in the group)/(total number of cells in the group). Each dot represents the transduction rate of neurons/Rbfox3+, astrocytes/S100b+, or oligodendrocytes/Olig2+ by AAV-PHP.eB or AAV-CAP-B10 in one animal. Histology data are averages across three brain slices per gene marker and animal. r indicates the Pearson correlation coefficient.*

2.4 Tropism profiling at transcriptomic resolution reveals AAV variant biases for neuronal subtypes

After validating our approach against the current standard of AAV tropism characterization (IHC imaging), we scrutinized the tropism of AAV-PHP.eB and AAV-CAP-B10 beyond the major cell types (Figure 3). Since AAV-CAP-B10 has increased neuronal bias relative to AAV-PHP.eB, we first sought to understand if there were neuronal subtypes that were differentially responsible for this bias. However, in-depth cell typing of transcriptomes collected from tissues with numerous and complex cell types, such as neurons in the brain, requires expert knowledge of the tissue composition, time to manually curate the data, and the availability of large datasets (Zeisel et al., 2018). To minimize the burden of manual annotation, computational tools have been developed that use previously-annotated single-cell databases to predict the cell type of cells in new, unannotated single-cell experiments, even across single-cell platforms (Cao et al., 2020; Tan and Cahan, 2019; Xu et al., 2021). We decided to leverage these tools and expanded our marker gene-based cell typing approach by having more complicated or well-established cell types be assigned based on annotations in a reference dataset (Supplemental Figure 4 A). To this end, we again employed scANVI to construct a joint model of cells from our samples and cells from an annotated reference

database. For this model, we used the Mouse Whole Cortex and Hippocampus 10x v2 dataset available from the Allen Brain Institute (Yao et al., 2021). Since this is a neuron-enriched dataset, we constructed the model using only the 109,992 cells in our dataset classified as neurons from our marker-based QC pipeline combined with the 561,543 neuronal cells from cortical regions from the reference database. We trained this model to predict to which of 14 neuron subtype groupings each cell belonged. We held out 10% of the data for testing: the model performed with 97.9% classification accuracy on the held-out data. We then applied the model to predict the neuron subtypes of our cells.

During our in-depth characterization, we discovered several previously unnoticed sub-cell-type biases for AAV-PHP.eB and AAV-CAP-B10 (Figure 3 A). Starting at the top of our neuronal hierarchy, the fraction of transduced cells that were glutamatergic neurons was markedly reduced for AAV-PHP.eB compared with AAV-CAP-B10 ($P = 0.03$, 2-sided student t-test, corrected for 2 neuron subtype comparisons). Furthermore, Pvalb+ and Sst+ inhibitory neurons both represented a larger fraction of transduced cells than Vip+ inhibitory neurons with both variants (adjusted $P < 0.0001$, $P = 0.10$, respectively, two-way ANOVA with multiple comparison correction for inhibitory neuron subtypes using Tukey's method).

To confirm these tropism biases in neuronal subtypes with a traditional technique, we performed FISH-HCR for glutamatergic and GABAergic gene markers (Figure 3 B) (Choi et al., 2014; Patriarchi et al., 2018). As indicated by our scRNA-seq data, AAV-CAP-B10, when compared with AAV-PHP.eB, has increased transduction efficiency of glutamatergic neurons (SLC17A7). Furthermore, FISH-HCR verified the downward trend in transduction efficiency from Pvalb+, to Sst+, to Vip+ neurons in both AAV variants (Figure 3 C).

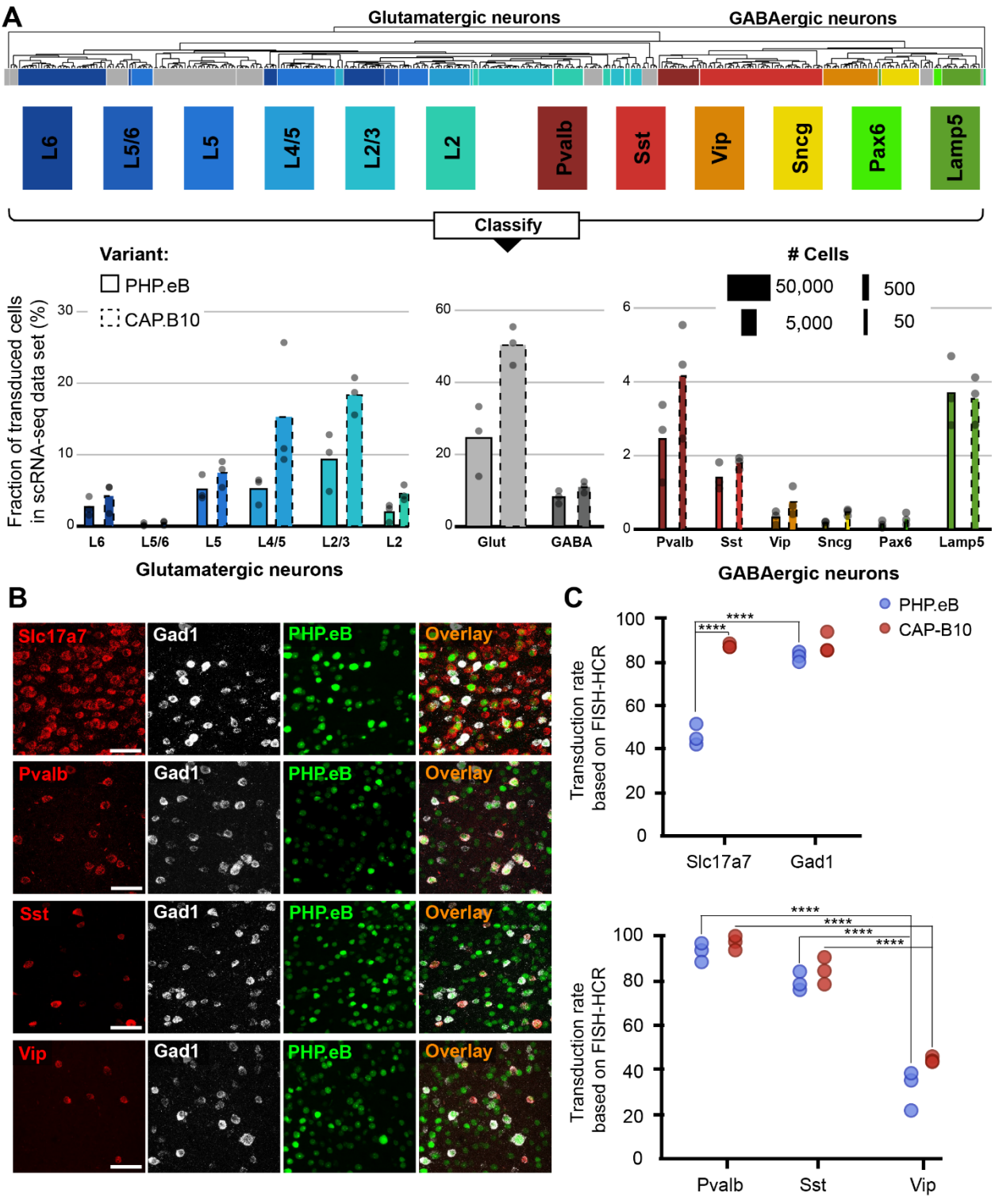


Figure 3. In-depth AAV tropism characterization of neuronal subtypes at transcriptomic resolution. *(A) Viral tropism profiling across neuronal sub types. Neuronal subtype*

*annotations are predicted by a model learned from the Allen Institute reference dataset using scANVI (Xu et al., 2021; Yao et al., 2021). Each dot represents data from one animal injected with AAV-PHP.eB and/or AAV-CAP-B10. Bar width indicates the total number of cells of a particular cell type present in our dataset. (B) Representative confocal images of cortical tissue from an animal injected with 1.5 × 10^{11} vg of AAV-PHP.eB. Tissue was labeled with FISH-HCR for gene markers of glutamatergic neurons (Slc17a7) and GABAergic neurons (Gad1, Pvalb, Sst, Vip). AAV-PHP.eB shows the endogenous fluorescence of mNeonGreen. Scale bar, 50 µm. (C) Confirmation of viral tropism biases across neuronal subtypes using FISH-HCR (3 mice per AAV variant, 1.5 × 10^{11} vg dose). Dots represent the average values across three brain slices from one animal. Results from a two-way ANOVA with correction for multiple comparisons using Tukey's test is reported with adjusted P-values (****P ≤ 0.0001; and P > 0.05 is not shown on the plot).*

2.5 Pooled AAVs packaging barcoded cargo recapitulate the non-neuronal tropism bias of PHP.V1

To enable profiling viral variants in parallel without needing distinct transgenes per variant, we established a barcoding strategy whereby we package AAV variants with the same transgene and regulatory elements but with short, distinguishing nucleotide sequences within the 3' UTR (Figure 4 A). To verify that this barcoding strategy can recover tropisms consistent with our previous transgene-based capsid-identification strategy, we performed a set of experiments to re-characterize the tropism of AAV-PHP.eB in parallel with that of the recently developed AAV-PHP.V1, which has increased specificity for vascular cells over AAV-PHP.eB (Ravindra Kumar et al., 2020).

We produced AAV-PHP.eB carrying CAG-mNeonGreen and AAV-PHP.V1 carrying either CAG-mRuby2 or CAG-tdTomato. Additionally, we produced AAV-PHP.eB and AAV-PHP.V1 both carrying CAG-mNeonGreen with 7-nucleotide barcodes 89 bp upstream of the polyadenylation start site such that they did not interfere with the WPRE. We ensured each barcode had equal G/C content, and that all barcodes were Hamming distance 3 from each other (Supplemental Table 5). Each of the barcoded variants was packaged with multiple

barcodes that were pooled together during virus production. Four animals received a retro-orbital co-injection of 1.5 x 10^{11} vg/each of AAV-PHP.V1 and AAV-PHP.eB. Two animals received viruses carrying separate fluorophores (cargo-based), and two animals received viruses carrying the barcoded cargo (barcode-based). For amplification of the viral cDNA in the animals receiving the barcoded cargo, we used primers closer to the polyA region such that the sequencing read covered the barcoded region (Supplemental Table 1). During the single-cell sequencing dissociation and recovery, one of our dissociations resulted in low recovery of neurons (Supplemental Figure 4 C); thus, we investigated only non-neuronal cells for this experiment.

Despite variability in the total transgene RNA content between barcodes of the same variant (Supplemental Figure 5 A), the estimated percent of cells expressing the transgene within each cell type was consistent between barcodes within a single animal, with standard deviations ranging from 0.003 to 0.058 (Supplemental Figure 6 A). Our analysis of both the barcode-based animals and cargo-based animals shows the same bias in non-neuronal tropism, with AAV-PHP.eB significantly preferring astrocytes over oligodendrocytes, vascular cells, and immune cells (Figure 4 D). Interestingly, our analysis also revealed that the variance between barcodes within an animal was less than the variance between animals, even when controlling for cargo and dosage ($P = 0.021$, Bartlett's test, P-values combined across all variants and cell types using Stouffer's method, weighted by transduced cell type distribution).

Next, we investigated the distribution of cells transduced by AAV-PHP.eB vs AAV-PHP.V1 in the major non-neuronal cell types across both barcode-based and cargo-based paradigms (Figure 4 E). The single-cell tropism data confirms the previously-established finding that AAV-PHP.V1 has a bias toward vascular cells relative to AAV-PHP.eB. Additionally, we uncovered that this is coupled with a bias away from astrocytes relative to AAV-PHP.eB, but that transduction of oligodendrocytes and immune cells did not differ between the variants. To investigate for a specific effect of the barcoding strategy, we performed a three-way ANOVA across the variant, cell type, and experimental paradigm factors. We found that

the cell type factor accounted for 89.25% of the total variation, the combined cell type + variant factor accounted for 7.7% of the total variation, and the combined cell type + experimental paradigm factor accounted for only 2.0% of the total variation, confirming our hypothesis that barcoded pools can recover tropism with minimal effect.

2.6 Relative tropism biases reveal non-neuronal subtypes with reduced AAV transduction

To further characterize the tropism biases of AAV-PHP.V1 and expand our method to less well-established cell hierarchies, we explored the non-neuronal cell types in our dataset. Since the Allen Brain Institute reference database that we used to investigate neuronal tropism was enriched for neurons, it does not contain enough non-neuronal cells to form a robust non-neuronal cell atlas. Our combined dataset consists of 169,265 non-neuronal cells, making it large enough to establish our own non-neuronal cell clustering. Thus, we performed an additional round of automatic clustering on the cells classified as non-neuronal in our combined dataset, and identified 12 non-neuronal cell subtypes based on previously established marker genes (Figure 4 B, C, Supplemental Table 2).

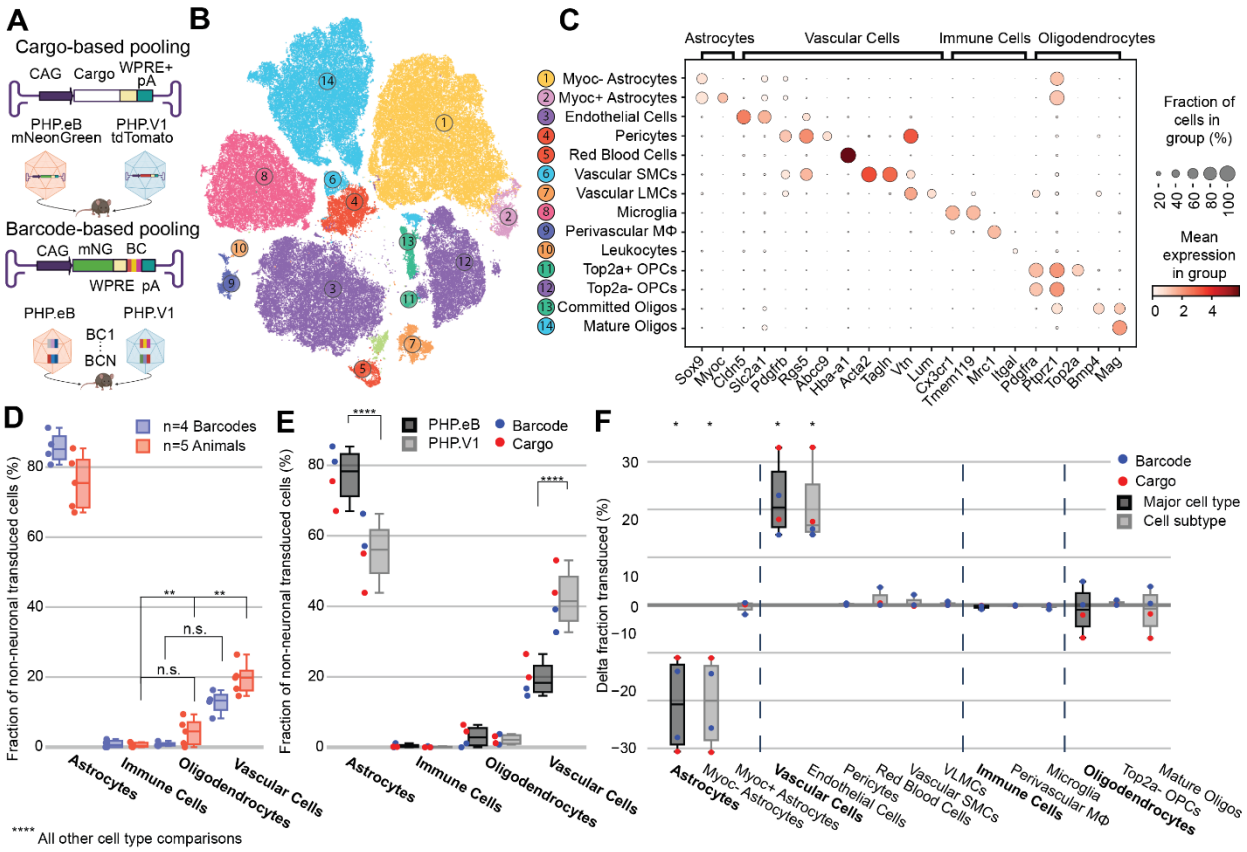


Figure 4. Barcoded co-injected rAAVs reveal the non-neuronal tropism bias of AAV-PHP.V1. *(A) Experimental design for comparing barcode vs cargo-based tropism profiling. Animals received dual injections of AAV-PHP.eB and AAV-PHP.V1, carrying either distinct fluorophores (cargo) or the same fluorophore with distinct barcodes. (B) t-SNE projection of the single-cell Variational Inference (scVI) latent space of cells and their cell type classification of the 169,265 non-neuronal cells across all our samples. Each number corresponds to the cell type labeled in (C). (C) Marker genes used to identify non-neuronal cell types. Darker colors indicate higher mean expression, and dot size correlates with the abundance of the gene in that cell type. (D) The distribution of non-neuronal cells expressing transcripts from AAV-PHP.eB across 4 barcodes within one animal (blue) and across 5 animals (red). All animals received dual injections, with one of the vectors being 1.5 x 10^{11} vg of PHP.eB carrying CAG-mNeonGreen. The y-axis represents the fraction of transduced non-neuronal cells that are of the specified cell type. Only the non-significant comparisons between cell types in a two-way ANOVA with correction for multiple comparisons using*

*Tukey's test are reported. All other cell-type comparisons within a paradigm were significant at $P \leq 0.0001$. (E) The distribution of non-neuronal cells expressing transcripts from AAV-PHP.eB (black) and AAV-PHP.V1 (gray). Results from the different experimental paradigms are combined. Results shown are from a two-way ANOVA with correction for multiple comparisons using Sidak's test comparing transduction by AAV-PHP.eB to AAV-PHP.V1 for each cell type, with adjusted P-values (****$P \leq 0.0001$ is shown; $P > 0.05$ is not shown). (F) Within-animal difference in the fraction of cells transduced with AAV-PHP.V1 relative to AAV-PHP.eB across four animals, two from each experimental paradigm. For each cell type in each sample, the combined 2-proportion z score for the proportion of that cell type transduced by AAV-PHP.V1 vs AAV-PHP.eB is reported. Cell types with fewer than 2 cells transduced by both variants were discarded. Z scores were combined across multiple animals using Stouffer's method and corrected for multiple comparisons. Cell-type differences with an adjusted P-value below 0.05 are indicated with *.*

Most cell subtypes had multiple clusters assigned to them, which suggested there may be additional subtypes of cells for which we did not find established marker genes. To determine whether any of these clusters delineated cell types with distinct transcriptional profiles, we investigated the probability of gene expression in each cluster compared to the other clusters of the same cell subtype (see Methods). Our approach determined two subclusters of pericytes, astrocytes, and oligodendrocyte precursor cells (OPCs). Both clusters of pericytes had strong expression of canonical pericytes marker genes Rgs5, Abcc9, and Higd1b. However, one of the clusters had no marker genes that made it distinct from the other pericyte cluster, nor from endothelial cells. Consistent with previous reports, this suggests that this cluster could be pericytes contaminated with endothelial cell fragments, and thus was not considered for further analysis (He et al., 2016; Vanlandewijck et al., 2018; Yang et al., 2021). Two distinct groups of astrocytes were detected, one of which had unique expression of Myoc and Fxyd6. Finally, one of the clusters of OPCs were uniquely expressing Top2a, Pbk, Spc24, Smc2, and Lmnb1. Using these new marker genes, we expanded our non-neuronal cell taxonomy to 14 cell types, now including Myoc+ and Myoc- astrocytes, and Top2a+ and Top2a- OPCs.

Given our finding that inter-sample variability exceeds intra-sample variability, we established a normalization method for comparing transduction biases between variants co-injected into the same animal. This normalization–calculating the difference in the fraction of transduced cells between variants–captures the relative bias between variants, instead of the absolute tropism of a single variant (see Methods). By considering the relative bias between variants, we are able to interrogate tropism in a way that is more robust to inter-sample variability that arises from different distributions of recovered cells, expression rate of delivered cargo, and success of the injection. Using this normalization method, we evaluated the non-neuronal cell type bias of AAV-PHP.V1 relative to AAV-PHP.eB in both the cargo-based animals and the barcode-based animals across our non-neuronal cell-type taxonomy (Figure 4 F). We discovered that the bias of AAV-PHP.V1 for vascular cells is driven by an increase in transduction of endothelial cells, but not pericytes. Similarly, AAV-PHP.V1's bias away from astrocytes is driven by a decrease in transduction of Myoc- astrocytes, but not Myoc+ astrocytes. Further inspection of the transduction of pericytes and Myoc+ astrocytes revealed that pericytes are not highly transduced by any of the AAVs tested in this work, and that Myoc+ astrocytes have both lower viral transcript expression and lower abundance than Myoc- astrocytes, and thus do not contribute significantly to tropism (Supplemental Figure 4, 7 A, B).

2.7 Single-cell RNA sequencing reveals early cell-type-specific responses to IV administration of AAV-PHP.eB that return to baseline by 3.5 weeks

To investigate the temporal cell-type-specific transcriptional effects of systemic AAV delivery and cargo expression, we performed a single-cell profiling experiment comparing animals injected with AAV to saline controls. We injected four male mice with AAV-PHP.eB (1.5×10^{11} vg) carrying mNeonGreen, and performed single-cell sequencing on two mice three days post-injection (3 DPI) and two mice twenty-five days post-injection (25 DPI). These time points were chosen based on previous work showing MHC presentation response peaking around day seven and transgene response peaking around day 30 (Lowenstein et al., 2007). The two saline control mice were processed 3 DPI. We then

analyzed differential gene expression for each cell type between injected animals and controls using DESeq2 (Supplemental Table 7). Of note, we excluded cell types with less than 50 cells in each sample, and excluded leukocytes and red blood cells given the risk of their presence due to dissociation rather than chemokine mediated infiltration. Additionally, we collapsed subtypes of excitatory neurons, inhibitory neurons, and OPCs to have greater than 50 cells for differential analysis. We estimated viral transduction rate of AAV-PHP.eB using its delivered cargo, mNeonGreen, across cell types and time points. We identified that Myoc- Astrocytes have significantly higher estimated transduction rate at 25 DPI compared to 3DPI (adjusted P-value = 0.0003, two-way ANOVA with multiple comparison correction using Sidak's method). It is also worth noting that endothelial cells have a similar transduction rate between the time points in both animals, while one of the animals at 25 DPI exhibited higher transduction in neurons (Figure 5A). The number of statistically relevant genes between the injected and control group (adjusted P-value < 0.05, DESeq2) were highest in pericytes (26 genes), endothelial cells (76 genes), and excitatory neurons (45 genes) at 3 DPI (Figure 5 B). At day twenty-five, only excitatory neurons had greater than 10 genes (14 genes total) differentially expressed (adjusted P-value < 0.05, DESeq2).

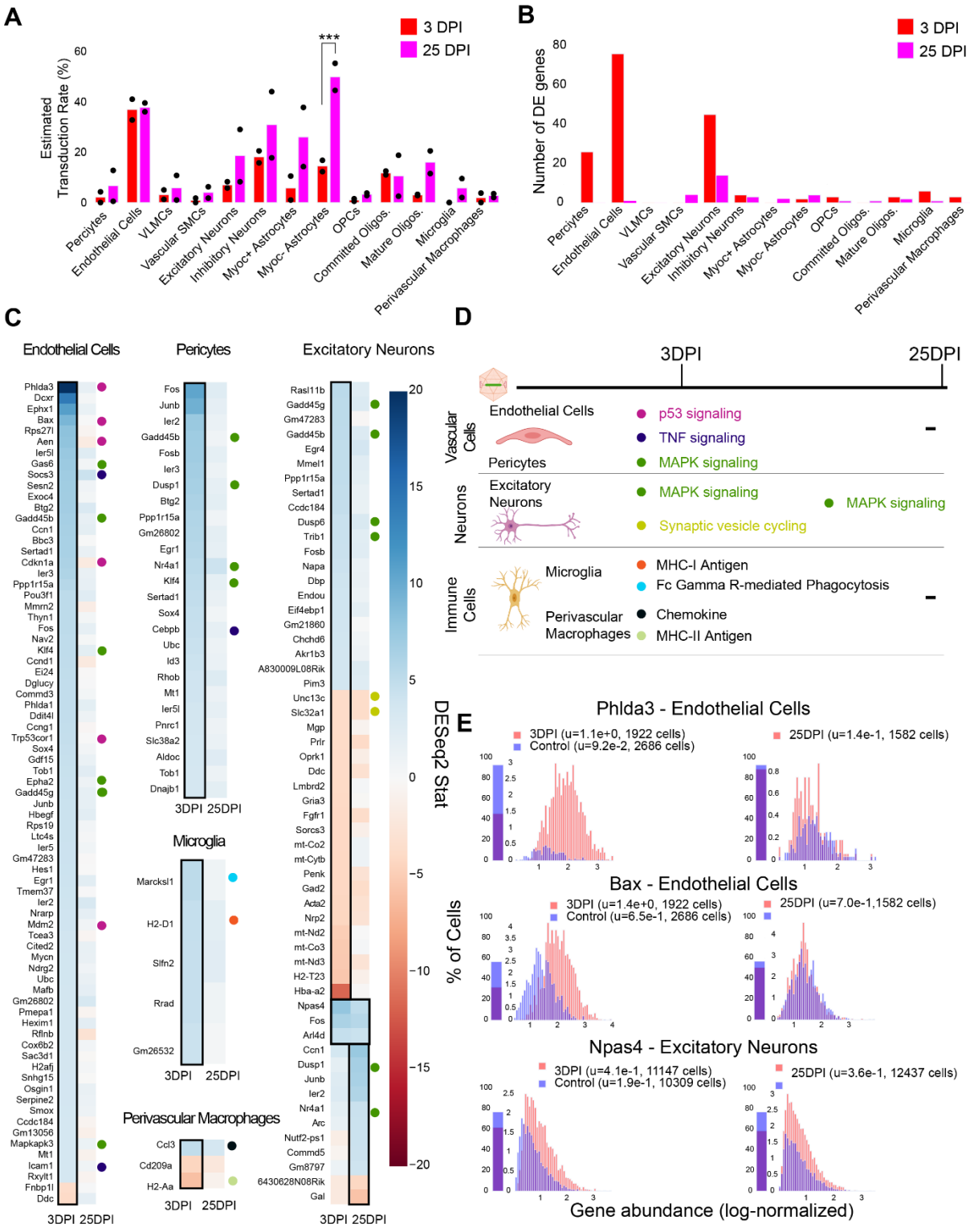

A
3 DPI
25 DPI
***
Estimated Transduction Rate (%)
Pericytes
Endothelial Cells
VLMCs
Vascular SMCs
Excitatory Neurons
Inhibitory Neurons
Myoc+ Astrocytes
Myoc- Astrocytes
OPCs
Committed Oligos.
Mature Oligos.
Microglia
Perivascular Macrophages
B
3 DPI
25 DPI
Number of DE genes
C
Endothelial Cells
Pericytes
Excitatory Neurons
Microglia
Perivascular Macrophages
3DPI 25DPI
DESeq2 Stat
% of Cells
D
3DPI
25DPI
Vascular Cells
Endothelial Cells
p53 signaling
TNF signaling
MAPK signaling
Pericytes
Neurons
Excitatory Neurons
MAPK signaling
Synaptic vesicle cycling
MAPK signaling
Immune Cells
Microglia
MHC-I Antigen
Fc Gamma R-mediated Phagocytosis
Perivascular Macrophages
Chemokine
MHC-II Antigen
E
Phlda3 - Endothelial Cells
3DPI (u=1.1e+0, 1922 cells)
Control (u=9.2e-2, 2686 cells)
25DPI (u=1.4e-1, 1582 cells)
Bax - Endothelial Cells
3DPI (u=1.4e+0, 1922 cells)
Control (u=6.5e-1, 2686 cells)
25DPI (u=7.0e-1,1582 cells)
Npas4 - Excitatory Neurons
3DPI (u=4.1e-1, 11147 cells)
Control (u=1.9e-1, 10309 cells)
25DPI (u=3.6e-1, 12437 cells)
Gene abundance (log-normalized)

Figure 5. Single-cell gene expression profiling finds cell-type-specific responses to AAV transduction in vascular cells and excitatory neurons. *(A) Estimated transduction rate (%) of mNeonGreen cargo at three and twenty-five days post-injection (DPI).* Results from a two-way ANOVA with correction for multiple comparisons using Sidak's method is reported with adjusted P-values (***$P \leq 0.001$; and $P > 0.05$ is not shown on the plot). *(B) Number of differentially expressed genes (adjusted P-value < 0.05, DESeq2) at 3 DPI and 25 DPI across 2 animals. (C) Differentially expressed genes across the two time points in endothelial cells, pericytes, microglia, perivascular macrophages, and excitatory neurons. Color indicates DESeq2 test statistic with red representing downregulation and blue representing upregulation. Genes outlined by a black rectangle are determined to have statistically significant differential expression compared to controls (adjusted P-value < 0.05, DESeq2). Colored circles adjacent to each gene indicate the corresponding pathway presented in (D). (D) A summary of corresponding pathways in which the differentially regulated genes in (C) are involved across the time points. (E) Distribution of p53 signaling transcripts in endothelial cells (animals are combined) and an example of a gene upregulated in both 3 and 25 DPI in excitatory neurons.*

We found that endothelial cells had the most acute response at 3 DPI with pathways such as p53, MAPK, and TNF signaling notably impacted. A significant upregulation of Phlda3 and its effectors Bax, Aen, Mdm2, and Cdkn1a, all involved in the p53/Akt signaling pathway, was present (Figure 5 C, E) (Ferreira and Nagai, 2019; Ghouzzi et al., 2016). Of relevance, we also detected Trp53cor1/LincRNA-p21, responsible for negative regulation of gene expression (Amirinejad et al., 2020), upregulated in endothelial cells at 3 DPI. Other examples of upregulated genes relevant to inflammation and stress response in vascular cells include the suppressor of cytokine signaling protein Socs3 (Baker et al., 2009), and Mmrn2, responsible for regulating angiogenesis in endothelial cells (Lorenzon et al., 2012). Expression of Socs3 and Icam1, which are upregulated in endothelial cells at 3 DPI, and Cepbp, which is upregulated in pericytes at 3 DPI, have all been linked to TNF signaling (Burger et al., 1997; Cao et al., 2018; Li et al., 2020). We have also observed genes linked to MAPK signaling upregulated in endothelial cells, such as Gas6, Epha2, and Mapkapk3, and

Klf4 in both endothelial cells and pericytes (Chen et al., 1997; Macrae et al., 2005; Riverso et al., 2017).

In brain immune cells, we observe a few substantial changes in genes pertaining to immune regulation at 3DPI which vanish at 25 DPI. For example, we observe an upregulation of MHC-I gene H2-D1 at 3 DPI in microglia, which then stabilizes back to control levels at 25 DPI (Figure 5 C). Marcksl1, previously reported as a gene marker for neuroinflammation induced by alpha-synuclein (Sarkar et al., 2020), also shows upregulation at 3 DPI. We did not observe significant differences in pro-inflammatory chemokines, Ccl2 and Ccl5, which are related to breakdown of the blood-brain barrier via regulation of tight-junction proteins and recruitment of peripheral leukocytes (Gralinski et al., 2009). Ccl3, responsible for infiltration of leukocytes and CNS inflammation (Chui and Dorovini-Zis, 2010), was upregulated in perivascular macrophages in 3 DPI and diminished back to control levels at 25 DPI (Figure 5 C). In contrast, Cd209a, a gene previously identified as critical for attracting and activating naïve T Cells (Franchini et al., 2019), was downregulated at 3 DPI.

Interestingly, we found that excitatory neurons had changes in genes across both 3 DPI and 25 DPI. MHC-Ib H2-T23, which is involved in the suppression of CD4+ T cell responses (Ohtsuka et al., 2008), is downregulated at 3 DPI. Meanwhile, the growth arrest genes, Gadd45g and Gadd45b (Vairapandi et al., 2002), are upregulated. Genes involved in synaptic vesicle cycling, such as Unc13c (Palfreyman and Jorgensen, 2017) and Slc32a1 (Taoufiq et al., 2020), are also downregulated at 3 DPI. Some genes remain upregulated throughout the study, such as Npas4, responsible for regulating excitatory-inhibitory balance (Spiegel et al., 2014). Genes implicated in MAPK signaling were upregulated – such as Gadd45b/g, Dusp6, and Trib1 at 3 DPI, and Dusp1 and Nr4a1 at 25 DPI (Muhammad et al., 2018; Ollila et al., 2012; Pérez-Sen et al., 2019; Salvador et al., 2013; Zhang and Yu, 2018). Gadd45b, Dusp1, Nr4a1 were also upregulated in pericytes and Gadd45b/g in endothelial cells (Figure 5 C).

Immediate early genes such as Ier2 (Kodali et al., 2020) were upregulated across pericytes, endothelial cells, inhibitory neurons, and OPCs at 3 DPI, while Fos, Ier2, Junb, and Arc were prominent in excitatory neurons at 25 DPI.

By investigating the gene expression differences in subpopulations of cells post-injection, we found that vascular cells such as endothelial cells and pericytes upregulate genes linked to p53, MAPK, and TNF signaling pathways at 3 DPI (Figure 5 D). Immune cells such as microglia and perivascular macrophages upregulate genes involved in chemokine signaling, MHCI/II antigen processing, and Fc Gamma R-Mediated Phagocytosis (Zhang et al., 2021) at 3 DPI (Figure 5 D). Excitatory neurons are the only cell type with genes implicated in the same pathway (MAPK signaling) upregulated across both of the time points (3 DPI, 25 DPI).

2.8 Larger pools of barcoded AAVs recapitulate complex tropism within a single animal

To showcase the capabilities of parallel characterization, we next designed a 7-variant barcoded pool that included the three previously characterized variants (AAV-PHP.eB, AAV-CAP-B10, and AAV-PHP.V1), AAV9 and AAV-PHP.B controls, and two additional variants, AAV-PHP.C1 and AAV-PHP.C2. For simplification of cloning and virus production, we designed a plasmid, UBC-mCherry-AAV-cap-in-cis, that contained both the barcoded cargo, UBC-mCherry, and the AAV9 capsid DNA (Supplemental Figure 1 B). We assigned three distinct 24 bp barcodes to each variant (Supplemental Table 5). Each virus was produced separately to control the dosage, and 1.5×10^{11} vg of each variant was pooled and injected into a single animal.

After 3 weeks of expression, we performed single-cell sequencing on extracted cortical tissue. To increase the number of cells available for profiling, we processed two aliquots of cells, for a total of 36,413 recovered cells. To amplify the viral transcripts, we used primers that bind near the 3’ end of mCherry such that the barcode was captured in sequencing (Supplemental Table 1).

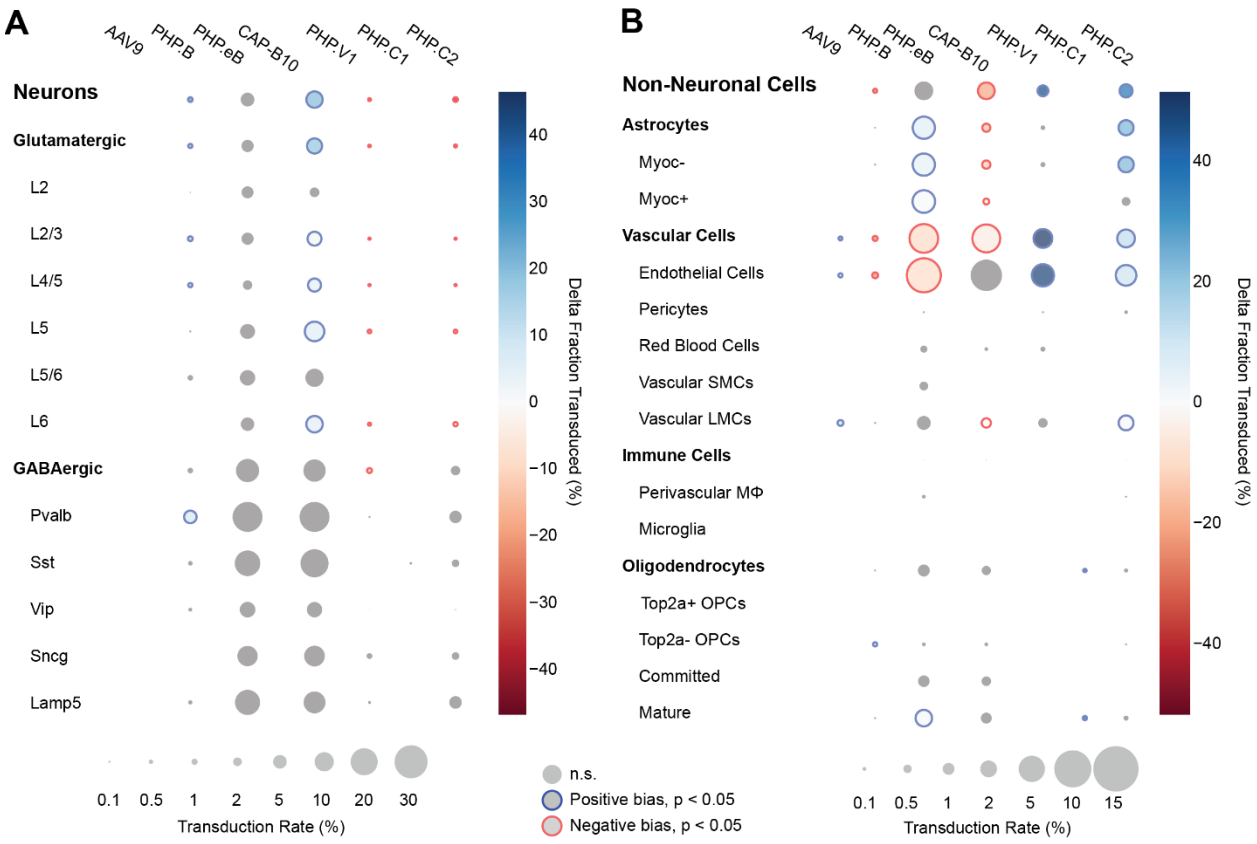


Figure 6. Single animal injections of multiple barcoded rAAVs enables deep, parallel characterization. *(A, B) Relative cell type tropism of 7 co-injected rAAVs for neuronal (A) and non-neuronal (B) cell types. The color scale indicates the difference in transduction bias of a variant relative to all other variants in the pool. The area of each circle scales linearly with the fraction of cells of that type with viral transcripts above background. For each variant and cell type, a 2-proportion z score was calculated to compare the number of cells of that type transduced by that variant relative to all other variants combined. Z scores were combined across two single-cell sequencing aliquots using Stouffer's method, and corrected for multiple comparisons. Cell types with fewer than 10 transduced cells in either the variant or variants compared against were discarded. Only cell-type biases at an adjusted P-value < 0.05 are colored; otherwise they are grayed out.*

Using our cell typing and viral transcript counting methods, we investigated the transcript counts and transduction bias of the variants in the pool. Compared with our previous profiling

experiments, the number of UBC-mCherry viral transcripts detected per cell was significantly lower than CAG-mNeonGreen-WPRE and CAG-tdTomato (adjusted P < 0.0001, P=0.0445, respectively, two-way ANOVA with multiple comparison correction using Tukey's method) and shifted towards vascular cells (adjusted P < 0.0001, P=0.0008, respectively, two-way ANOVA with multiple comparison correction using Tukey's method) (Supplemental Figure 5 B, C). Next, we looked at the transduction rate difference for each variant compared with the rest of the variants in the pool for each cell type in our taxonomy (Figure 6 A, B). Despite the lower expression rate and bias shift, the transduction rate difference metric captured the same tropism biases for AAV-CAP-B10 and AAV-PHP.V1 as determined from our previous experiments. AAV-CAP-B10 showed enhanced neuronal targeting relative to other variants in the pool, with this bias coming specifically from an increase in the transduction of glutamatergic neurons. All five variants with transcripts detected in neurons showed a decreased transduction rate in Vip+ neurons relative to other GABAergic neuronal subtypes (Supplemental Figure 7 C). AAV-PHP.eB showed enhanced targeting of astrocytes (+6.2%, $P = 1.4 \times 10^{-8}$, 2-proportion z-test, multiple comparison corrected with Benjamini/Hochberg correction), and AAV-PHP.V1 showed strong bias for vascular cells (+51.6%, $p = 1.7 \times 10^{-43}$). In addition to confirming all our existing hypotheses, we were able to identify biases for the previously reported AAV-PHP.C2, which has not been characterized in depth. This variant, which was reported as having a non-neuronal bias similar to AAV-PHP.V1, showed significant transduction bias not only toward vascular cells (+13.6%, $P = 8.3 \times 10^{-6}$), but also toward astrocytes (+24.0%, $P = 1.6^{-30}$), and a bias away from neurons (−38%, $p = 4.1 \times 10^{-32}$).

2.9 Discussion

The advent of NGS has enabled screening of large libraries of AAV capsids in vivo by extracting viral DNA from relevant tissue followed by sequencing of capsid gene inserts or DNA barcodes corresponding to defined capsids. To date, NGS-based screening has been successfully applied to libraries created by peptide insertions (Davidsson et al., 2019; Körbelin et al., 2016b), DNA shuffling of capsids (De Alencastro et al., 2020; Herrmann et

al., 2019; Paulk et al., 2018), and site-directed mutagenesis (Adachi et al., 2014). Although these NGS-based strategies allow the evolution of new AAV variants with diverse tissue tropisms, it has been difficult to obtain a comprehensive profiling for multiple variants across cell types, which is of utmost importance in organs with complex cell-type compositions, such as the brain (Deverman et al., 2016; Ravindra Kumar et al., 2020; Tasic et al., 2016, 2018; Zeisel et al., 2018). Towards this end, techniques such as IHC, fluorescent in situ RNA hybridization (Chen et al., 2015; Choi et al., 2014; Femino et al., 1998; Lubeck et al., 2014; Shah et al., 2016a, 2016b) or in situ RNA sequencing (Ke et al., 2013; Lee et al., 2014; Wang et al., 2018) can be employed. Several limitations make it challenging to apply these techniques as high-throughput, post-selection AAV tropism profiling methods. First, the limits of optical resolution and the density of transcripts in single cells pose challenges for full in situ transcriptome analysis and, until recently, have restricted the total number of simultaneously measured genes in single cells within tissue to several hundred (Ke et al., 2013; Lee et al., 2014; Liao et al., 2020; Shah et al., 2016a; Wang et al., 2018). By contrast, scRNA-seq with the 10x Genomics Chromium system enables detection of over 4000 genes per cell (Yao et al., 2021), fast transcriptomic analysis, and multiplexing across different tissue types (McGinnis et al., 2019; Stoeckius et al., 2018). Furthermore, the method is already widely used by the research community which can help with adoption of our proposed pipelines. Although droplet-based scRNA-seq methods lose spatial information during the dissociation procedure, analysis packages have been developed that can infer single-cell localization by combining scRNA-seq data with pre-existing information from ISH-based labeling for specific marker genes (Achim et al., 2015; Durruthy-Durruthy et al., 2015; Halpern et al., 2017; Nitzan et al., 2019; Satija et al., 2015; Stuart et al., 2019). Therefore, scRNA-seq techniques have great potential to rapidly profile the tropism of multiple AAV variants in parallel across several thousand cells defined by their entire transcriptome.

Here, we established an experimental and data-analysis pipeline that leverages the capabilities of scRNA-seq to achieve simultaneous characterization of several AAV variants across multiplexed tissue cell types within a single animal. To differentiate multiple AAV capsid variants in the sequencing data, we packaged variants with unique transgenes or the same transgene with unique barcodes incorporated at the 3' end. We added standard Illumina sequencing primer recognition sites (Read 2) to the viral transcripts using PCR amplification such that the barcoded region could be consistently read out from the Illumina sequencing data. Our computational pipeline demultiplexes viral reads found in the transcriptome according to which matching sequence is most abundant in a separate amplified viral transgene library. Comparing the distribution of viral transcripts by cell type to a null model of empty droplets, we could then determine the cell-type biases.

Our platform has corroborated the tropism of several previously characterized AAV variants and has provided more detailed tropism information beyond the major cell types. The fraction of transduced cells that are glutamatergic neurons was found to be markedly reduced for AAV-PHP.eB when compared with AAV-CAP-B10. Furthermore, within all the variants we tested, both Pvalb+ and Sst+ inhibitory neurons have greater transduction rates than Vip+ neurons. This bodes well for delivery to Pvalb+ neurons, which have been implicated in a wide range of neuro-psychiatric disorders (Ruden et al., 2021), and suggests Vip+ interneurons, which have recently been identified as being a sufficient delivery target for induction of Rett syndrome-like symptoms, as a target for optimization (Mossner et al., 2020). Awareness of neuronal subtype biases in delivery vectors is critical both for neuroscience researchers and for clinical applications. Dissection of neural circuit function requires understanding the roles of neuronal subtypes in behavior and disease and relies on successful and sometimes specific delivery of transgenes to the neuronal types under study (Bedbrook et al., 2018).

We further discovered that the vascular bias of AAV-PHP.V1 originates from its transduction bias towards endothelial cells. Interestingly, this is the only cell type we detected expressing Ly6a (Supplemental Figure 8), a known surface receptor for AAV variants in the PHP.B family (Batista et al., 2020; Hordeaux et al., 2019; Huang et al., 2019). Given AAV-PHP.V1's sequence similarity to AAV-PHP.B and its tropism across mouse strains, this pattern suggests that AAV-PHP.V1 transduction may also be Ly6a-mediated. Finding such associations between viral tropism and cell-surface membrane proteins also suggests that full transcriptome sequencing data may hold a treasure trove of information on possible mechanisms of transduction of viral vectors.

We also revealed that AAV-PHP.C2 has a strong, broad non-neuronal bias toward both vascular cells and astrocytes. AAV-PHP.C2 also transduces BALB/cJ mice, which do not contain the Ly6a variant that mediates transduction by PHP.B family variants (Hordeaux et al., 2019). This suggests that PHP.C2 may be the most promising candidate from this pool for researchers interested in delivery to non-neuronal cells with minimal neuronal transduction both in C57BL/6J mice and in strains and organisms that do not have the Ly6a variant.

All our tested variants with non-neuronal transduction have lower expression in Myoc+ astrocytes and pericytes. Astrocytes expressing Myoc and Gfap, which intersect in our data (Supplemental Figure 8), have been previously identified as having reactive behavior in disease contexts, making them a target of interest for research on neurological diseases (Perez-Nievas and Serrano-Pozo, 2018; Wu et al., 2017). Similarly, pericytes, whose dysfunction has been shown to contribute to multiple neurological diseases, may be an important therapeutic target (Blanchard et al., 2020; Liu et al., 2020; Montagne et al., 2020). Both of these cell types may be good candidates for further AAV optimization, but may have been missed with marker gene-based approaches. In both AAV characterization and neuroscience research efforts, different marker genes are often used for astrocyte classification – sometimes more restrictive genes such as Gfap, and other times more broadly expressing genes such as S100b or Aldh1l1 (Yang et al., 2011; Zhang et al., 2019). Similarly,

defining marker genes for pericytes is still an active field (He et al., 2016; Yang et al., 2021). Given the constraints of having to choose specific marker genes, it is difficult for staining-based characterizations to provide tropism profiles that are relevant for diverse and changing research needs. This highlights the importance of using unbiased, full transcriptome profiling for vector characterization.

We have shown that our combined experimental and computational platform is able to recover transduction biases and profile multiple variants in a single animal, even amidst the noise of ambient RNA. We have further shown that our method is robust to the variability inherent in delivery and extraction from different animals, with different transgenes, and with different regulatory elements. For example, we discovered lower overall expression from vectors carrying UBC-mCherry compared with CAG-mNeonGreen-WPRE. Such differences are not surprising since the WPRE is known to increase RNA stability and therefore transcript abundance (Johansen et al., 2003). Furthermore, the shift in cell-type bias may come from the UBC promoter, as even ubiquitous promoters such as CAG and UBC have been shown to have variable levels of expression in different cell types (Qin et al., 2010). Despite these biases, looking at the differences in transduction between variants delivering the same construct within an individual animal reveals the strongest candidate vectors for on-target and off-target cell types of interest. While we show that our method can profile AAVs carrying standard fluorescent cargo, caution is needed when linking differences in absolute viral tropism to changes in capsid composition alone without considering the contribution of the transgene and regulatory elements. Therefore, for more robust and relative tropism between variants, we found it beneficial to use small barcodes and co-injections of pools of vectors. Our scRNA-seq-based approach is not restricted to profiling capsid variants, but can be expanded in the future to screen promoters (Chuah et al., 2014; Jüttner et al., 2019; Rincon et al., 2015), enhancers (Hrvatin et al., 2019; Mich et al., 2020), or transgenes (Gustafsson et al., 2004; Shirley et al., 2020), all of which are essential elements requiring optimization to improve gene therapy.

Finally, we have used scRNA-seq to understand how intra-orbital administration of AAV-PHP.eB affects the host cell transcriptome across distinct time points. Results from our study show genes pertaining to the p53 pathway in endothelial cells are differentially expressed 3 days after injection. The highest number of differentially expressed genes being in endothelial cells suggests that vascular cells could be the initial responders to viral transduction and expression of the transgene. This is supported by Kodali et al., who have shown that endothelial cells are the first to elicit a response to peripheral inflammatory stimulation by transcribing genes for proinflammatory mediators and cytokines (Kodali et al., 2020). With regards to p53 differentially expressed genes, Ghouzzi, et al. have also shown that the genes Phdla3, Aen, and Cdkn1a were upregulated in cells infected with ZIKA virus, signifying genotoxic stress and apoptosis induction (Ghouzzi et al., 2016). Upregulation of genes such as Bax and Cdkn1a in our data hint at an initiation of apoptosis and cell cycle arrest, respectively, in response to cellular stress induced by viral transduction (Ferreira and Nagai, 2019; Zamagni et al., 2020). The reduction in the number of differential expressed genes across all cells (Figure 5 B) at day twenty-five imparts that the initial inflammatory responses did not escalate. Downregulation of Cd209a gene in perivascular macrophages in our data further implies that the AAV-PHP.eB infection did not necessitate a primary adaptive immune response. Additionally, antigen presenting genes, such as H2-D1, returning back to control expression levels and a lack of proinflammatory cytokines being upregulated supports that the event of infiltration of peripheral leukocytes is unlikely, in agreement with prior studies (Chamberlin et al., 1998; McCown et al., 1996). Upregulation of genes such as Gadd45g, Gadd45b, and Ppp1r15a suggest that neurons are turning on stress-related programs as an early response to encountering the virus. Genes such as Nr4a1 and Dusp1, which play a role in the MAPK pathway, indicate sustained stress response even at day 25. Based on prior studies, we speculate that the genes that are differentially expressed at day 25 in the excitatory neurons are due to transgene expression and not due to the virion (Lowenstein et al., 2007). It is important to note that the findings discussed here are specific to the rAAV, transgene, and dosage. Our results highlight the power of single-cell profiling in being able to ascertain cell-type-specific responses at an early time point post-injection.

In summary, our platform could aid the gene therapy field by allowing more thorough characterization of existing and emerging recombinant AAVs by helping uncover cellular responses to rAAV-mediated gene therapy, and by guiding the engineering of novel AAV variants.

2.10 Acknowledgements

We thank the Gradinaru and Thomson labs for helpful discussions, Allan-Hermann Pool for advice on the mouse brain tissue dissociation procedure, Jeff Park for advice on 10X Genomics Chromium single-cell library preparation, Min Jee Jang for help in designing probes and troubleshooting FISH-HCR, and Ben Deverman and Ken Chan for early discussions on strategy. This work was supported by the NIH Pioneer DP1OD025535, Beckman Institute for CLARITY, Optogenetics and Vector Engineering Research at Caltech, the Single-Cell Profiling and Engineering Center (SPEC) in the Beckman Institute at Caltech, and the Curci Foundation. V.G. and M.T. are Heritage Principal Investigators supported by the Heritage Medical Research Institute.

Author contributions

D.B., M.A., T.D., and V.G. conceived the project and designed the experiments. S.C. and M.T. provided critical single-cell RNA sequencing expertise. T.D., M.A., and D.B. prepared the DNA constructs and produced virus. M.A. performed the injections, tissue dissociation, histology, imaging and image quantification. D.B. and T.D. performed the single-cell library preparation and prepared samples for sequencing. D.B. and M.A. built the data processing pipeline. D.B., M.A., T.D., and A.W. performed the analysis. All authors contributed to the MS as drafted by D.B., M.A., and V.G.. M.T. supervised single-cell RNA sequencing computational pipelines while V.G. supervised the overall project.

2.11 Methods

Animals

Animal husbandry and all experimental procedures involving animals were performed in accordance with the California Institute of Technology Institutional Animal Care and Use Committee (IACUC) guidelines and approved by the Office of Laboratory Animal Resources at the California Institute of Technology (animal protocol no. 1650). Male C57BL/6J mice (Stock No: 000664) used in this study were purchased from the Jackson Laboratory (JAX). AAV variants were injected i.v. into the retro-orbital sinus of 6–7 week old mice.

Plasmids

In vivo vector characterization of AAV variant capsids was conducted using single-stranded (ss) rAAV genomes. pAAV:CAG-NLS-mNeonGreen, pAAV:CAG-NLS-mRuby2, pAAV:CAG-tdTomato, and pAAV:CAG-NLS-tdTomato constructs were adapted from previous publications (Chan et al., 2017; Ravindra Kumar et al., 2020). To introduce barcodes into the polyA region of CAG-NLS-mNeonGreen, we digested the plasmid with BglII and EcoRI, and performed Gibson assembly (E2611, NEB) to insert synthesized fragments with 7bp degenerate nucleotide sequences 89 bp upstream of the polyadenylation site. We then seeded bacterial colonies and selected and performed Sanger sequencing on the resulting plasmids to determine the corresponding barcode.

The UBC-mCherry-AAV-cap-in-cis plasmid was adapted from the rAAV-Cap-in-cis-lox plasmid from a previous publication (Deverman et al., 2016). We performed a restriction digest on the plasmid with BsmbI and SpeI to remove UBC-mCherry and retain the AAV9 cap gene and remaining backbone. We then circularized the digested plasmid using a gblock joint fragment to get a plasmid containing AAV2-Rep, AAV9-Cap, and the remaining backbone via T4 ligation. In order to insert UBC-mCherry with the desired orientation and location, we amplified its linear segment from the original rAAV-Cap-in-cis-lox plasmid.

The linear UBC-mCherry-polyA segment and circularized AAV2-Rep,AAV9-cap plasmid were then both digested with HindIII and ligated using T4 ligation. In order to get the SV40 PolyA element in the proper orientation with respect to the inserted UBC-mCherry, we removed the original segment from the plasmid using AvrII and AccI enzymes and inserted AvrII, AccI treated SV40 gblock using T4 ligation to get the final plasmid.

To insert barcodes into UBC-mCherry-AAV-cap-in-cis, we obtained 300 bp DNA fragments containing the two desired capsid mutation regions for each variant and the variant barcode, flanked by BsrGI and XbaI cut sites. The three segments of the fragment were separated by BsaI Type I restriction sites. We digested the UBC-mCherry-AAV-cap-in-cis plasmid with BsrGI and XbaI, and ligated each variant insert to this backbone. Then, to reinsert the missing regions, we performed Golden Gate assembly with two inserts and BsaI-HF.

Viral production

To produce viruses carrying *in trans* constructs, we followed established protocols for the production of rAAVs (Challis et al., 2019). In short, HEK293T cells were triple transfected using polyethylenimine (PEI) with three plasmids: pAAV (see Plasmids), pUCmini-iCAP-PHP.eB (Chan et al., 2017), pUCmini-iCAP-CAP-B10 (Flytzanis et al., 2020), or pUCmini-iCAP-PHP.V1 (Ravindra Kumar et al., 2020), and pHelper. After 120 h, virus was harvested and purified using an iodixanol gradient (Optiprep, Sigma). For our 7-variant pool, we modified the protocol to be a double transfection using PEI with two plasmids: UBC-mCherry-AAV-cap-in-cis and pHelper.

Tissue processing for single-cell suspension

Three to four weeks after the injection, mice (9-10 weeks old) were briefly anesthetized with isoflurane (5%) in an isolated plexiglass chamber followed by i.p. injection of euthasol (100 mg/kg). The following dissociation procedure of cortical tissue into a single-cell suspension was adapted with modifications from a previous report (Pool et al., 2020). Animals were transcardially perfused with ice-cold carbogenated (95% O_2 and 5% CO_2) NMDG-HEPES-

ACSF (93 mM NMDG, 2.5 mM KCl, 1.2 mM NaH_2PO_4, 30 mM $NaHCO_3$, 20 mM HEPES, 25 mM glucose, 5 mM Na L-ascorbate, 2 mM thiourea, 3 mM Na-pyruvate, 10 mM $MgSO_4$, 1 mM $CaCl_2$, 1 mM kynurenic acid Na salt, pH adjusted to 7.35 with 10N HCl, osmolarity range 300–310 mOsm). Brains were rapidly extracted and cut in half along the anterior-posterior axis with a razor blade. Half of the brain was used for IHC histology while the second half of the brain was used for scRNA-seq. Tissue used for scRNA-seq was immersed in ice-cold NMDG-HEPES-ACSF saturated with carbogen. The brain was sectioned into 300-µm slices using a vibratome (VT-1200, Leica Biosystems, IL, USA). Coronal sections from Bregma −0.94 mm to −2.80 mm were collected in a dissection dish on ice containing NMDG-HEPES-ACSF. Cortical tissue from the dorsal surface of the brain to ~3.5 mm ventral was cut out and further sliced into small tissue pieces. NMDG-HEPES-ACSF was replaced by trehalose-HEPES-ACSF (92 mM NaCl, 2.5 mM KCl, 1.2 mM NaH_2PO_4, 30 mM $NaHCO_3$, 20 mM HEPES, 25 mM glucose, 2 mM $MgSO_4$, 2 mM $CaCl_2$, 1 mM kynurenic acid Na salt, 0.025 mM D-(+)-trehalose dihydrate*$2H_2O$, pH adjusted to 7.35, osmolarity ranging 320–330 mOsm) containing papain (60 U/ml; P3125, Sigma Aldrich, pre-activated with 2.5 mM cysteine and a 0.5–1 h incubation at 34°C, supplemented with 0.5 mM EDTA) for the enzymatic digestion. Under gentle carbogenation, cortical tissue was incubated at 34°C for 50 min with soft agitation by pipetting every 10 min. 5 µl 2500 U/ml DNase I (04716728001 Roche, Sigma Aldrich) was added to the single-cell suspension 10 min before the end of the digestion. The solution was replaced with 200 µl trehalose-HEPES-ACSF containing 3 mg/ml ovomucoid inhibitor (OI-BSA, Worthington) and 1 µl DNase I. At room temperature, the digested cortical tissue was gently triturated with fire-polished glass Pasteur pipettes for three consecutive rounds with decreasing pipette diameters of 600, 300, and 150 µm. 800 µl of trehalose-HEPES-ACSF with 3 mg/ml ovomucoid inhibitor was added. The uniform single-cell suspension was pipetted through a 40 µm cell strainer (352340, Falcon) into a new microcentrifuge tube followed by centrifugation at 300 g for 5 min at 4°C. The supernatant was discarded and cell pellet was resuspended in 1 ml of trehalose-HEPES-ACSF. After mixing using a Pasteur pipette with a 150 µm tip diameter, the single-cell suspension was centrifuged again. Supernatant was replaced with fresh trehalose-HEPES-ACSF and the resuspended cell pellet was strained with a 20 µm nylon net

filter (NY2004700, Millipore). After resuspension in trehalose-HEPES-ACSF, cells were pelleted again and resuspended in 100 µl of ice-cold resuspension-ACSF (117 mM NaCl, 2.5 mM KCl, 1.2 mM NaH_2PO_4, 30 mM $NaHCO_3$, 20 mM HEPES, 25 mM glucose, 1 mM $MgSO_4$, 2 mM $CaCl_2$, 1 mM kynurenic acid Na salt and 0.05% BSA, pH adjusted to 7.35 with Tris base, osmolarity range 320–330 mOsm). Cells were counted with a hemocytometer and the final cell densities were verified to be in the range of 400–2,500 cells/µl. The density of single-cell suspension was adjusted with resuspension-ACSF if necessary.

Transcriptomic library construction

Cell suspension volumes containing 16,000 cells–expected to retrieve an estimated 10,000 single-cell transcriptomes–were added to the 10x Genomics RT reaction mix and loaded to the 10x Single Cell Chip A (230027, 10x Genomics) for 10x v2 chemistry or B (2000168, 10x Genomics) for 10x v3 chemistry per the manufacturer's protocol (Document CG00052, Revision F, Document CG000183, Revision C, respectively). We used the Chromium Single Cell 3' GEM and Library Kit v2 (120237, 10x genomics) or v3 (1000075, 10x Genomics) to recover and amplify cDNA, applying 11 rounds of amplification. We took 70 ng to prepare Illumina sequencing libraries downstream of reverse transcription following the manufacturer's protocol, applying 13 rounds of sequencing library amplification.

Viral library construction

We selectively amplified viral transcripts from 15 ng of cDNA using a cargo-specific primer binding to the target of interest and a primer binding the partial Illumina Read 1 sequence present on the 10x capture oligos (Supplemental Table 1). For animals injected with a single cargo, amplification was performed only once using the primer for the delivered cargo; for animals with distinct cargo sequences per variant, amplification was performed in parallel reactions from the same cDNA library using different cargo-specific primers for each reaction. We performed the amplification using 2x KAPA HiFi HotStart ReadyMix (KK2600) for 28 cycles at an annealing temperature of 53°C. Afterwards, we performed a left-sided SPRI cleanup with a concentration dependent on the target amplicon length, in

accordance with the manufacturer's protocol (SPRISelect, Beckman Coulter B23318). We then performed an overhang PCR on 100 ng of product with 15 cycles using primers that bind the cargo and the partial Illumina Read 1 sequence and appending the P5/P7 sequences and Illumina sample indices. We performed another SPRI cleanup, and analyzed the results via an Agilent High Sensitivity DNA Chip (Agilent 5067-4626).

Sequencing

Transcriptome libraries were pooled together in equal molar ratios according to their DNA mass concentration and their mean transcript size as determined via bioanalyzer. Sequencing libraries were processed on Novaseq 6000 S4 300-cycle lanes. The run was configured to read 150 bp from each end. Sequencing was outsourced to Fulgent Genetics and the UCSF Center for Advanced Technology.

All viral transcript libraries except barcoded UBC-mCherry were pooled together in equal molar ratios into a 4 nM sequencing library, then diluted and denatured into a 12 pM library as per the manufacturer's protocol (Illumina Document #15039740v10). The resulting library was sequenced using a MiSeq v3 150-cycle reagent kit (MS-102-3001), configured to read 91 base pairs for Read 2 and 28 base pairs for Read 1. To characterize the effect of sequencing depth, one viral transcript library was additionally processed independently on a separate MiSeq run.

The UBC-mCherry viral transcript library, which was recovered with primers near the polyadenylation site, consisted of fragments ~307 bp long. Since this length is within the common range for an Illumina NovaSeq run, this viral transcript library was pooled and included with the corresponding transcriptome library.

Transcriptome read alignment

For transcriptome read alignment and gene expression quantification, we used 10x Cell Ranger v5.0.1 with default options to process the FASTQ files from the transcriptome

sequencing library. The reads were aligned against the mus musculus reference provided by Cell Ranger (mm10 v2020-A, based on Ensembl release 98).

To detect viral transcripts in the transcriptome, we ran an additional alignment using 10x Cell Ranger v5.0. 1 with a custom reference genome based on mm10 v2020-A. We followed the protocol for constructing a custom Cell Ranger reference as provided by 10x Genomics. This custom reference adds a single gene containing all the unique sequences from our delivered plasmids in the study, delineated as separate exons. Sequences that are common between different cargo are provided only once, and annotated as alternative splicings.

Viral transcript read alignment

For viral read alignment, we aligned each Read 2 to a template derived from the plasmid, excluding barcodes. The template sequence was determined by starting at the ATG start site of the XFP cargo and ending at the AATAAA polyadenylation stop site. We used a Python implementation of the Striped Smith-Waterman algorithm from scikit-bio to calculate an alignment score for each read, and normalized the score by dividing by the maximum possible alignment score for a sequence of that length, minus the length of the barcode region. For each Read 2 that had a normalized alignment score of greater than 0.7, we extracted the corresponding cell barcode and UMI from Read 1, and any insertions into the template from Read 2.

Constructing the variant lookup table

For co-injections with multiple templates and injections of barcoded templates, we constructed a lookup table to identify which variant belongs to each cell barcode/UMI. For each template, we counted the number of reads for each cell barcode/UMI. For reads of barcoded cargo, we only counted reads where the detected insertion in the barcode region unambiguously aligned to one of the pre-defined variant barcodes. Due to sequencing and PCR amplification errors, most cell barcode/UMI combinations had reads associated with multiple variants. Thus, we identified the variant with the largest count for each cell

barcode/UMI. We discarded any cell barcode/UMIs that had more than one variant tied for the largest count. Finally, each cell barcode/UMI that was classified as a viral transcript in the transcriptome (see Transcriptome read alignment) was converted into the virus detected in the variant lookup table, or was discarded if it did not exist in the variant lookup table.

Estimating transduction rate

To determine an estimate of the percent of cells within a group expressing viral cargo above background, we compared the viral transcript counts in that group of cells to a background distribution of viral transcript counts in debris (see Droplet type classification). First, we obtained the empirical distribution of viral transcript counts by extracting the viral counts for that variant in cell barcodes classified as the target cell type as well as cell barcodes classified as debris. Next, we assumed a percentage of cells containing debris. For each viral transcript count, starting at 0, we calculated the number of cells that would contain this transcript count, if the assumed debris percentage was correct. We then calculated an error between this estimate and the number of cells with this transcript count in the cell type of interest. We tallied this error over all the integer bins in the histogram, allowing the error in a previous bin to roll over to the next bin. We repeated this for all possible values of percentage of debris from 0 to 100 in increments of 0.25, and the value that minimized the error was the estimated percentage of cells whose viral transcript count could be accounted for by debris. The inverse of this was our estimate of the number of cells expressing viral transcripts above background.

To validate that this method reliably recovers an estimate of transduction rate, we performed a series of simulations using models of debris viral transcript counts added to proposed cell type transcript count distributions across a range of parameterizations. To get estimates of the background distribution of debris, we used diffxpy (https://github.com/theislab/diffxpy) to fit the parameters of a negative binomial distribution to the viral transcript counts in debris droplets within a sample. We then postulated 1,000 different parameterizations of the negative binomial representing transcript counts in groups of cells, with 40 values of r ranging from 0.1 to 10, spaced evenly apart, and 25 values of p ranging from 0.001 to 0.99, spaced evenly apart. For each proposed negative binomial model, we drew 1,000 random

samples of viral counts from the learned background distribution, and 1,000 random samples from the proposed cell distribution, and summed the two vectors. This summed vector was then used in our transduction rate estimation function, along with a separate 1,000 random samples of background viral transcripts for the function to use as an estimate of the background signal. We calculated the true probability of non-zero expression in our proposed cell negative binomial model (1 – P(X = 0)), and compared this value with the estimated value from the transduction rate estimation method.

Calculating viral tropism

For each variant v_n and cell type of interest c_i, we estimated the percentage of cells expressing viral cargo. To calculate tropism bias, we used this estimated expression rate, t_{c_i,v_n}, to estimate the number of cells expressing viral transcripts in that cell type, T_{c_i,v_n} out of the total number of cells of that type, N_{c_i}. $T_{c_i,v_n} = t_{c_i,v_n} N_{c_i}$. Cell type bias, b_{c_i,v_n}, within a sample was then calculated as the ratio of the number of cells of interest divided by the total number of transduced cells, $b_{c_i,v_n} = \frac{T_{c_i,v_n}}{\sum_j T_{c_j v_n}}$. Finally, to calculate the difference in transduction bias for a particular variant relative to other variants in the sample, δ_{c_i,v_n}, we subtracted the bias of the variant from the mean bias across all other variants, $\delta_{c_i,v_n} = \frac{T_{c_i,v_n}}{\sum_j T_{c_j v_n}} - \frac{\sum_{m \neq n} T_{c_i,v_m}}{\sum_{m \neq n} \sum_j T_{c_j,v_m}}$.

Histology

Immunohistochemistry

The immunohistochemistry procedure was adapted from a previous publication (Oikonomou et al., 2019). Brain tissue was fixed in 4% paraformaldehyde (PFA) at 4°C overnight on a shaker. Samples were immersed in 30% sucrose in 1x phosphate buffered saline (PBS) solution for >2 days and then embedded in Tissue-Tek O.C.T. Compound (102094-104, VWR) before freezing in dry ice for 1 h. Samples were sectioned into 50 µm coronal slices on a cryostat (Leica Biosystems). Brain slices were washed once with 1x phosphate buffered saline (PBS) to remove O.C.T. Compound. Samples were then incubated overnight at 4°C

on a shaker in a 1x PBS solution containing 0.1% Triton X-100, 10% normal goat serum (NGS; Jackson ImmunoResearch, PA, USA), and primary antibodies. Sections were washed three times for 15 min each in 1x PBS. Next, brain slices were incubated at 4°C overnight on a shaker in a 1x PBS solution containing 0.1% Triton X-100, 10% NGS, and secondary antibodies. Sections were washed again three times for 15 min each in 1x PBS. Finally, slices were mounted on glass microscope slides (Adhesion Superfrost Plus Glass Slides, #5075-Plus, Brain Research Laboratories, MA, USA). After the brain slices dried, DAPI-containing mounting media (Fluoromount G with DAPI, 00-4959-52, eBioscience, CA, USA) was added before protecting the slices with a cover glass (Cover glass, #4860-1, Brain Research Laboratories, MA, USA). Confocal images were acquired on a Zeiss LSM 880 confocal microscope (Zeiss, Oberkochen, Germany). The following primary antibodies were used: rabbit monoclonal to NeuN (Rbfox3) (1:500; ab177487; Abcam, MA, USA), rabbit monoclonal to S100 beta (1:500; ab52642; Abcam, MA, USA), and rabbit monoclonal to Olig2 (1:500; ab109186; Abcam, MA, USA). The following secondary antibody was used: goat anti-rabbit IgG H&L Alexa Fluor 647 (1:500; ab150079; Abcam, MA, USA).

Fluorescent in situ hybridization chain reaction

FISH-HCR was conducted as previously reported (Patriarchi et al., 2018). Probes targeting neuronal markers were designed using custom-written software (https://github.com/GradinaruLab/HCRprobe). Probes contained a target sequence of 20 nucleotides, a spacer of 2 nucleotides, and an initiator sequence of 18 nucleotides. Criteria for the target sequences were: (1) a GC content between 45%–60%, (2) no nucleotide repeats more than three times, (3) no more than 20 hits when blasted, and (4) the ΔG had to be above –9 kcal/mol to avoid self-dimers. Last, the full probe sequence was blasted and the Smith-Waterman alignment score was calculated between all possible pairs to prevent the formation of cross-dimers. In total, we designed 26 probes for *Gad1*, 20 probes for *Vip*, 22 probes for *Pvalb*, 18 probes for *Sst,* and 28 probes for *Slc17a7*. Probes were synthesized by Integrated DNA Technologies.

Droplet type identification

scRNA-seq datasets were analyzed with custom-written scripts in Python 3.7.4 using a custom fork off of scVI v0.8.1, and scanpy v1.6.0. To generate a training dataset for classifying a droplet as debris, multiplets, neuronal, or non-neuronal cells, we randomly sampled cells from all 27 cortical tissue samples. We sampled a total of 200,000 cells, taking cells from each tissue sample proportional to the expected number of cells loaded into the single-cell sequencing reaction. Within each sample, cells were drawn randomly, without replacement, weighted proportionally by their total number of detected UMIs. For each sample, we determined a lower bound on the cutoff between cells and empty droplets by constructing a histogram of UMI counts per cell from the raw, unfiltered gene count matrix. We then found the most prominent trough preceding the first prominent peak, as implemented by the scipy peak_prominences function. We only sampled from cells above this lower bound. Using these sampled cells, we trained a generative neural network model via scVI with the following parameters: 20 latent features, 2 layers, and 256 hidden units. These parameters were chosen from a coarse hyperparameter optimization centered around the scVI default values (Supplemental Table 3). We included the sample identifier as the batch key so that the model learned a latent representation with batch correction.

After training, Leiden clustering was performed on the learned latent space as implemented by scanpy. We used default parameters except for the resolution, which we increased to 2 to ensure isolation of small clusters of cell multiplets. Using the learned generative model, we draw 5000 cells from the posterior distribution based on random seed cells in each cluster. We draw an equal number conditioned on each batch. From these samples, we then calculated a batch-corrected probability of each cluster expressing a given marker gene (see Cluster marker gene determination). For this coarse cell typing, we chose a single marker gene for major cell types expected in the cortex (Supplemental Table 2). If a cluster was expressing the neuron marker gene Rbfox3, it was labeled as “Neurons”. If a cluster was expressing any of the other non-neuronal marker genes, it was labelled as “Non-neurons”. Next, we ran Scrublet on the training cells to identify potential multiplets. Scrublet was run on each sample independently, since it is not designed to operate on combined datasets with potential batch-specific confounds. We then calculated the percentage of droplets in each

cluster of the combined data that were identified as multiplets by Scrublet. We found a percentage threshold for identifying a cluster as containing predominantly multiplets by using Otsu's threshold, as implemented by scikit-image. All droplets in any cluster above the multiplet percentage threshold were labelled as "Multiplets". All other clusters were labelled as "Debris".

Next, we trained a cell-type classifier using scANVI on the droplets labeled as training data. We used the weights from the previously trained scVI model as the starting weights for scANVI. Rather than using all cells for every epoch of the trainer, we implemented an alternative sampling scheme that presented each cell type to the classifier in equal proportions. Once the model was trained, all cells above the UMI lower noise bound were run through the classifier to obtain their cell-type classification. Droplets classified as "Neurons" or "Non-neurons" were additionally filtered by their scANVI-assigned probability. We retained only cells above an FDR threshold of 0.05, corrected for multiple comparisons using the Benjamini-Hochberg procedure. Finally, since the original run of Scrublet for multiplet detection was performed on only the training data, and thus did not take advantage of all the cells available, we ran Scrublet on all droplets classified as cells, and removed any identified multiplets.

Cluster marker gene determination

To identify which clusters are expressing marker genes, we determined an estimated probability of a marker gene being expressed by a random cell in that cluster. For each cluster, we randomly sampled 5,000 cells, with replacement. We used scVI to project each cell into its learned latent space, and then used scVI's posterior predictive sampling function to generate an example cell from this latent representation, and tallied how many times the gene is expressed. We repeated this for each batch, conditioning the posterior sample on that batch, to account for technical artifacts such as sequencing depth. Once we obtained a probability of expression of a marker gene for each cluster, we find a threshold for expression using Otsu's method, as implemented by scikit-image. Clusters that have a probability of expression above the threshold are considered positive for that marker gene.

Neuronal subtype classification

Cells classified as neurons were further subtyped using annotations from a well-curated reference dataset. We used the Mouse Whole Cortex and Hippocampus 10x dataset from the Allen Institute for Brain Science as our reference dataset (Yao et al., 2021). First, we filtered the reference dataset to contain only cell types that are found within the brain regions collected for our experiments. To ensure that, overall, enough cells per cell type were present in our datasets, we merged cell types with common characteristics, such as expression of key marker genes. We re-aligned our cell transcriptome reads to the same pre-mRNA reference used to construct the reference dataset, so that the gene count matrices had a 1:1 mapping. We then trained a joint scANVI model with all cells identified as neurons from our samples and the reference database to learn a common latent space between them. The model was trained to classify cells based on the labels provided in the reference dataset. Cells were sampled from each class in equal proportions during training. After the model was trained, all neurons from our sample were run through the model to obtain their cell type classification.

Non-neuronal subtype classification

Cells classified as non-neuronal were further subtyped using automatic clustering and marker gene identification. We trained an scVI model using only the non-neuronal cells and performed Leiden clustering as implemented by scanpy on the latent space. We determined which clusters were expressing each of 31 marker genes across 13 cell subtypes. Marker genes were identified from a review of existing scRNA-seq, bulk RNA-seq, or IHC studies of mouse brain non-neuronal subtypes (Supplemental Table 2). Each cluster was assigned to a cell subtype if it was determined positive for all the marker genes for that cell subtype (see Cluster marker gene determination). If a cluster contained all the marker genes for multiple cell subtypes, the cluster was assigned to the cell subtype with the greatest number of marker genes. Clusters that did not express all the marker genes for any cell subtype were labeled as “Unknown”. Clusters that expressed all the marker genes for multiple cell subtypes with the same total number of marker genes were labeled as “Multiplets”. For cell types that contained

multiple clusters, we then calculated the probability of every gene being zero in each cluster (see Cluster marker gene determination). We then compared gene presence between clusters of the same cell type to see if there were any subclusters that had a dominant marker gene (present in > 50% of samples), that was not present in any of the other clusters (< 10% of samples). For the three cell types that had unique marker genes, we named the cluster after the gene with the highest 2-proportion z-score between the sampled gene counts in that cluster vs the rest.

Quantification of images

Quantitative data analysis of confocal images was performed blind with regard to AAV capsid variant. Manual quantification was performed using the Cell Counter plugin, present in the Fiji distribution of ImageJ (National Institutes of Health, Bethesda, MD) (Schindelin et al., 2012). Transduction rate was calculated as the total number of double positive cells (i.e. viral transgene and cell type marker) divided by the total number of cell type marker labeled cells. For each brain slice, at least 100 cells positive for the gene markers of interest were counted in the cortex.

Differential expression

To calculate differential expression within cell types between groups of animals, we used the DESeq2 R package (Love et al., 2014). For each cell type, the gene counts are summed across all cells of that type and treated as a pseudo-bulk sample. The summed gene counts from each animal are then included as individual columns for a DESeq2 differential expression analysis. We performed 3 DPI DE and 25 DPI separately, testing each sample against saline-injected controls. For each cell type, only genes that were present in all samples of at least one condition are included.

Marker gene dot plots

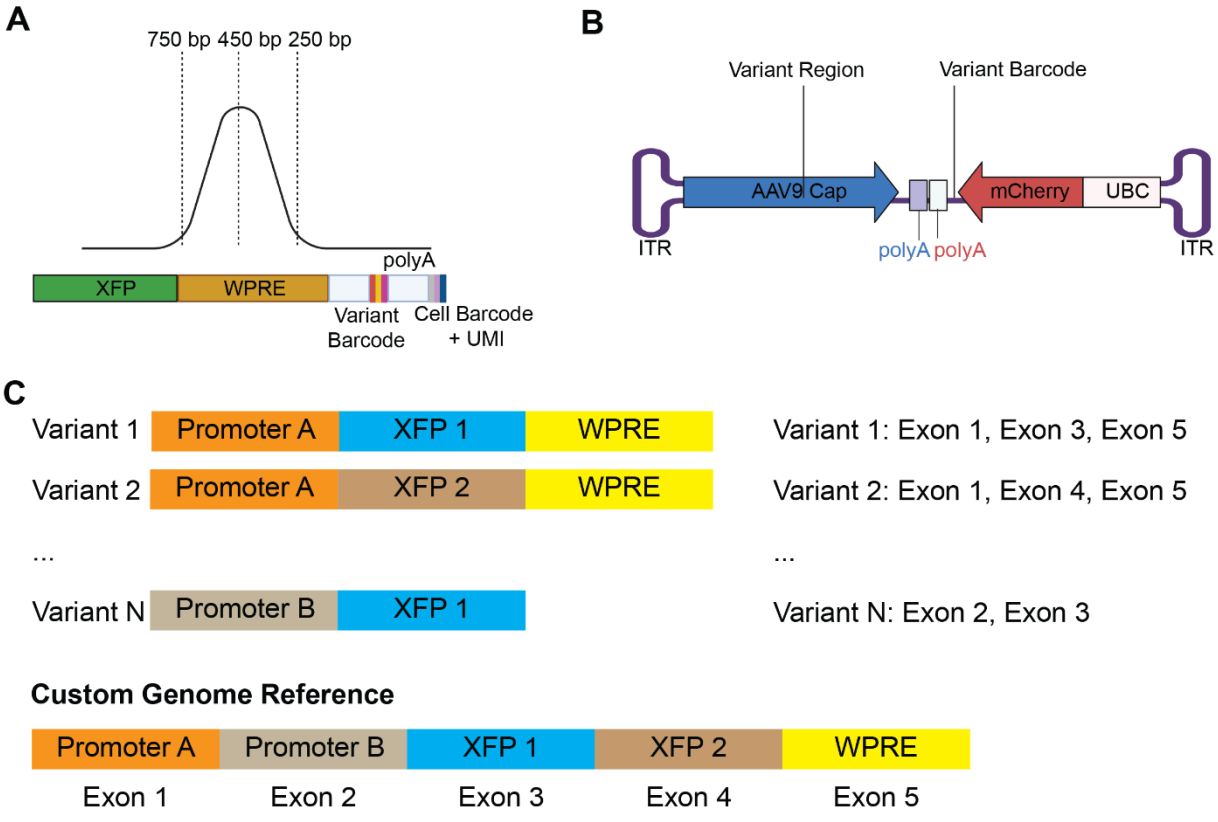
To generate dot plots for marker genes, we used scanpy's dotplot function (Wolf et al., 2018). Gene counts were normalized to the sum of the total transcript counts per cell using scanpy's

normalize_total function. Normalized gene expression values are log-transformed as part of the plotting function.

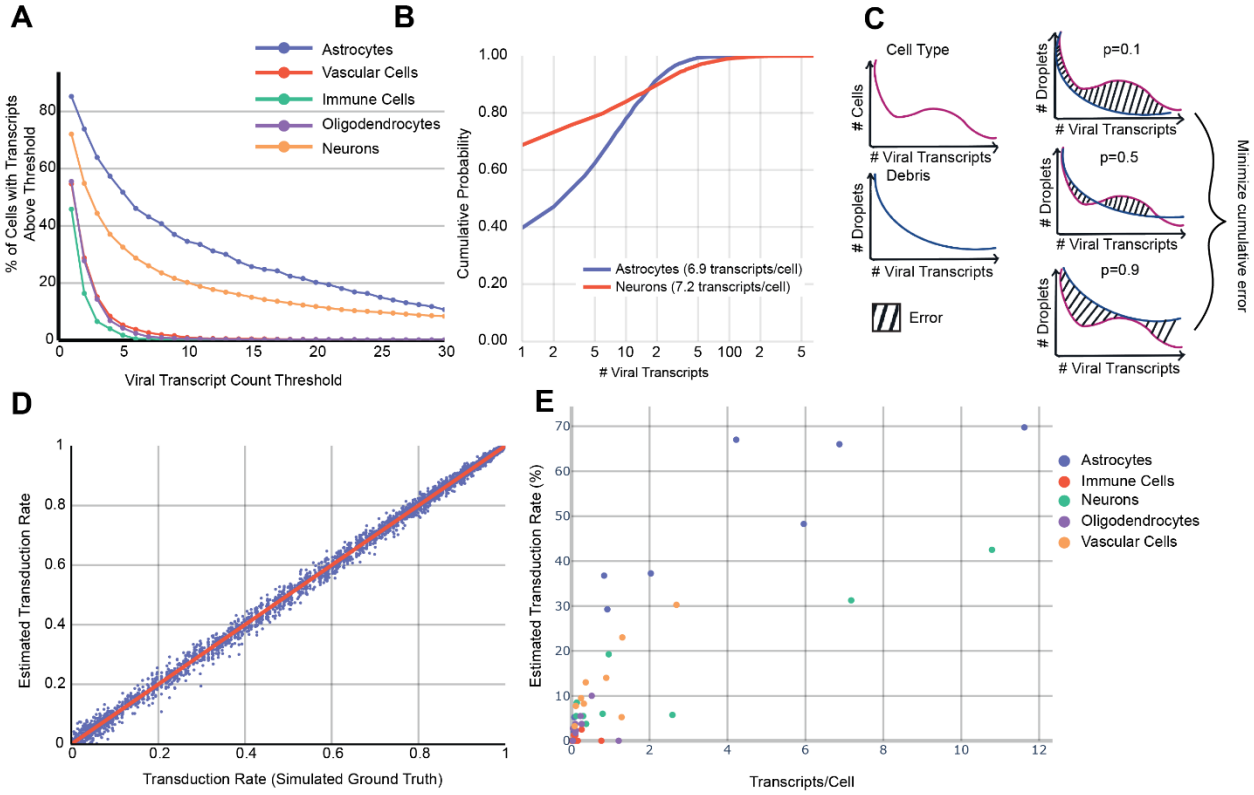
Statistics

Statistical analyses comparing the fraction of transduced cells in different cell types for Figures 2, 3, and 4 C were conducted using GraphPad Prism 9. Statistical analyses comparing proportions of transduced cells within an animal in Figure 4 E and Figure 6 were performed using the Python statsmodels library v0.12.1. No statistical methods were used to predetermine sample sizes. The statistical test applied, sample sizes, and statistical significant effects are reported in each figure legend. The significance threshold was defined as a = 0.05.

2.12 Supplemental Material

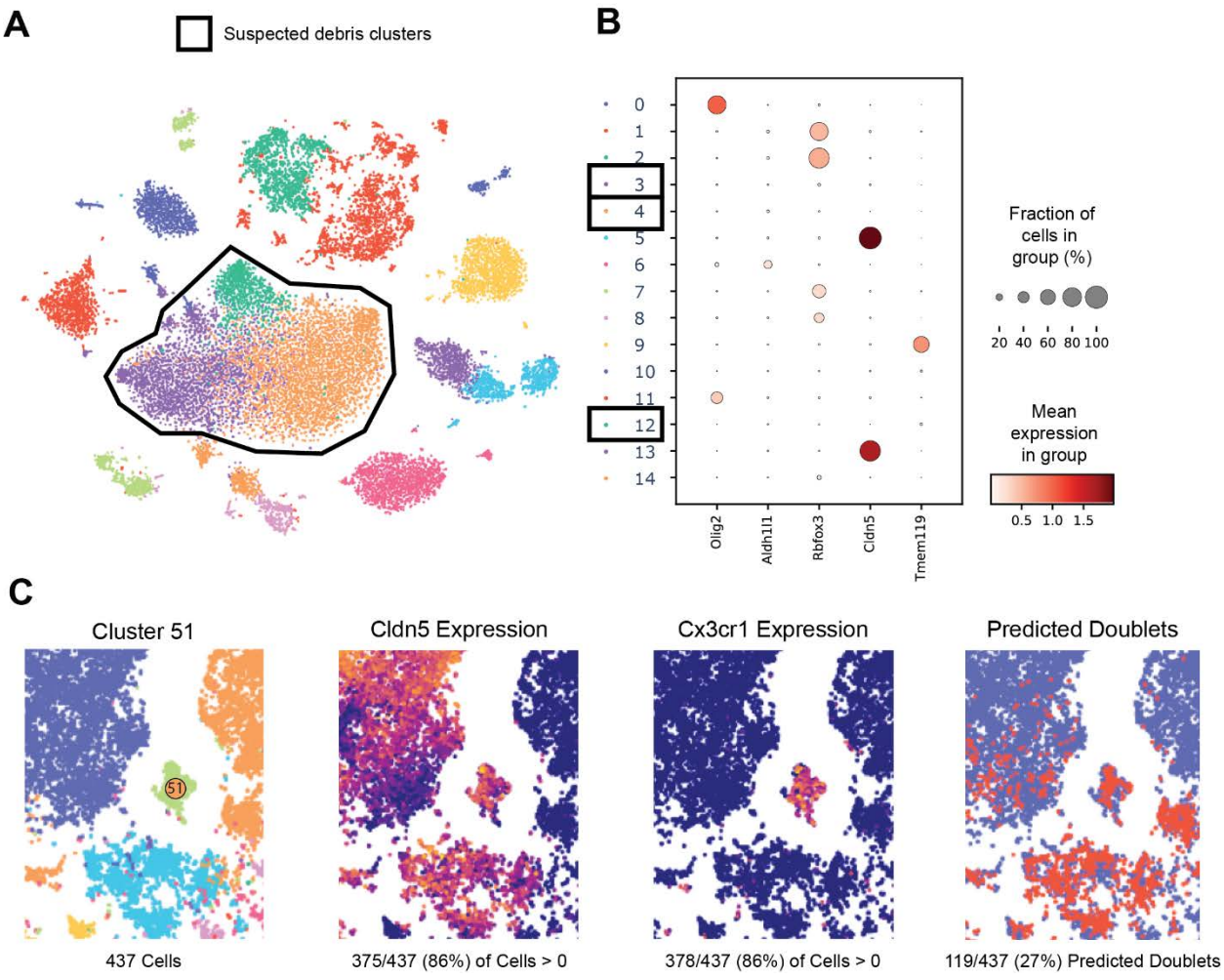


Supplemental Figure 1. Plasmid details. *(A) Size of typical transcriptome cDNA library post-fragmentation. Both distinguishing XFPs and variant barcodes fall outside the typical capture region of single-cell RNA sequencing workflows. (B) UBC-mCherry-AAV-cap-in-cis plasmid used for 7-variant barcoded pool. (C) Visualization of the construction procedure for the custom genome reference. Variant cargos are segmented into common and uncommon regions, and each unique segment is concatenated together as a contiguous gene. Variants are defined as different splicings of the custom AAV gene.*



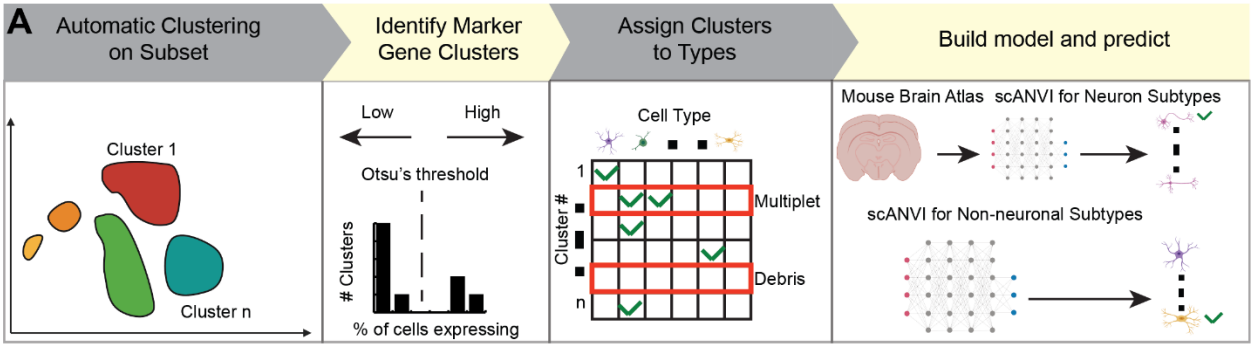
Supplemental Figure 2. Expression rate estimation. *(A) Percent of cells expressing AAV-PHP.eB cargo transcripts above a fixed threshold in a single sample. (B) An example of the distribution of viral transcript counts in a single animal from AAV-PHP.eB carrying CAG-mNeonGreen-WPRE in neurons and astrocytes. (C) Visualization of our expression-rate estimation algorithm. The distribution of the cell type of interest and background debris is*

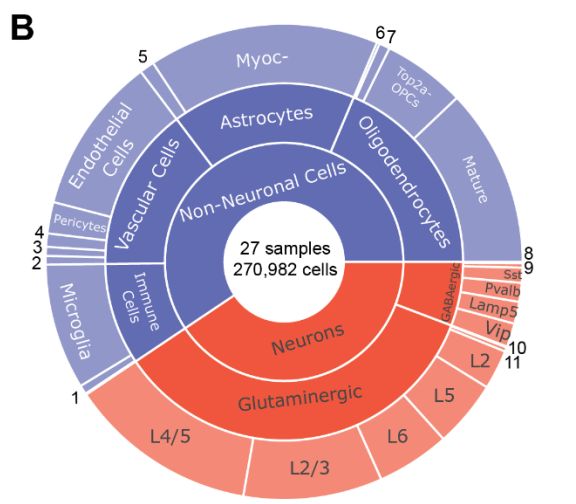
obtained. An error is calculated for different estimates of the percent of the cells that express background levels of transcripts. This error is minimized to find the best fit. (D) Performance of the expression rate estimation algorithm on simulated data consisting of negative binomial distributions with parameters r between 0.1 and 10 and p between 0.001 and 0. 99, spaced evenly apart. (E) Comparison between mean transcripts/cell (x) and the estimated transduction rate (y) in major cell types for AAV-PHP.eB across 9 samples.



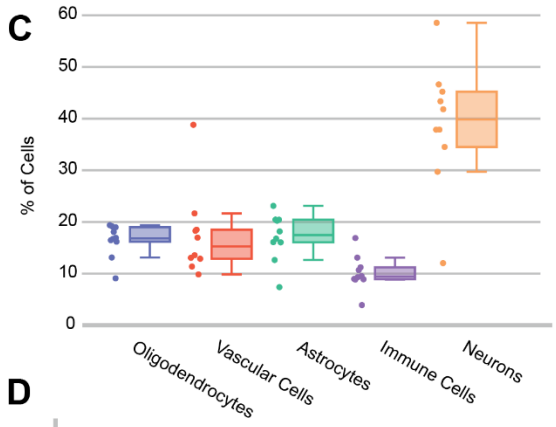
Supplemental Figure 3: Noise from debris and doublets. *(A) An example of a Cell Ranger filtered dataset. This is a t-SNE projection of the log-normalized gene expression space. Suspected debris clusters are outlined. (B) Marker gene expression for the major cell types in the brain—Oligodendrocytes/Olig2, Astrocytes/Aldh1l1, Neurons/Rbfox3, Vascular*

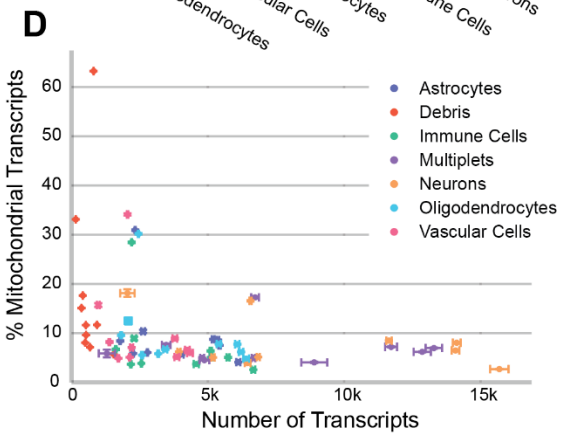
Cells/Cldn5, Immune Cells/Tmem119—for each cluster. Darker colors indicate higher mean expression, and dot size correlates with the abundance of the gene in that cluster. (C) An example of a multiplet cluster from the joint scVI space of all training samples, projected via t-SNE. Cluster 51 is annotated, and raw gene expression of Cldn5 and Cx3cr1 are shown. The percentage of cells in cluster 51 expressing each marker gene is displayed. (right) Predicted doublets from Scrublet are overlaid in red.



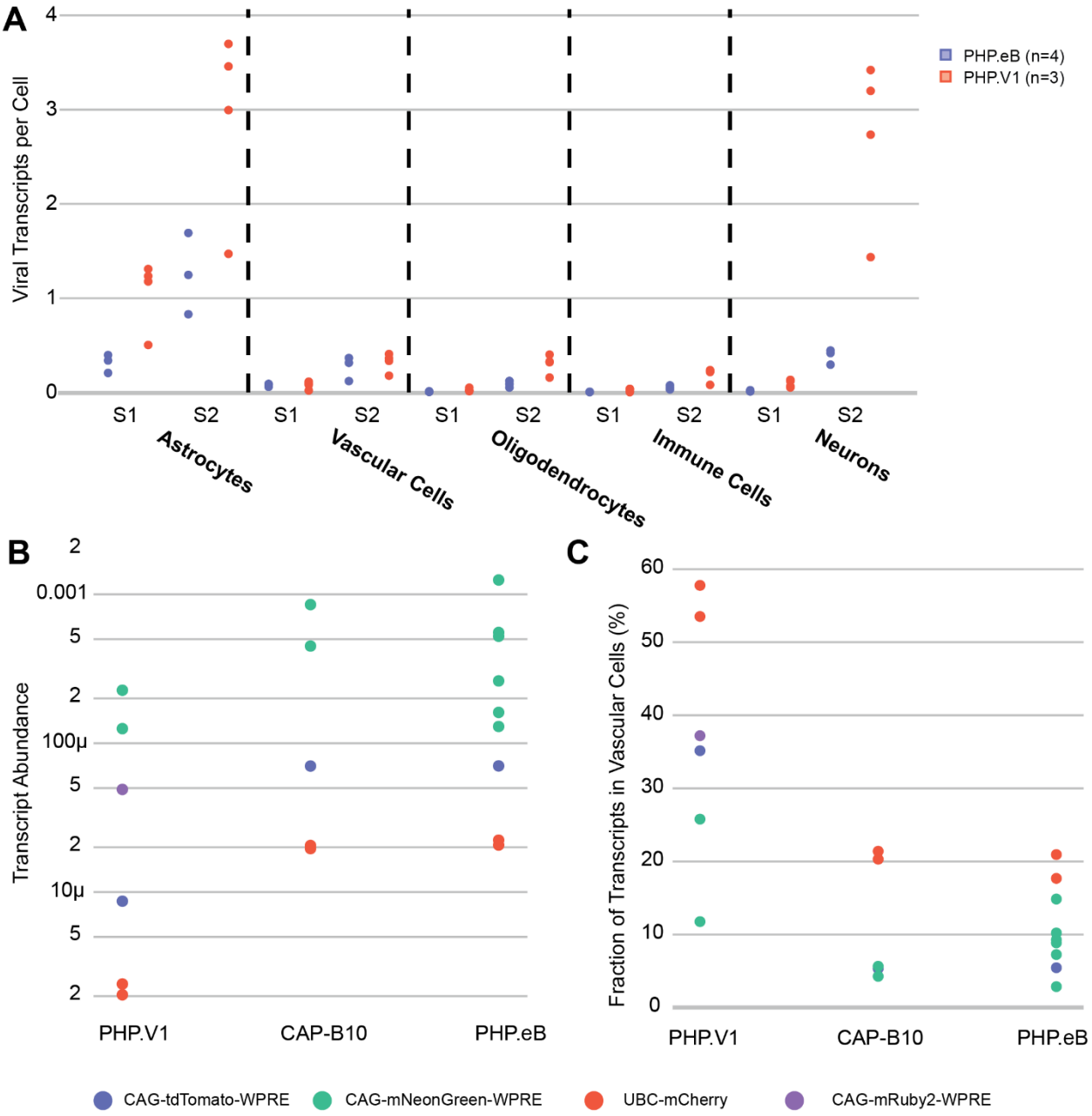


1. Perivascular Macrophages
2. Red Blood Cells
3. Vascular SMCs
4. VLMCs
5. Myoc+
6. Top2a+ OPCs
7. Committed Oligodendrocytes
8. Pax6
9. Sncg
10. L3
11. L5/6

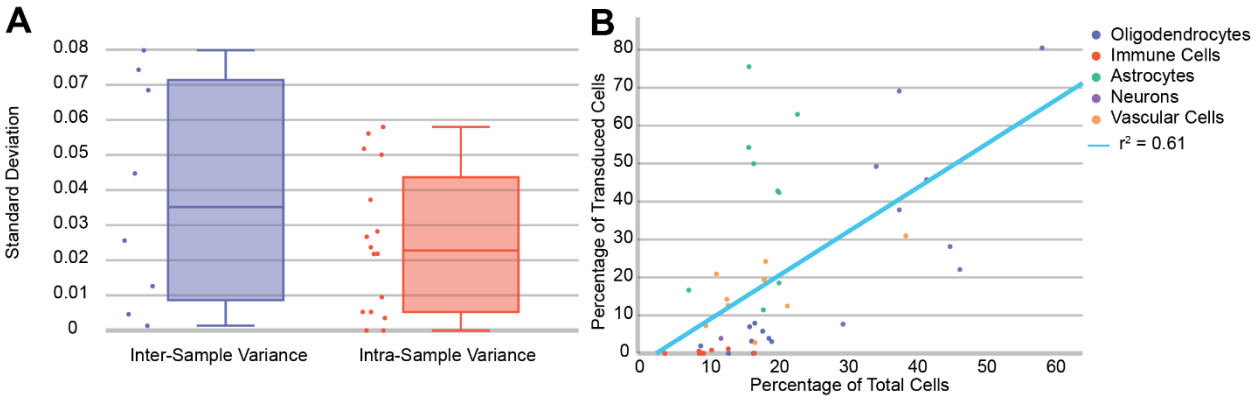




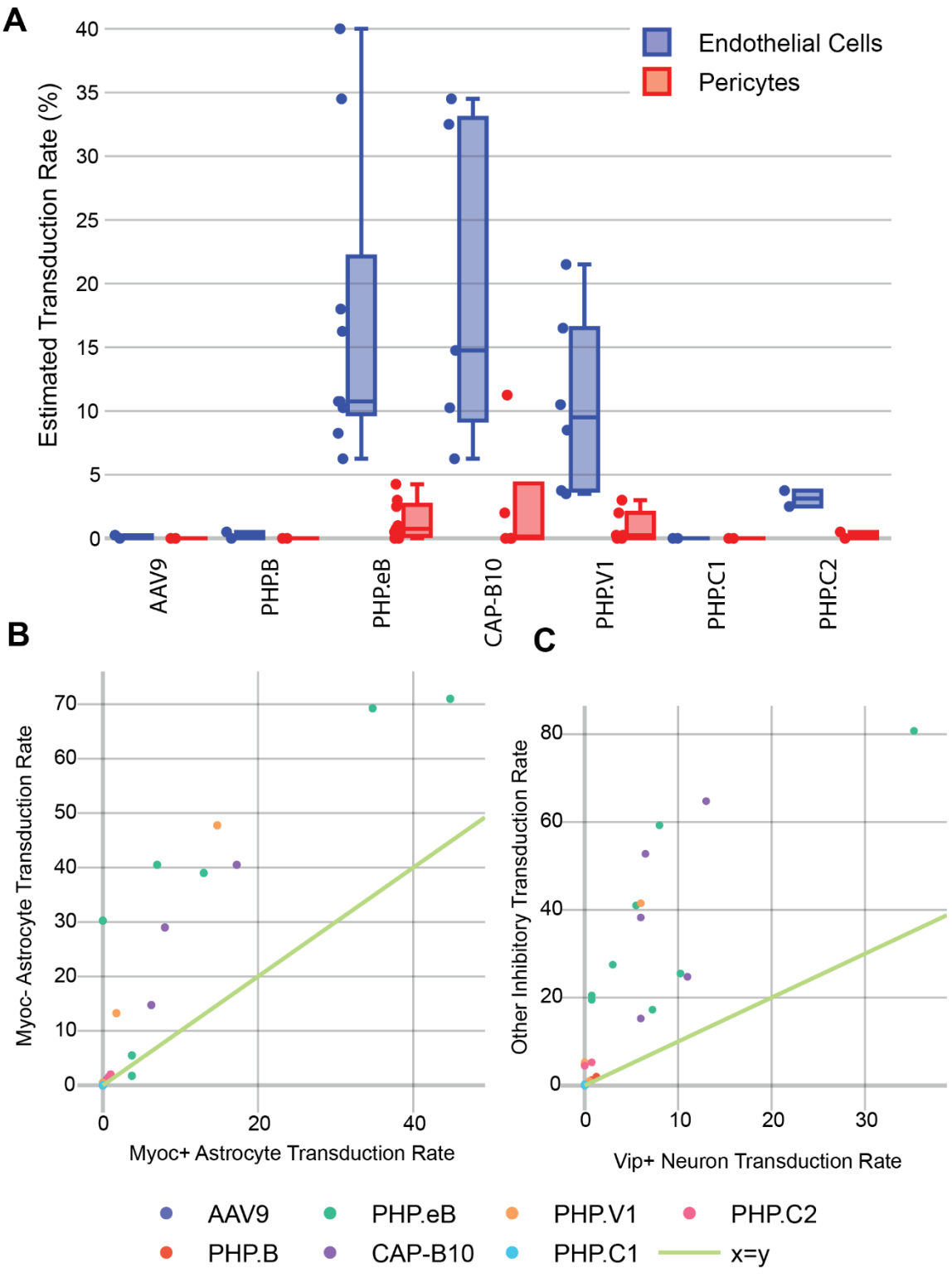
Supplemental Figure 4. Cell typing. *(A) Cell typing workflow. A subset of cells are used for training. For each marker gene, clusters expressing that marker gene are identified. Clusters that have no marker genes (debris) or are determined to be multiplets via Scrublet are marked for removal. Training data used to train a scANVI model to predict the remaining cells. A reference database can be used instead of manually labeled cells, as we did for neuronal subtypes. (B) Cell-type distribution of all identified cells from our combined cell-type taxonomy. This includes samples described in the study as well as additional controls and animals used for troubleshooting and prototyping. (C) Cell-type percentages across the major cell types in the ten samples used for AAV tropism characterization. One of the samples, BC1, had dramatically fewer neurons than any other sample and correspondingly higher percentages of non-neurons. (D) Mitochondrial gene ratio and total transcript counts of the major cell type clusters in the ten samples used for tropism characterization.*



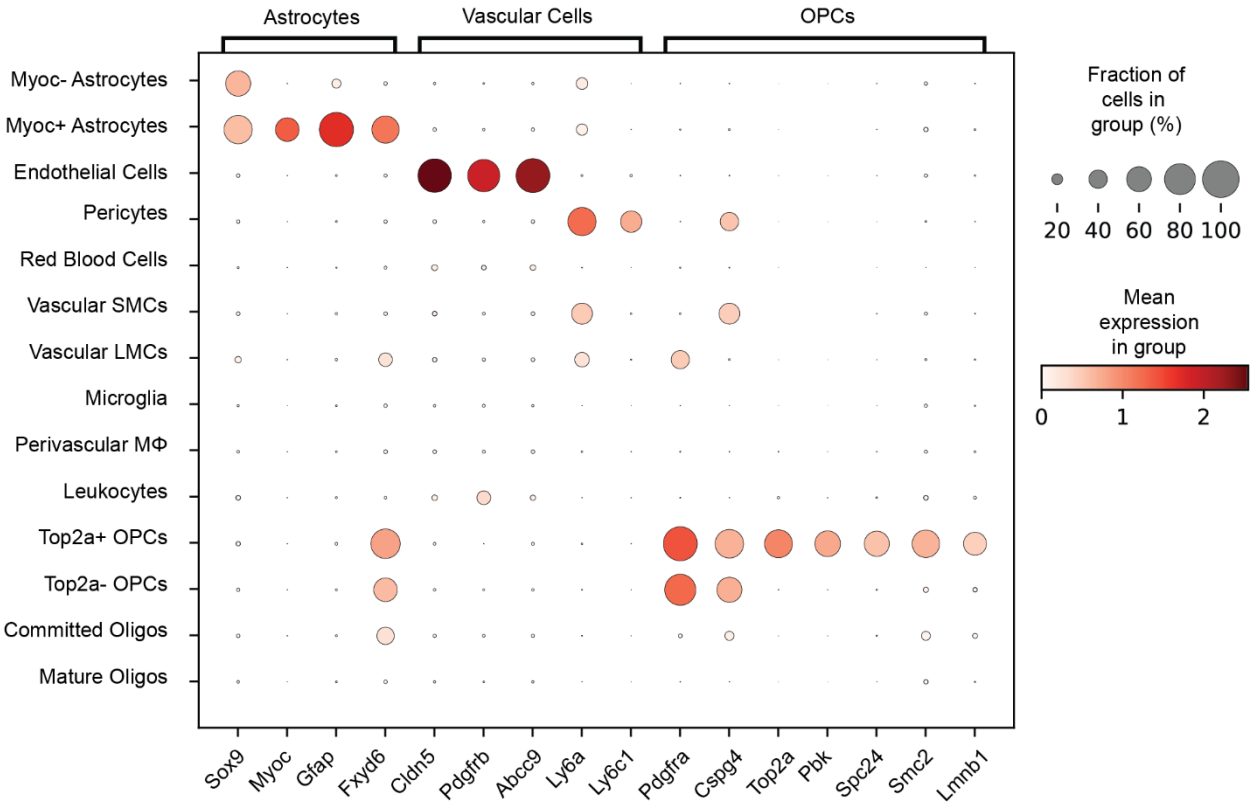
Supplemental Figure 5. Transcript expression. *(A) Viral transcript expression of different barcodes across two samples (S1, S2). Each point is a distinct barcode. (B) Viral transcript abundance in entire samples (viral transcripts / total transcripts) across different variants carrying different cargo. (C) Fraction of transcripts detected in vascular cells vs all other cell types.*



Supplemental Figure 6. Inter-sample variability. *(A) The standard deviations between measurements of the fraction of transduced cells in all major non-neuronal cell types in AAV-PHP.V1 and AAV-PHP.eB. Inter-sample variance (left) refers to the standard deviation between animals, and intra-sample variance (right) refers to the standard deviation between barcodes within the same animal. (B) The distribution of recovered cell types compared to the distribution of transduced cells across nine samples injected with AAV-PHP.eB.*


A
Endothelial Cells
Pericytes
Estimated Transduction Rate (%)
40
35
30
25
20
15
10
5
0
AAV9
PHP.B
PHP.eB
CAP-B10
PHP.V1
PHP.C1
PHP.C2
B
Myoc- Astrocyte Transduction Rate
70
60
50
40
30
20
10
0
0
20
40
Myoc+ Astrocyte Transduction Rate
C
Other Inhibitory Transduction Rate
80
60
40
20
0
0
10
20
30
Vip+ Neuron Transduction Rate
AAV9
PHP.B
PHP.eB
CAP-B10
PHP.V1
PHP.C1
PHP.C2
x=y

Supplemental Figure 7. Cell subtype inspection. *(A) Estimated transduction rate of endothelial cells vs pericytes across all samples and variants. (B) Pairwise transduction rate of Myoc+ and Myoc- astrocytes across all variants and samples. Each point is a single variant in a different sample. (C) Pairwise transduction rate of Vip+ neurons vs all other inhibitory neurons across all variants and samples.*



Supplemental Figure 8. Cell subtype markers. *Gene expression of additional marker genes for astrocyte and OPC subtypes.*

Supplemental Table 1. Primers. *Primers used for round 1 and round 2 amplification of viral transcripts. Primers with TC1 and TC2 in the amplicon name indicate they were used only for those samples.*

Amplicon	Read	Round	Sequence (Ns indicate Illumina sample index)
All Viruses	1	1	CTACACGACGCTCTTCCGATCT

All Viruses	1	2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT
mNeonGreen TC1	2	1	TTCAAGGAGTGGCAAAAGGCCTTTACCGATGTGAT
mRuby2	2	1	CAACGGGAACATGCAGTTGCCAAGTTTGCTGG
mNeonGreen	2	1	TAACTATCTGAAGAACCAGCCGATGTAC
tdTomato TC2	2	1	AGGACTACACAATTGTCGAACAGTATGAG
tdTomato	2	1	ACAACGAGGACTACACCATCGTGG
mCherry	2	1	CATCGTGGAACAGTACGAACG
WPRE	2	2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGACGAGTCGGATCTCCCT
mNeonGreen	2	2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTCAAGGAGTGGCAAAAGGC
mRuby2	2	2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAACGGGAACATGCAGTTGC
tdTomato TC2	2	2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCATGGACGAGCTGTACAAG
tdTomato	2	2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTCTTTCTCTATGGGATGGATGA
mCherry	2	2	CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGGCATGGACGAGCTGT

Supplemental Table 2. Marker Genes.

Cell Type	Marker Gene(s)
Astrocytes	Aldh1l1 *(Cahoy et al., 2008)*, Sox9 *(Sun et al., 2017)*
Neurons	Rbfox3 *(Lin et al., 2016)*
Vascular Cells	Cldn5 *(Song et al., 2020)*
Endothelial Cells	Slc2a1 *(Veys et al., 2020)*
Pericytes	Pdgfrb *(Winkler et al., 2010)*, Rgs5, Abcc9 *(He et al., 2016)*
Red Blood Cells	Hba-a1, Hba-a2 *(Capellera-Garcia et al., 2016)*
Vascular SMCs	Acta2, Myh11, Tagln *(Chasseigneaux et al., 2018)*
Vascular LMCs	Fam180a, Slc6a13, Dcn, Ptgds *(Marques et al., 2016)*
Microglia	Cx3cr1, Tmem119 *(Jordão et al., 2019)*
Leukocytes	Itgal, Gzma *(Huang and Sabatini, 2020)*
Perivascular Macropages	Mrc1 *(Jordão et al., 2019)*
Oligodendrocytes	Olig2 *(Dai et al., 2015)*
OPCs	Pdgfra, Cspg4 *(Suzuki et al., 2017)*
Mature Oligos	Mog, Mbp *(Miron et al., 2011)*
Committed Oligos	Ptprz1, Bmp4, Nkx2-2, Vcan *(Marques et al., 2016)*

Supplemental Table 3. scVI Hyperparameter Tuning

Dispersion	Latent Lib Size	# Latent	# Layers	# Hidden	Test KL Divergence
Gene	False	10	1	128	5366.4

Gene-batch	True	10	1	128	5406.1
Gene	True	10	1	128	5391.0
Gene-batch	True	50	2	512	5362.6
Gene-batch	False	10	1	128	5378.7
Gene	False	25	1	128	5354.6
Gene	False	25	2	256	5337.8
Gene	False	20	2	256	5336.7
Gene	False	40	4	1024	5338.0

Supplemental Table 4. Sample Metadata. *Supplemental file contains the following fields.*

Field Name	***Description***
10X Version	Whether the sample was processed using 10X V2 or V3 chemistry
Animal ID	A unique animal identifier. Some animals provided multiple samples
Target # Cells	The target number of cells for extraction. 1.6X this number is loaded into the 10X Chromium instrument
# Recovered Cells	The number of cells recovered, after debris and multiplet filtering
Cell Ranger # Cells	The number of cells as predicted by Cell Ranger
Predicted Multiplets	The number of predicted multiplets
Transcriptome Sequencing Depth	The number of reads

Transcriptome Reads/Cell	The number of reads divided by the number of recovered cells
Median UMIs/Cell	Of the recovered cells, the median total UMI count
Median Genes/Cell	Of the recovered cells, the median number of genes detected with at least one transcript
Variants Recovered	Which variants were recovered from this sample. Samples labeled “Cell Typing Only” were not used for tropism analysis, but were included in the cell type classifier
Virus Sequencing Depth	The number of reads of the amplified viral transcripts across all templates
Virus Reads/Cell	The read depth of the amplified viral transcripts
Age at Extraction (Days)	The age of the animal at extraction time
Virus Incubation Time (Days)	How many days prior to extraction the animal was injected
Percent of Virus UMIs Determined	What percent of transcriptome reads that aligned to the virus gene were disambiguated from the amplified lookup table

Supplemental Table 5. Variant Barcodes

Variant	Cargo	Barcodes
AAV-PHP.eB	pAAV:CAG-NLS-mNeonGreen	
AAV-PHP.V1	pAAV:CAG-NLS-mRuby2	
AAV-PHP.eB	pAAV:CAG-NLS-mNeonGreen	CCTGACA, GGACAGA, GCACAGA, CGAGAGA
AAV-PHP.V1	pAAV:CAG-tdTomato	

AAV-PHP.V1	pAAV:CAG-NLS-mNeonGreen	CAGTGTC, GAGAGTG, GTGTGAG
AAV-CAP-B10	pAAV:CAG-NLS-mNeonGreen	
AAV-PHP.eB	pAAV:CAG-NLS-tdTomato	
AAV-CAP-B10	pAAV:CAG-NLS-tdTomato	
AAV9	UBC-mCherry-AAV-cap-in-cis	CGTCTCAGCTATAACTTCCAA CGAGGTCGTAAGGTCGGCATT TGATTATCATGCCTGCTCAGG
AAV-PHP.B	UBC-mCherry-AAV-cap-in-cis	TATACCCAACCACTCAGTCCC CGGTTTTAGCACGGCCATAGA AAGCGATGTCTCTACACGATA
AAV-PHP.eB	UBC-mCherry-AAV-cap-in-cis	TACAGCTTTTTGACTGGAGGT CTGGCATTAATACGCGGGTCA TACAGGTCCTAGACAGGTGAT
AAV-CAP-B10	UBC-mCherry-AAV-cap-in-cis	GCTGGGCGTTAAAGTACTCGC GCAACTGGGATAATCGTAGTC AACGGAGTGAACGGACCCTAG
AAV-PHP.V1	UBC-mCherry-AAV-cap-in-cis	GTGGCGGGTTTCCGAAAAAGT TCGTCGGCACTCTCTTAGAGC CATGTGATAGTGAAGCACGCC
AAV-PHP.C1	UBC-mCherry-AAV-cap-in-cis	TCTGTGCTGCTCTTCTAACAA TCTGACGGCGGGTAAACACTG TGGCCACCCGCAGAGTATACT
AAV-PHP.C2	UBC-mCherry-AAV-cap-in-cis	GACTAGGGTAAGTGAGCTATG CGAATTTCTTCCATACCTCCT TAGTGCCAACAACGGAGAAGA

Supplemental Table 6. Differentially Expressed Genes. *Supplemental file contains one tab for astrocytes, pericytes, and OPCs, with the following fields.*

Field Name	Description
Gene ID	The Ensembl Gene ID
Gene name	The canonical gene name
P Non Zero	The probability of a cell expressing this gene in this cluster
P Non Zero Rest	The probability of a cell expressing this gene in the other cell subtype clusters

Supplemental Table 7. Differentially Expressed Genes Across Time Points. *Supplemental file contains one tab per cell type, with the following fields.*

Field Name	Description
Gene ID	The Ensembl Gene ID
Gene name	The canonical gene name
Mean expression	The mean expression of this gene in this group
L2FC	The log fold change of this gene
L2FC SE	The standard error of the L2FC
Stat	The stat, as reported by DESeq2
P-value	The unadjusted P-value
Adjusted P-value	The adjusted P-value

2.13 References

Adachi, K., Enoki, T., Kawano, Y., Veraz, M., and Nakai, H. (2014). Drawing a high-resolution functional map of adeno-associated virus capsid by massively parallel sequencing. Nat. Commun. *5*, 3075.

Amirinejad, R., Rezaei, M., and Shirvani-Farsani, Z. (2020). An update on long intergenic noncoding RNA p21: a regulatory molecule with various significant functions in cancer. Cell Biosci *10*, 82.

Arruda, V.R., Fields, P.A., Milner, R., Wainwright, L., De Miguel, M.P., Donovan, P.J., Herzog, R.W., Nichols, T.C., Biegel, J.A., Razavi, M., et al. (2001). Lack of germline transmission of vector sequences following systemic administration of recombinant AAV-2 vector in males. Mol. Ther. *4*, 586–592.

Baker, B.J., Akhtar, L.N., and Benveniste, E.N. (2009). SOCS1 and SOCS3 in the control of CNS immunity. Trends in Immunology *30*, 392–400.

Bartlett, J.S., Kleinschmidt, J., Boucher, R.C., and Samulski, R.J. (1999). Targeted adeno-associated virus vector transduction of nonpermissive cells mediated by a bispecific F(ab'γ) 2 antibody. Nat. Biotechnol. *17*, 181–186.

Batista, A.R., King, O.D., Reardon, C.P., Davis, C., Shankaracharya, Philip, V., Gray-Edwards, H., Aronin, N., Lutz, C., Landers, J., et al. (2020). *Ly6a* Differential Expression in Blood–Brain Barrier Is Responsible for Strain Specific Central Nervous System Transduction Profile of AAV-PHP.B. Human Gene Therapy *31*, 90–102.

Bedbrook, C.N., Deverman, B.E., and Gradinaru, V. (2018). Viral strategies for targeting the central and peripheral nervous systems. Annu. Rev. Neurosci. *41*, 323–348.

Berto, S., Liu, Y., and Konopka, G. (2020). Genomics at cellular resolution: insights into cognitive disorders and their evolution. Human Molecular Genetics *29*, R1–R9.

Blanchard, J.W., Bula, M., Davila-Velderrain, J., Akay, L.A., Zhu, L., Frank, A., Victor, M.B., Bonner, J.M., Mathys, H., Lin, Y.-T., et al. (2020). Reconstruction of the human blood-brain barrier in vitro reveals a pathogenic mechanism of APOE4 in pericytes. Nat Med *26*, 952–963.

Burger, D., Lou, J., Dayer, J.-M., and Grau, G.E. (1997). Both soluble and membrane-associated TNF activate brain microvascular endothelium: relevance to multiple sclerosis. Mol Psychiatry *2*, 113–116.

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L., Christopherson, K.S., Xing, Y., Lubischer, J.L., Krieg, P.A., Krupenko, S.A., et al. (2008). A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. Journal of Neuroscience *28*, 264–278.

Calcedo, R., Chichester, J.A., and Wilson, J.M. (2018). Assessment of humoral, innate, and T-cell immune responses to adeno-associated virus vectors. Hum. Gene Ther. Methods *29*, 86–95.

Cao, L., Wang, Z., and Wan, W. (2018). Suppressor of Cytokine Signaling 3: Emerging Role Linking Central Insulin Resistance and Alzheimer's Disease. Front. Neurosci. *12*, 417.

Cao, Z.-J., Wei, L., Lu, S., Yang, D.-C., and Gao, G. (2020). Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. Nat. Commun. *11*, 3458.

Capellera-Garcia, S., Pulecio, J., Dhulipala, K., Siva, K., Rayon-Estrada, V., Singbrant, S., Sommarin, M.N.E., Walkley, C.R., Soneji, S., Karlsson, G., et al. (2016). Defining the Minimal Factors Required for Erythropoiesis through Direct Lineage Conversion. Cell Reports *15*, 2550–2562.

Challis, R.C., Kumar, S.R., Chan, K.Y., Challis, C., Beadle, K., Jang, M.J., Kim, H.M., Rajendran, P.S., Tompkins, J.D., Shivkumar, K., et al. (2019). Systemic AAV vectors for widespread and targeted gene delivery in rodents. Nat. Protoc. *14*, 379–414.

Chamberlin, N.L., Du, B., de Lacalle, S., and Saper, C.B. (1998). Recombinant adeno-associated virus vector: use for transgene expression and anterograde tract tracing in the CNS. Brain Research *793*, 169–175.

Chan, K.Y., Jang, M.J., Yoo, B.B., Greenbaum, A., Ravi, N., Wu, W.-L., Sánchez-Guardado, L., Lois, C., Mazmanian, S.K., Deverman, B.E., et al. (2017). Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. Nat. Neurosci. *20*, 1172–1179.

Chang, H., Lim, J., Ha, M., and Kim, V.N. (2014). TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3′ End Modifications. Molecular Cell *53*, 1044–1052.

Chasseigneaux, S., Moraca, Y., Cochois-Guégan, V., Boulay, A.-C., Gilbert, A., Le Crom, S., Blugeon, C., Firmo, C., Cisternino, S., Laplanche, J.-L., et al. (2018). Isolation and differential transcriptome of vascular smooth muscle cells and mid-capillary pericytes from the rat brain. Sci Rep *8*, 12272.

Chen, J., Carey, K., and Godowski, P.J. (1997). Identification of Gas6 as a ligand for Mer, a neural cell adhesion molecule related receptor tyrosine kinase implicated in cellular transformation. Oncogene *14*, 2033–2039.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*.

Choi, H.M.T., Beck, V.A., and Pierce, N.A. (2014). Next-generation in situ hybridization chain reaction: Higher gain, lower cost, greater durability. ACS Nano *8*, 4284–4294.

Chuah, M.K., Petrus, I., De Bleser, P., Le Guiner, C., Gernoux, G., Adjali, O., Nair, N., Willems, J., Evens, H., Rincon, M.Y., et al. (2014). Liver-specific transcriptional modules identified by genome-wide in silico analysis enable efficient gene therapy in mice and non-human primates. Mol. Ther. *22*, 1605–1613.

Chui, R., and Dorovini-Zis, K. (2010). Regulation of CCL2 and CCL3 expression in human brain endothelial cells by cytokines and lipopolysaccharide. J Neuroinflammation *7*, 1.

Colella, P., Ronzitti, G., and Mingozzi, F. (2018). Emerging issues in AAV-mediated in vivo gene therapy. Mol. Ther. Methods Clin. Dev. *8*, 87–104.

Dai, J., Bercury, K.K., Ahrendsen, J.T., and Macklin, W.B. (2015). Olig1 Function Is Required for Oligodendrocyte Differentiation in the Mouse Brain. Journal of Neuroscience *35*, 4386–4402.

Dalkara, D., Byrne, L.C., Klimczak, R.R., Visel, M., Yin, L., Merigan, W.H., Flannery, J.G., and Schaffer, D.V. (2013). In vivo-directed evolution of a new adeno-associated virus for therapeutic outer retinal gene delivery from the vitreous. Sci. Transl. Med. *5*, 189ra76.

Davidsson, M., Wang, G., Aldrin-Kirk, P., Cardoso, T., Nolbrant, S., Hartnor, M., Mudannayake, J., Parmar, M., and Björklund, T. (2019). A systematic capsid evolution approach performed in vivo for the design of AAV vectors with tailored properties and tropism. Proc. Natl. Acad. Sci. U.S.A. *116*, 27053–27062.

Davis, A.S., Federici, T., Ray, W.C., Boulis, N.M., O'Connor, D., Clark, K.R., and Bartlett, J.S. (2015). Rational design and engineering of a modified adeno-associated virus (AAV1)-based vector system for enhanced retrograde gene delivery. Neurosurgery *76*, 216–225.

Daya, S., and Berns, K.I. (2008). Gene therapy using adeno-associated virus vectors. Clin. Microbiol. Rev. *21*, 583–593.

De Alencastro, G., Pekrun, K., Valdmanis, P., Tiffany, M., Xu, J., and Kay, M.A. (2020). Tracking adeno-associated virus capsid evolution by high-throughput sequencing. Hum. Gene Ther. *31*, 553–564.

Deleage, C., Wietgrefe, S.W., Del Prete, G., Morcock, D.R., Hao, X.P., Piatak, M., Bess, J., Anderson, J.L., Perkey, K.E., Reilly, C., et al. (2016). Defining HIV and SIV reservoirs in lymphoid tissues. Pathog. Immun. *1*, 68–106.

Deleage, C., Chan, C.N., Busman-Sahay, K., and Estes, J.D. (2018). Next-generation in situ hybridization approaches to define and quantify HIV and SIV reservoirs in tissue microenvironments. Retrovirology *15*, 4.

Deverman, B.E., Pravdo, P.L., Simpson, B.P., Kumar, S.R., Chan, K.Y., Banerjee, A., Wu, W.L., Yang, B., Huber, N., Pasca, S.P., et al. (2016). Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. Nat. Biotechnol. *34*, 204–209.

Deverman, B.E., Ravina, B.M., Bankiewicz, K.S., Paul, S.M., and Sah, D.W.Y. (2018). Gene therapy for neurological disorders: Progress and prospects. Nat. Rev. Drug Discov. *17*, 641–659.

Duan, D. (2018). Systemic AAV micro-dystrophin gene therapy for Duchenne muscular dystrophy. Mol. Ther. *26*, 2337–2356.

Excoffon, K.J.D.A., Koerber, J.T., Dickey, D.D., Murtha, M., Keshavjee, S., Kaspar, B.K., Zabner, J., and Schaffer, D.V. (2009). Directed evolution of adeno-associated virus to an infectious respiratory virus. Proc. Natl. Acad. Sci. U.S.A. *106*, 3865–3870.

Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of single RNA transcripts in situ. Science *280*, 585–590.

Ferreira, M.P., and Nagai, M.A. (2019). PHLDA3 (Pleckstrin Homology-Like Domain, family A, member 3). Atlas of Genetics and Cytogenetics in Oncology and Haematology.

Finn, J.D., Hui, D., Downey, H.D., Dunn, D., Pien, G.C., Mingozzi, F., Zhou, S., and High, K.A. (2010). Proteasome inhibitors decrease AAV2 capsid derived peptide epitope presentation on MHC class I following transduction. Mol. Ther. *18*, 135–142.

Flytzanis, N.C., Goeden, N., Goertsen, D., Cummins, A., Pickel, J., and Gradinaru, V. (2020). Broad gene expression throughout the mouse and marmoset brain after intravenous delivery of engineered AAV capsids. BioRxiv 2020.06.16.152975.

Gaj, T., Epstein, B.E., and Schaffer, D.V. (2016). Genome engineering using adeno-associated virus: Basic and clinical research applications. Mol. Ther. *24*, 458–464.

Gao, G., Wang, Q., Calcedo, R., Mays, L., Bell, P., Wang, L., Vandenberghe, L.H., Grant, R., Sanmiguel, J., Furth, E.E., et al. (2009). Adeno-associated virus-mediated gene transfer to nonhuman primate liver can elicit destructive transgene-specific T cell responses. Hum. Gene Ther. *20*, 930–942.

George, L.A., Sullivan, S.K., Giermasz, A., Rasko, J.E.J., Samelson-Jones, B.J., Ducore, J., Cuker, A., Sullivan, L.M., Majumdar, S., Teitel, J., et al. (2017). Hemophilia B gene therapy with a high-specific-activity factor IX variant. N. Engl. J. Med. *377*, 2215–2227.

Ghouzzi, V.E., Bianchi, F.T., Molineris, I., Mounce, B.C., Berto, G.E., Rak, M., Lebon, S., Aubry, L., Tocco, C., Gai, M., et al. (2016). ZIKA virus elicits P53 activation and genotoxic stress in human neural progenitors similar to mutations involved in severe forms of genetic microcephaly and p53. Cell Death Dis *7*, e2440–e2440.

Gokce, O., Stanley, G.M., Treutlein, B., Neff, N.F., Camp, J.G., Malenka, R.C., Rothwell, P.E., Fuccillo, M.V., Südhof, T.C., and Quake, S.R. (2016). Cellular Taxonomy of the Mouse Striatum as Revealed by Single-Cell RNA-Seq. Cell Reports *16*, 1126–1137.

Grabinski, T.M., Kneynsberg, A., Manfredsson, F.P., and Kanaan, N.M. (2015). A method for combining RNAscope in situ hybridization with immunohistochemistry in thick free-floating brain sections and primary neuronal cultures. PLoS One *10*, e0120120.

Gralinski, L.E., Ashley, S.L., Dixon, S.D., and Spindler, K.R. (2009). Mouse Adenovirus Type 1-Induced Breakdown of the Blood-Brain Barrier. JVI *83*, 9398–9410.

Grimm, D., Lee, J.S., Wang, L., Desai, T., Akache, B., Storm, T.A., and Kay, M.A. (2008). In vitro and in vivo gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. J. Virol. *82*, 5887–5911.

Gustafsson, C., Govindarajan, S., and Minshull, J. (2004). Codon bias and heterologous protein expression. Trends Biotechnol. *22*, 346–353.

He, L., Vanlandewijck, M., Raschperger, E., Andaloussi Mäe, M., Jung, B., Lebouvier, T., Ando, K., Hofmann, J., Keller, A., and Betsholtz, C. (2016). Analysis of the brain mural cell transcriptome. Sci Rep *6*, 35108.

Herrmann, A.-K., Bender, C., Kienle, E., Grosse, S., El Andari, J., Botta, J., Schürmann, N., Wiedtke, E., Niopek, D., and Grimm, D. (2019). A robust and all-inclusive pipeline for shuffling of adeno-associated viruses. ACS Synth. Biol. *8*, 194–206.

Herzog, R.W., Cooper, M., Perrin, G.Q., Biswas, M., Martino, A.T., Morel, L., Terhorst, C., and Hoffman, B.E. (2019). Regulatory T cells and TLR9 activation shape antibody formation to a secreted transgene product in AAV muscle gene transfer. Cellular Immunology *342*, 103682.

Hinderer, C., Katz, N., Buza, E.L., Dyer, C., Goode, T., Bell, P., Richman, L.K., and Wilson, J.M. (2018). Severe toxicity in nonhuman primates and piglets following high-dose intravenous administration of an adeno-associated virus vector expressing human SMN. Hum. Gene Ther. *29*, 285–298.

Hirsch, M.L., and Samulski, R.J. (2014). AAV-Mediated Gene Editing via Double-Strand Break Repair. In Gene Correction: Methods and Protocols, F. Storici, ed. (Totowa, NJ: Humana Press), pp. 291–307.

Hordeaux, J., Yuan, Y., Clark, P.M., Wang, Q., Martino, R.A., Sims, J.J., Bell, P., Raymond, A., Stanford, W.L., and Wilson, J.M. (2019). The GPI-Linked Protein LY6A Drives AAV-PHP.B Transport across the Blood-Brain Barrier. Molecular Therapy *27*, 912–921.

Hösel, M., Broxtermann, M., Janicki, H., Esser, K., Arzberger, S., Hartmann, P., Gillen, S., Kleeff, J., Stabenow, D., Odenthal, M., et al. (2012). Toll-like receptor 2–mediated innate immune response in human nonparenchymal liver cells toward adeno-associated viral vectors. Hepatology *55*, 287–297.

Hrvatin, S., Tzeng, C.P., Nagy, M.A., Stroud, H., Koutsioumpa, C., Wilcox, O.F., Assad, E.G., Green, J., Harvey, C.D., Griffith, E.C., et al. (2019). A scalable platform for the development of cell-type-specific viral drivers. ELife *8*, e48089.

Huang, K.W., and Sabatini, B.L. (2020). Single-cell analysis of neuroinflammatory responses following intracranial injection of G-deleted rabies viruses. Front. Cell. Neurosci. *14*, 65.

Huang, Q., Chan, K.Y., Tobey, I.G., Chan, Y.A., Poterba, T., Boutros, C.L., Balazs, A.B., Daneman, R., Bloom, J.M., Seed, C., et al. (2019). Delivering genes across the blood-brain barrier: LY6A, a novel cellular receptor for AAV-PHP.B capsids. PLoS ONE *14*, e0225206.

Hunter, J.E., Gurda, B.L., Yoon, S.Y., Castle, M.J., and Wolfe, J.H. (2019). In situ hybridization for detection of AAV-mediated gene expression. Methods Mol. Biol. *1950*, 107–122.

Johansen, J., Tornøe, J., Møller, A., and Johansen, T.E. (2003). Increased *in vitro* and *in vivo* transgene expression levels mediated through *cis* -acting elements: *Cis* Elements Increased *Ex Vivo* Gene Expression. J. Gene Med. *5*, 1080–1089.

Jordão, M.J.C., Sankowski, R., Brendecke, S.M., Sagar, Locatelli, G., Tai, Y.-H., Tay, T.L., Schramm, E., Armbruster, S., Hagemeyer, N., et al. (2019). Single-cell profiling identifies myeloid cell subsets with distinct fates during neuroinflammation. Science *363*, eaat7554.

Jüttner, J., Szabo, A., Gross-Scherf, B., Morikawa, R.K., Rompani, S.B., Hantz, P., Szikra, T., Esposti, F., Cowan, C.S., Bharioke, A., et al. (2019). Targeting neuronal and glial cell

types with synthetic promoter AAVs in mice, non-human primates and humans. Nat. Neurosci. *22*, 1345–1356.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. Nat. Methods *10*, 857–860.

Kodali, M.C., Chen, H., and Liao, F.-F. (2020). Temporal unsnarling of brain's acute neuroinflammatory transcriptional profiles reveals panendothelitis as the earliest event preceding microgliosis. Mol Psychiatry.

Körbelin, J., Sieber, T., Michelfelder, S., Lunding, L., Spies, E., Hunger, A., Alawi, M., Rapti, K., Indenbirken, D., Müller, O.J., et al. (2016a). Pulmonary targeting of adeno-associated viral vectors by next-generation sequencing-guided screening of random capsid displayed peptide libraries. Mol. Ther. *24*, 1050–1061.

Körbelin, J., Dogbevia, G., Michelfelder, S., Ridder, D.A., Hunger, A., Wenzel, J., Seismann, H., Lampe, M., Bannach, J., Pasparakis, M., et al. (2016b). A brain microvasculature endothelial cell-specific viral vector with the potential to treat neurovascular and neurological diseases. EMBO Mol. Med. *8*, 609–625.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods *16*, 1289–1296.

Kotterman, M.A., and Schaffer, D.V. (2014). Engineering adeno-associated viruses for clinical gene therapy. Nat. Rev. Genet. *15*, 445–451.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol *21*, 31.

Lee, E.J., Guenther, C.M., and Suh, J. (2018). Adeno-associated virus (AAV) vectors: Rational design strategies for capsid engineering. Curr. Opin. Biomed. Eng. *7*, 58–63.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly multiplexed subcellular RNA sequencing in situ. Science *343*, 1360–1363.

Li, J., Lv, H., and Che, Y. (2020). microRNA-381-3p Confers Protection Against Ischemic Stroke Through Promoting Angiogenesis and Inhibiting Inflammation by Suppressing Cebpb and Map3k8. Cell Mol Neurobiol *40*, 1307–1319.

Liao, J., Lu, X., Shao, X., Zhu, L., and Fan, X. (2020). Uncovering an organ's molecular architecture at single-cell resolution by spatially resolved transcriptomics. Trends Biotechnol. *39*, 43–58.

Lin, Y., Ghazanfar, S., Wang, K.Y.X., Gagnon-Bartsch, J.A., Lo, K.K., Su, X., Han, Z.-G., Ormerod, J.T., Speed, T.P., Yang, P., et al. (2019). scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. Proc Natl Acad Sci USA *116*, 9775–9784.

Lin, Y.-S., Wang, H.-Y., Huang, D.-F., Hsieh, P.-F., Lin, M.-Y., Chou, C.-H., Wu, I.-J., Huang, G.-J., Gau, S.S.-F., and Huang, H.-S. (2016). Neuronal Splicing Regulator RBFOX3 (NeuN) Regulates Adult Hippocampal Neurogenesis and Synaptogenesis. PLoS ONE *11*, e0164164.

Liu, Q., Yang, Y., and Fan, X. (2020). Microvascular pericytes in brain-associated vascular disease. Biomedicine & Pharmacotherapy *121*, 109633.

Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat Methods *15*, 1053–1058.

Lorenzon, E., Colladel, R., Andreuzzi, E., Marastoni, S., Todaro, F., Schiappacassi, M., Ligresti, G., Colombatti, A., and Mongiat, M. (2012). MULTIMERIN2 impairs tumor angiogenesis and growth by interfering with VEGF-A/VEGFR2 pathway. Oncogene *31*, 3136–3147.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Lowenstein, P., Mandel, R., Xiong, W., Kroeger, K., and Castro, M. (2007). Immune Responses to Adenovirus and Adeno-Associated Vectors Used for Gene Therapy of Brain Diseases: The Role of Immunological Synapses in Understanding the Cell Biology of Neuroimmune Interactions. CGT *7*, 347–360.

Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M., and Cai, L. (2014). Single-cell in situ RNA profiling by sequential hybridization. Nat. Methods *11*, 360–361.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., Marioni, J.C., and participants in the 1st Human Cell Atlas Jamboree (2019). EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Macrae, M., Neve, R.M., Rodriguez-Viciana, P., Haqq, C., Yeh, J., Chen, C., Gray, J.W., and McCormick, F. (2005). A conditional feedback loop regulates Ras activity through EphA2. Cancer Cell *8*, 111–118.

Maheshri, N., Koerber, J.T., Kaspar, B.K., and Schaffer, D.V. (2006). Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. Nat. Biotechnol. *24*, 198–204.

Manno, C.S., Pierce, G.F., Arruda, V.R., Glader, B., Ragni, M., Rasko, J.J.E., Ozelo, M.C., Hoots, K., Blatt, P., Konkle, B., et al. (2006). Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. Nat. Med. *12*, 342–347.

Marques, S., Zeisel, A., Codeluppi, S., van Bruggen, D., Mendanha Falcao, A., Xiao, L., Li, H., Haring, M., Hochgerner, H., Romanov, R.A., et al. (2016). Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science *352*, 1326–1329.

Martino, A.T., Suzuki, M., Markusic, D.M., Zolotukhin, I., Ryals, R.C., Moghimi, B., Ertl, H.C.J., Muruve, D.A., Lee, B., and Herzog, R.W. (2011). The genome of self-complementary adeno-associated viral vectors increases Toll-like receptor 9–dependent innate immune responses in the liver. Blood *117*, 6459–6468.

McCown, T.J., Xiao, X., Li, J., Breese, G.R., and Jude Samulski, R. (1996). Differential and persistent expression patterns of CNS gene transfer by an adeno-associated virus (AAV) vector. Brain Research *713*, 99–107.

McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019). MULTI-seq: Sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626.

Mével, M., Bouzelha, M., Leray, A., Pacouret, S., Guilbaud, M., Penaud-Budloo, M., Alvarez-Dorta, D., Dubreil, L., Gouin, S.G., Combal, J.P., et al. (2020). Chemical modification of the adeno-associated virus capsid to improve gene delivery. Chem. Sci. *11*, 1122–1131.

Miao, C.H., Nakai, H., Thompson, A.R., Storm, T.A., Chiu, W., Snyder, R.O., and Kay, M.A. (2000). Nonrandom transduction of recombinant adeno-associated virus vectors in

mouse hepatocytes in vivo: Cell cycling does not influence hepatocyte transduction. J. Virol. *74*, 3793–3803.

Mich, J.K., Graybuck, L.T., Hess, E.E., Mahoney, J.T., Kojima, Y., Ding, Y., Somasundaram, S., Miller, J.A., Weed, N., Omstead, V., et al. (2020). Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. BioRxiv 555318.

Mingozzi, F., and High, K.A. (2013). Immune responses to AAV vectors: Overcoming barriers to successful gene therapy. Blood *122*, 23–36.

Mingozzi, F., Maus, M.V., Hui, D.J., Sabatino, D.E., Murphy, S.L., Rasko, J.E.J., Ragni, M.V., Manno, C.S., Sommer, J., Jiang, H., et al. (2007). CD8 + T-cell responses to adeno-associated virus capsid in humans. Nat. Med. *13*, 419–422.

Mingozzi, F., Meulenberg, J.J., Hui, D.J., Basner-Tschakarjan, E., Hasbrouck, N.C., Edmonson, S.A., Hutnick, N.A., Betts, M.R., Kastelein, J.J., Stroes, E.S., et al. (2009). AAV-1–mediated gene transfer to skeletal muscle in humans results in dose-dependent activation of capsid-specific T cells. Blood *114*, 2077–2086.

Miron, V.E., Kuhlmann, T., and Antel, J.P. (2011). Cells of the oligodendroglial lineage, myelination, and remyelination. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease *1812*, 184–193.

Montagne, A., Nation, D.A., Sagare, A.P., Barisano, G., Sweeney, M.D., Chakhoyan, A., Pachicano, M., Joe, E., Nelson, A.R., D'Orazio, L.M., et al. (2020). APOE4 leads to blood–brain barrier dysfunction predicting cognitive decline. Nature *581*, 71–76.

Mossner, J.M., Batista-Brito, R., Pant, R., and Cardin, J.A. (2020). Developmental loss of MeCP2 from VIP interneurons impairs cortical function and behavior. ELife *9*, e55639.

Muhammad, K.A., Nur, A.A., Nurul, H.S., Narazah, M.Y., and Siti, R.A.R. (2018). Dual-specificity phosphatase 6 (DUSP6): a review of its molecular characteristics and clinical relevance in cancer. Cancer Biology & Medicine *15*, 14.

Müller, O.J., Kaul, F., Weitzman, M.D., Pasqualini, R., Arap, W., Kleinschmidt, J.A., and Trepel, M. (2003). Random peptide libraries displayed on adeno-associated virus to select for targeted gene therapy vectors. Nat. Biotechnol. *21*, 1040–1046.

Naso, M.F., Tomkowicz, B., Perry, W.L., and Strohl, W.R. (2017). Adeno-associated virus (AAV) as a vector for gene therapy. BioDrugs *31*, 317–334.

Nathwani, A.C., Tuddenham, E.G.D., Rangarajan, S., Rosales, C., McIntosh, J., Linch, D.C., Chowdary, P., Riddell, A., Pie, A.J., Harrington, C., et al. (2011). Adenovirus-associated virus vector–mediated gene transfer in hemophilia B. N. Engl. J. Med. *365*, 2357–2365.

Nathwani, A.C., Reiss, U.M., Tuddenham, E.G.D., Rosales, C., Chowdary, P., McIntosh, J., Della Peruta, M., Lheriteau, E., Patel, N., Raj, D., et al. (2014). Long-term safety and efficacy of factor IX gene therapy in hemophilia B. N. Engl. J. Med. *371*, 1994–2004.

Ogden, P.J., Kelsic, E.D., Sinai, S., and Church, G.M. (2019). Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. Science *366*, 1139–1143.

Ohtsuka, M., Inoko, H., Kulski, J.K., and Yoshimura, S. (2008). Major histocompatibility complex (Mhc) class Ib gene duplications, organization and expression patterns in mouse strain C57BL/6. BMC Genomics *9*, 178.

Oikonomou, G., Altermatt, M., Zhang, R., Coughlin, G.M., Montz, C., Gradinaru, V., and Prober, D.A. (2019). The serotonergic raphe promote sleep in zebrafish and mice. Neuron *103*, 686-701.e8.

Ojala, D.S., Sun, S., Santiago-Ortiz, J.L., Shapiro, M.G., Romero, P.A., and Schaffer, D.V. (2018). In vivo selection of a computationally designed SCHEMA AAV library yields a novel variant for infection of adult neural stem cells in the SVZ. Mol. Ther. *26*, 304–319.

Ollila, H.M., Utge, S., Kronholm, E., Aho, V., Van Leeuwen, W., Silander, K., Partonen, T., Perola, M., Kaprio, J., Salomaa, V., et al. (2012). TRIB1 constitutes a molecular link between regulation of sleep and lipid metabolism in humans. Transl Psychiatry *2*, e97–e97.

Palfreyman, M.T., and Jorgensen, E.M. (2017). Unc13 Aligns SNAREs and Superprimes Synaptic Vesicles. Neuron *95*, 473–475.

Patriarchi, T., Cho, J.R., Merten, K., Howe, M.W., Marley, A., Xiong, W.-H., Folk, R.W., Broussard, G.J., Liang, R., Jang, M.J., et al. (2018). Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors. Science *360*, eaat4422.

Paulk, N. (2020). Gene Therapy: It Is Time to Talk about High-Dose AAV: The deaths of two children with X-linked myotubular myopathy in the ASPIRO trial prompts a reexamination of vector safety. Genetic Engineering & Biotechnology News *40*, 14–16.

Paulk, N.K., Pekrun, K., Zhu, E., Nygaard, S., Li, B., Xu, J., Chu, K., Leborgne, C., Dane, A.P., Haft, A., et al. (2018). Bioengineered AAV capsids with combined high human liver transduction in vivo and unique humoral seroreactivity. Mol. Ther. *26*, 289–303.

Pekrun, K., De Alencastro, G., Luo, Q.-J., Liu, J., Kim, Y., Nygaard, S., Galivo, F., Zhang, F., Song, R., Tiffany, M.R., et al. (2019). Using a barcoded AAV capsid library to select for clinically relevant gene therapy vectors. JCI Insight *4*, e131610.

Perez-Nievas, B.G., and Serrano-Pozo, A. (2018). Deciphering the Astrocyte Reaction in Alzheimer's Disease. Front. Aging Neurosci. *10*, 114.

Pérez-Sen, Queipo, Gil-Redondo, Ortega, Gómez-Villafuertes, Miras-Portugal, and Delicado (2019). Dual-Specificity Phosphatase Regulation in Neurons and Glial Cells. IJMS *20*, 1999.

Pien, G.C., Basner-Tschakarjan, E., Hui, D.J., Mentlik, A.N., Finn, J.D., Hasbrouck, N.C., Zhou, S., Murphy, S.L., Maus, M.V., Mingozzi, F., et al. (2009). Capsid antigen presentation flags human hepatocytes for destruction after transduction by adeno-associated viral vectors. J. Clin. Invest. *119*, 1688–1695.

Polinski, N.K., Gombash, S.E., Manfredsson, F.P., Lipton, J.W., Kemp, C.J., Cole-Strauss, A., Kanaan, N.M., Steece-Collier, K., Kuhn, N.C., Wohlgenant, S.L., et al. (2015). Recombinant adenoassociated virus 2/5-mediated gene transfer is reduced in the aged rat midbrain. Neurobiol. Aging *36*, 1110–1120.

Polinski, N.K., Manfredsson, F.P., Benskey, M.J., Fischer, D.L., Kemp, C.J., Steece-Collier, K., Sandoval, I.M., Paumier, K.L., and Sortwell, C.E. (2016). Impact of age and vector construct on striatal and nigral transgene expression. Mol. Ther. Methods Clin. Dev. *3*, 16082.

Pool, A.-H., Wang, T., Stafford, D.A., Chance, R.K., Lee, S., Ngai, J., and Oka, Y. (2020). The cellular basis of distinct thirst modalities. Nature *588*, 112–117.

Pulicherla, N., Shen, S., Yadav, S., Debbink, K., Govindasamy, L., Agbandje-McKenna, M., and Asokan, A. (2011). Engineering liver-detargeted AAV9 vectors for cardiac and musculoskeletal gene transfer. Mol. Ther. *19*, 1070–1078.

Puray-Chavez, M., Tedbury, P.R., Huber, A.D., Ukah, O.B., Yapo, V., Liu, D., Ji, J., Wolf, J.J., Engelman, A.N., and Sarafianos, S.G. (2017). Multiplex single-cell visualization of nucleic acids and protein during HIV infection. Nat. Commun. *8*, 1882.

Qin, J.Y., Zhang, L., Clift, K.L., Hulur, I., Xiang, A.P., Ren, B.-Z., and Lahn, B.T. (2010). Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. PLoS ONE *5*, e10611.

Ravindra Kumar, S., Miles, T.F., Chen, X., Brown, D., Dobreva, T., Huang, Q., Ding, X., Luo, Y., Einarsson, P.H., Greenbaum, A., et al. (2020). Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types. Nat. Methods *17*, 541–550.

Rincon, M.Y., Sarcar, S., Danso-Abeam, D., Keyaerts, M., Matrai, J., Samara-Kuko, E., Acosta-Sanchez, A., Athanasopoulos, T., Dickson, G., Lahoutte, T., et al. (2015). Genome-wide computational analysis reveals cardiomyocyte-specific transcriptional Cis-regulatory motifs that enable efficient cardiac gene therapy. Mol. Ther. *23*, 43–52.

Riverso, M., Montagnani, V., and Stecca, B. (2017). KLF4 is regulated by RAS/RAF/MEK/ERK signaling through E2F1 and promotes melanoma cell growth. Oncogene *36*, 3322–3333.

Rogers, G.L., Shirley, J.L., Zolotukhin, I., Kumar, S.R.P., Sherman, A., Perrin, G.Q., Hoffman, B.E., Srivastava, A., Basner-Tschakarjan, E., Wallet, M.A., et al. (2017). Plasmacytoid and conventional dendritic cells cooperate in crosspriming AAV capsid-specific CD8+ T cells. Blood *129*, 3184–3195.

Rossi, A., Dupaty, L., Aillot, L., Zhang, L., Gallien, C., Hallek, M., Odenthal, M., Adriouch, S., Salvetti, A., and Büning, H. (2019). Vector uncoating limits adeno-associated viral vector-mediated transduction of human dendritic cells and vector immunogenicity. Sci. Rep. *9*, 3631.

Ruden, J.B., Dugan, L.L., and Konradi, C. (2021). Parvalbumin interneuron vulnerability and brain disorders. Neuropsychopharmacol. *46*, 279–287.

Salvador, J.M., Brown-Clay, J.D., and Fornace, A.J. (2013). Gadd45 in Stress Signaling, Cell Cycle Control, and Apoptosis. In Gadd45 Stress Sensor Genes, D.A. Liebermann, and B. Hoffman, eds. (New York, NY: Springer New York), pp. 1–19.

Samulski, R.J., and Muzyczka, N. (2014). AAV-mediated gene therapy for research and therapeutic purposes. Annu. Rev. Virol. *1*, 427–451.

Sarkar, S., Dammer, E.B., Malovic, E., Olsen, A.L., Raza, S.A., Gao, T., Xiao, H., Oliver, D.L., Duong, D., Joers, V., et al. (2020). Molecular Signatures of Neuroinflammation Induced by αSynuclein Aggregates in Microglial Cells. Front. Immunol. *11*, 33.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: An open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Sen, D. (2014). Improving clinical efficacy of adeno associated vectors by rational capsid bioengineering. J. Biomed. Sci. *21*, 103.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016a). In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron *92*, 342–357.

Shah, S., Lubeck, E., Schwarzkopf, M., He, T.F., Greenbaum, A., Sohn, C.H., Lignell, A., Choi, H.M., Gradinaru, V., Pierce, N.A., et al. (2016b). Single-molecule RNA detection at depth by hybridization chain reaction and tissue hydrogel embedding and clearing. Development *143*, 2862–2867.

Shao, W., Earley, L.F., Chai, Z., Chen, X., Sun, J., He, T., Deng, M., Hirsch, M.L., Ting, J., Samulski, R.J., et al. (2018). Double-stranded RNA innate immune response activation from long-term adeno-associated virus vector transduction. JCI Insight *3*.

Shin, S., and Park, J. (2016). Characterization of sequence-specific errors in various next-generation sequencing systems. Mol. BioSyst. *12*, 914–922.

Shirley, J.L., Jong, Y.P. de, Terhorst, C., and Herzog, R.W. (2020). Immune Responses to Viral Gene Therapy Vectors. Mol. Ther. *28*, 709–722.

Somanathan, S., Breous, E., Bell, P., and Wilson, J.M. (2010). AAV vectors avoid inflammatory signals necessary to render transduced hepatocyte targets for destructive T cells. Mol. Ther. *18*, 977–982.

Song, H.W., Foreman, K.L., Gastfriend, B.D., Kuo, J.S., Palecek, S.P., and Shusta, E.V. (2020). Transcriptomic comparison of human and mouse brain microvessels. Sci Rep *10*, 12358.

Spiegel, I., Mardinly, A.R., Gabel, H.W., Bazinet, J.E., Couch, C.H., Tzeng, C.P., Harmin, D.A., and Greenberg, M.E. (2014). Npas4 Regulates Excitatory-Inhibitory Balance within Neural Circuits through Cell-Type-Specific Gene Programs. Cell *157*, 1216–1229.

Srivastava, A. (2020). AAV vectors: Are they safe? Hum. Gene Ther. *31*, 697–699.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*, 224.

Sun, W., Cornwell, A., Li, J., Peng, S., Osorio, M.J., Aalling, N., Wang, S., Benraiss, A., Lou, N., Goldman, S.A., et al. (2017). SOX9 Is an Astrocyte-Specific Nuclear Marker in the Adult Brain Outside the Neurogenic Regions. J. Neurosci. *37*, 4493–4507.

Suzuki, N., Sekimoto, K., Hayashi, C., Mabuchi, Y., Nakamura, T., and Akazawa, C. (2017). Differentiation of Oligodendrocyte Precursor Cells from Sox10-Venus Mice to Oligodendrocytes and Astrocytes. Sci Rep *7*, 14133.

Tan, Y., and Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. Cell Systems *9*, 207-213.e2.

Taoufiq, Z., Ninov, M., Villar-Briones, A., Wang, H.-Y., Sasaki, T., Roy, M.C., Beauchain, F., Mori, Y., Yoshida, T., Takamori, S., et al. (2020). Hidden proteome of synaptic vesicles in the mammalian brain. Proc Natl Acad Sci USA *117*, 33586–33596.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. *19*, 335–346.

Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature *563*, 72–78.

Tervo, D.G., Hwang, B.Y., Viswanathan, S., Gaj, T., Lavzin, M., Ritola, K.D., Lindo, S., Michael, S., Kuleshova, E., Ojala, D., et al. (2016). A designer AAV variant permits efficient retrograde access to projection neurons. Neuron *92*, 372–382.

Vairapandi, M., Balliet, A.G., Hoffman, B., and Liebermann, D.A. (2002). GADD45b and GADD45g are cdc2/cyclinB1 kinase inhibitors with a role in S and G2/M cell cycle checkpoints induced by genotoxic stress. J. Cell. Physiol. *192*, 327–338.

Vandenberghe, L.H., Wang, L., Somanathan, S., Zhi, Y., Figueredo, J., Calcedo, R., Sanmiguel, J., Desai, R.A., Chen, C.S., Johnston, J., et al. (2006). Heparin binding directs activation of T cells against adeno-associated virus serotype 2 capsid. Nat. Med. *12*, 967–971.

Vanlandewijck, M., He, L., Mäe, M.A., Andrae, J., Ando, K., Del Gaudio, F., Nahar, K., Lebouvier, T., Laviña, B., Gouveia, L., et al. (2018). A molecular atlas of cell types and zonation in the brain vasculature. Nature *554*, 475–480.

Veys, K., Fan, Z., Ghobrial, M., Bouché, A., García-Caballero, M., Vriens, K., Conchinha, N.V., Seuwen, A., Schlegel, F., Gorski, T., et al. (2020). Role of the GLUT1 Glucose

Transporter in Postnatal CNS Angiogenesis and Blood-Brain Barrier Integrity. Circ Res *127*, 466–482.

Wang, S.K., Lapan, S.W., Hong, C.M., Krause, T.B., and Cepko, C.L. (2020). In situ detection of adeno-associated viral vector genomes with SABER-FISH. Mol. Ther. Methods Clin. Dev. *19*, 376–386.

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science *361*, eaat5691.

Wilson, J.M., and Flotte, T.R. (2020). Moving forward after two deaths in a gene therapy trial of myotubular myopathy. Hum. Gene Ther. *31*, 695–696.

Winkler, E.A., Bell, R.D., and Zlokovic, B.V. (2010). Pericyte-specific expression of PDGF beta receptor in mouse models with normal and deficient PDGF beta receptor signaling. Mol Neurodegeneration *5*, 32.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol *19*, 15.

Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. *8*, 281-291.e9.

Wu, Y.E., Pan, L., Zuo, Y., Li, X., and Hong, W. (2017). Detecting Activated Cell Populations Using Single-Cell RNA-Seq. Neuron *96*, 313-329.e6.

Wu, Z., Asokan, A., and Samulski, R.J. (2006). Adeno-associated virus serotypes: Vector toolkit for human gene therapy. Mol. Ther. *14*, 316–327.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol *17*.

Yang, A.C., Vest, R.T., Kern, F., Lee, D.P., Maat, C.A., Losada, P.M., Chen, M.B., Agam, M., Schaum, N., Khoury, N., et al. (2021). A human brain vascular atlas reveals diverse cell mediators of Alzheimer's disease risk (Neuroscience).

Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol *21*, 57.

Yang, Y., Vidensky, S., Jin, L., Jie, C., Lorenzini, I., Frankl, M., and Rothstein, J.D. (2011). Molecular comparison of GLT1+ and ALDH1L1+ astrocytes in vivo in astroglial reporter mice. Glia *59*, 200–207.

Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell *184*, 3222-3241.e26.

Ying, Y., Müller, O.J., Goehringer, C., Leuchs, B., Trepel, M., Katus, H.A., and Kleinschmidt, J.A. (2010). Heart-targeted adeno-associated viral vectors selected by in vivo biopanning of a random viral display peptide library. Gene Ther. *17*, 980–990.

Zaiss, A.K., Cotter, M.J., White, L.R., Clark, S.A., Wong, N.C.W., Holers, V.M., Bartlett, J.S., and Muruve, D.A. (2008). Complement is an essential component of the immune response to adeno-associated virus vectors. J. Virol. *82*, 2727–2740.

Zamagni, A., Pasini, A., Pirini, F., Ravaioli, S., Giordano, E., Tesei, A., Calistri, D., Ulivi, P., Fabbri, F., Foca, F., et al. (2020). CDKN1A upregulation and cisplatin-pemetrexed resistance in non-small cell lung cancer cells. Int J Oncol.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular architecture of the mouse nervous system. Cell *174*, 999-1014.e22.

Zhang, Z., and Yu, J. (2018). NR4A1 Promotes Cerebral Ischemia Reperfusion Injury by Repressing Mfn2-Mediated Mitophagy and Inactivating the MAPK–ERK–CREB Signaling Pathway. Neurochem Res *43*, 1963–1977.

Zhang, X., Lu, W., Zheng, Y., Wang, W., Bai, L., Chen, L., Feng, Y., Zhang, Z., and Yuan, Z. (2016). In situ analysis of intrahepatic virological events in chronic hepatitis B virus infection. J. Clin. Invest. *126*, 1079–1092.

Zhang, X., Hu, W., Lei, Z., Wang, H., Kang, X., Graduate School of Xinjiang Medical University, Urumqi 830000, China, and Department of Dermatology, Xinjiang Uygur Autonomous Region People's Hospital, Urumqi 830001, China (2021). Identification of key genes and evaluation of immune cell infiltration in vitiligo. Mathematical Biosciences and Engineering *18*, 1051–1062.

Zhang, Z., Ma, Z., Zou, W., Guo, H., Liu, M., Ma, Y., and Zhang, L. (2019). The Appropriate Marker for Astrocytes: Comparing the Distribution and Expression of Three Astrocytic Markers in Different Mouse Cerebral Regions. BioMed Research International *2019*, 1–15.

Zhao, J., Yue, Y., Patel, A., Wasala, L., Karp, J.F., Zhang, K., Duan, D., and Lai, Y. (2020). High-resolution histological landscape of AAV DNA distribution in cellular compartments and tissues following local and systemic injection. Mol. Ther. Methods Clin. Dev. *18*, 856–868.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.

Zhu, J., Huang, X., and Yang, Y. (2009). The TLR9-MyD88 pathway is critical for adaptive immune responses to adenoassociated virus gene therapy vectors in mice. J. Clin. Investig. *119*, 2388–2398.

Chapter 3

APPROACHES TOWARDS ENGINEERING THE NEURONAL ARC CAPSID FOR TARGETED RNA EXPORT

3.1 Introduction

Efficient and targeted RNA delivery could have numerous applications, from trafficking undesired RNA to degradation sites and studying intracellular trafficking, to perturbation studies of intercellular communication. In terms of disease, it has been shown that transcription of non-coding DNA repeats generates expanded RNA transcripts that are toxic to the cells, causing disruption of multiple downstream cellular processes and resulting in multisystem health disorders (Ranum, et al. 2006). Myotonic Dystrophy 1 (DM1) is considered to be one of such multisystem disorders in which there is an abnormal expansion of trinucleotide repeat sequences. Another undesired form of RNA that can be present in mammalian cells is that of latent RNA viruses, such as that of HIV and Human T-cell leukemia virus type 1 (HTLV1) (Speck, et al. 2010). A method enabling removal of these disease-inducing RNAs from cells would allow for an alternative strategy which does not include integration of permanent rescue genes. Additionally, current RNA therapies suffer from shortcomings such as; limited targeting of a wide range of spliced RNAs, unmodified siRNA is susceptible to degradation by RNAse, and the potential induction of immune system response (Damase, et al. 2021). Furthermore, introducing intervention at the gene expression level over one that targets the toxic protein itself is that potentially negative downstream cellular pathogenic effects that arise from abnormal functioning of that single protein could be reduced or eliminated (Ghosh, et al. 2017).

Recently, the neuronal immediate early gene, Arc, has been found to assemble into virus-like particles and transfer RNA intracellularly in neurons (Pastuzyn, et al. 2018). Inspired by this finding, I have designed and tested Arc fusion constructs to target specific RNA and export them out of the cell. Deactivated Cas13 (dCas13) enzymes, such as dPspCas13b and dLwaCas13a, have been shown to bind RNA without cleaving the target (Abudayyeh, et al. 2017, Cox, et al. 2017). Cox, et al. have also demonstrated that dCas13 enzymes fused to

ADAR2, an RNA-binding and editing protein, edit RNAs specifically. Accordingly, I have generated and tested fusions of Arc and deactivated Cas13 enzymes in order to target Arc to a specific RNA. I observed capsid formation of dLwaCas13a-Arc fusions in bacteria and characterized their RNA encapsulation. Additionally, I have developed a functional assay to measure export of target RNA. Overall, I demonstrate that it is possible for Arc to maintain its capsid-forming ability and to encapsulate RNA when fused with a dCas13 enzyme. These preliminary results support the possibility of engineering Arc as a scientific tool for RNA delivery, as well as capture of disease relevant targets such as amyloid-beta precursor mRNA.

3.2 Arc RNA targeting strategies

Strategies to achieve targeted RNA binding and export are illustrated in Figure 1.

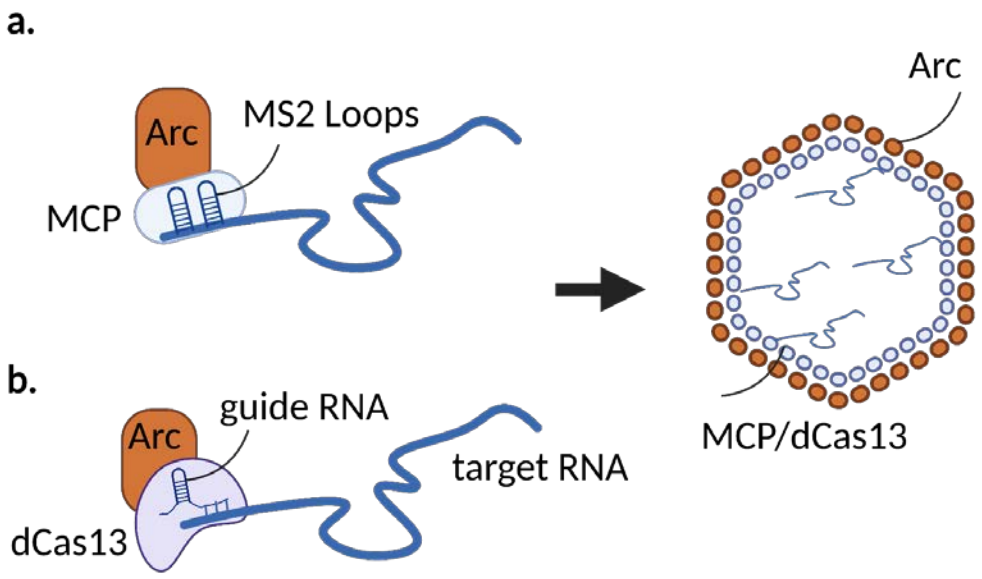


Figure 1. Arc engineering strategies for purpose of targeted RNA export (A) Arc fused to MCP would capture target RNA containing MS2 loops. This method requires modification of target RNA (B) Arc fused to dCas13 enzymes captures target RNA using guide RNA. Guide RNA needs to be made and optimized for target RNA. Ideally, both monomers oligomerize to form capsids, protect encapsulated RNA, and get exported into extracellular space.

An MS2-MCP system is composed of bacteriophage MS2 multimerized RNA stem loops, named MS2 for short, and a coat protein which binds the MS2 stem loops, MCP (MS2

Coat Protein). The system is commonly used to image live mRNAs to study RNA dynamics and has been further optimized to ensure that the MS2 loops binding strength is sufficient to allow for RNA degradation (Tutucci, et al. 2017). Binding strength of MCP to MS2 tagged RNA is an important property of the M2S-MCP system (Han, et al. 2020). For example, in the case of fusing Arc to MCP, the binding of MCP to target RNA tagged with MS2 might be desirable such that the RNA survives the entire process of oligomerization and cell expulsion. The disadvantage of using the MS2-MCP system in conjunction with Arc is that it requires modification of target RNA. Additionally, given that MS2 was derived from a bacteriophage, immune cells could be more likely to interact with the complex, either causing its degradation and/or confounding readout.

Cas13 is a type of enzyme containing nucleotide-binding endoRNase domains that allow for targeted RNA cleavage (Shmakov, et al. 2015). Three families of Cas13 enzymes have been found thus far; Cas13a, Cas13b, and Cas13c. Gootenberg JS, et al. have shown that Cas13a can be used for nucleic acid detection, while Abudayyeh, et al. have used Cas13a for transcript tracking. In order to disable the cleaving function of Cas13 enzymes, Abudayyeh, et al. have created a catalytically dead variant of Cas13a (dLwaCas13a) by mutating catalytic arginine residues. Similarly, Cox, et al. have created catalytically dead variant of Cas13b (dPspCas13b) by mutating the conserved catalytic residues in the Higher Eukaryote and Prokaryote Nucleotide-binding (HEPN) domains. The dPspCas13b was then fused to Adenosine deaminases acting on RNA (ADAR) enzyme to achieve targeted RNA editing. The concept of fusing Arc with dCas13a/b enzymes to achieve oligomerization and export of targeted RNAs was inspired by dCas13b-ADAR work. The biggest advantage of the dCas13 system for targeting RNAs is that it does not require modification of target RNA. However, the optimization and testing of guide RNAs that would allow for sufficient binding to the RNA of interest takes time and the system in which the guide RNA is tested may not always translate between different cell types.

Though MCP-Arc and mCherry-MS2 plasmid constructs were generated, I have chosen to work with the dCas13 RNA targeting system due to its flexibility and different variants of the enzyme being available. The Elowitz lab has kindly provided shortened versions of dLwaCas13a and dPspCas13b enzymes to be fused with Arc for targeted RNA trafficking.

Given that Pastuzyn, et al. have shown that GFP fused to Arc allows for functional RNA transfer between donor and recipient cells, I have chosen to focus on characterizing dCas13 fused to N-terminus of Arc (dCas13-Arc). Additionally, given the smaller size of shortened dLwaCas13a enzyme (1092 bp) versus shortened dPspCas13b enzyme (2379 bp), much of the characterization was focused on dLwaCas13a-Arc fusion.

3.3 dCas13-Arc fusions produce monomers in mammalian cells

Figure 2 illustrates the functional concept for the dCas13-Arc system in which dCas13-Arc monomers bind to target RNA using respective guide RNA, oligomerize into a capsid-like structure, get packaged into an exosome, and get expelled into extracellular space. Constructs of dCas13-Arc fusions are illustrated in Figure 2.b, in which dCas13 enzyme is fused with Arc at its N terminus. As mentioned in the previous section, this design was selected based on demonstration of functional RNA transfer by GFP-Arc as seen in Pastuzyn, et al. It is important to acknowledge the potential risk of placing dCas13 at Arc's N terminus. It is not understood whether the dCas13 holding RNA of interest via the guide RNA would end up inside the capsid-like structure with the RNA well protected or whether some and/or all dCas13 would have its RNA exposed externally, risking RNAse degradation. To first determine whether dCas13-Arc monomers are produced by mammalian HEK293 cells and their respective size, I have performed fluorescent western blot analysis on cell lysate of cells transfected with dCas13-Arc fusions. Arc antibodies were used in conjunction with secondary fluorescent antibodies to mark monomer fusions. Arc is natively a 45.3 kDa protein (396 amino acids). Shortened dLwaCas13a is similarly sized as Arc at 361 amino acids, with an expected size of roughly 43.55 kDa, while shortened dPspCas13b is 92.13 kDa (https://www.bioinformatics.org/sms/prot_mw.html). This places the estimated size of dLwaCas13a-Arc at 88.85 kDa and dPspCas13b-Arc at 137.43 kDa.

As expected, mammalian HEK293 cells express dLwaCas13-Arc at roughly below 100 kDa and dPspCas13b-Arc between 100 kDa and 150 kDa. Negative control in which no transfection occurred showed no clear band of detection for Arc.

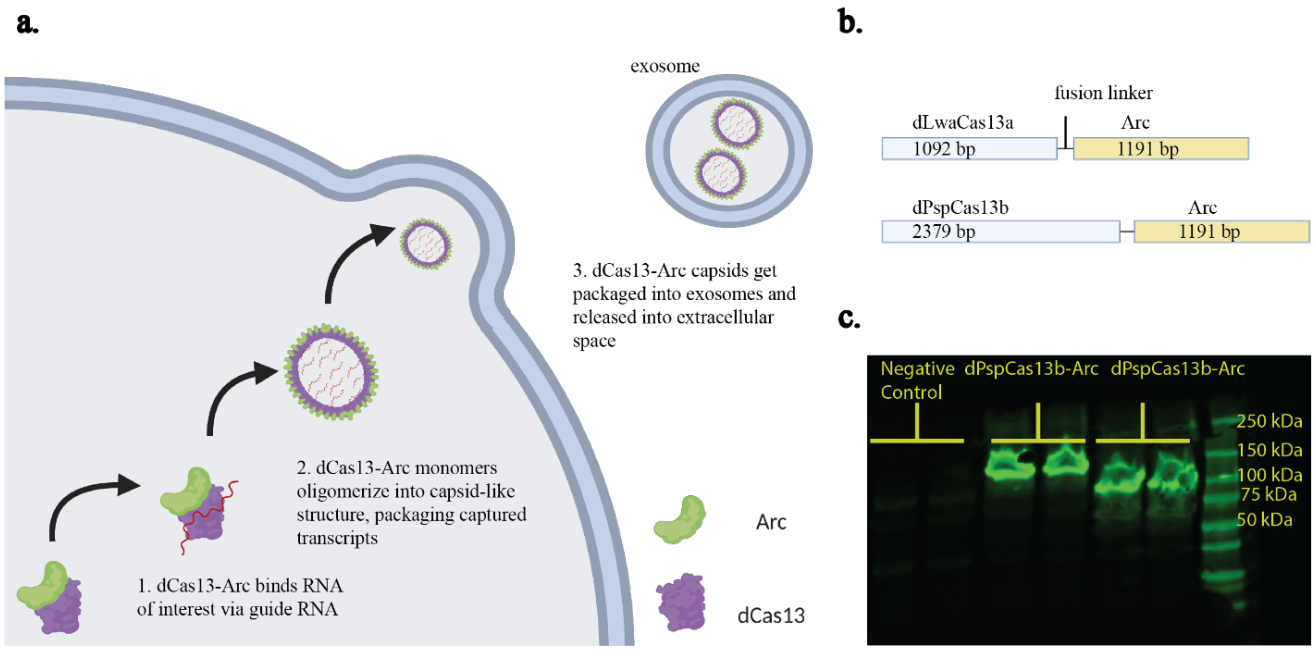


Figure 2. Determining dCas13-Arc monomer size in human HEK293 cells (A) Conceptualization of Arc RNA export system using dCas13 enzyme and guide against target RNA (B) dCas13-Arc configuration and size (C) Fluorescent Western Immunoblot stained against Arc antibody in transfected dCas13-Arc HEK293 cells.

Additionally, HEK293 cells were chosen due to them not natively expressing Arc protein, eliminating potential confounding effects that would otherwise be present with cells such as neurons.

3.4 Methods for characterizing formation of dLwaCas13a-Arc capsids

In order to determine whether dCas13-Arc fusion is capable of making capsids I have purified cell medium from HEK293 cells transfected with dLwaCas13a-Arc, GFP-Arc, and myc-Arc. The transfections were done using FUGENE reagent and the medium was changed 12 hours post transfection to eliminate free floating plasmid DNA that would potentially confound the signal. After 18-24 hours, the cell medium from ten 150 mm plates were concentrated using Amicon Ultra-15 mL centrifugal filters with molecular weight cutoff 30 kDa. Concentrating the cell medium is necessary to get sufficient amount of Arc EVs given their poor release efficiency. The concentrated cell medium was then passed through Izon's qEV10 gravity column which flows through extracellular vesicles (EV) such exosomes, while trapping smaller proteins (Figure 3.a). The fractions collected from

the column were then run on western blot with fluorescent immunostaining against Arc to determine whether denatured dCas13-Arc capsids produce same sized bands seen in cell lysate of Figure 2.c. GFP-Arc and native Arc were used as positive controls. Fractions with brightest western blot lanes were then used for Transmission Electron Microscopy (TEM).

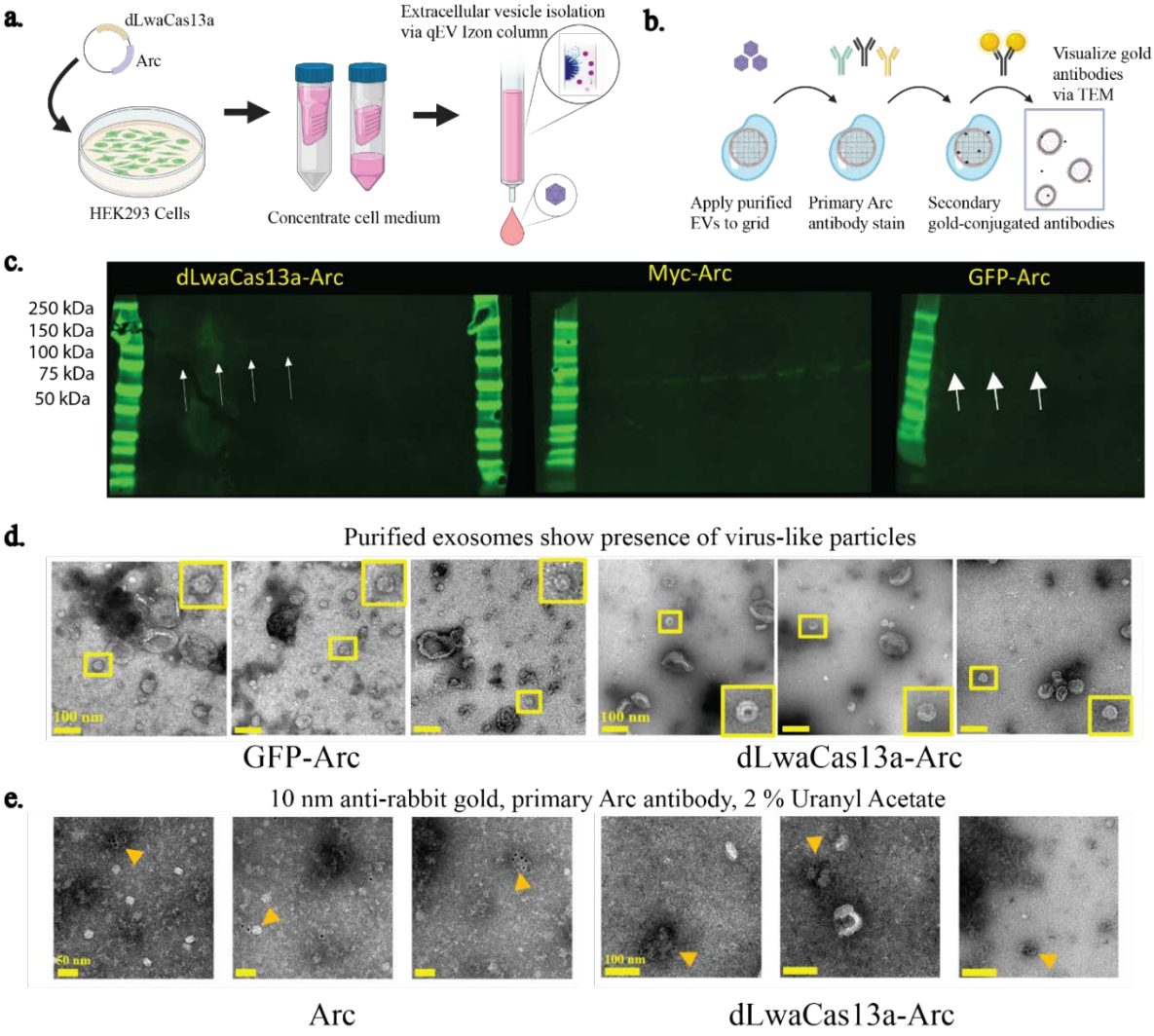


Figure 3. Characterization of dLwaCas13a-Arc capsid formation in HEK293 cells (A) Visualization of protocol for purification of extracellular vesicles (EVs) from mammalian cells. Large molecules, like exosomes and capsids, get passed through while small proteins get stuck to the column. (B) Protocol for staining purified EVs and capsids against Arc using gold antibodies for EM. (C) Fluorescent Western Immunoblot stained against Arc antibody to determine EV fractions most abundant in Arc (D) Representative EM images of purified GFP-Arc and dLwaCas13a-Arc EVs showing presence of capsid-like structures (E) Representative EM images of purified EVs stained in accordance to protocol illustrated in (B).

Figure 3.c shows four faint bands around 100 kDa for dLwaCas13a-Arc. Band at 50 kDa and 75 kDa was observed respectively for native Arc and GFP-Arc as expected. Electron microscopy (EM) images of dLwaCas13a-Arc and GFP-Arc show presence of exosome and capsid-like structures as illustrated in unstained purified EVs (Figure 3.d). In order to verify that the capsids observed in the EM images are Arc capsids, the purified EVs were stained using Arc antibodies. As per protocol outlined in Korkut et, al., purified fractions were then fixed with 2 % PFA, transferred to EM grid, washed with 0.1M Tris to quench the fixation, blocked with saponin to permeabilize the exosomes, and then stained with Arc antibodies and secondary antibodies with conjugated 10 nm gold beads (Figure 3.b). I performed the protocol across numerous condition of antibody incubation wait times, wash steps, and reagents. However, as can be seen in Figure 3.e, much of the exosome and capsid-like structures ended up being broken down into its oligomer components or washed away. An array of block dots, the 10 nm gold particles attached to Arc antibody, surrounding oligomer chunks can be seen in Figure 3.e. I presume that the Arc-gold antibody conjugate are binding Arc structures that did get expelled into medium by the cell for the purpose of extracellular RNA transfer. However due to a lack of clear capsid-like structures being tagged with the Arc gold antibodies across both dLwaCas13a-Arc and the positive Arc control, it is challenging to confirm and characterize dLwaCas13a-Arc ability to form capsid-like structures resembling those of Arc in HEK293 cells.

3.5 dLwaCas13a-Arc fusions generate capsid-like structure and encapsulate Arc RNA

Given the challenges posed by purification and characterization of Arc capsids in HEK293 cells, such as poor release efficiency of Arc and heterogeneous exosome content generated by HEK cells, BL21 bacterial cells were used for Arc capsid purification, characterization,

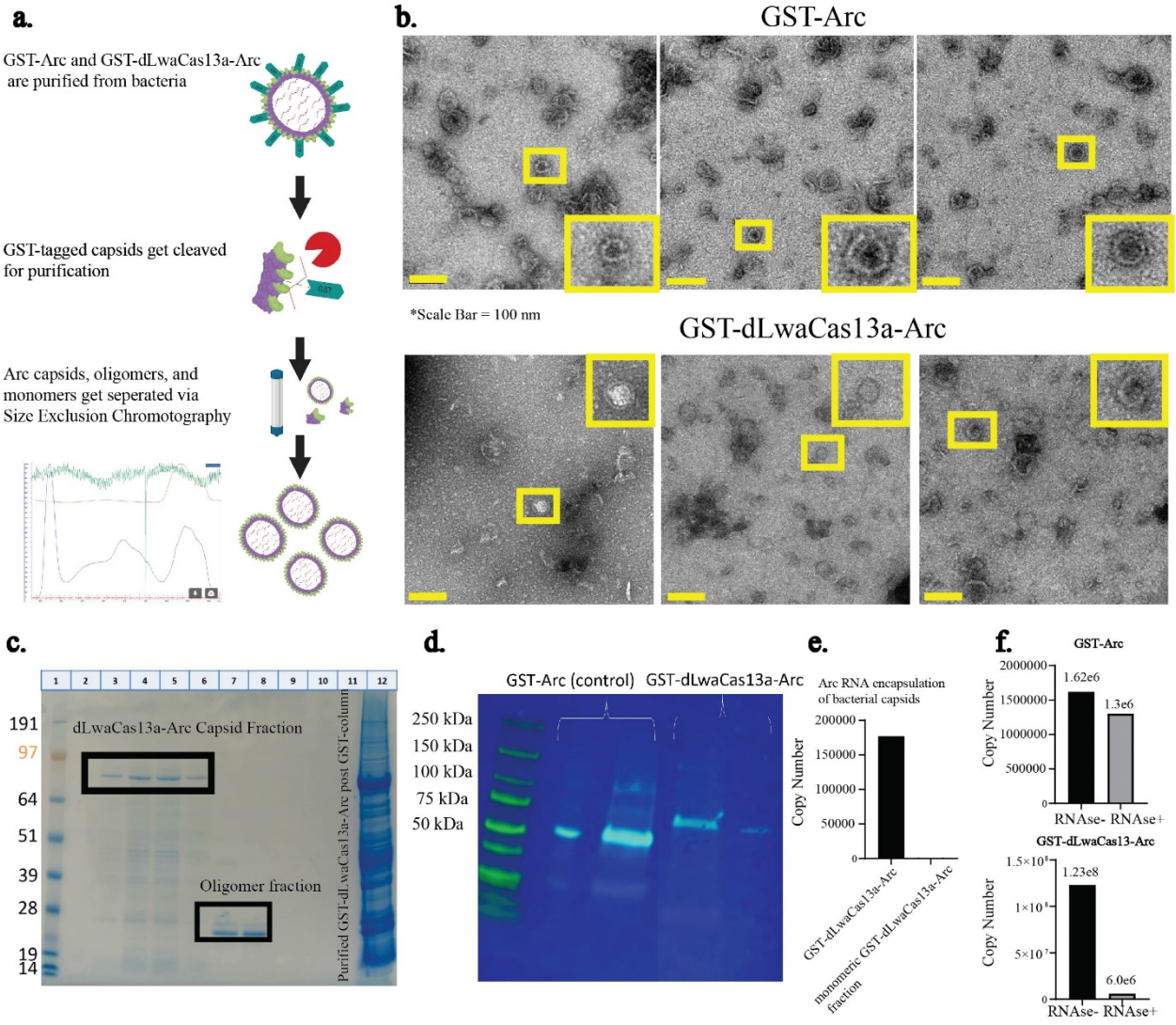


Figure 4. dLwaCas13a-Arc forms capsid structures and encapsulated Arc RNA (A) Visualization of protocol for purifying GST-tagged Arc capsids from BL21 bacterial cells. SEC is used to filter our capsids from monomers and tetramers (B) Representative EM images of purified Arc and dLwaCas13a-Arc capsids from bacterial BL21 cells (C) Coomassie gel of fractions purified from SEC. Lane 12 is purified dLwaCas13a-Arc post GST-column and pre-SEC (D) Fluorescent Western Immunoblot against Arc antibody for purified Arc and dLwaCas13a-Arc capsid fractions (E) RT-qPCR of dLwaCas13a-Arc capsid fraction and monomeric dLwaCas13a-Arc fraction using Arc probes (F) RT-qPCR of Arc and dLwaCas13a-Arc pre- and post- RNAse treatment.

and validation of RNA encapsulation. Plasmids with Glutathione S-transferase (GST) tag fused to N-terminus of Arc and dLwaCas13a-Arc were transformed, corresponding capsids were cleaved in GST columns via PreScission Protease, and separated into capsid, oligomer, and monomeric fractions using size exclusion chromatography (SEC) (Figure 4.a). Neither GST nor Arc are natively expressed in BL21 bacterial cells. As such given that GST is fused with Arc there is certainty in that whatever is captured by the GST column is part of Arc. The protocol for bacterial purification of Arc capsids was adapted from Pastuzyn, et al. and modified with the assistance of protein expression center (PEC) at Caltech. Fully formed spherical shells sized around 50 nm with shell structures can be observed in representative images of both Arc and dLwaCas13a-Arc capsids purified from BL21 cells in Figure 4.b. It also appears that partially formed capsids lacking the shell structure are present, as can be seen in the second image of dLwaCas13a-Arc (Figure 4.b). Coomassie gel shows a prominent band between 64 and 97 kDa for dLwaCas13a-Arc capsid fraction, in agreement with the expected size of 88.5 kDa for a dLwaCas13a-Arc fusion (Figure 4.c). Fluorescent Western Immunoblot of Arc and dLwaCas13a-Arc capsid fractions show Arc near the 50 kDa band and dLwaCas13a-Arc between 75 kDa and 100 kDa (Figure 4.d). A lack of strong band in Coomassie gel and Western blot for dLwaCas13a-Arc for other sizes suggests that dLwaCas13a-Arc monomers predominantly make up the capsids seen in the EM images. It is also interesting to note that the dLwaCas13a-Arc do not seem to be bigger than Arc, hinting at the capsid needing only a specific amount of monomers to oligomerize properly. Purification of dLwaCas13a-Arc has also been performed by Sheperd Lab at University of Utah for replication purposes where capsids with shell structures were observed (Supplementary Figure 7).

Similar to an Arc capsid, I expected that a dLwaCas13a-Arc capsid would encapsulate its own RNA and protect it from RNAse treatment. To demonstrate this, I have extracted RNA (Direct-zol RNA kit) from Arc and dLwaCas13a-Arc capsid fraction, and from monomer fraction of dLwaCas13a-Arc purified product as negative control. The extracted RNA was converted to cDNA and measured via RT-qPCR using Arc probes mentioned in Pastuzyn, et al. Figure 4.e shows that RNA was detected in one of the dLwaCas13a-Arc capsid fractions (post RNAse treatment) as compared to monomeric fraction of dLwaCas13a-Arc,

hinting that capsid structure composed of dLwaCas13a-Arc monomers are necessary to protect the encapsulated RNA. In a separate purification run, I have compared the dLwaCas13a-Arc and Arc fraction pre- and post RNAse treatment. Arc's titer did not change much with RNA treatment, while not treating dLwaCas13a-Arc capsids with RNAse showed a stark increase in titer (Figure 4.f). It is important to note that with RNAse treatment, dLwaCas13a-Arc's titers are similar to that of Arc, though making solid claims would be unfounded due to lack of repeated experiments. This stark increase could potentially be explained by having some of the dLwaCas13a-Arc forming the capsid hold on to RNA on the outside of the capsid, instead of properly encapsulating it and protecting the captured RNA.

3.6 Methods for validating functional and targeted export of RNA via dLwaCas13a-Arc constructs

In prior sections, I presented data showing that dLwaCas13a-Arc forms capsid structures and protects Arc RNA in BL21 bacterial cells, with capsid-like structures observed in EVs purified from mammalian HEK293 cells. The next step was to determine whether dLwaCas13a-Arc capsids are capable of preferentially picking up target RNA of interest via guide RNA and shuttling it outside the cell. The Elowitz lab has kindly provided stable Citrine HEK293 cells that are induced via tetracycline, as well as the appropriate guide RNA to Citrine. Having stable Citrine HEK293 cells also minimizes the number and variables for transient DNA transfections. If the hypothesis that dLwaCas13a-Arc can preferentially export target RNA of interest is true, then I would expect to see higher titers of Citrine RNA in the cell medium transfected with dLwaCas13a-Arc and Citrine guide RNA compared to dLwaCas13a-Arc co-transfected with pUC (pUC19-common cloning vector). Figure 5.b shows different construct designs for dLwaCas13a and Arc placement. Membrane localizer (ML) constructs with and without Arc was kindly provided by Dr. Bjorkman's lab. The hypothesis for ML-Arc constructs was that higher levels of Arc could be expelled from the cell if more of them localized to the membrane, as such exporting more RNA. A construct in which the dLwaCas13a was fused to C terminus of Arc was also explored due to data presented in prior sections, with the idea being that if dLwaCas13a is

in this configuration then it would fold inside the Arc monomers and protect captured RNA more efficiently. ML-dLwaCas13a constructs were made to explore whether Arc is necessary for expulsion of RNA into the cell medium. Citrine HEK293 cells were transfected in 6-well plates with relevant combination of plasmids specified in Figure 5.c, the cell medium was changed 6 hours post transfection, and then collected following day 18-24 hours post transfection to be concentrated down to 100 ul (triplicates). The concentrated cell medium was then treated with RNAse to remove free floating RNA and to leave only capsids that are protecting encapsulated RNA. The RNA was then extracted (Direct-zol RNA kit), converted to cDNA, and quantified via RT-qPCR using dLwaCas13a and Citrine probes. As can be seen in Figure 5.c, titers for Citrine expelled into cell medium

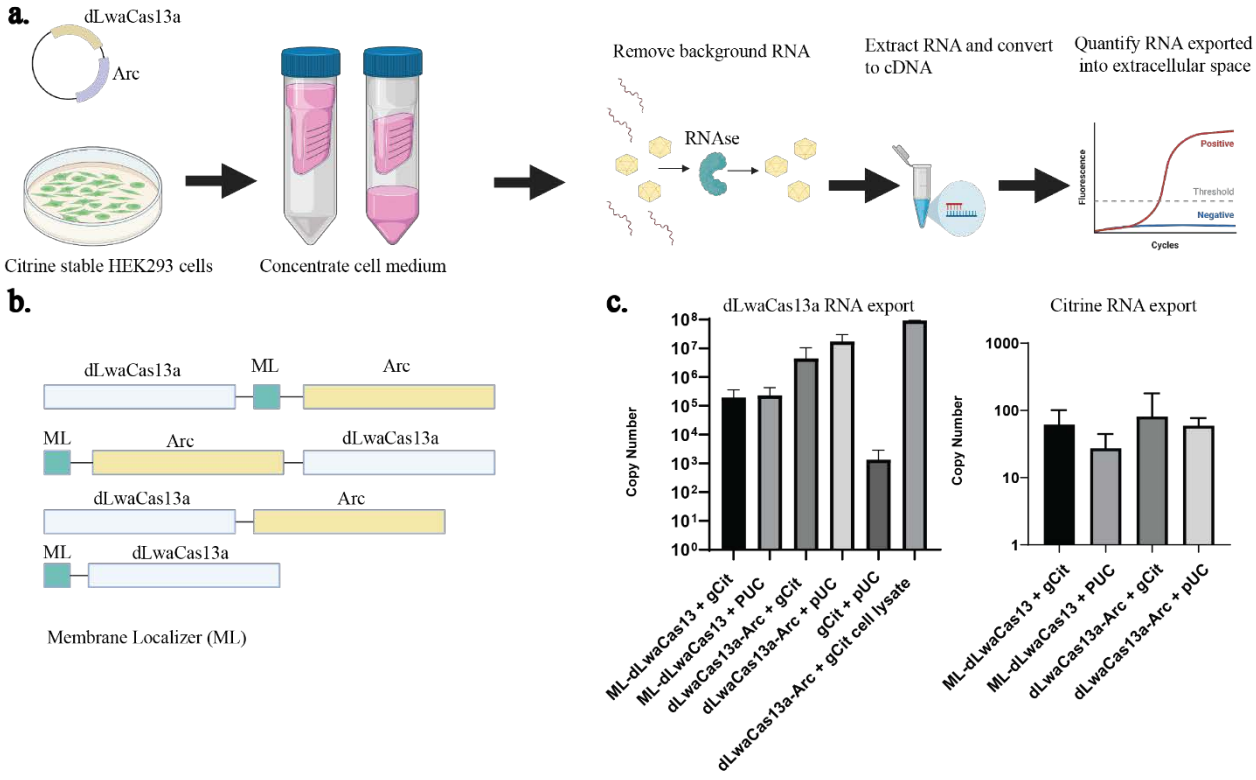


Figure 5. Characterization of targeted RNA export via dLwaCas13-Arc constructs (A) Visualization of RNA export assay, starting with transfection of Citrine stable HEK293 cells, cell medium concentration, RNA extraction, cDNA conversion, and transcript quantification via RT-qPCR (B) dLwaCas13a and Arc constructs used for testing targeted RNA export of Citrine transcripts (C) RT-qPCR results of constructs presented in (B) titered against dLwaCas13a probes. Similarly, results for titer against Citrine transcripts is shown on the right-hand panel (gCit + pUC and cell-lysate from dLwaCas13a-Arc + gCit are not included).

were low across all conditions. However, dLwaCas13a titers were high, expect in the negative control case of guide RNA against citrine (gCit) and pUC, suggesting there is a release and protection of dLwaCas13a transcripts by the dLwaCas13a-Arc capsid, but not Citrine.

3.7 Limitations and Challenges

Getting Arc to target RNA of interest and export it out of the cell is a proof of concept project with numerous limitations, challenges, and risky components. First, the shortened dLwaCas13a and dPspCas13b were still in development and characterization phase when I obtained them from Elowitz lab. There is likely more optimization and validation left to confirm their functionality, especially when fused to another protein. Ideally, this sort of validation would be done with an RNA immunoprecipitation protocol (RIP) in which Arc antibodies can be used to capture the dCas13-Arc capsids and protein, RNA precipitated, converted to cDNA, and quantified using RT-qPCR. If the guide RNA is more effective in capturing target transcripts then those transcripts would be enriched in cells transfected with dLwaCas13a-Arc/dPspCas13b-Arc and guide RNA versus cells transfected with dLwaCas13a-Arc/dPspCas13-Arc and pUC. I have performed one such RIP experiment (Magna RIP Kit, 17-700, Sigma-Aldrich) in which I had two conditions, one of which was HEK293 cells transfected with dPspCas13b-Arc, guide against Citrine, and Citrine, and the other dPspCas13b-Arc, pUC, and Citrine. I have chosen to do this experiment with dPspCas13b fusion because prior art, as referenced in section 3.2, has shown that dPspCas13b is more effective at targeting transcript of interest compared to dLwaCas13a. HEK293 cells were transfected using FUGENE transfection reagent and cell lysates were processed 24 hours post transfection. Arc antibodies were used to capture dPspCas13b-Arc proteins, while normal rabbit IgG served as control. RNA extracted from the two conditions was quantified using RT-qPCR against Citrine and Arc probes, as seen in Figure 6. RIP assays are expensive and known to be noisy requiring multiple repetitions, as such it is important to keep in mind that experiment presented is an N=1. The first two bar graphs show RT-qPCR titer results for Citrine and Arc probes respectively. In both cases the signal for transcripts captured by Arc antibody is several fold higher than the control samples,

suggesting that Arc antibody captured dPspCas13b-Arc protein fragments and their respective transcripts. However, as also seen in the first bar plot, there is not a large fold difference of Citrine transcripts with guide RNA against Citrine (gCit) versus pUC. The intent of the third bar plot is to determine how well dPspCas13b-Arc is detargeted from encapsulating its own RNA, it appears that there is a log fold difference in Citrine transcripts captured compared to Arc. However, without further experimental repeats it is difficult to pinpoint whether that effect is driven by noise or intended biological phenomenon. It is also interesting to observe that even with the lack of a guide RNA, dPspCas13b-Arc captured more Citrine transcripts with pUC co-transfected compared to Arc transcripts. Though all plasmids were transfected in identical ratios, it is possible that the cells produce more Citrine transcripts. For example, do cells produce more Citrine if you feed them oranges from Citrus tree? To answer that, I could vary the amount of Citrine transfected to determine whether the amount of Citrine transcripts picked by dPspCas13b-Arc correlate to the amount of Citrine transfected. It would also be more ideal to perform this kind of experiment in Citrine stable cells to eliminate a transfection variable. Similarly, the membrane localizer was received from Bjorkman Lab as a plug-and-play component

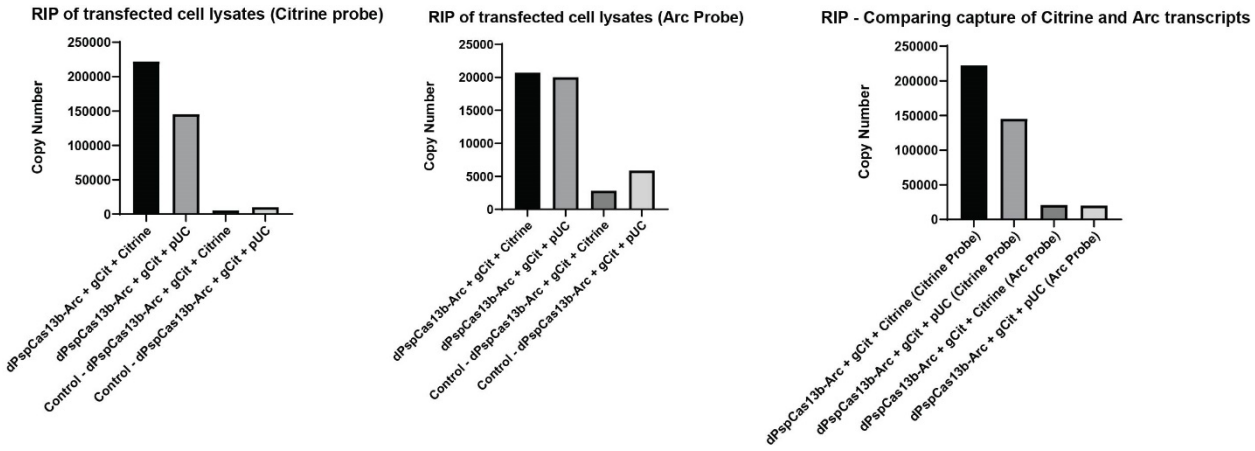


Figure 6. RNA Immunoprecipitation (RIP) RT-qPCR characterization of dPspCas13b-Arc constructs

for the dCas13-Arc system and not optimized accordingly. Second, though Pastuzyn, et al. have demonstrated successful transfer of both Arc and GFP RNA in the donor-recipient assay as seen in their Figure 4 for HEK293 cells, the efficiency of this transfer is low as

suggested by their Supplementary Figure 4. Low transfer rate could be due to poor release efficiency of Arc into the cell medium, uptake efficiency of Arc exosomes by recipient cells, or a combination of both. Performing donor-recipient and release assays is also a challenge in neuronal cells given the large amounts of cells needed, Arc being natively expressed in neurons, and constant need of extracting primary neurons from either rat or mouse pups. Given that Arc is a well-known early response gene in neuronal cells and responsible for mediating plasticity in learning tasks, it likely has numerous mechanisms regulating its release and transfer efficiency. For example, Zhang, et al. have shown the complexity of Arc's function where while Arc's N- and C-lobe control high-order oligomerization of Arc, they are not responsible for Arc association with vesicle proteins in order to be exported from the cell. My suspicion is that the lack of a clear signal seen in Figure 5 is likely due to poor release efficiency of Arc, making it difficult to assess dCas13-Arc's targeted RNA export. Third, in addition to Arc's role in neurons still being rigorously researched, its function in immortalized cell lines such as HEK293 cells is also not well understood. Thus, using HEK293 cells to develop a specific function for a protein that is designed to work in complex cells such as neurons is not ideal.

3.8 Discussion

I was able to demonstrate that dLwaCas13a-Arc fused to N-terminus of Arc is able to create spherical shelled capsids similar to that of Arc, encapsulate its own RNA, and protect it from degradation. Additionally, I have validated a method for purifying Arc and dLwaCas13a-Arc EVs from HEK293 cells that recover monomers of expected size. It is still not clear whether the dLwaCas13a enzyme is facing the outside of the capsid holding on to the transcripts it has captured via guide RNA, with Arc packaging other nearby transcripts inside the capsid. Generating a construct in which dLwaCas13a is on the C-terminus of Arc could answer this curiosity. Lack of clear signal demonstrating that dLwaCas13a-Arc prefers packaging RNA of interest could have been confounded by much of its captured transcripts being lost to RNAse degradation if they are external to the capsid. It would also elucidate whether capsid formation is possible with proteins and/or enzyme being placed on C-terminus of Arc.

Given the challenges and limitations encountered in engineering Arc to target specific RNA for export, I will discuss strategies going forward. One of the biggest challenges in working with Arc is its poor release efficiency, as such this function could be optimized via directed evolution. Since the release of Pastuzyn, et al. work showing that Arc forms capsids and transports RNA, numerous works have emerged further unveiling information on how Arc's sequence and structure correlates to its function. In 2019, Nielsen, et al. have identified that Arc interacts directly with several neuroreceptors, such as GluN2A and GluN2B, and that these interactions interfere with oligomerization of Arc into capsids. Binding to these ligands stabilizes monomeric state of Arc and inhibits formation of higher order Arc oligomers. In 2018, Hallin et al. have shown that the interaction between Arc and the cellular domain is mediated by the positively charged N-terminal Domain (NTD) binding to phospholipid head groups. Additionally, both Hallin, et al. and Nielsen, et al. have shown that homo-oligomerization of Arc capsids is absent without the NTD. Zhang, et al. have also introduced mutations into Arc to identify sections responsible for Arc oligomerization and their data suggest that N-terminal half of Arc could play a role in association with vesicle proteins. Given these findings, NTD site on Arc (AA 17-154) could be a good candidate for error prone PCR mutagenesis to optimize association with cell membrane and vesicle proteins. Mutating regions in N-lobe and C-lobe could also optimize oligomerization efficiency. In order to select an Arc mutant which increases Arc's release efficiency, libraries at site of interest could be transfected into HEK293 cells, cell medium collected and concentrated, and RNA extracted and converted to cDNA. Arc variants could then be amplified from the cDNA and sequenced to determine the most abundant mutant.

In 2019, Fukuda, et al. have constructed guide RNAs which induce A-to-I mutations by guiding adenosine deaminase acting on RNA (ADAR), RNA editing enzyme, to target site. Similarly, given Arc's tendency to bind nearby RNA, such designer guide RNAs could be created to guide Arc to target RNA without having to fuse it with an enzyme or another protein.

Though Arc is an appealing option for targeting and exporting RNA given its natural presence in humans, other systems could be used to accomplish similar task that are more

well characterized. For example, Group-specific antigen (Gag) domain of HIV and Equine Infectious Anemia Virus (EIAV) contain p6 and p9 motifs, respectively, which bind ESCRT proteins and assist in viral budding (Bhaskar, et al. 2017, Jalaguier, et al. 2011). Fusion of HIV's GAG domain with dCas13 enzymes has the potential to achieve similar targeted RNA export and have a more efficient release due its viral properties.

Another impact that a system capable of exporting RNA to extracellular space and neighboring cells could have is to amplify a target therapeutic gene. For example, if there is a synthetic receptor that gets triggered by ligand, it could activate the production of Arc that can bind and deliver RNA of interest from the source cell to neighbors. This could reduce the need for highly efficient transduction by gene delivery methods, as the content that is already delivered could be further spread by the RNA export system (Supplementary Figure 8).

In conclusion, engineering a system to selectively pick up RNA of interest within a cell and traffick it either to extracellular space or to its neighbors remains a challenge. A lot of components necessary to make such a system are still being studied today, and development of new methods is necessary to readout both the export efficiency as well as disturbance caused to the cell state post RNA removal. However, results discussed in this chapter demonstrate glimpses of promise and exciting engineering avenues to achieve such an RNA export and trafficking system.

3.9 Acknowledgements

dLwaCas13a and dPspCas13b constructed were provided by Michael Flynn in Elowitz lab. Membrane Localization constructs were provided by Magnus Hoffman in Bjorkman lab. All Arc plasmids were provided by Jason Shepherd lab. Purification of dLwaCas13a-Arc capsid was also replicated by members of Jason Sheperd lab. Protein Expression Center at Caltech performed bacterial purifications of dLwaCas13a-Arc and Arc. All electron microscopy images were taken at Caltech's EM Imaging Facilities.

3.10 Methods

Size of Arc monomers and its respective fusions were calculated using https://www.bioinformatics.org/sms/prot_mw.html

Fluorescent Western Immunoblotting

HEK293 Cells were transfected using FUGENE transfection reagent with dLwaCas13a-Arc and dPspCas13b-Arc plasmids in 100 mm TC-treated plates. 18 hours post-transfection the cells were spun down to rid of cell medium and 300 ul of RIPA buffer (1X RIPA Lysis Buffer, 1M Magnesium Chloride, 0.5 % DNAse, 0.02 % Pierce Universal Nuclease) was added. The cells resuspended in RIPA buffer were then incubated at 37 C while mixing at 950 RPM for 30 minutes. Spin down the cell lysate and combine equal volume sample with sample buffer (10 % 2-Mercaptoethanol in 2X Laemmli Sample Buffer). Samples were ran on gel using appropriate directions in accordance with the vendor (i.e. Bio-Rad, Running Buffer: 10 % 10X Tris/Glycine/SDS Buffer in Nanopure Water for 35 minutes at 150 V). The gel was removed from the cassette and activated via chemical crosslinking using UV (stain free Bio-Rad Imager). Blotting membrane paper was prepared by soaking it in 100 % Methanol for 5 minutes. Gel was transferred onto blotting paper (e.g. Bio-Rad Immun-Blot PVDF membrane) and placed accordingly unto a blot sandwich configuration. The blot was ran in accordance with manufacturer's instructions (Blotting Buffer: 10 % Methanol, 10 % 10X Tris/Glycine Buffer, 80 % Nanopure water, 100 V for 30 minutes). Blotting paper was blocked in non-fat dry milk (NFDM) for 1 hour at room temperature (RT). Primary antibody staining was performed in blocking buffer (1 % NFDM in TBST) overnight at 4 C (Recommended antibody dilution 1:500). Blotting paper was washed in TBST three times in order to remove unbound antibodies. Secondary fluorescent antibody was incubated in TBST for 2 hours at RT and then washed three times to remove unbound antibodies. Blots were detected using Bio-Rad Imager.

Extracellular Vesicle Purification from HEK293 Cells

Ten 150 mm plates with HEK293 cells cultured to 80 % confluence were transfected with dLwaCas13a-Arc, Arc, and GFP-Arc respectively. Cell medium (5 % FBS, DMEM) was

changed 12 hours post transfection. 24 hours post-transfection, exchanged medium was collected, spun down to remove debris and free floating cells, and concentrated using Amicon Ultra-15 mL centrifugal filters. Izon's qEV10/70 nm columns were used per vendor's instructions. Fractions were collected and processed through western immunoblotting to determine fractions with highest Arc content.

BL21 Bacterial Purification of Arc Capsids

BL21 bacterial cultures transformed with dLwaCas13a-Arc and Arc plasmids were grown with Carbenicillin. Terrific Broth was prepared with glycerol and autoclaved for each protein purification (1 L). The culture were then shaken at 250 RPM for 1.5 – 2 hours. When optical density (OD) reached 0.6 IPTG was added to induce the bacterial cultures and then the cultures were shaken overnight at 250 RPM at 27 C. Bacterial pellet was spun down 6000 RPM for 15 minutes and weight recorded to determine lysis buffer amount (1 mg = 100 ul lysis buffer). Pellets were lysed in 1X TBS, PMSF and Roche tablets. Samples were collected throughout the protocol for quality assurance. Samples were spun down for 30 minutes at 6000 RPM to get a clear cell lysates. The lysates were loaded unto GST-trap columns, washed with 20CV 1XTBS, and then 12CV 1XTBS, 1 mM EDTA, 1 mM DTT buffer. To cleave GST-captured proteins, 4.5-5 mL of PreScission Protease (PP) in PP buffer (1XTBS, 1mM EDTA, 1 mM DTT) was loaded into columns using column adapter. The columns were left overnight and eluted with 1XTBS, 1 mM EDTA, 1 mM DTT buffer the following day. Prior to feeding the samples through Size Exclusion Column (SEC) the samples were buffer exchanged with 50 mM Tris, 150 mM NaCl, pH 8.0 buffer.

RNA Cell Medium Export Assay and RT-qPCR

Citrine stable HEK293 cells were transfected in 6-wells (in triplicates) with DNA plasmids using FUGENE reagent. Prior to transfection, Citrine expression was induced using Tetracycline (200 ng/mL recommended). Cell medium (5 % FBS, DMEM) was exchanged 6 hours post-transfection. 18-24 hours post-transfection the cell medium is spun down to remove dead cells and debris, concentrated using Amicon Ultra-4 mL to 100 ul and treated with RNAse to remove free floating RNA for 15 minutes at 37 C. Direct-zol RNA Microprep kit was used to extract RNA and ThermoFisher's High-Capacity cDNA Reverse

Transcription Kit was used to convert RNA to cDNA. The extracted cDNA was then quantified using RT-qPCR (2 ul of DNA was loaded per well, 9.5 ul Ultra-Pure Water, 12.5 ul FastStart Universal SYBR Green Master, 0.5 ul of 5 uM titer primers). Cycling parameters: Step 1 - 95 C for 10 min, Step 2 - 95 C for 15 seconds, Step 3 – 60 C for 60 seconds, Repeat Step 2 – 4 40X. Arc primers adapted from Pastuzyn, et al. 2018

Cit-forward: ATCTTCTTCAAGGACGACGG

Cit-reverse: TTGTGGCTGTTGTAGTTGTA

dLwaCas13a-forward: GACCAAGGTCGACGGCATCA

dLwaCas13a-reverse: TCCAGCCGGATGCTCAGCAG

3.11 Supplementary Material

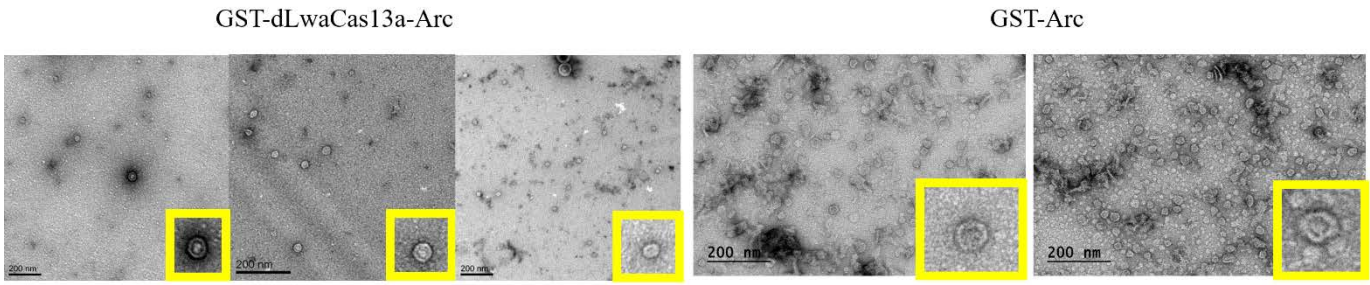


Figure 7. Bacteria (BL21) purification of dLwaCas13a-Arc and Arc performed by Shepherd Lab for scientific replication purposes

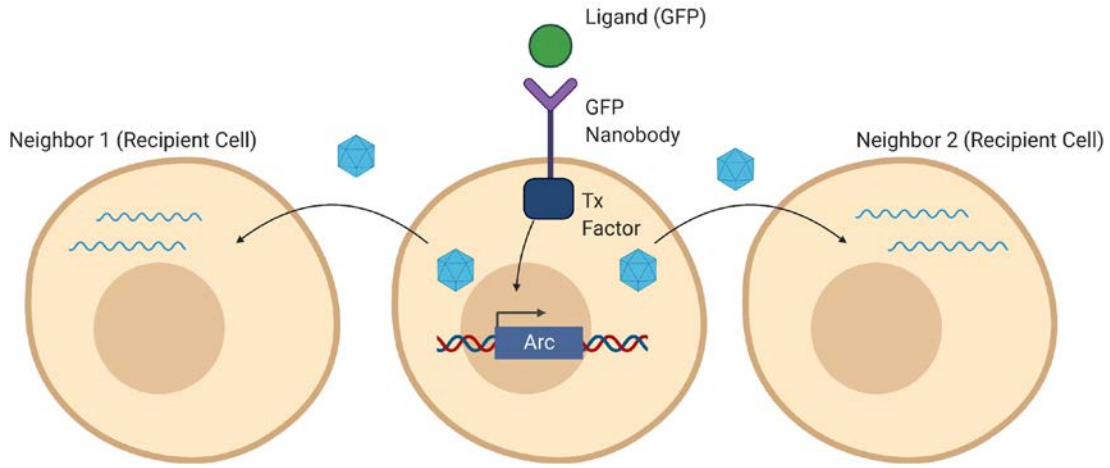


Figure 8. Concept of using Arc RNA export system to amplify target therapeutic gene in neighboring cells upon ligand stimulation

3.12 References

Ranum, L. P. W. & Cooper, T. A. RNA-MEDIATED NEUROMUSCULAR DISORDERS. Annu. Rev. Neurosci. 29, 259–277 (2006).

Speck, S. H. & Ganem, D. Viral Latency and Its Regulation: Lessons from the γ-Herpesviruses. Cell Host & Microbe 8, 100–115 (2010).

Ghosh, R. & Tabrizi, S. J. Gene suppression approaches to neurodegeneration. Alz Res Therapy 9, (2017).

Damase, T. R. et al. The Limitless Future of RNA Therapeutics. Front. Bioeng. Biotechnol. 9, (2021).

Pastuzyn, E. D. et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. Cell 172, 275-288.e18 (2018).

Abudayyeh, O. O. et al. RNA targeting with CRISPR–Cas13. Nature 550, 280–284 (2017).

Cox, D. B. T. et al. RNA editing with CRISPR-Cas13. Science 358, 1019–1027 (2017).

Tutucci, E. et al. An improved MS2 system for accurate reporting of the mRNA life cycle. Nat Methods 15, 81–89 (2017).

Han, S. et al. RNA–protein interaction mapping via MS2- or Cas13-based APEX targeting. Proc Natl Acad Sci USA 117, 22068–22079 (2020).

Shmakov, S. et al. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Molecular Cell 60, 385–397 (2015).

Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science 356, 438–442 (2017).

Korkut, C. et al. Regulation of Postsynaptic Retrograde Signaling by Presynaptic Exosome Release. Neuron 77, 1039–1046 (2013).

Zhang, W. et al. Arc Oligomerization Is Regulated by CaMKII Phosphorylation of the GAG Domain: An Essential Mechanism for Plasticity and Memory Formation. Molecular Cell 75, 13-25.e5 (2019).

Nielsen, L. D., Pedersen, C. P., Erlendsson, S. & Teilum, K. The Capsid Domain of Arc Changes Its Oligomerization Propensity through Direct Interaction with the NMDA Receptor. Structure 27, 1071-1081.e5 (2019).

Hallin, E. I. et al. Structure of monomeric full-length ARC sheds light on molecular flexibility, protein interactions, and functional modalities. J. Neurochem. 147, 323–343 (2018).

Fukuda, M. et al. Construction of a guide-RNA for site-directed RNA mutagenesis utilising intracellular A-to-I RNA editing. Sci Rep 7, (2017).

Bhaskar, S. & Lim, S. Engineering protein nanocages as carriers for biomedical applications. NPG Asia Mater 9, e371–e371 (2017).

Jalaguier, P., Turcotte, K., Danylo, A., Cantin, R. & Tremblay, M. J. Efficient Production of HIV-1 Virus-Like Particles from a Mammalian Expression Vector Requires the N-Terminal Capsid Domain. PLoS ONE 6, e28314 (2011).

Chapter 4

ENGINEERING TRANSCRIPTION FACTOR EB (TFEB) TO MODULATE MACROPHAGE STATE

4.1 Introduction

Macrophages are the immune system's adaptive soldiers, capable of switching between pro-inflammatory state (M1) and anti-inflammatory state (M2). M1 macrophages are commonly associated with phagocytizing bacteria and secreting proinflammatory cytokines. M2 macrophages are often associated with clearing apoptotic cells, controlling inflammatory response, and promoting wound healing. M2 macrophages have also been subcategorized into M2a, M2b, and M2c divisions based on stimuli and induced transcriptional state. M2a macrophages are known for their high secretion of IL-10 and TGB-beta for purposes of wound healing and taming inflammatory responses. M2b, on one hand, are responsive to immune complexes and bacterial lipopolysaccharides (LPS). M2c are sensitive to glucocorticoids and TGF-beta and are often referred to as deactivated macrophages, macrophages turned from M1 state (Roszer, et al. 2015). Macrophages are also known to be modulated by tumor micro-environment into M2 macrophages in order to minimize inflammatory response necessary for immune cells, such as T cells, to destroy cancerous cells.

Given macrophages' important role in innate immune response and flexibility, they are implicated in numerous immune-related diseases such as rheumatoid arthritis, neurodegeneration, and cancer (Ardura, et al 2019). This makes them both a challenging and attractive candidate for cellular reprogramming. Majority of the work on modulating macrophages has been limited to introducing combinations of cytokines to push them into a particular state. In 2019, Zhang, et al. delivered an mRNA interferon factor 5 to change M2 phenotype within tumor associated macrophages (TAMs) to M1 phenotype macrophage. In 2020, Klichinsky, et al. engineered human macrophages with chimeric antigen receptors (CAR) to direct the macrophages' phagocytic activity against cancerous cells. They were able to show that these engineered CAR-M macrophages express pro-inflammatory signals,

convert TAMs into M1 macrophages, and resist effects of immunosuppressive cytokines. Despite these incredible advances, there remains a need to generate personalized and optimized cell-state specific engineering strategies which can leverage the complex signaling pathways of these chameleon-like cells. We believe that in order to achieve that, we ought to look at engineering transcription factors within the macrophages.

Transcription factors regulate numerous downstream molecules, making them an engineering challenge for approaches that readout only a single or a few variables, such as directed evolution and rational protein design. However, recent advent of single-cell RNA sequencing (scRNA-seq) allows one to readout single cells' transcriptomic state and assess presence of thousands of genes. Our hypothesis is that, similar to how enzymatic activity can be optimized through mutations in its catalytic domains, desired cellular states can be enriched for through mutation of transcription factors via single-cell transcriptomic readout. Certainly, this approach comes with its fair share of challenges. To name a few, cellular states vary between different cell types and organisms, and transcriptomic states in different contexts are still being researched and catalogued. In order to assess our hypothesis, we have selected transcription factor EB (TFEB) as a mutation candidate to modulate macrophage's polarization states.

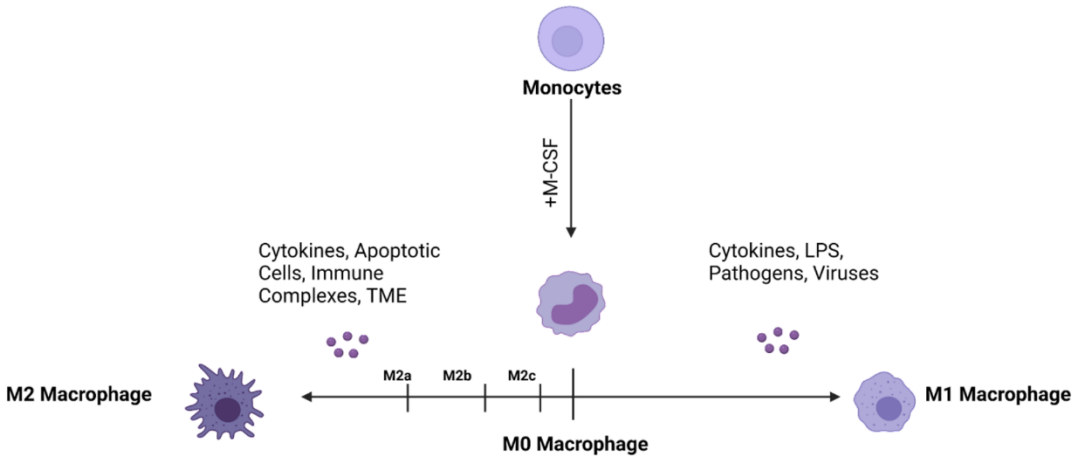


Figure 1. "Macrophage State Ruler" showing the range of states in which macrophages polarize to achieve context specific immune function

4.2 Transcription Factor EB (TFEB)

TFEB is part of MiTF/TFE (microphthalmia) family and is responsible for regulating lysosomal function and autophagy. In particular, TFEB controls lysosomal biogenesis, lysosomal exocytosis and promotes cellular clearance. Prior art has shown that TFEB expression plays an important role in regulating macrophage state (Puertollano, et al. 2018). Fang, et al. 2017 mention that TFEB downregulation results in M2 polarization and tumor-promoting capacity, while TFEB expression is implicated in M1 polarization potentially response for inhibiting tumor growth. Pastore, at al. 2016 have shown that cells deficient in TFEB resulted in reduced inflammatory response. This tie between autophagy and macrophage state makes this duo an ideal candidate for tuning macrophage states.

4.3 Transcription Factor EB (TFEB) mutation strategies

We have selected activation domain (AD) and DNA binding domain (DBD) of TFEB as mutation candidates to determine whether we could optimize an M1 or M2 state of macrophage, or generate a new state.

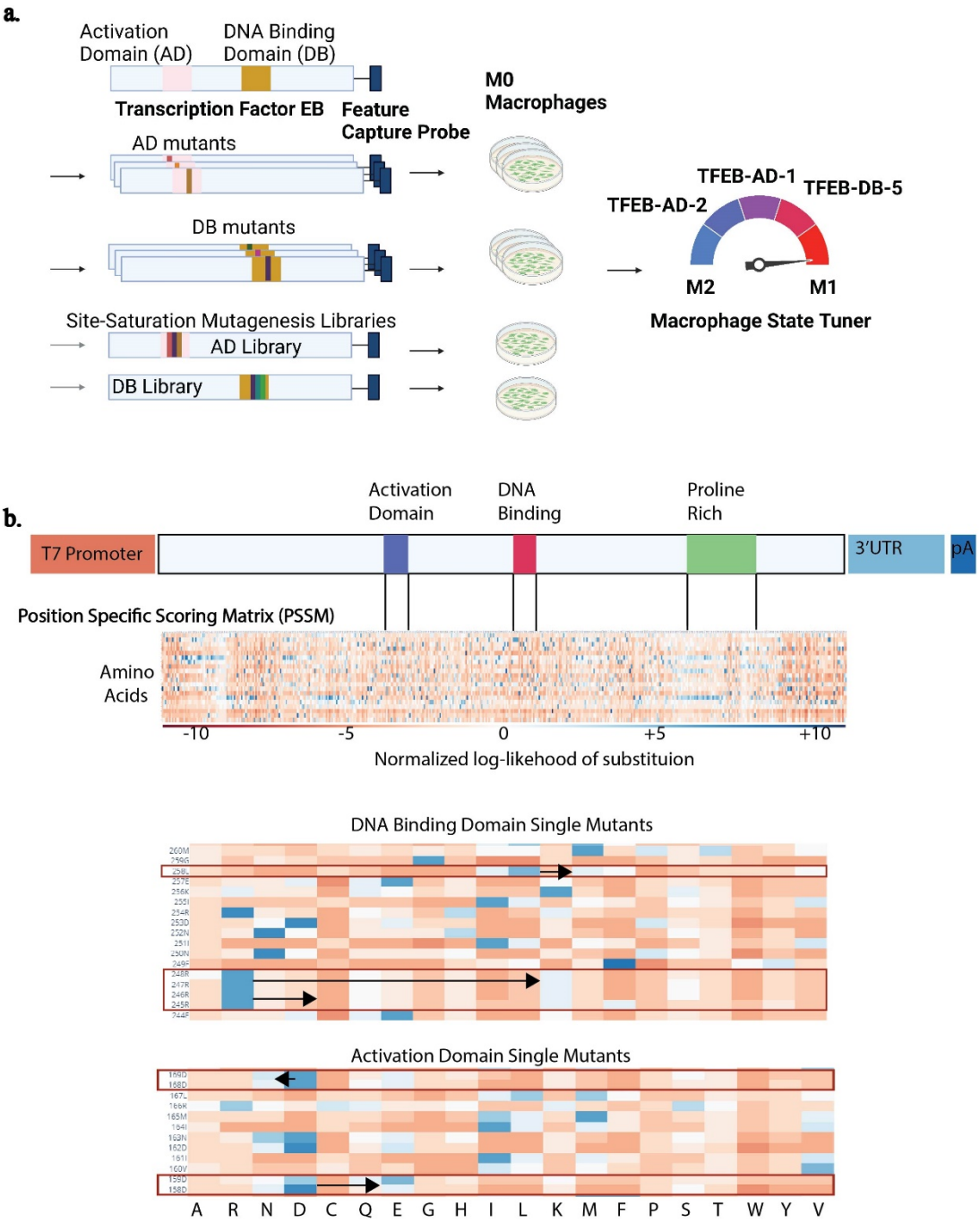


Figure 2. TFEB Mutation Experimental Design and Strategy (A) An experimental concept for rationally designed TFEB mutants and libraries at activation domain (AD) and DNA binding domain (DB) (B) Design strategy for rationally selected mutants using Position Specific Scoring Matrix (PSSM). Arrows indicate what the current amino acids at those positions are and point to the amino acids which will be substituted for those positions. High positive normalized log-likehood indicates that the amino acid is more likely at the given position.

The rationale for choosing those domains is their functional importance, in particular AD is required for its transcriptional activation function and DNA binding domain is critical to have TFEB bind to DNA in the form of homodimers and heterodimers (Napolitano, et al. 2016, Hemesath et al., 1994, Pogenberg et al., 2012). Site-saturated mutagenesis libraries and rationally-chosen mutants were generated in AD and DB domain of TFEB. In order to rationally choose mutants in AD and DB domains, we fed the TFEB sequence into Protein Structure Prediction (ProSPr) to compute normalized log-like likehood of substitution at a given position (Billings, et al. 2019). The Position Specific Scoring Matrix (PSSM) heat map is shown in Figure 2.b, where positive blue colors indicate that the amino acid's presence is more likely to be in that position, whereas red colors indicate that amino acid's presence is unlikely to be at that position. Three mutants were created in DNA binding domain and two in Activation Domain. Arginine in positions 245-248 was changed to the next most likely predicted amino acid, Lysine. Another mutant that would likely have a disabling effect on the DNA binding domain was generated by changing Arginine in 245-248 to one of the least likely predicted amino acids, Cysteine. We have also chosen a mutant that creates a single-point mutation at position 258 from Leucine to Methionine (Figure 2.b). For AD, at positions 158 and 159 Aspartic Acid was mutated to the next most likely amino acid, Glutamic Acid. Similarly, at positions 168 and 169 Aspartic Acid was mutated to Asparagine. In order to make RNA libraries to transfect into cultured macrophages, TFEB was cloned into an RNA backbone containing T7 promoter, 3'untranslated region (UTR), and a poly-A tail via overlap polymerase chain reaction (PCR). PCR fragments were then made into RNA using in vitro transcription, 5'capping, and poly-A tailing kits. In vitro transcription kit with modified nucleosides was used for production of RNA to ensure macrophages can be transfected and minimize their immune response to the delivered cargo. The RNA was capped post transcription and poly-A tail extended to ensure stability.

4.4 Exploring DNA transfection for introduction of TFEB libraries

In order to generate classical M1 and M2 states in cultured human macrophages and explore whether DNA transfection is a potential route for transfecting TFEB mutants, we

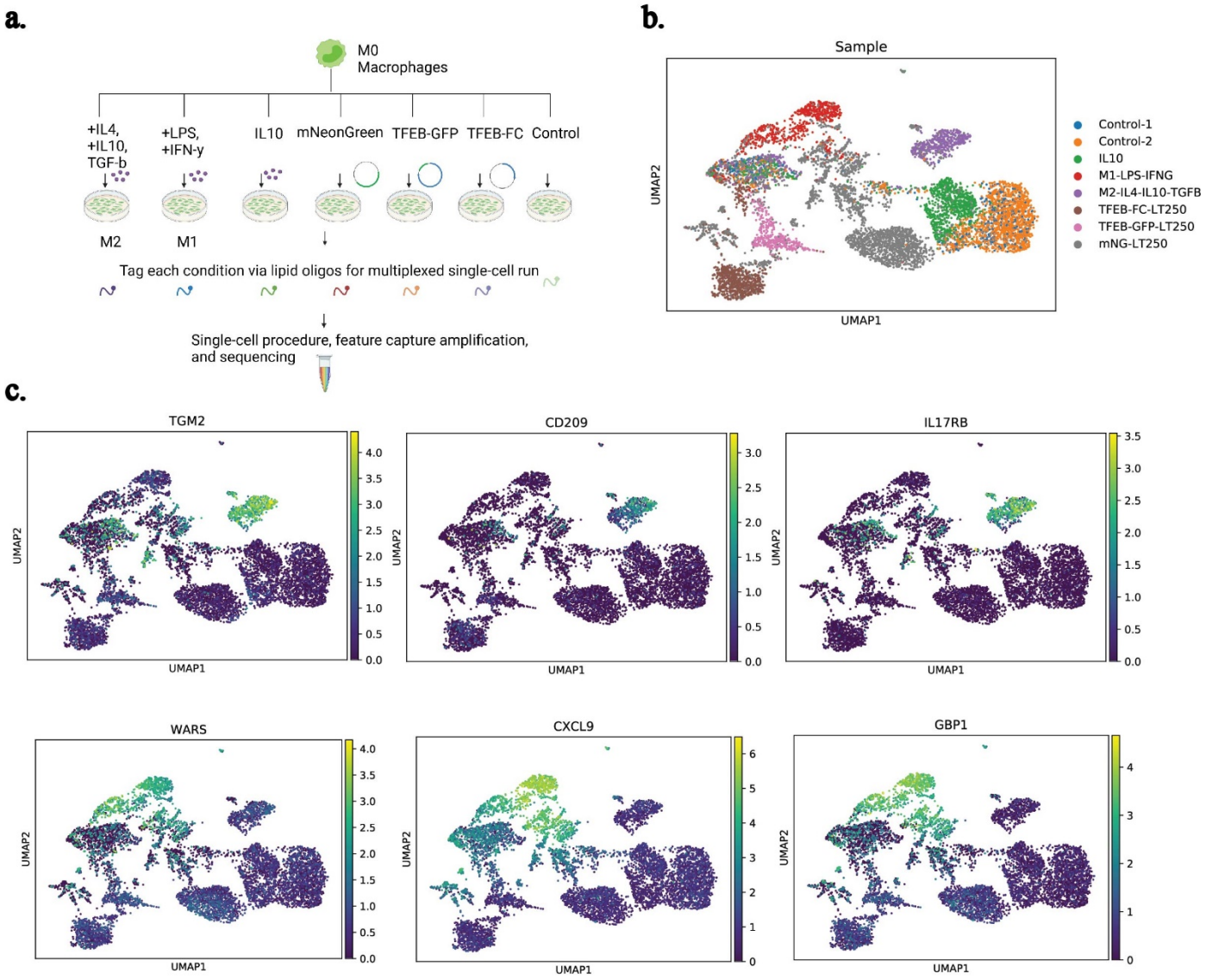


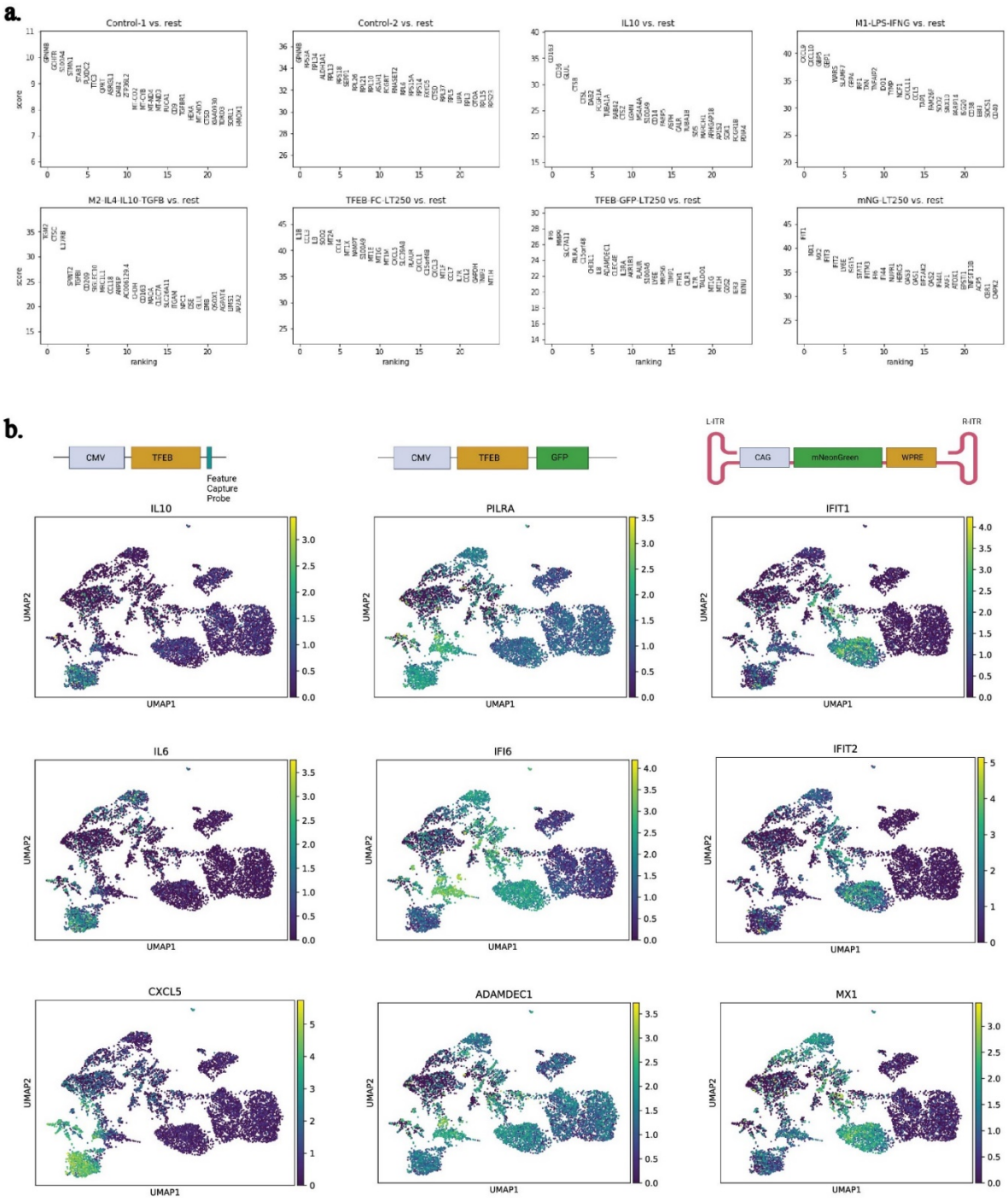
Figure 3. Experimental results of DNA transfection and M1/M2 state induction in cultured human macrophages (A) Experimental design in which M1/M2 induction and plasmid transfection conditions are multiplexed into a single scRNA-seq experiment (B) Macrophage gene expression data projected into UMAP (Scanpy), colors indicate the different conditions and their source of induction (C) Top panel shows expression of markers pertaining to M2 macrophage state, whereas bottom panel shows expression of markers pertaining to M1 macrophage state.

performed a small pilot experiment as shown in Figure 3.a. One of the reasons DNA transfection is appealing is that the transfected plasmids can stick around, whereas RNA libraries can degrade, and later captured via feature capture probe on beads containing the complementary sequence (Supplementary Figure 5). We induced classical M2 state using IL4, IL10, and TGF-β (20 ng/mL) and M1 state using LPS (50 ng/mL), IFN-γ (20 ng/mL). An IL-10 (20 ng/mL) condition was included given prior literature suggesting that IL-10 can inhibit starvation induced autophagy in macrophages (Hun-Jung Park, et al. 2011).

Three plasmids were transfected in macrophages using PromoFectin transfection reagent: TFEB with a feature capture probe (cloned between TFEB stop codon and SV40 poly-A signal), TFEB fused with GFP (TFEB-GFP) (pEGFP-N1-TFEB, addgene #38119), and mNeonGreen (pAAV-CAG-mNeongreen, addgene #99134). Plasmid structures for these are shown in Figure 4.b. TFEB-GFP was used to visualize transfection and presence of TFEB in macrophages, while mNeonGreen was used as a control for transfection. Two control wells of macrophages with no stimuli were also processed to serve as a baseline for M0 macrophage state. Upon transfecting the plasmids, no fluorescence was observed in either TFEB-GFP or mNeonGreen in macrophages. However, we still processed these conditions for scRNA-seq. All conditions were multiplexed into a single 10X chromium lane using lipid-tagged indices from MULTI-seq protocol (McGinnis, et al. 2019). In order to attempt capturing the TFEB containing the feature capture probe, we created a primer to amplify the user introduced TFEB. Capture of transfected TFEB is important for determining which TFEB libraries got expressed since macrophages produce TFEB transcripts natively. scRNA-seq data was aligned as outlined in the Methods section. Samples were demultiplexed using MUTLI-seq indices and Scanpy was used to normalize, analyze, and visualize the data. The control samples demonstrate an overlap of cells as expected, suggesting a lack of variability in states between unperturbed macrophages (Figure 3.b). We first verified that we observed expected M1 and M2 markers in the respective conditions. TGM2, CD209, and IL17RB were all prominently expressed in M2 cluster (Martinez, et al. 2013, Buchacher, et al. 2015, Gratchev, et al. 2021). WARS, CXCL9, and GBP1 had prominent expression in M1 condition cluster (Brown, Wallet, et al. 2010; Orecchioni, et al. 2019; Qiu, et al. 2018). This confirmed delineation between the

conditions is useful in serving as a ruler on which we can project the cell states of the mutants proposed in Section 4.3 (Figure 3). Next, we used Scanpy's function (rank gene groups via Wilcoxon Rank-Sum) to determine genes most prominent in each condition of Figure 4.a. Macrophages are notoriously challenging to transfect with DNA and we did not observe any transfection via fluorescence, however we were curious to see if there were any interesting responses induced by the different plasmids shown in Figure 4.b. mNeonGreen is seen producing prominent antiviral proteins; MX1 as pictured in Figure 4.b and MX2 as seen in Figure 4.a (Braun, et al. 2015), and interferon-induced genes IFIT1, IFIT2 whose expression is enhanced following viral infection (Pidugu, et al. 2019). These genes are also observed in the M1 cluster, potentially hinting at the viral element present on the mNeonGreen as being responsible for inducing a pro-inflammatory response. In particular, mNeonGreen plasmid contains a CAG promoter which is composed of cytomegalovirus (CMV) early enhancer, chicken beta-actin gene, and a splice acceptor of the rabbit beta-globin gene (Miyazaki, et al. 1989, Niwa, et al. 1991). Additionally, it has inverted terminal repeats (ITR) from Adeno-Associated Viruses and woodchuck hepatitis virus posttranscriptional regulatory element (WPRE). Given that no expression was seen from mNeonGreen, it is highly likely that one or a combination of these elements contributed to macrophages starting to move into pro-inflammatory anti-viral state. Macrophages transfected with TFEB-FC display an ambiguous state between pro- and anti-inflammatory. Macrophages in this condition expressed IL6, which can act as both a pro-inflammatory and anti-inflammatory cytokine, and IL10, which is known as an anti-inflammatory cytokine. Additionally, they expressed CXCL5, which is known recruit neutrophils and regulate inflammatory response. Interestingly, TFEB-GFP appears to highly express genes that are also observed in M1 macrophages and cells infected with mNeonGreen. Our hypothesis as to why TFEB plasmid did not induce a similar effect is due to presence of GFP in the TFEB-GFP plasmid. TFEB is natively expressed in macrophages, while GFP is foreign, making it more likely to raise an alarm. Given that the purpose of TFEB is to induce autophagy, we have screened four known autophagy markers, ULK1, ATG7, LAMP1, and DRAM1, and did not see a distinct expression of these genes in TFEB conditions (Bednarczyk, et al. 2018) as seen in Supplementary Figure 6. Overall,

the experiment confirmed challenges of macrophage DNA transfection, yielded interesting data on macrophage response to different plasmids, and provided clean M1, M2



transcriptomic data.

Figure 4. Prevalent genes expressed in macrophages transfected with TFEB and mNeonGreen DNA plasmids (A) Most prominent genes ranked for each condition using Scanpy's rank gene algorithm (B) Transfected vectors and some of the most prominent genes expressed in those conditions visualized on UMAP

4.5 Future Work

Given a lack of success with DNA transfection in macrophages, we have generated the RNA TFEB mutant libraries described in Figure 2. We have confirmed that RNA transfection yields good results (> 80% fluorescence) in Supplementary Figure 7. The next steps are to analyze the impact of rationally designed mutants, and site-saturation mutagenesis libraries, and understand how the method of transcription factor engineering via scRNA-seq can be improved.

4.6 Acknowledgements

David Brown provided a script to de-multiplex MULTI-seq tags. Jong Hwee Park helped with the MULTI-seq portion of the protocol and provided feedback on experimental design. Elowitz lab provided RNA backbone and Citrine RNA for our development and testing.

4.7 Methods

Transcript Alignment

For transcriptome read alignment and gene expression quantification, we used 10x Cell Ranger v5.0.1 with default options to process the FASTQ files from the transcriptome sequencing library. The reads were aligned against the human reference provided by Cell Ranger (hg19-1.2.0).

Macrophage Culturing

Primary human monocytes were purchased from Stem Cell Technologies and cultured in RPMI 1640 supplemented with 10 vol % FBS with human M-CSF (Miltenyi Biotec, #130-

096-491) to induce macrophage state (50 ng/mL). Cell medium was changed every third day and the cells were washed with pre-warmed medium to remove non-adherent cells. Cells were transfected with DNA on the seventh day post-seeding and culturing with M-CSF. DNA was transfected using PromoFectin-Macrophage (PromoCell, # PK-CT-2000-MAC-10).

RNA Production

Site-saturation mutagenesis primer libraries were ordered from Twist Biosciences. Individually selected primer mutants were ordered from IDT. Both libraries and individual mutants were created using overlap PCR and then 1 ug of DNA was fed into Cell Script's INCOGNITO T7 Ψ-RNA In Vitro Transcription kit (#C-ICTY110510). The 5' capping and poly-A extension was done using Cell Script's T7 mScript Standard mRNA Production System kit (#C-MSC11610). All intermediate reactions were purified using MEGAclear transcription clean-up kit (ThermoFisher, #AM1908).

4.8 Supplementary Material

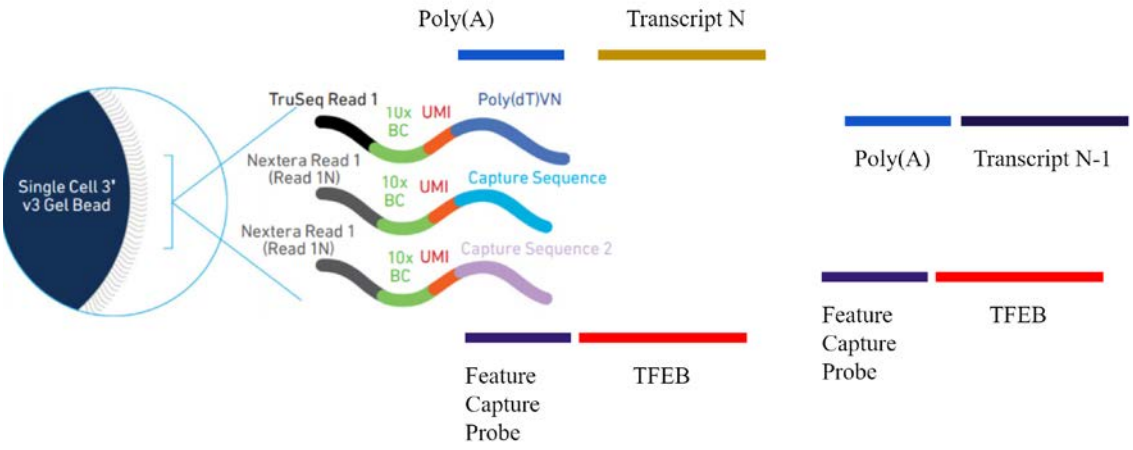


Figure 52. Conceptual drawing demonstrating how TFEB mutants introduced into macrophages would be captured using 10X Genomics' Capture Sequence 2 on the Gel Bead along with the rest of the cell transcriptome

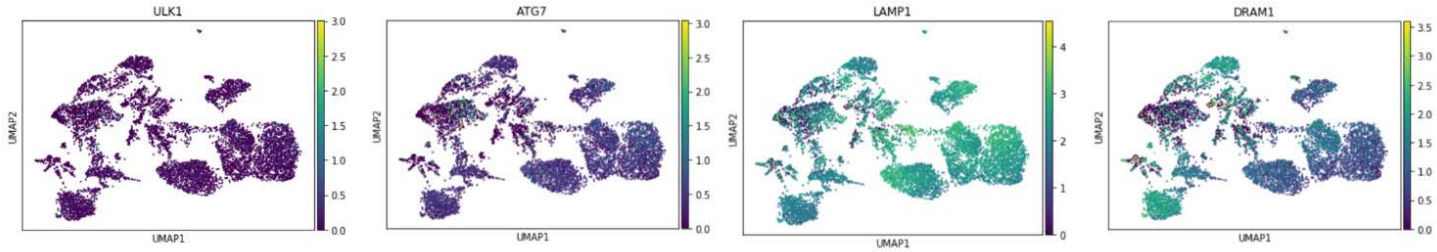


Figure 6. UMAP projection showing expression of known autophagy markers across the various conditions outlined in Figure 3.

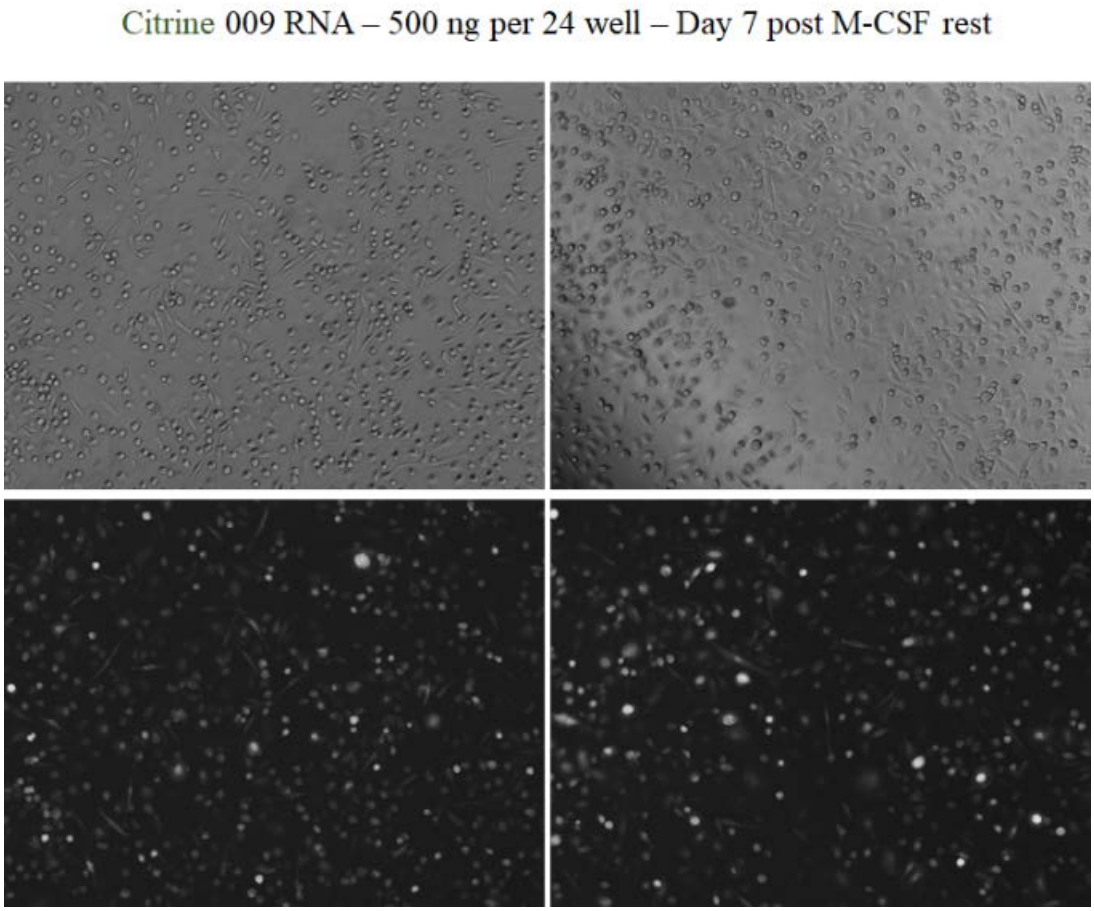


Figure 3. Images of cultured macrophages (day 7 post M-CSF rest) showing transfection with Citrine RNA (bottom panel). Top shows morphology of the macrophages at day 7 and post RNA transfection in bright field view.

4.9 References

Rőszer, T. Understanding the Mysterious M2 Macrophage through Activation Markers and Effector Mechanisms. Mediators of Inflammation 2015, 1–16 (2015).

Ardura, J. A. et al. Targeting Macrophages: Friends or Foes in Disease? Front. Pharmacol. 10, (2019).

Klichinsky, M. et al. Human chimeric antigen receptor macrophages for cancer immunotherapy. Nat Biotechnol 38, 947–953 (2020).

Zhang, F. et al. Genetic programming of macrophages to perform anti-tumor functions using targeted mRNA nanocarriers. Nat Commun 10, (2019).

Puertollano, R., Ferguson, S. M., Brugarolas, J. & Ballabio, A. The complex relationship between TFEB transcription factor phosphorylation and subcellular localization. EMBO J 37, (2018).

Fang, L. et al. Transcriptional factor EB regulates macrophage polarization in the tumor microenvironment. OncoImmunology 6, e1312042 (2017).

Pastore, N. et al. TFEB and TFE3 cooperate in the regulation of the innate immune response in activated macrophages. Autophagy 12, 1240–1258 (2016).

Napolitano, G. & Ballabio, A. TFEB at a glance. Journal of Cell Science (2016) doi:10.1242/jcs.146365.

Hemesath, T. J. et al. microphthalmia, a critical factor in melanocyte development, defines a discrete transcription factor family. Genes & Development 8, 2770–2780 (1994).

Pogenberg, V. et al. Restricted leucine zipper dimerization and specificity of DNA recognition of the melanocyte master regulator MITF. Genes & Development 26, 2647–2658 (2012).

Billings, W. M., Hedelius, B., Millecam, T., Wingate, D. & Corte, D. D. ProSPr: Democratized Implementation of Alphafold Protein Distance Prediction Network. (2019) doi:10.1101/830273.

Park, H.-J. et al. IL-10 inhibits the starvation induced autophagy in macrophages via class I phosphatidylinositol 3-kinase (PI3K) pathway. Molecular Immunology 48, 720–727 (2011).

McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods 16, 619–626 (2019).

Martinez, F. O. et al. Genetic programs expressed in resting and IL-4 alternatively activated mouse and human macrophages: similarities and differences. Blood 121, e57–e69 (2013).

Buchacher, T., Ohradanova-Repic, A., Stockinger, H., Fischer, M. B. & Weber, V. M2 Polarization of Human Macrophages Favors Survival of the Intracellular Pathogen Chlamydia pneumoniae. PLoS ONE 10, e0143593 (2015).

Gratchev, A. et al. Activation of a TGF-β-Specific Multistep Gene Expression Program in Mature Macrophages Requires Glucocorticoid-Mediated Surface Expression of TGF-β Receptor II. J Immunol 180, 6553–6565 (2008).

Brown, J. N., Wallet, M. A., Krastins, B., Sarracino, D. & Goodenow, M. M. Proteome bioprofiles distinguish between M1 priming and activation states in human macrophages. Journal of Leukocyte Biology 87, 655–662 (2009).

Orecchioni, M., Ghosheh, Y., Pramod, A. B. & Ley, K. Macrophage Polarization: Different Gene Signatures in M1(LPS+) vs. Classically and M2(LPS–) vs. Alternatively Activated Macrophages. Front. Immunol. 10, (2019).

Qiu, X. et al. Down-regulation of guanylate binding protein 1 causes mitochondrial dysfunction and cellular senescence in macrophages. Sci Rep 8, (2018).

Braun, B. A., Marcovitz, A., Camp, J. G., Jia, R. & Bejerano, G. Mx1 and Mx2 key antiviral proteins are surprisingly lost in toothed whales. Proc Natl Acad Sci USA 112, 8036–8040 (2015).

Pidugu, V. K., Pidugu, H. B., Wu, M.-M., Liu, C.-J. & Lee, T.-C. Emerging Functions of Human IFIT Proteins in Cancer. Front. Mol. Biosci. 6, (2019).

Jun-ichi, M. et al. Expression vector system based on the chicken β-actin promoter directs efficient production of interleukin-5. Gene 79, 269–277 (1989).

Hitoshi, N., Ken-ichi, Y. & Jun-ichi, M. Efficient selection for high-expression transfectants with a novel eukaryotic vector. Gene 108, 193–199 (1991).

Bednarczyk, M., Zmarzły, N., Grabarek, B., Mazurek, U. & Muc-Wierzgoń, M. Genes involved in the regulation of different types of autophagy and their participation in cancer pathogenesis.