








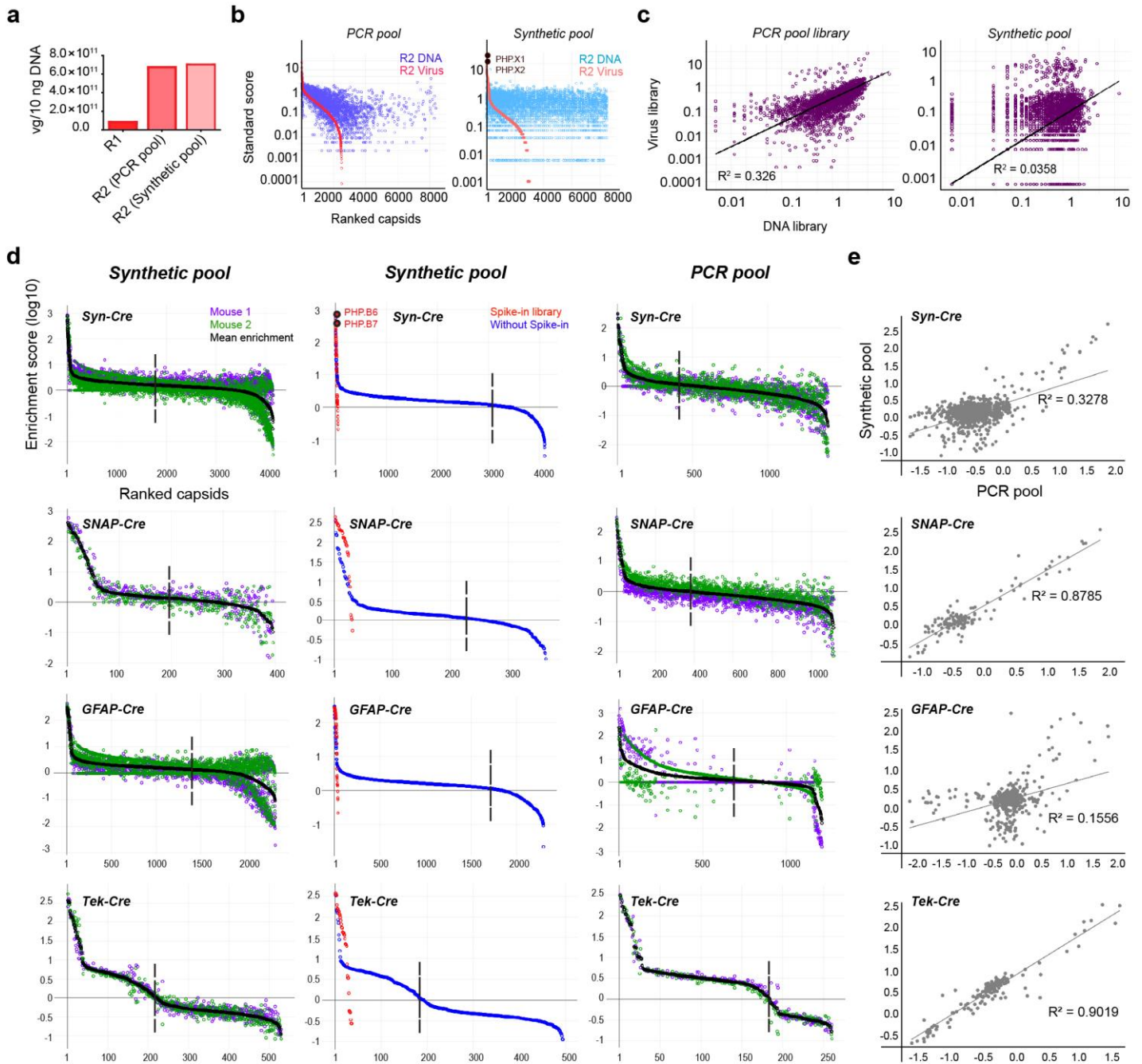
In the format provided by the authors and unedited.

Multiplexed Cre-dependent selection yields systemic AAVs for targeting distinct brain cell types

Sripriya Ravindra Kumar ¹, Timothy F. Miles ^{1,5}, Xinhong Chen ^{1,5}, David Brown¹, Tatyana Dobreva¹, Qin Huang^{1,2}, Xiaozhe Ding ¹, Yicheng Luo ¹, Pétur H. Einarsson¹, Alon Greenbaum^{1,3,4}, Min J. Jang¹, Benjamin E. Deverman^{1,2} and Viviana Gradinaru ¹ 

¹Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. ²Present address: Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA. ³Present address: Joint Department of Biomedical Engineering, North Carolina State University, Raleigh, NC, USA. ⁴Present address: University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. ⁵These authors contributed equally: Timothy F. Miles, Xinhong Chen. ✉e-mail: viviana@caltech.edu

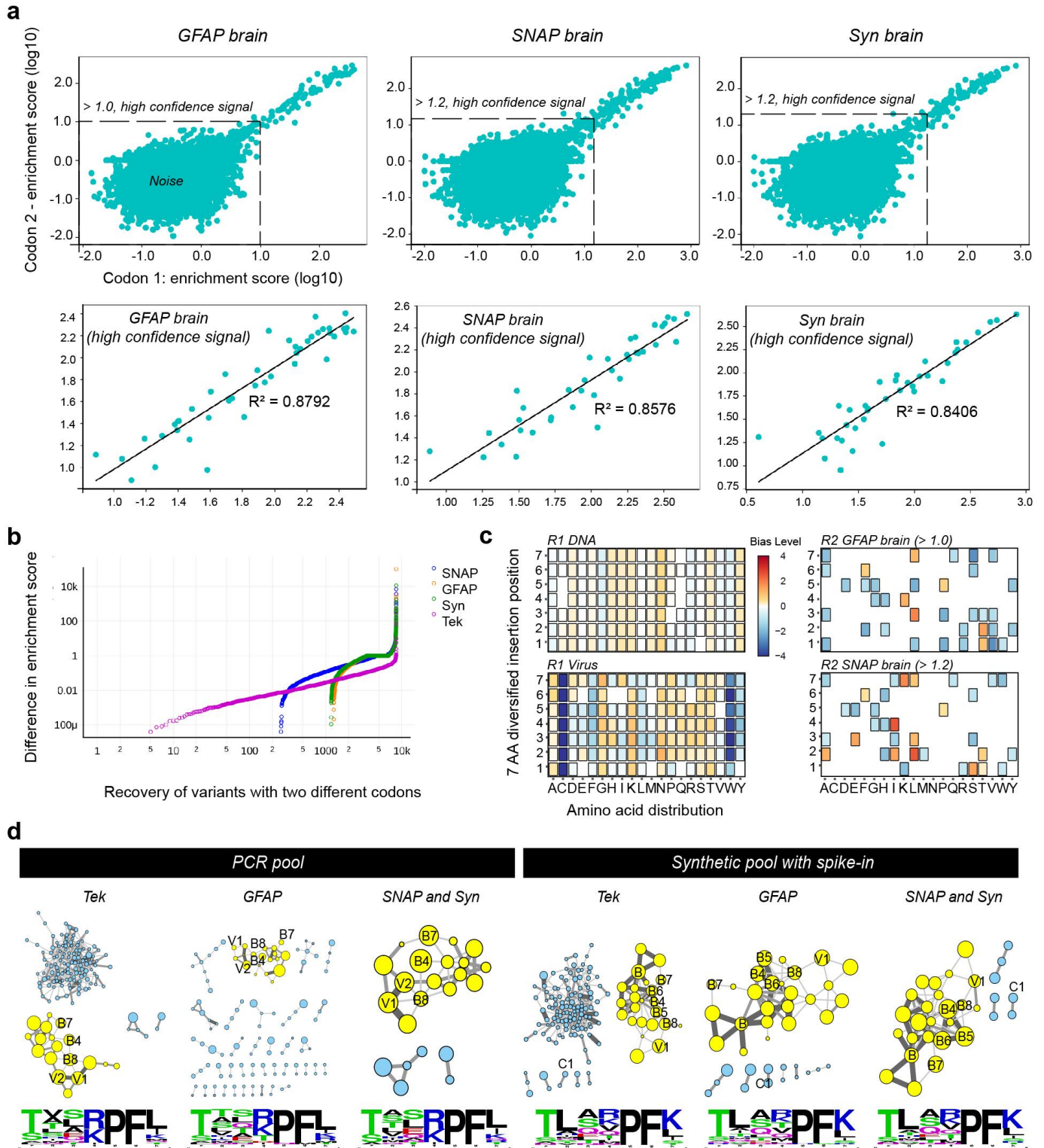
480 bp in *cap* gene in addition to the stop codons that are designed to prevent synthesis of VP1, VP2, and VP3 proteins. AAP protein translation is unaffected by these modifications. **b**, Schematic of the protocol to selectively recover rAAV genomes from the target population using the Cre-Lox flipping strategy and preparation of the sample for deep sequencing. **c**, The library coverage for R1 DNA and virus libraries obtained from specific sequencing depths. **d**, The percentage of variant overlap within the sampled DNA and virus, or across different Cre lines within tissues, or across tissues from R1 selection. **e**, The distributions of AAV capsid read counts for libraries recovered by NGS from brain tissue across different Cre transgenic mice post R1 selection. The dotted line is illustrative only and roughly separates the signal from noise (see Methods for estimation of signal v.s. noise) where signal in this context represents the input for the R2 selection. **f**, rAAV genome recovery from tissues using different treatments are shown with total rAAV genome recovery from 0.1 g of liver, **g**, Percentage of rAAV genomes recovered per ng of total extracted DNA, and **h**, The CT value (cycle threshold from qPCR) of rAAV genome extracted by trizol that were treated with SmaI restriction enzyme or untreated and **i**, CT value of mitochondrial DNA (internal control for smaller genome recovery, fold change = $10.79 (2^{\Delta CT})$) recovered from 1 ng of total DNA from liver tissue. In **f-i**, $n = 4$ mice; 2 from GFAP-Cre line and 2 from Tek-Cre line, each data point is drawn from the mean of three technical replicates, error bar is mean \pm S.E.M., Mann-Whitney test, two-tailed (exact P-value of 0.0286 (* $P \leq 0.05$), in **f**, **g**, **i**, and 0.1143 (n.s., $P > 0.05$, CI 95%) in **h**). The data reported **f**, **g**, **i** are from one independent trial, and **h**, from three independent trials.



Supplementary Figure 2

Analysis of 7-mer-i rAAV Libraries From Round-2 Selections.

a, The vector yields obtained per 10 ng of capsid DNA library across R1 and R2 vector productions. **b**, Distributions of the DNA and virus libraries produced by the *synthetic pool* and *PCR pool* methods by the standard score of NGS read counts. The variants in virus libraries are sorted by the decreasing order of standard score and their scores from respective DNA libraries are mapped onto them. **c**, Correlations between the standard scores of read counts for the DNA and virus libraries ($n = 1$ per library) produced by the *synthetic pool* and *PCR pool* methods is determined by linear least-squares regression, and the regression line (best fit) and R^2 representing the coefficient of determination is shown. **d**, Distributions of capsid libraries from brain tissue of two mice (purple and green) used in each Cre line selection, as produced by the *synthetic pool* (left) and *PCR pool* (right) designs. The distribution of *spike-in* library introduced in the *synthetic pool* library design is shown in red (center). **e**, Correlations of enrichment scores of variants from the brain libraries ($n = 2$ per Cre line, mean is plotted) produced by *synthetic pool* and *PCR pool* methods is determined by the same method described in **c**.

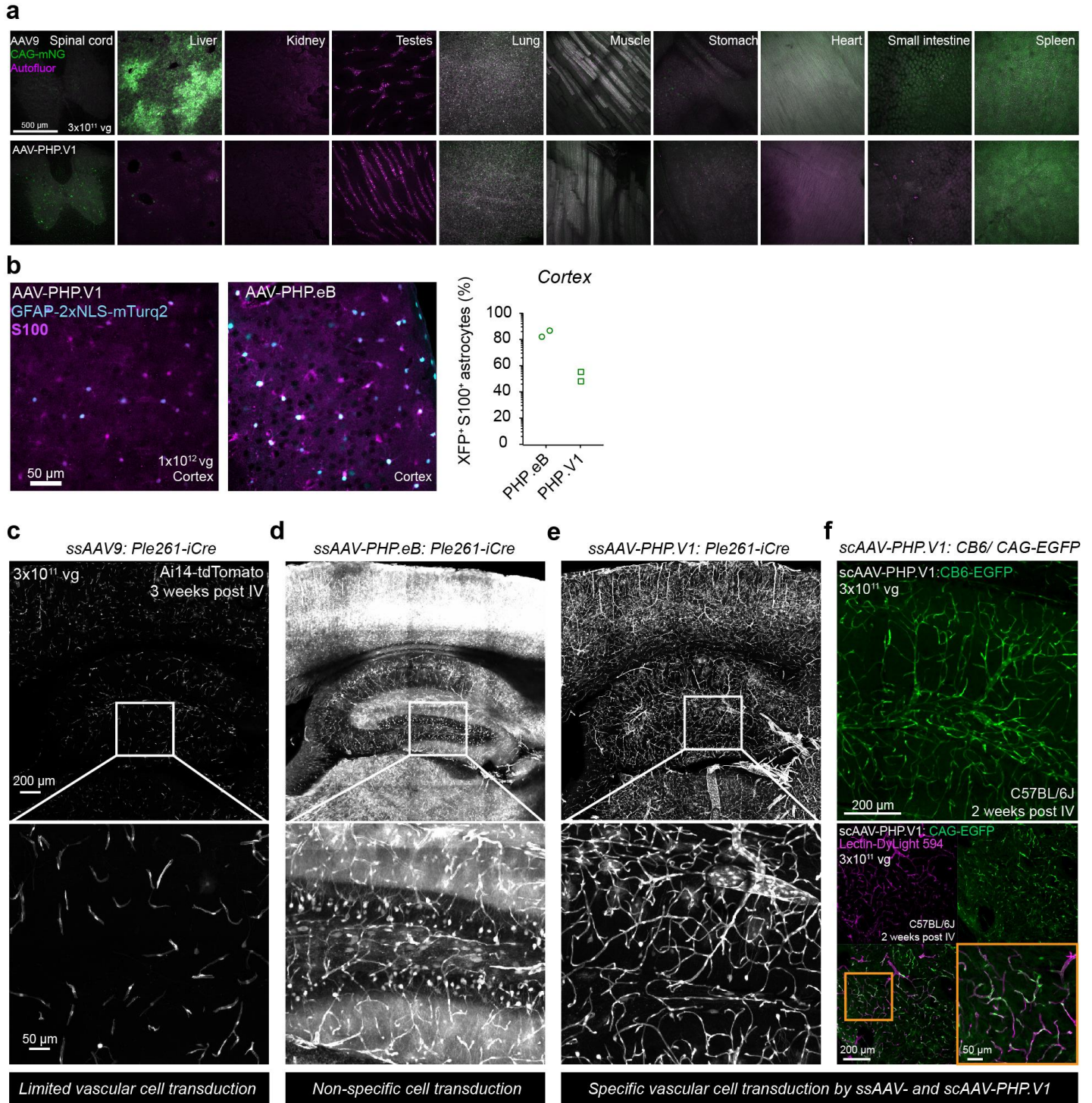


Supplementary Figure 3

Analysis of Round-2 7-mer-i Tissue Libraries From *Synthetic Pool* And *PCR Pool* Methods.

a, Correlation analysis between the enrichment score (log₁₀) of two alternate codon replicates of variants from the GFAP-Cre (left), SNAP-Cre (center), and Syn-Cre (right) brain libraries by linear least-squares regression ($n = 2$ per Cre line, mean is plotted). The dotted line separates the high-confidence signal from noise. High confidence signal (below) is assessed by a linear regression line (best

fit) and R^2 represents the coefficient of determination. **b**, The difference in enrichment score between the two codon replicates of a variant, across different brain libraries, with over 8000 variants recovered in replicates. **c**, Heatmaps represent the magnitude (\log_2 fold change) of AA bias in "output" library 1 normalized to "input" library 2 that reach statistical significance (boxed if $P\text{-value} \leq 0.0001$, two-sided, two-proportion z-test, except in R1 DNA normalized to known NNK template where one-proportion z-test was performed, and P-values corrected for multiple comparisons using Bonferroni correction) is shown. R1 DNA library normalized to NNK template (top left, ~9 million sequences), R1 virus normalized to R1 DNA libraries (bottom left, ~10 million sequences), R2 GFAP library with enrichment score above 1.0 in brain normalized to R2 virus (top right, 20 sequences,) and R2 SNAP library with enrichment score above 1.2 normalized to R2 virus (bottom right, 17 sequences) are shown ($n = 1$ for DNA, virus, and $n = 2$ for brain libraries). **d**, Clustering analysis of positively enriched variants from Tek, GFAP, and combined neuron brain libraries (SNAP and Syn) by *PCR pool* design, and by *synthetic pool* design with *spike-in* library are shown with size of nodes representing their relative enrichment in brain, and the thickness of edges (connecting lines) representing the extent of shared AA identity between nodes. A distinct family is highlighted in yellow with the corresponding AA frequency logo below (AA size reflects prevalence and color coded based on AA properties).

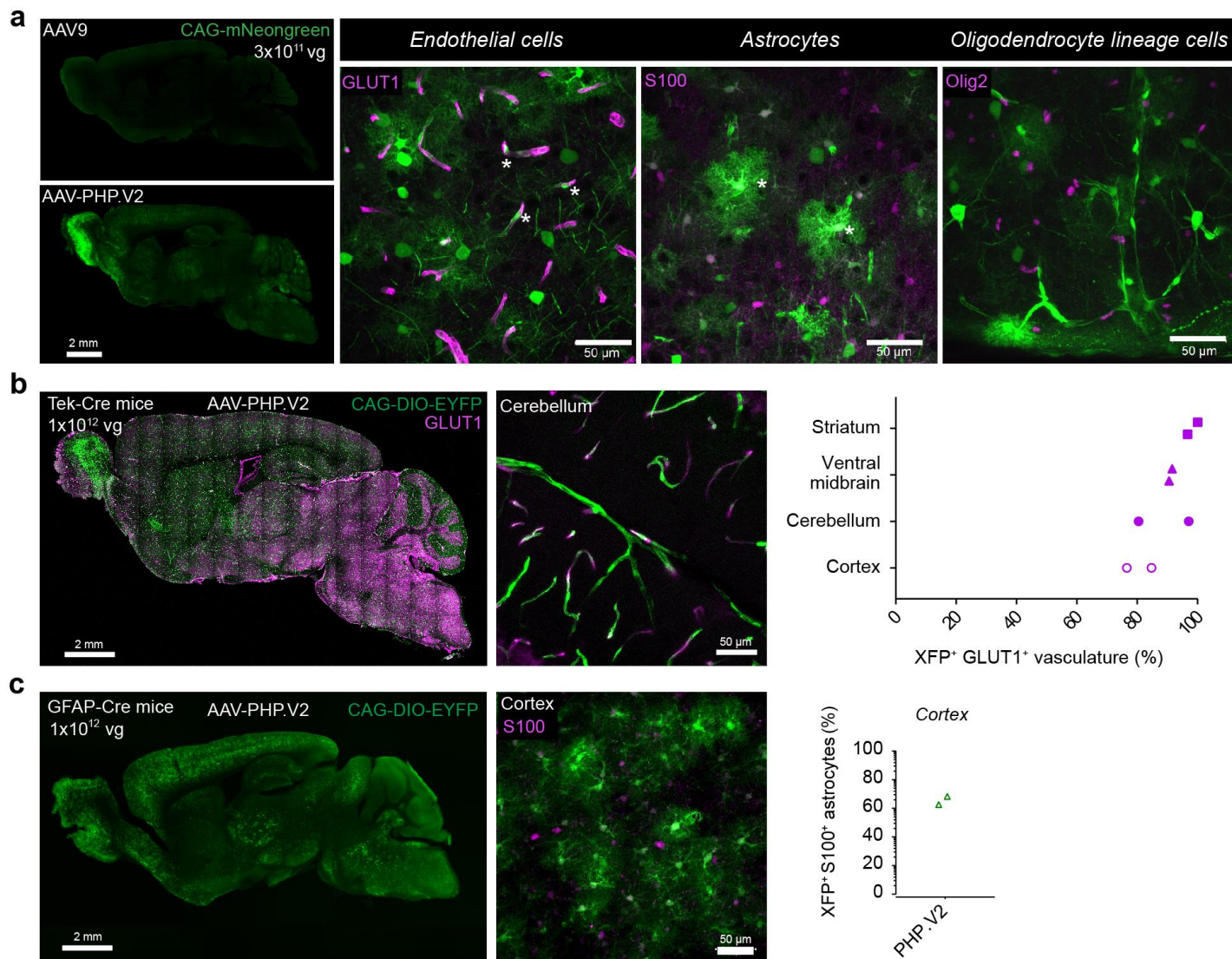


Supplementary Figure 4

AAV-PHP.V1 Efficiently Targets the Brain Vasculature.

a, Expression of AAV9 (above) and AAV-PHP.V1 (below) packaging ssAAV:CAG-mNeonGreen across all organs is shown ($n=3$, 3×10^{11} vg dose per adult C57BL/6J mouse, 3 weeks of expression). The background auto fluorescence is represented in magenta. **b**, Expression in cortical astrocytes (S100⁺) after IV delivery of AAV-PHP.V1 (left) and AAV-PHP.eB (right) capsids carrying ssAAV:GfABC1D-2xNLS-mTurquoise2 (1×10^{12} vg dose per adult mouse, 4 weeks of expression). Percentage of cortical S100⁺ cells that overlapped with mTurquoise2 expression is quantified ($n = 2$, each data point is mean from 3 images per mouse). **c**, Expression of

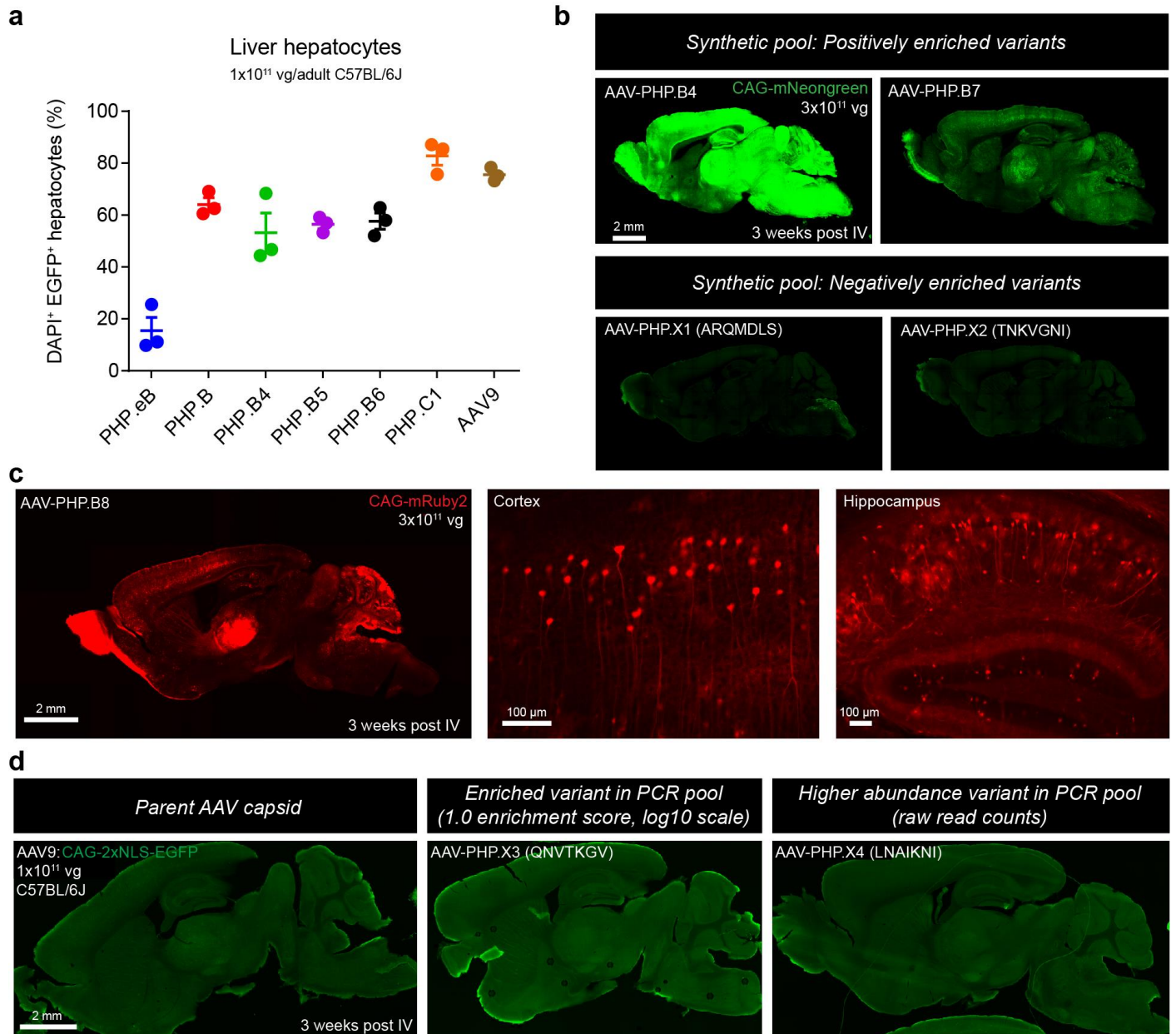
AAV9, AAV-PHP.eB **(d)**, AAV-PHP.V1 **(e)** packaging ssAAV:Plc261-iCre in Ai14-tdTomato reporter adult mouse (n=2-3 per group, 3×10^{11} vg dose per adult mouse, 3 weeks of expression). **f**, Expression of AAV-PHP.V1 carrying self-complementary (sc) scAAV:CB6-EGFP (above) and scAAV:CAG-EGFP (below). Magenta represents the lectin DyLight 594 staining (n=2-3, 3×10^{11} vg dose per adult C57BL/6J mouse, 2 weeks of expression). Experiments in **c-e** are reported from one independent trial from a fresh batch of viruses, and titered in the same assay for dosage consistency, **e** and **f** validated in two independent trials (n = 2 per group).



Supplementary Figure 5

AAV-PHP.V2 Variant Exhibits Biased Transduction Towards Brain Vascular Cells.

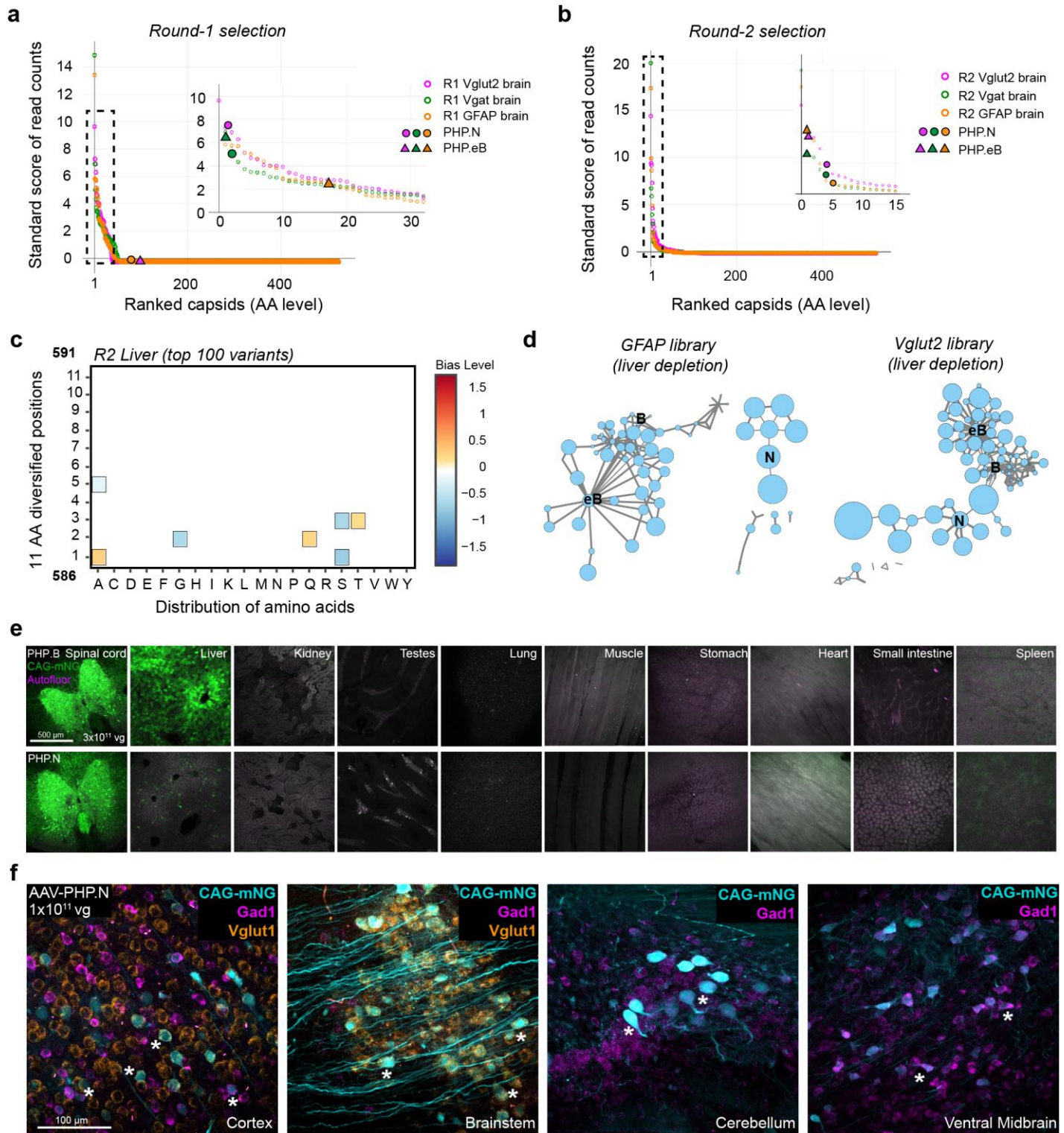
a, Transduction of mouse brain by the AAV-PHP.V2 variant and control AAV9, carrying the ssAAV:CAG-mNeonGreen ($n = 3$, 3×10^{11} vg IV dose per C57BL/6J adult mouse, 3 weeks of expression) is shown. The sagittal brain images (left) are matched in fluorescence intensity to the sagittal brain images in Fig. 3c. Higher magnification images of AAV-PHP.V2 transduced brain sections stained with α GLUT1 or α S100 or α Olig2 (magenta) are shown. **b**, Transduction of brain vasculature by AAV-PHP.V2 carrying ssAAV:CAG-DIO-EYFP (green) in Tek-Cre adult mice (left, 1×10^{12} vg IV dose per mouse, 4 weeks of expression) is shown, and its efficiency (right) is determined by the overlap of α GLUT1 staining (red) with EYFP expression across different brain areas ($n = 2$, mean of 3 images per brain region per mouse). **c**, Transduction of astrocytes by AAV-PHP.V2 in GFAP-Cre adult mouse (1×10^{12} vg IV dose per mouse, 4 weeks of expression) is shown. Percentage of cortical S100+ cells that overlapped with EYFP expression is quantified ($n = 2$, mean of 3 images per mouse).



Supplementary Figure 6

Further Validation of Synthetic Pool and PCR Pool Variants Demonstrates Higher Confidence in Synthetic Pool NGS Data.

a, Transduction levels of liver hepatocytes quantified as the percentage of DAPI⁺ cells that are EGFP⁺ ($n = 3$, vectors packaged with ssAAV:CAG-2xNLS-EGFP, 1x10¹¹ vg IV dose/adult C57BL/6J mouse, 3 weeks of expression, mean±S.E.M, 4 images per mouse per group. One-way ANOVA, non-parametric Kruskal-Wallis test gave an approximate P-value of 0.0088). **b**, Transduction of brain tissue by AAV-PHP.B4, B7, AAV-PHP.X1 (ARQMDLS), and AAV-PHP.X2 (TNKVGNI) packaging ssAAV:CAG-mNeonGreen genome ($n = 3$, 1x10¹¹ vg IV dose/adult C57BL/6J mouse, 3 weeks of expression), matched in fluorescence intensity to AAV9 and AAV-PHP.V1 sagittal brain images in Fig. 3c. **c**, Transduction of the brain by AAV-PHP.B8 using the ssAAV:CAG-mRuby2 genome ($n = 3$, 3x10¹¹ vg IV dose/adult C57BL/6J mouse, 3 weeks of expression). **d**, Transduction of AAV9 (left), AAV-PHP.X3 (QNVTKGV) (middle) and AAV-PHP.X4 (LNAIKNI) (right) vectors packaging ssAAV:CAG-2xNLS-EGFP ($n = 2$, 1x10¹¹ vg IV dose/adult C57BL/6J mouse, 3 weeks of expression). **a-d** data is reported from one independent trial.

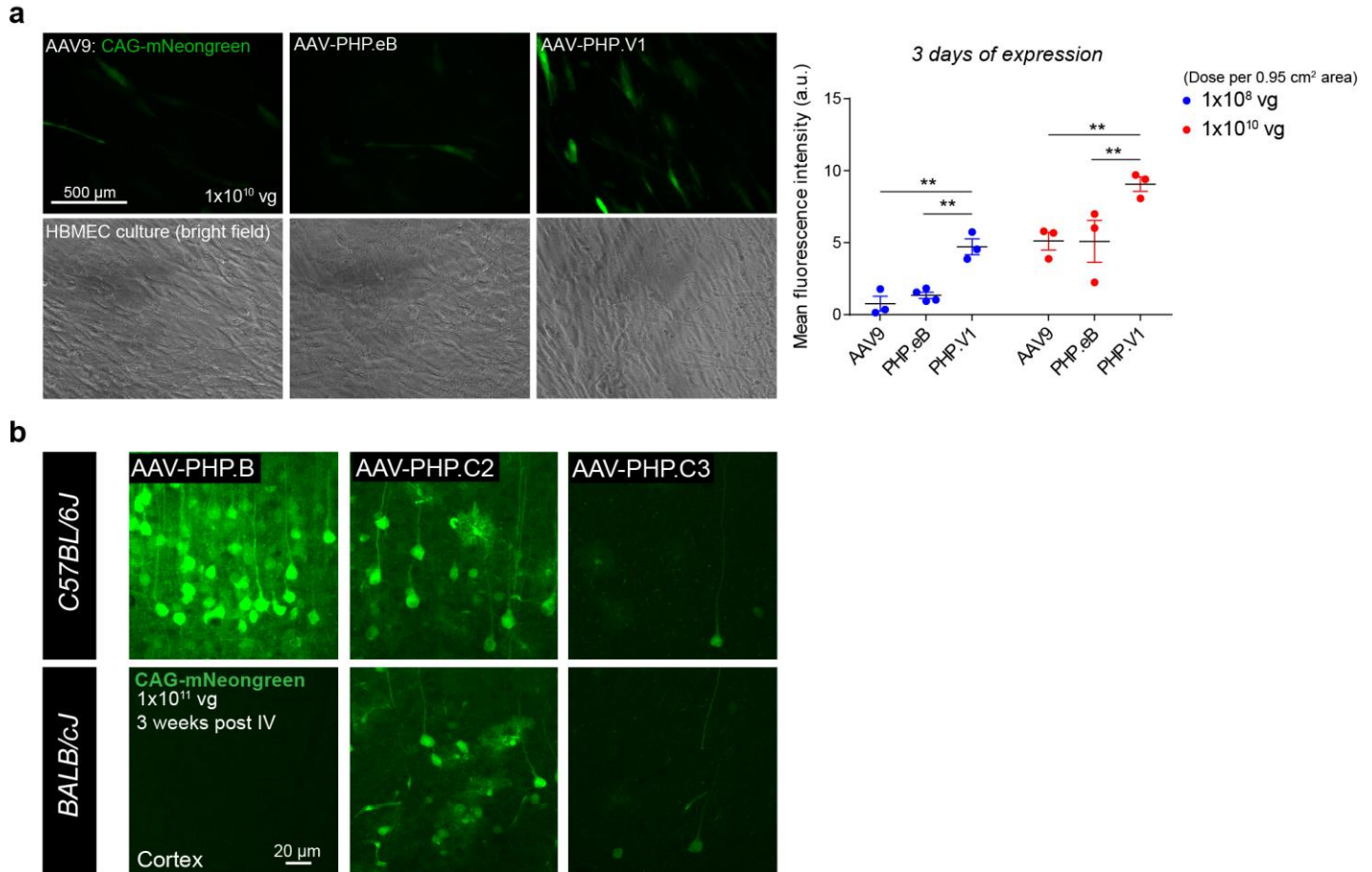


Supplementary Figure 7

Evolution of the AAV-PHP.B Capsid by Diversifying Amino Acid Positions 587-597.

a, Distributions of R1 and **b**, R2 brain libraries (at AA level, standard score (SS) of RCs sorted in decreasing order of scores) is shown. The SS for AAV-PHP.N and AAV-PHP.eB across libraries are mapped on the zoomed-in view of this plot (dotted line box). **c**, Heatmap of AA distributions across the diversified region of the enriched variants from R2 liver library (top 100 sequences) normalized to the R2

virus (input library). **d**, Clustering analysis of enriched variants from GFAP and Vglut2 brain libraries are shown with size of nodes representing their relative depletion in liver, and the thickness of edges (connecting lines) representing their relative identity between nodes. **e**, Expression of AAV-PHP.B (above) and AAV-PHP.N (below) packaged with ssAAV:CAG-mNeonGreen across all organs is shown ($n = 3$, 3×10^{11} vg IV dose per adult C57BL/6J mouse, 3 weeks of expression). The background auto fluorescence is in magenta. **f**, Transduction of mouse brain by the AAV-PHP.N variant, carrying the CAG promoter that drives the expression of mNeonGreen ($n = 3$, 1×10^{11} vg IV dose per C57BL/6J adult mouse, 3 weeks of expression) is shown. Fluorescence *in situ* hybridization chain reaction (FITC-HCR) was used to label excitatory neurons with Vglut1 and inhibitory neurons with Gad1. Few cells where EGFP expression co-localized with specific cell markers are highlighted by asterisks symbol.



Supplementary Figure 8

Investigation of AAV-PHP Variants Across Different Mouse Strains and In Vitro Human Brain Microvascular Endothelial Cells.

a, Transduction of AAV9, AAV-PHP.eB and AAV-PHP.V1 in human brain microvascular endothelial cell culture (HBMEC) is shown. The vectors were packaged with ssAAV:CAG-mNeongreen. The mean fluorescence intensity across the groups were quantified ($n=3$ tissue culture wells of 0.95 cm² surface area per group, 3 images per well per group per dose was imaged after three days of expression, doses 1x10⁸ vg and 1x10¹⁰ vg per 0.95 cm² surface area). A two-way ANOVA with correction for multiple comparisons using Tukey's test gave adjusted P-value of 0.0051 for AAV9 vs PHP.V1, 0.0096 for PHP.eB vs PHP.V1, 0.8222 for AAV9 vs PHP.eB for 1x10⁸ vg, and 0.0052 for AAV9 vs PHP.V1, 0.0049 for PHP.eB vs PHP.V1, 0.9996 for AAV9 vs PHP.eB for 1x10¹⁰ vg (** $P \leq 0.01$, is shown and $P > 0.05$ is not shown on the plot; mean \pm S.E.M., 95% CI). **b**, The transduction of cortex brain region by AAV-PHP.B, AAV-PHP.C2 and AAV-PHP.C3 across two different mouse strains: C57BL/6J and BALB/cJ are shown. The vectors were packaged with ssAAV:CAG-mNeongreen ($n = 2-3$ per group, 1x10¹¹ vg IV dose/ adult mouse, 3 weeks of expression). The fluorescence intensity is matched across all the images. The data reported in **a,b** are from one independent trial where all viruses were freshly prepared and titered in the same assay for dosage consistency, with additional validation for AAV-PHP.C2 and AAV-PHP.C3 in an independent trial for BALB/cJ.

SUPPLEMENTARY TABLES 1-4:

Design Parameters	<i>Synthetic pool</i> design	<i>PCR pool</i> design
Carryover of R1 selection bias among variants	No, likelihood of false positives is low	Yes, potential to minimize by normalization
Carryover of R1 selection induced mutants	No	Yes
Confidence in library performance	High, using alternate codon replicates	Low
Customize library or add internal controls	Yes, in an unbiased manner	Yes, with greater risk of bias
Control library size	Yes, without reducing libraries or pooling	Yes, with libraries reduced for pooling
Cost for R2 library generation	High	Low

Supplementary Table 1: Comparison between the two methods for R2 selection.

The table summarizes the pros and cons of selection design parameters by the *synthetic pool* and *PCR pool* R2 selection methods.

AAV Variants	<i>Synthetic pool</i> enrichment rank	<i>PCR pool</i> enrichment rank	<i>PCR pool</i> read count rank
PHP.V1	1	4	3
PHP.V2	2	1	1
PHP.B4	4	10	56
PHP.B7	6	13	36
PHP.B8	3	7	23
PHP.C1	13	34	74
PHP.C2	12	20	293
PHP.C3	16	Not recovered	Not recovered

Supplementary Table 2: Ranking of AAV-PHP capsids across methods.

Ranks of selected variants among all capsids recovered from R2 Tek-Cre selection by *synthetic pool* enrichment score (representing M-CREATE), *PCR pool* enrichment score (representing closer to M-CREATE), or *PCR pool* read counts (representing CREATE), the highest ranks of which starts from 1, and “Not recovered” represent absence of the variant from R2 sequencing data.

AAV Variants	Reference / Selection method	Tropism	Production	Rounds of evolution from parent capsid
PHP.B, B2, B3	Deverman et al, 2016 / CREATE	Broad CNS transduction	Good	1 round from AAV9
PHP.A	Deverman et al, 2016 / CREATE	Astrocyte transduction	Poor; prone to precipitate upon storage at 4°C.	1 round from AAV9
PHP.eB	Chan et al, 2017 / CREATE	Enhanced Broad CNS transduction	Good	2 rounds from AAV9 or 1 round from PHP.B
PHP.S	Chan et al, 2017 / CREATE	Sensory neuron transduction	Good	1 round from AAV9
PHP.V1, V2	Current study / M-CREATE	BBB Vascular cells and astrocytes transduction	Good	1 round from AAV9
PHP.B4, B7, B8,	Current study / M-CREATE	Broad CNS transduction	Good	1 round from AAV9
PHP.B5, B6	Current study / M-CREATE and CREATE	Broad CNS transduction	Good	2 rounds from AAV9 or 1 round from PHP.B
PHP.C1, C2, C3	Current study / M-CREATE	Broad CNS transduction across mouse strains	Good; PHP.C1 prone to precipitate upon storage at 4°C.	1 round from AAV9
PHP.N	Current study / M-CREATE and CREATE	Neuron transduction	Average	2 rounds from AAV9 or 1 round from PHP.B

**PHP variants – named in memory of late Professor Paul H. Patterson, Caltech.*

Supplementary Table 3: AAV-PHP vectors identified by CREATE and M-CREATE.

The table provides a summary of the variants that have been identified so far using CREATE and M-CREATE, along with their tropism and the evolutionary steps from the parent capsid that was involved in their discovery.

Primer name	Sequence (5'-3')	Forward or reverse direction
XF	ACTCATCGACCAATACTTGACTATCTCTCTAGAAC	Forward
7xMNN-588i	GTATTCCTTGGTTTTGAACCAACCGGTCTGCGCCTGTGCMNNMNNMNNMNNMNNMNNM NNTTGGGCACTCTGGTGGTTTGTG	Reverse
588-R2lib-F	CACTCATCGACCAATACTTGACTATCTCTCT	Forward
588-R2lib-R	GTATTCCTTGGTTTTGAACCAACCG	Reverse
11-mer-588i	GTATTCCTTGGTTTTGAACCAACCGGTCTGCGCXXXXXXXXMNNMNNMNNMNNMNNMNNM NXXXXXXACTCTGGTGGTTTGTG	Reverse
71F	CTTCCAGTTCAGCTACGAGTTTGAGAAC	Forward
CDF/R	CAAGTAAACCTCTACAAATGTGGTAAATCG	Forward/Reverse, see Methods
588i-lib-PCR1-6bpUID-F	CACGACGCTCTCCGATCTAANNNNNNAGTCCTATGGACAAGTGGCCACA	Forward
588i-lib-PCR1-R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCCTTGGTTTTGAACCAACCG	Reverse
1527	ACACTCTTCCCTACACGACGCTCTTCCGATCTGACAAGTGGCCACAAACCACCG	Forward
1532	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTTGGTTTTGAACCAACCG	Reverse
mNeonGreen-F	CGACACATGAGTTACACATCTTTGGCTC	Forward
mNeonGreen-R	GGAGGTCACCCTTGGTGGACTTC	Reverse
Mito-F	CCCAGCTACTACCATCATTCAAGT	Forward
Mito-R	GATGTTTTGGGAGATTGGTTGATGT	Reverse
CapF-56	ATTGGCACCAGATACCTG ACTCGTAA	Forward
Cre-R-57	GTCCAAACTCATCAATGTATCTTATCATGTCTG	Reverse
NGS-QC-F	AATGATACGGCGACCACCGAG	Forward
NGS-QC-R	CAAGCAGAAGACGGCATA CGA	Reverse

Supplementary Table 4: Primers used in M-CREATE selection.

The table provides a list of primers used in M-CREATE across the different steps of the selection process as described in the Methods.

Variant Name	Amino acid motif of variants with "AQ" overhangs from AAV9: AA587-88 and AA589-90. (11-mer)	Nucleotide sequence of 11-mer	Forward primer (5'-3')	Reverse primer (5'-3')
AAV-PHP.V1 (Addgene ID: 127847)	AQTALKPFLAQ	GCCCAAACCGCCCT CAAACCTTCCTCG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACCGCCCTCAA ACCC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGA GGAAGGGTTTGAGGGCGGT TTGGGC
AAV-PHP.V2 (Addgene ID: 127848)	AQTTLKPFLAQ	GCCCAAACCACCCT CAAACCTTCCTCG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACCACCCTCAA CCC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGA GGAAGGGTTTGAGGGTGGT TTGGGC
AAV-PHP.B4 (Addgene ID: 127849)	AQTLQIPFKAQ	GCCCAAAGTTGCA GATTCTTTAAGG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACGTTGCAGATT CCT	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCCTT AAAAGGAATCTGCAACGTTT GGGC
AAV-PHP.B5	AQTLQLPFKAQ	GCCCAAACCTCCA ACTCCCTTCAAAG CCCAA	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACCTCCAATC CCC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCTTGGGCTTT GAAGGGAGTTGGAGGGTT TGGGC
AAV-PHP.B6	AQTLQQPFKAQ	GCCCAAACCTTTGCA CAGCCGTTAAGG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACCTTTGCAGCAG CCG	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCCTT AAACGGCTGTGCAAGTTT GGGC
AAV-PHP.B7	AQSIERPFKAQ	GCCCAAAGCATCGA AAGACCTTCAAAG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAAGCATCGAAAG ACCC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCCTT GAAGGGTCTTTCGATGCTTT GGGC
AAV-PHP.B8	AQTMQKPFIAQ	GCCCAAACCATGCA AAAACCTTCAATCG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACCATGCAAAAA CCC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGA TGAAGGGTTTTGCATGGTT TGGGC
AAV-PHP.C1	AQRYQGDSVAQ	GCCCAAAGGTATCA GGGTGATTCTGTTG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAAGGTATCAGGG TGAT	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCAA CAGAATCACCTGATACCTT TGGGC
AAV-PHP.C2	AQWSTNAGYAQ	GCCCAATGGTCGAC AAACGCTGGTTACG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAATGGTCGACAAA CGCT	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGT AACCAGCGTTTGTGACCAT TGGGC
AAV-PHP.C3	AQERVGFQAQ	GCCCAAGAGCGTGT AGGTTTCGCACAGG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAGAGCGTGTAGG TTTC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGT GTGCGAACCTACACGCTCT TGGGC
AAV-PHP.N (Addgene ID: 127851)	AQTLAVPFSNP	GCGCAGACCCTAG CTGTCCCTTTTTTCG AACCCT	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCGAGACCCTAGCTGT CCCT	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCAGGGTTTCG AAAAGGGACAGCTAGGGTC TGCGC
AAV-PHP.X1	AQARQMDLSAQ	GCCCAAGCCAGACA AATGGACCTCAGCG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAGCCAGACAAAT GGAC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGC TGAGGTCCATTTGTCTGGCT TGGGC
AAV-PHP.X2	AQTNKVGNIQAQ	GCCCAAACCAACAA AGTCGGCAACATCG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAAACCAACAAAGTC GGC	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGA TGTTGCCGACTTTGTGGTT TGGGC
AAV-PHP.X3	AQQNVTKGVAQ	GCCCAACAGAACGT AACGAAGGGTGTG GCACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAACAGAACGTAAC GAAG	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCCA CACCCTCGTTACGTTCTGT TGGGC
AAV-PHP.X4	AQLNAIKNIAQ	GCCCAACTCAACGC TATCAAGAATCG CACAG	GGAGTCCTATGGACAAGT GGCCACAAACCACCAGAG TGCCCAACTCAACGCTATC AAG	TTCCTTGGTTTTGAACCCAA CCGGTCTGCGCCTGTGCGA TGTTCTGATAGCGTTGAGT TGGGC

Supplementary Table 5: Sequence motifs of AAV-PHP variants

The table provides the sequence motif information for all the new AAV-PHP variants identified in this study, and the primer information to clone them into pUCmini-iCAP-PHP.B backbone (Addgene ID: 103002) as described in Methods. The variants from 7-mer-i and 3-mer-s libraries were cloned as 11-mer substitution.

SUPPLEMENTARY NOTES 1-29:

Supplementary Note 1: *Spike-in* and *PCR pool* library

AAV9 and ~50 additional variants (and their alternative codon duplicates), identified in previously described work^{1,2} (includes well characterized variants like AAV-PHP.B or AAV-PHP.eB as well as many variants identified using the previous methodology but uncharacterized *in vivo*) act as internal selection controls and standards for the relative performance of the new variants (see Supplementary Dataset 1). The *PCR pool* library was generated by pooling the amplicons from the R1 brain selections after normalization based on their relative diversity (see Methods).

Supplementary Note 2: Experimental noise

The degree of enrichment at which correlation breaks down appears to vary with Cre-line. A downside of *PCR pool* is that there is no way to tell whether it or *synthetic pool* is the more 'true' enrichment score, or even that there may be cause for concern regarding certain enrichment values. The correlation among positively enriched variants between the two methods were found to improve with the magnitude of positive enrichment. For each experiment there is a level of enrichment below which the scores become irreproducible, or noisy.

Supplementary Fig. 3a demonstrates that neither PCR pool nor Synthetic pool is inherently more 'true' at lower enrichment scores. This is because Synthetic pool methodology with its codon replicates has a self-contained control to determine an enrichment level below which enrichment value has no further predictive power. In the main text we use the term 'noise' to refer to regions of enrichment in a particular experiment below which values lose their reproducibility and predictive power. Being able to experimentally determine enrichment signal above noise allows researchers to focus their attention and data analyses on enrichment levels that are internally reproducible and thereby avoid selecting false positive variants or drawing invalid conclusions.

Thus, if one is interested in only the highest enriched variants for a particular tissue, *PCR pool* design coupled with enrichment normalization to virus library may not drastically differ from *synthetic pool* design over one additional round of selection for a subset of *in vivo* selections (such as Tek-Cre or SNAP-Cre). Without additional validation, however, it is difficult to predict whether a given *in vivo* system will perform akin to Tek-Cre. This becomes critical in a multiplexed selection study where target-specific variants may not garner the highest enrichments in one particular *in vivo* selection.

Supplementary Note 3: The dominance of PHP.B-like motif

The ability to twice recover the AAV-PHP.B sequence family from completely independently constructed and selected libraries confirms that the viral library's sequence space coverage was broad enough to recover a family of variants sharing a common motif. Unlike CREATE which identified only one variant, AAV-PHP.B, M-CREATE yielded a diverse PHP.B-like family that hints toward important chemical features of this motif. The sequence diversity within this family

suggests that isolating AAV-PHP.B was not simply good fortune in our prior study (considering a theoretical starting library size of ~1.3 billion), and that this is a dominant family for this particular experiment.

Supplementary Note 4: AAV-PHP.V2 capsid

AAV-PHP.V2 was found at high abundance in R1 selection across all brain libraries and was highly enriched in R2 (Fig. 1d, 2g (right panel), 3a,b, Supplementary Fig. 1e). Given its sequence similarity, we predicted similar tropism to that of AAV-PHP.V1. We validated this *in vivo* in C57BL/6J adult mice (ssAAV-PHP.V2:CAG-mNeongreen genome, 3×10^{11} vg dose per adult mice, $n = 3$, Supplementary Fig. 5a), in Tek-Cre mice (ssAAV-PHP.V2:CAG-DIO-EYFP genome, 1×10^{12} vg dose per adult mouse, $n = 2$, Supplementary Fig. 5b), and in GFAP-Cre mice (ssAAV-PHP.V2:CAG-DIO-EYFP, 1×10^{12} vg dose per adult mouse, $n = 2$, Supplementary Fig. 5c).

Supplementary Note 5: Spike-in library validation *in vivo*

To evaluate the performance of the *spike-in* library, we chose two highly enriched variants similarly placed in sequence space: AAV-PHP.B6 – TLQLPFK and AAV-PHP.B7 – TLQQPFK (Supplementary Fig. 2d (middle panel), 3d) that were previously identified in the 3-mer-s PHP.B library² but never validated *in vivo*. At a modest dose of 1×10^{11} vg in C57BL/6J adult mice, these variants also display PHP.B-like tropism (Fig. 4a,b, Supplementary Fig. 6a).

Supplementary Note 6: 3-mer-s PHP.B library design

Briefly, the re-investigated 3-mer-s PHP.B library diversified positions 587-597 of the AAV-PHP.B capsid (equivalent of 587-590 AA on AAV9) in portions of three consecutive AAs, (~40,000 total variants) (Fig. 5a). Selections were performed in three Cre-transgenic lines: Vglut2-IRES-Cre for glutamatergic neurons, Vgat-IRES-Cre for GABAergic neurons, and GFAP-Cre for astrocytes.

Supplementary Note 7: Lack of neuronal subtype-specific selection

While Vglut2-Cre and Vgat-Cre mice were used for *in vivo* selection, we didn't find variants that stood out for neuronal subtype-specific transduction of excitatory and inhibitory populations from our initial investigations on the NGS dataset. It is possible that a biological solution to this (stringent) selection was not present in this library.

Supplementary Note 8: Plasmids for library generation

The rAAV- Δ Cap-in-cis-Lox2 plasmid consists of three major elements that are flanked by AAV2 ITRs.

(i) UBC ubiquitous promoter driving the expression of fluorescent protein, mNeonGreen, followed by a synthetic polyadenylation sequence. The mCherry expression cassette of the previous version of the plasmid was replaced by mNeonGreen cassette.

(ii) A portion of AAV2 rep gene that has the splicing sequences and AAV5 p41 promoter (1680-1974 residues of GenBank AF085716.1) followed by AAV9 cap gene. The prior version of this plasmid, rAAV- Δ Cap-in-cis-Lox, has a short 12 bp sequence between restriction sites XbaI and AgeI at AA 450 and 592 of the AAV9 Cap gene. This was replaced by a 723 bp sequence of mRuby2 gene in-frame (acts as filler DNA) in the newer version of the plasmid.

(iii) SV40 polyadenylation sequence that is flanked by lox71 and lox66 sites. The minor changes were introduced to the prior version of the plasmid to facilitate ease of cloning and to visualize mammalian cell transfection. The Lox sites in these rAAV plasmids show modest levels of Cre-independent flipping. This was minimized during PCR-based capsid recovery by lowering the number of amplification cycles to a point where we cannot recover any rAAV capsids from the control DNA extracted from wild-type mice (i.e., lacking Cre expression) that were injected with the library.

The pCRII-9Cap-XE plasmid contains the AAV9 capsid gene sequence from AAs 450-592 and is flanked by XbaI and AgeI restriction sites.

The AAV2/9 REP-AAP- Δ Cap plasmid has the five previously existing stop codons of AAV2/9 REP-AAP in addition to the deletion of AAs 450-592 of the AAV9 capsid sequence. These modifications did not affect vector production. The deletion of the overlapping fragment between the REP-AAP and rAAV- Δ Cap-in-cis-Lox2 plasmids minimizes recombination between plasmids that could potentially generate AAV9 wild-type capsids during co-transfection in vector production.

Supplementary Note 9: Plasmids for capsid characterization

pAAV:CAG-mNeonGreen² genome consists of a ubiquitous CMV- β -Actin-intron- β -Globin (CAG) hybrid promoter driving the expression of a fluorescent protein, mNeonGreen (equivalent plasmid, pAAV: CAG-eYFP³; Addgene ID: 104055). pAAV:CAG-NLS-EGFP¹ consists of NLS sequences at the N- and C-termini of EGFP and is driven by the CAG promoter. An equivalent version with one NLS is on Addgene (ID 104061). pAAV:CAG-DIO-EYFP³ (Addgene ID: 104052) consists of a EYFP gene built in the reverse direction of the CAG promoter, and it is flanked by a pair of Cre-Lox sites (Lox P and Lox 2272) on either ends. In cells expressing Cre, the Cre-lox pair inverts EYFP enabling transcription and translation, followed by excision in the lox site to prevent re-inversion. pAAV: GfABC1D-2xNLS-mTurquoise2³, referred to elsewhere as pAAV:GFAP-2xNLS-mTurquoise2 (Addgene ID: 104053), consists of NLS sequences at the N- and C-termini of mTurquoise2 and is driven by the astrocyte-specific promoter GfABC1D⁴. pAAV:Ple261-iCre⁵ (Addgene ID 49113) contains an endothelial-cell-specific promoter driving the expression of iCre.

We packaged pAAV:CAG-XFP (mNeonGreen) for characterizing AAV variants. However, when performing quantification of cell-types: neurons, astrocytes and oligodendrocytes, we use CAG-NLS-EGFP to restrict the expression to nucleus for easier quantification using microscope

images. GFAP-NLS-mTurq2 is used to quantify astrocytes. CAG-DIO-EYFP is used for Cre driver lines, due to the presence of lox sites in this plasmid.

The self-complementary genome from Dr. Guangping Gao, scAAV:CB6-EGFP genome has a hybrid ubiquitous CB6 promoter (975 bp) comprising a CMV enhancer (cytomegalovirus immediate early enhancer), a chicken- β -actin promoter and hybrid intron, that drives the expression of EGFP. The genome has a rabbit globin poly A (127 bp) following the EGFP gene. The scAAV:CAG-EGFP⁶ (Addgene ID:83279), vector uses a ubiquitous CMV- β -Actin-intron- β -Globin (CAG) hybrid promoter to drive the expression of EGFP.

Supplementary Note 10: AAV capsid library generation

To avoid PCR-induced biases resulting from point mutations, recombination, and template switching, PCR amplification of the library was limited to 15 – 20 cycles and the reactions were scaled up to get the required yield. The resulting PCR products were run on a 1% agarose gel and extracted with a Zymoclean Gel DNA Recovery kit (Zymo Research; D4007). It is critical to avoid AAV contamination during this step by taking precautionary measures like using a clean gel-running box and freshly prepared 1x TAE buffer.

Supplementary Note 11: Quality control for AAV capsid library generation

The following four steps were carried out to ensure successful capsid library generation.

1. To validate successful assembly of the library, 1 ng of the final assembled library was transformed into *E. coli* SURE 2 Supercompetent Cells (Integrated Sciences; 200152). We checked for colonies on an LB/Agar plate containing carbenicillin antibiotic after overnight incubation at 37°C.
2. The DNA library was sequenced around the insertion site (Laragen; Sanger Sequencing). A non-biased library may match the diversity of the NNK/MNN motif (where N = 25% each of A, T, G, C; K = 50% each of G, T; M = 50% each of A, C) with some fluctuations across the diversified region.
3. To verify that the ITRs were intact, SmaI digestion was carried out as per the NEB recommended protocol (NEB; R0141S).
4. To validate successful transfection and assess the vector-production yield per 150 mm dish, 10 ng of 7-mer-i library was used to transfect 293T producer cells (293T; ATCC CRL 3216). Uniform expression of mNeonGreen protein across HEK cells was observed, and an average yield of 0.1 – 1×10^{11} vg was obtained per 150 mm dish. Using the average yield per dish, we scaled up the vector production for *in vivo* selection (see Supplementary Fig. 2a).

Supplementary Note 12: Estimation of high-confidence variants from round-1 selection

This process involved estimation of the diversity precluding noise and consideration of amplification of this diversity across samples by determining the area under the curve for the interval of high-confidence variants that falls in the higher RC range. We estimated the area

under the curve (AUC) using the composite Simpson's rule by plotting all the recovered variants in a library (X-coordinate) to their read counts (RCs or copy number from deep sequencing data, Y-coordinate) (see Supplementary Fig. 1e). To determine the definite intervals for AUC, we sorted the data based on the decreasing order of the RCs. Noticeably, the distribution has two phases, with a steadier slope of variants in the higher RC range, followed by a steep drop in the slope of the curve (~50-1000 fold lower RCs). By observation, this steeper side of the curve is predominant in sequencing errors/ PCR mutations, hence we precluded this error dominant slope otherwise called noise from our AUC estimation. When comparing composite Simpson's rule with another function, such as composite trapezoidal rule, the difference was miniscule.

This area is then used to determine the fraction of an individual library that needs to be pooled into PCR pool library using the formula: [Area under the curve/ total number of libraries pooled].

Supplementary Note 13: R2 library generation and vector production

The R1 libraries used to build R2 were the Cre-Lox flipped rAAV DNA from half of the mouse brains (~0.3 g) and portion of spinal cords (0.1-0.2g) from all Cre lines. The amount of tissue processed here was sufficient for complete capsid library recovery. The differentially pooled and amplified libraries (by *PCR pool* or *synthetic pool*) were assembled using gibson assembly with a follow-up PS or Exonuclease V treatment (as described in R1 library generation). We validated successful library generation by transformation, Sanger sequencing, and an ITR SmaI digest.

For vector production, about 10 ng of the purified and assembled library was used to transfect each 150 mm dish of 293T cells, and we obtained a yield of about 6×10^{11} vg per 150 mm dish (i.e. the R2 yield was six times that of R1, unsurprisingly given these sequences have already produced well enough to survive R1 selection).

Supplementary Note 14: AAV virus library production

In addition to the 10 ng of library transfection per 150 mm dish of 293T producer cells, we transfected three plasmids: AAV2/9 REP-AAP- Δ Cap, pUC18 and pHelper (genes encoding adenoviral proteins for AAV replication) at a ratio of 1:1:2. The plasmid pUC18 acts as a filler DNA to compensate for the low amount of library DNA in order to maintain the N:P ratio required for optimal transfection using polyethylenimine (PEI, Polysciences; 24765-1) transfection). The cells and culture media were harvested at 60 h post-transfection to collect the viral particles. rAAV harvest and purification were performed as per the protocol³. The small amount of library DNA per plate and early cell harvest time are critical for reducing the possibility of mosaic capsid assemblies during vector production (similar considerations seen in prior reports⁷⁻⁹).

For *7-mer-i* library, the production was scaled up to 60 dishes ($\sim 1.8 \times 10^7$ cells/dish) and with ~10% transfected with the library, resulted in $\sim 1 \times 10^8$ total transformants. For an NNK 7mer library with $\sim 1 \times 10^8$ total transformants, the number of unique variants is 9.99×10^7 (See Bosley & Ostermeier (2004)¹⁰, section 2.1.2 for mathematics).

Supplementary Note 15: rAAV genome extraction by proteinase K

In order to degrade any contaminating DNA from the purified library, it was treated with DNase I enzyme (5 μ l of 10 U/ μ l) (Sigma-Aldrich; 4716728001) in 100 μ l of DNase I buffer and incubated for 1 h at 37°C. The enzyme was inactivated by adding 5 μ l of 0.5 M EDTA at 70°C for 10 min. Following DNase I treatment, the capsid protein shell was digested by adding 120 μ l of proteinase solution containing 5 μ l of 20 μ g/ μ l of proteinase K and incubated at 50°C overnight. To inactivate the proteinase K, the mixture was boiled at 95°C. The extracted rAAV library DNA was then concentrated and purified using phenol chloroform and ethanol. An equal volume of Phenol:Chloroform:Isoamyl Alcohol 25:24:1, pH8.0 (~250 μ l; ThermoFisher Scientific; 15593031) was added and vortexed for 30 s. The mixture is incubated for 5 min at room temperature (RT) before centrifugation at 15,000 rpm for 10 min at 4°C. The upper aqueous phase was separated and mixed with an equal volume of chloroform and vortexed for 30 s. Following 5 min incubation at RT, centrifuge at 15,000 rpm for 10 min at 4°C. The upper aqueous phase was separated and one-tenth volume of 3M sodium acetate (pH 5.2) along with 2 μ l Co-Precipitant Pink (Bioline; BIO-37075) and 2.5 volumes of ice cold 100% ethanol was added before vortexing for 30 s. The mixture was incubated for at least 1 hr at -20°C before centrifugation at 15,000 rpm for 15 min at 4°C. The pellet was air dried and resuspended in TE buffer. The DNA concentration was determined using the Qubit ssDNA assay.

Supplementary Note 16: Optimization of rAAV genome extraction from tissue

For rAAV genome extraction from tissues, we used both the Trizol method (Life Technologies; 15596) and the QIAprep Spin Miniprep kit (Qiagen, Inc; 27104) according to the manufacturers' recommended protocols, and found the Trizol method to be more efficient (see Supplementary Fig. 1f,g,i). The total rAAV genome recovery from 0.1 g of mouse liver was quantified by quantitative PCR using the primers mNeonGreen-F and mNeonGreen-R, which binds to the mNeonGreen gene of the ssAAV- Δ Cap-in-cis-Lox2 genome (see Supplementary Table 4). As an internal control, we also quantified the amount of mitochondrial DNA (a surrogate for the recovery of circular, episomal genomes) using primers Mito-F and Mito-R (see Supplementary Table 4). Although the percentage of viral DNA per 1 ng total extracted DNA was about 1.5 fold higher with the QIAprep kit than with the Trizol method, the overall recovery was lower with the QIAprep kit.

The extracted viral genome was digested with a restriction enzyme, such as SmaI (found within the ITRs), as it appeared to help improve rAAV genome recovery by PCR with ~4 fold change (see Supplementary Fig. 1h, Δ CT ~ 2, 1 CT = 2-fold difference in DNA copies, fold change = $(2^{\Delta$ CT) = 4). This was analyzed by quantitative PCR with Cre- primers, CapF-56 and Cre-R-57 (see Supplementary Table 4).

Supplementary Note 17: The optional UID feature

The primer, 588i-lib-PCR1-6bpUID-F: 5' CACGACGCTCTTCCGATCTAANNNNNNAGTCCTATGGACAAGTGGCCACA-3' used to minimally amplify DNA and virus libraries for NGS has 6 nucleotides long UID (unique identifier) "NNNNNN" that sits after 19 nucleotides of Read-1 sequence used in NGS "5'-

CACGACGCTCTTCCGATCT” and linker “AA”. The sequence after UID “AGTCCTATGGACAAGTGGCCACA” is the region that anneals to the AAV9 capsid. UID is an optional feature for NGS data analysis to identify potential PCR amplification errors. However, this feature wasn't utilized in the NGS data analysis in this study to maintain consistency with the primers used in rAAV genome recovery from tissues which lacks this UID feature (primers 71F and CDF/R, See Supplementary Table 4). The UID or any kind of overhangs seemed to affect the PCR based recovery from tissue. Presumably, the primer thermostability have a key role to play in very low amount of extracted rAAV genomes from tissues.

Supplementary Note 18: Additional processing and consideration for NGS sample preparation

The PCR products post indices addition were run on a freshly prepared 2% low-melting-point agarose gel (ThermoFisher Scientific; 16520050) for better separation and recovery of the approx. 120 bp DNA band on the gel. Before sending the sample for NGS, the nucleotide diversity at the randomized 7-mer position was verified by Sanger sequencing. If needed, an optional PCR was carried out to send sufficient sample for Sanger sequencing using 15 – 20 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 10 s with the primers NGS-QC-F and NGS-QC-R (see Supplementary Table 4). Upon validation, the libraries were sent for deep sequencing using the Illumina HiSeq 2500 System (Millard and Muriel Jacobs Genetics and Genomics Laboratory, Caltech; Integrative Genomics Core, City of Hope).

Supplementary Note 19: AAV dosage and expression time

The AAV doses are determined by the experimental needs. CAG-NLS-GFP related experiments for quantification were done at medium dose of 1×10^{11} vg given this was the dose previously determined for AAV-PHP.eB characterization. Otherwise, the non-NLS genome related experiments were done at 3×10^{11} vg, with the exception of Cre-driver lines (GFAP-Cre or Tek-Cre), or a lower strength promoter containing genome (GFAP-NLS-mTurq) where the dose was 1×10^{12} vg. The high dose was chosen to understand the full potential of the new vectors in these systems.

All experiments with vectors carrying CAG, a strong ubiquitous promoter, were incubated for 3 weeks. The 4 week incubations are those that involved expression from Cre driver lines or cell-type specific promoter where it is generally recommended for a longer wait time. The 2 week incubations are those where the vectors carried self-complementary genomes with strong ubiquitous promoters.

Supplementary Note 20: Immunohistochemistry

Tissue sections, typically 100- μ m thick, were first incubated in blocking buffer (10% normal donkey serum, 0.1% Triton X-100, and 0.01% sodium azide in 0.1 M PBS, pH 7.4) with primary antibodies at appropriate dilutions for 24 h at RT on a rocker. The primary antibodies used in this study were rabbit S100 (1:400, Abcam, ab868), rabbit Olig2 (1:400; Abcam, ab109186), rabbit NeuN (1:400, Abcam, ab177487), and rabbit GLUT-1 (1:400; Millipore Sigma, 07-1401).

After primary antibody incubation, the tissues were washed 1 – 3 times with wash buffer 1 (0.1% Triton X-100 in 0.1 M PBS buffer, pH 7.4) over a period of 5 – 6 h in total. The tissues were then incubated in blocking buffer with the secondary antibodies at appropriate dilutions for 12 – 24 h at RT and then washed in three times in 0.1 M PBS, pH 7.4 over a total duration of 5 – 6 h. The secondary antibody was Alexa Fluor 647 AffiniPure donkey anti-rabbit IgG (H+L) (Jackson ImmunoResearch Lab, 711-605-152). When performing DNA staining, 4',6-Diamidino-2-phenylindole dihydrochloride (DAPI, Sigma Aldrich, 10236276001) is used at a 1:1000 dilution in 0.1 M PBS, pH 7.4 and incubated with tissues for 15 minutes followed by a single wash for 10 minutes in 0.1 M PBS, pH 7.4. The DAPI and/or antibody-stained tissue sections were mounted with ProLong Diamond Antifade Mountant (ThermoFisher Scientific, P36970).

Supplementary Note 21: Tissue processing for hybridization chain reaction

To characterize the AAV capsid variant AAV-PHP.N in brain tissue, HCR method was sought to label excitatory and inhibitory neurons. Fluorescence *in situ* hybridization chain reaction (FITC-HCR) was used to label excitatory neurons with VGLUT1 and inhibitory neurons with GAD1. Adapting the third-generation HCR¹¹, we designed 13 probe sets for each target by using custom-made software (<https://github.com/GradinaruLab/HCRprobe>). After 3 weeks of expression, the mice were transcardially perfused and fixed as described earlier (Section D. Tissue processing). To minimize RNase enzyme exposure in fixed tissues, following overnight fixation in 4% PFA, the tissues were washed and stored at 4°C in 0.1 M RNase-free PBS and 0.05% sodium azide. The harvested brains were henceforth handled with care to avoid exposure to RNase using reagents such as RNAlater stabilization solution/RNase-free PBS/RNaseZap (ThermoFisher Scientific, AM7021, AM9624, AM9780). Once the harvested brains were sagittally sliced to 100-µm thick sections, we performed FITC-HCR to detect both genes. We permeabilized tissue slices with 0.1% Triton X-100 in 0.1 M RNase-free PBS for 1 h at RT and pre-hybridized them in hybridization solution (10% dextran sulfate and 10% ethylene carbonate in 2xSSC buffer (saline-sodium citrate)) for >30 min at 37°C. The designed probes were diluted in hybridization solution to get a final concentration of 2 nM. The tissue sections were then subjected to hybridization with the probes overnight at 37°C. Following this, the sections were washed with pre-warmed wash buffer (10% ethylene carbonate in 2xSSC) at 37°C for 30 min twice, followed by 2xSSC at RT for 30 min twice. Amplification with hairpin pairs (Molecular Technologies, CA) were performed in amplification buffer (10x dextran sulfate in 2xSSC); hairpins were snap-cooled at 95°C for 90 s, followed by RT for 30 min, and diluted with amplification buffer (60 nM). Tissues were then incubated in this amplification buffer with hairpins overnight at RT with gentle agitation. Once the amplification was done, samples were briefly washed with 2xSSC and mounted in Prolong Diamond for imaging.

Supplementary Note 22: Measures to prevent imaging artifacts

To prevent any imaging artifacts resulting from multiple fluorescence spectral overlap, the fluorescence excitation and emission spectra were kept distinct following the recommended linear unmixed acquisition of individual colors. A far-red fluorescent dye was chosen for any additional marker staining to keep the imaging parameters distinct from *in vivo* fluorescent expression thereby preventing any spectral overlap across detector channels. The tissues were routinely monitored for auto fluorescence or imaging artifacts before acquisition, and imaging

parameters were adjusted if needed. The imaging parameters were cross-checked with tissues lacking *in vivo* transduction to avoid any imaging artifacts. The regions used for the images were closely matched across experimental groups to minimize bias during comparisons.

Supplementary Note 23: Tissue clearing, optical clearing and imaging of thick tissues

To demonstrate the ability of PHP.V1 to transduce the vasculature across thick tissues, such as a mouse-brain hemisphere, we assessed tissue from Tek-Cre mice 4 weeks' post administration. The brain hemisphere was stained with the primary antibody, Anti-GFP (1:200, Aves Labs, GFP-1020), and the secondary antibody, goat anti-Chicken IgY, Alexa Fluor 633 (1:200, ThermoFisher Scientific, A-21103), and cleared via the iDISCO protocol¹². For imaging, a commercial light-sheet microscope (Lavisoin BioTec) with a custom objective lens (4x) was used¹³. The resulting image files were reorganized by a custom MATLAB script to allow stitching with TeraStitcher¹⁴. For 3-D visualization, Imaris (Bitplane) was used.

For images of peripheral organs such as heart, muscle, spleen that were over 500 microns in thickness, optical clearing was performed by incubating the tissues in 5 ml of *ScaleS4(0)*¹⁵ solution overnight at RT with gentle agitation, then mounted the tissues in fresh *ScaleS4(0)* solution and imaged under the confocal microscope.

Supplementary Note 24: Tissue processing and imaging for quantification of rAAV transduction *in vivo*

For quantification of PHP.B-family variant transduction in tissues, the images were acquired using 25x objective with 1x digital zoom on a Zeiss LSM 880 confocal microscope. With n=3 mice per variant, images were acquired across 4 brain regions – cortex, striatum, ventral midbrain and thalamus and tissues were stained with 3 cell type markers (NeuN, Olig2, and S100). For each mouse, 2 images per brain region per cell type marker were acquired, and the mean were plotted.

For PHP.N transduction analysis, the images were acquired using 20x objective on Keyence BZ-X700 microscope. With n=3 mice, images across 4 brain regions - cortex, striatum, ventral midbrain and thalamus were acquired to cover the entire brain regions for 3 cell type markers (NeuN, Olig2, and S100). This involved 6-8 images to cover cortex, thalamus and striatum, and 2 images to cover ventral midbrain per mouse per cell-type marker. For each mouse, across each region, the mean from the images were plotted.

For PHP.V GLUT1⁺ transduction analysis, the images were acquired using 25x objective with 1x digital zoom on a Zeiss LSM 880 confocal microscope. Each distinct blood vessel in the image with GLUT1⁺ staining and XFP expression was determined as positive for transduction. Quantification of expression from the CAG-mNeonGreen vector was performed across the cortex (n = 3 per group). Each data point is drawn from the mean of 3-2 images per mouse. Different brain regions were quantified for Tek-Cre and Ai14 mouse experiments with n = 2 per group. For cortex, cerebellum, striatum and ventral midbrain, the mean was plotted from 3-4 images per mouse per region.

Supplementary Note 25: Quantification of rAAV vector transduction

Manual counting was performed for GLUT-1-stained blood vessels and expression of the ssAAV:CAG-mNeonGreen and ssAAV:CAG-DIO-EYFP, where the efficiency was calculated as the percentage of XFP⁺ vessels relative to the GLUT-1 staining. Manual counting was also performed to quantify nuclear or soma stained cells, including NeuN-, Olig2-, and S100-stained cells. The efficiency was calculated as the percentage of XFP⁺ cells relative to cell-marker⁺ cells.

Keyence Hybrid Cell Count software (BZ-H3C) was used to quantify expression of nuclear localized AAV genomes in liver hepatocytes that co-localized with the DNA stain, DAPI; and also for the study involving ssAAV:GFAP-2xNLS-mTurquoise2 genomes with S100 cell marker.

The mean fluorescence intensity across microscopic images were quantified using ImageJ software. The images were processed for background subtraction and using the Threshold operation, the mean fluorescence intensity was measured.

Supplementary Note 26: NGS raw data processing

The pipeline to process these datasets involved filtering the dataset to remove the low-quality reads by using the deep sequencing quality score for each sequence. The variant sequences were then recovered from the sequencing reads by searching for the flanking template sequences, and extracting the nucleotides of the diversified region (perfect string match algorithm). The quality of the aligned data was further investigated to remove any erroneous sequences (such as ones with stop codons). The raw data was plotted (as shown in Supplementary Fig. 1e) to study the quality of recovery across every library. Based on the RC distribution, we adapted a thresholding method to remove plausible erroneous mutants that may have resulted from PCR or NGS based errors. The assumption is that if there is a PCR mutation or NGS error on the recovered parent sequence, the parent must have existed at least one round earlier than the erroneous sequence, and thus a difference in RCs should exist.

For R1 tissue libraries, we observed a steep drop in the slope of the distribution curve following a long tail of low count sequences, and were found to be rich in sequences that are variations of the parents in the higher counts range. We manually setup a threshold for RCs to remove such erroneous mutants. The thresholded data were then processed differently based on the experimental needs as described elsewhere using custom Python based scripts.

For R2 tissue libraries from *PCR pool* and *synthetic pool*, given the smaller library size compared to R1, we thresholded the data in two steps. We only considered the tissue recovered sequences that were present in the respective input DNA and virus library (after removing lower count variants from input libraries following the same principle as R1 tissue libraries). This step partially removed the long tail of low count reads. As a second step, we applied the thresholding that was described for R1 tissue libraries.

While it is plausible that true variants may be lost during thresholding, this method minimized false positives as the low count mutants in tissue and virus libraries often seemed to have very high enrichment score (as RCs are normalized to input library). In other words, thresholding allowed selective investigation on enriched variants that had a higher-confidence in their NGS RCs.

As an alternative to our manual thresholding method, an optional error correction method called “Collapsing” was built to further validate the outcome from filtered datasets. This method starts at the lowest count variants (variants of count 1) and searches for potential parent variants that are off by one nucleotide but have at least 2-fold higher counts (fold change = $(2^{\Delta CT})$ where CT is PCR cycle threshold). This error correction method then transfers the counts of these potential erroneous sequences to their originating sequences and repeats recursively until all sequences have been considered. On applying this error correction to our thresholded data, an additional ~0.002-0.03% of sequences were captured (compared to >19% captured by thresholding), confirming that our thresholding strategy was largely successful.

Supplementary Note 27: Estimation of enrichment score for R1 selected variants.

Since the DNA and virus libraries were not completely sampled unlike the tissue libraries, we assigned an estimated RC for variants that were not present in the input library but were present in the output library. For instance, R1 virus library is the input library to the R1 tissue libraries. The estimated RC is defined as a number that is lower than the lowest RC in the library with the assumption that these variants were found at a relatively lower abundance than the variants recovered from the deep sequencing. In virus libraries, since RC of 1.0 was the lowest, we assigned all missing variants an estimated RC of 0.9. We use this method to calculate the enrichment score of the R1 tissue libraries which is normalized to R1 virus library (Fig. 1d). This was done to represent libraries across two selection rounds consistently. Although, the individual enrichment score among R1 variants didn't add a significant value to the variants selected for R2 selection as described in the criteria to separate signal vs noise in R1 using the RCs.

Supplementary Note 28: Heatmap for amino acid bias

Each heatmap uses both an expected (input) distribution of amino acid sequences and an output distribution. The output distribution must be a list of sequences and their count, and the input distribution can be either a list of sequences and their count, or an expected amino acid frequency from a template, such as NNK. For both input and output, the total count of amino acids in each position is tallied in accordance to each sequence's count and then divided by the total sum of counts, giving a frequency of each amino acid at each position. Then, the log2 fold change is calculated between the output and the input. For amino acids with a count of 0 in either the input or output, no calculation is performed. In order to distinguish between statistically significant amino acid biases, a statistical test was performed using the statsmodels Python library. For the case where there are two amino acid counts, a two-sided, two-proportion z-test was performed; for comparing the output amino acid count to an expected input frequency from a template, a one-proportion z-test was performed. All p-values were then corrected for multiple comparisons using Bonferroni correction. Only bias differences below a significance threshold of $1e-4$ are then outlined on the heatmap; all other (insignificant) squares are left empty.

Supplementary Note 29: Clustering analysis

The reverse Hamming distances (representing the number of shared AAs between two peptides) was determined for all unique capsid variants with greater than 10 count and greater than 2.5-fold enrichment after R2 selection. This process iteratively compares each variant with all other variants within the group. Capsid variants were then clustered by their reverse Hamming distances using Cytoscape. The minimum reverse Hamming distance for visualization was chosen manually based on sequence similarity.

For the amino acid frequency plots, the number on the bottom represents the position of the diversified motif starting from 1. The size of the amino acid in the stack reflects the proportion of unique clones in which the AA appears at that specific position in the motif. The color code is based on the AA properties. The positively charged residues K, R, and H are in blue. The negatively charged residues D and E are in red. The amide containing polar residues Q, and N are in magenta. The polar residues T, and S, are in green. The hydrophobic residues A, L, V, I, P, F, M, and W are in black.

REFERENCES:

1. Deverman, B. E. *et al.* Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nat. Biotechnol.* **34**, 204–209 (2016).
2. Chan, K. Y. *et al.* Engineered AAVs for efficient noninvasive gene delivery to the central and peripheral nervous systems. *Nat. Neurosci.* **20**, 1172–1179 (2017).
3. Challis, R. C. *et al.* Systemic AAV vectors for widespread and targeted gene delivery in rodents. *Nat. Protoc.* **1** (2019) doi:10.1038/s41596-018-0097-3.
4. Lee, Y., Messing, A., Su, M. & Brenner, M. GFAP promoter elements required for region-specific and astrocyte-specific expression. *Glia* **56**, 481–93 (2008).
5. de Leeuw, C. N. *et al.* rAAV-compatible MiniPromoters for restricted expression in the brain and eye. *Mol. Brain* **9**, (2016).
6. Paulk, N. K. *et al.* Bioengineered Viral Platform for Intramuscular Passive Vaccine Delivery to Human Skeletal Muscle. *Mol. Ther. - Methods Clin. Dev.* **10**, 144–155 (2018).
7. Maheshri, N., Koerber, J. T., Kaspar, B. K. & Schaffer, D. V. Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nat. Biotechnol.* **24**, 198–204 (2006).

8. Schmit, P. F. *et al.* Cross-Packaging and Capsid Mosaic Formation in Multiplexed AAV Libraries. *Mol. Ther. - Methods Clin. Dev.* **17**, 107–121 (2020).
9. Koerber, J. T., Jang, J.-H. & Schaffer, D. V. DNA Shuffling of Adeno-associated Virus Yields Functionally Diverse Viral Progeny. *Mol. Ther.* **16**, 1703–1709 (2008).
10. Bosley, A. D. & Ostermeier, M. Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomol. Eng.* **22**, 57–61 (2005).
11. Choi, H. M. T. *et al.* Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018).
12. Renier, N. *et al.* iDISCO: A Simple, Rapid Method to Immunolabel Large Tissue Samples for Volume Imaging. *Cell* **159**, 896–910 (2014).
13. Dodt, H.-U. *et al.* Ultramicroscopy: development and outlook. *Neurophotonics* **2**, 041407 (2015).
14. Bria, A. & Iannello, G. TeraStitcher - a tool for fast automatic 3D-stitching of teravoxel-sized microscopy images. *BMC Bioinformatics* **13**, 316 (2012).
15. Hama, H. *et al.* ScaleS: an optical clearing palette for biological imaging. *Nat. Neurosci.* **18**, 1518–1529 (2015).