nature biotechnology

SUPPLEMENTARY INFORMATION

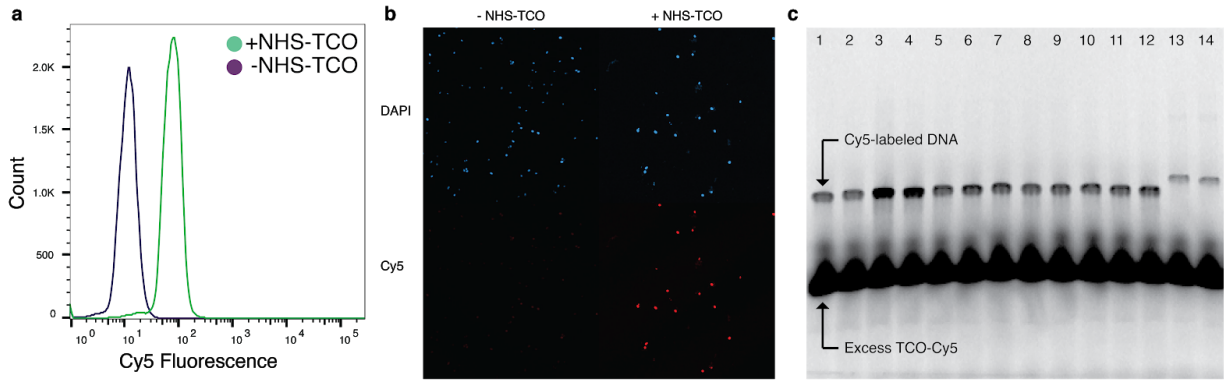
https://doi.org/10.1038/s41587-019-0372-z

In the format provided by the authors and unedited.

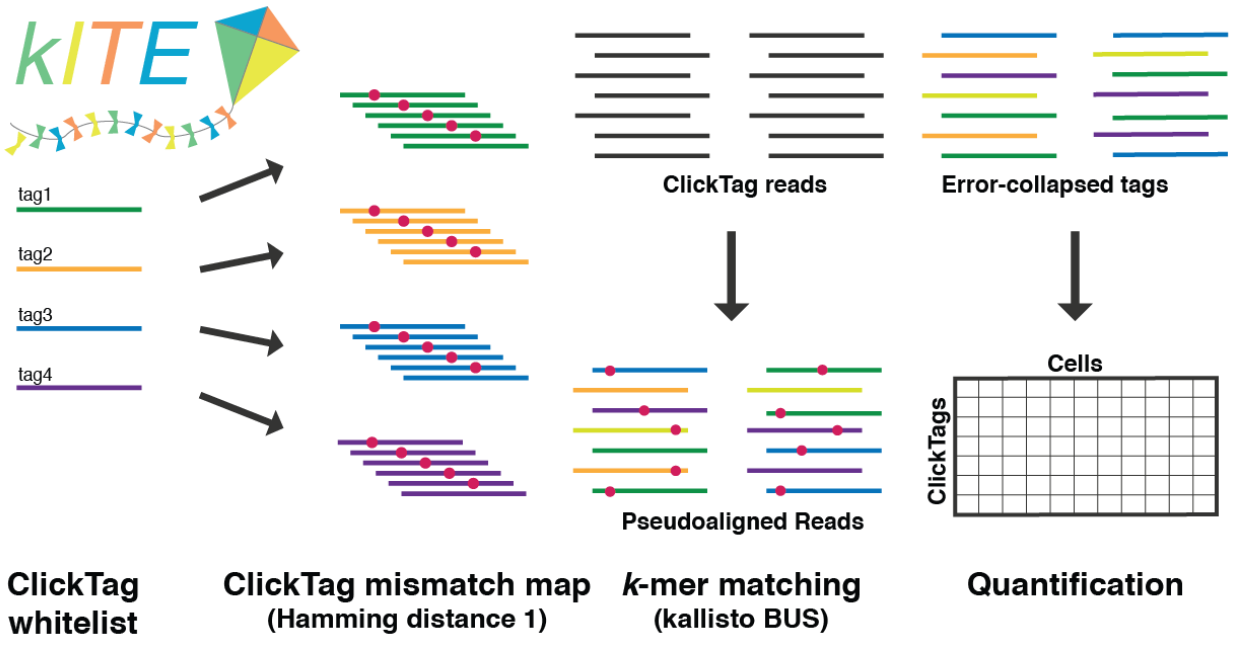
Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins

Jase Gehring[1,2], Jong Hwee Park[2], Sisi Chen[2], Matthew Thomson[2] and Lior Pachter[2,3]*

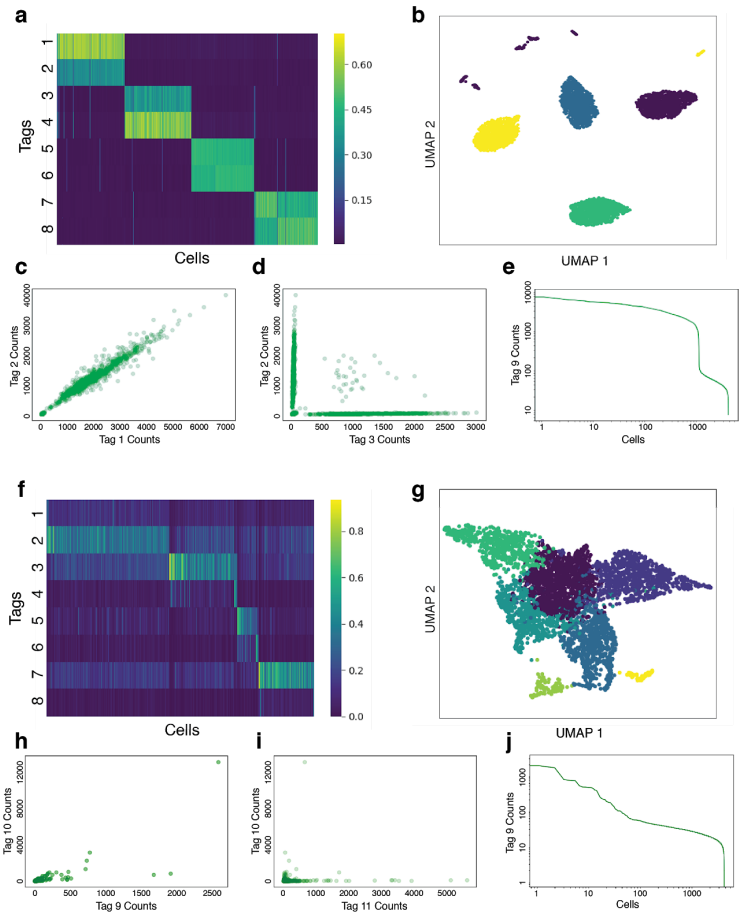
[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [3]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. *e-mail: lpachter@caltech.edu



Supplementary Figure 1: Direct labeling with Inverse Electron-Demand Diels-Alder (IEDDA) chemistry. (**a**) Yeast cells were fluorescently labeled in a one-pot, two-step reaction with NHS-TCO and MTZ-Cy5. Control reactions omitted NHS-TCO. (**b**) Fluorescence microscopy of yeast cells labeled with NHS-TCO and MTZ-Cy5 show labeling only in the presence of NHS-TCO cross-linker. (**c**) Activity assay for panels of methyltetrazine-activated ClickTags. MTZ-DNAs were reacted with TCO-Cy5 and the products separated by polyacrylamide gel electrophoresis. Lanes 1-12 are 3'-amine modified, while lanes 13 and 14 are 5'-amine modified. Data shown are representative results from one of three independent experiments.

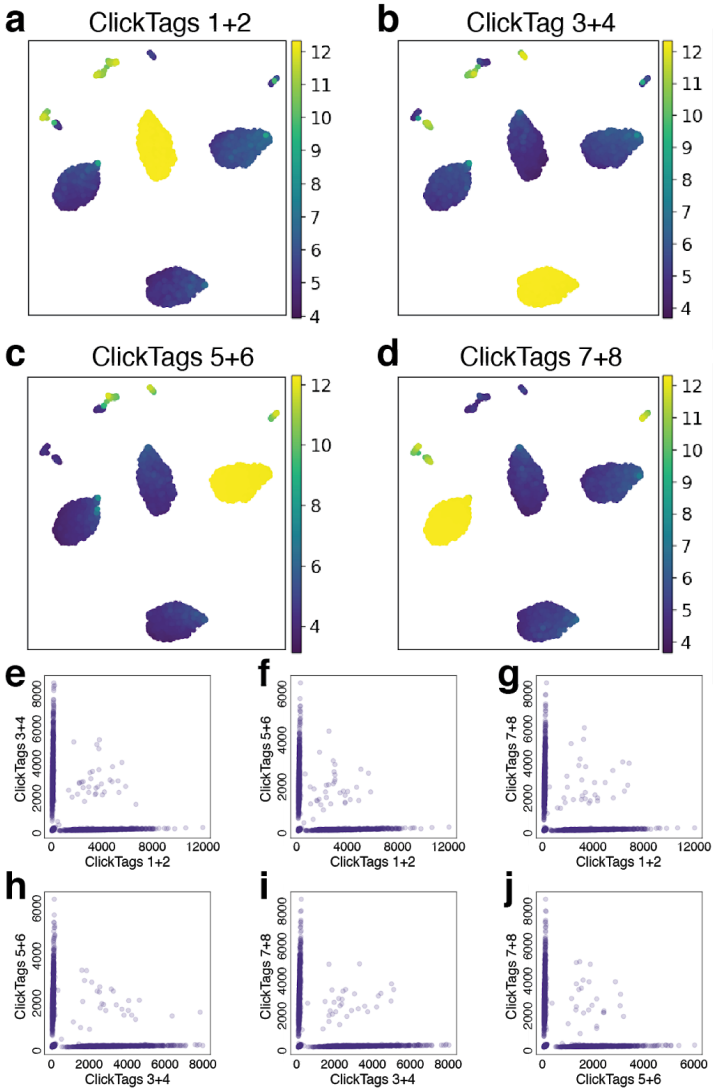


Supplementary Figure 2 Overview of the kITE (kallisto Indexing and Tag Extraction) workflow. In kITE, the kallisto RNA-seq pseudoalignment algorithm is used for fast matching of sequencing reads to ClickTag barcodes. To account for the occurrence of errors in sequencing data, a whitelist of ClickTag barcode sequences is converted to a mismatch map containing the correct barcodes as well as all of their Hamming distance 1 variations. The mismatch map is used to create a kallisto index, and `kallisto bus` commands are run without modification, producing a BUS file where each record contains a unique 10x cell barcode/UMI combination and the identity of the matched sequence. Finally, the mismatch map is used to collapse the BUS file into a *ClickTags x Cells* matrix which can be analyzed with standard scRNA-seq software.

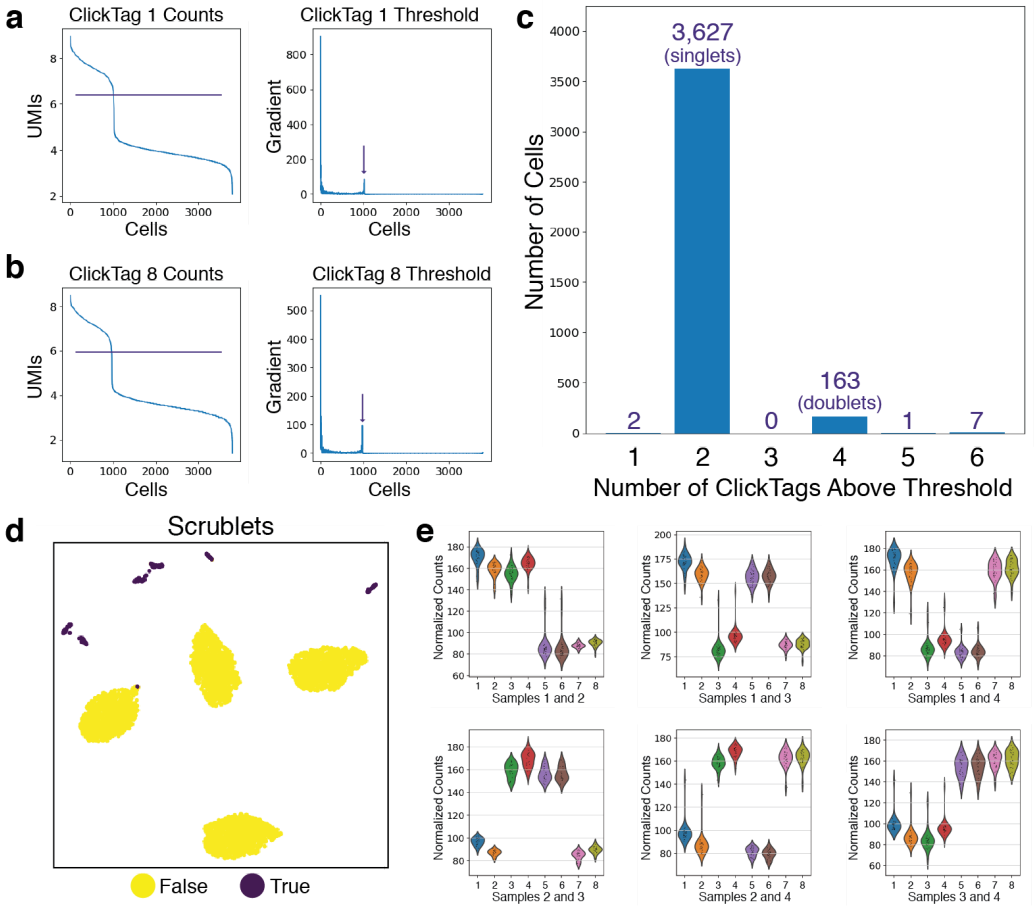


Supplementary Figure 3 Proof-of-concept ClickTag labeling experiment. (**a**) Heatmap showing n=3,800 detected cells originating from four methanol-fixed samples, each labeled with a pair of sample-specific ClickTags. (**b**) UMAP visualization of ClickTag data from panel **a** (n=3,800) colored by Louvain community detection. Four main clusters are observed, corresponding to the four individual samples as well as $\binom{4}{2} = 6$ small clusters corresponding to each possible combination of cell doublet originating from two different samples. (**c**) Scatter plot of counts for ClickTags 1 and 2, which were used to label the same sample, from panel **a**. The low-count population (bottom-left) is background from droplets not containing cells from the sample, while the high-count population corresponds to positive cells from the sample, and displays a strong correlation between the two ClickTag counts (Pearson's correlation coefficient r = 0.96, n=3,800). (**d**) "Barnyard plot" showing two ClickTags from separate samples.

ClickTag labeling is orthogonal, with doublets identifiable as points away from the axes (n=3,800). (**e**) Counts for ClickTag 1 from each cell in the experiment (n=3,800), ordered from highest to lowest and showing an inflection point between ClickTag 1 (+) and ClickTag 1 (-) cells. (**f-j**) Similar analysis for four samples of live cells labeled using the same procedure (n=3,800 filtered cell barcodes). Unlike the case with methanol-fixed cells, poor signal-to-noise precludes highly accurate sample assignment for aqueous-labeled cells.

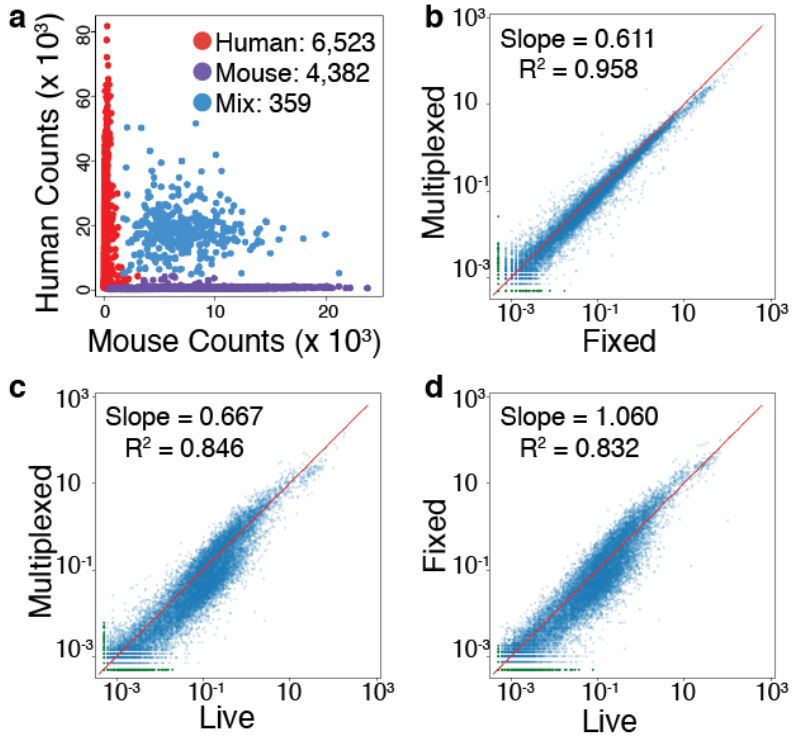


Supplementary Figure 4 Analysis of ClickTag "expression" from the methanol-fixed four-sample multiplexing experiment shown in **Supplementary Fig. 3a-e**. Panels **a-d** show the sum of normalized, log-transformed ClickTag counts for each of the four samples. Each pair of unique ClickTags labels one cluster and exactly three sub-clusters (doublets). Panels **e-j** show barnyard plots for pairs of ClickTags corresponding to the experimental design.

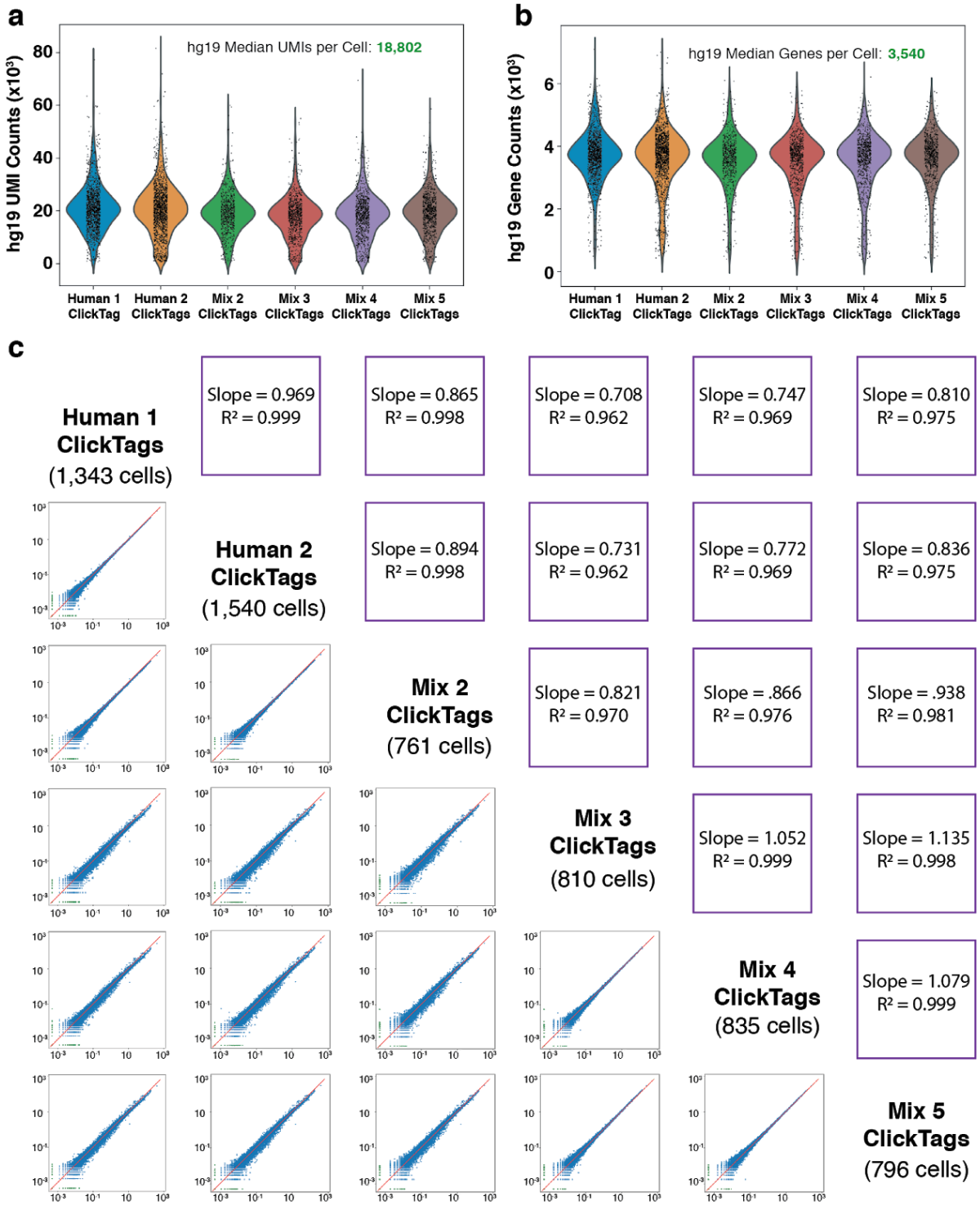


Supplementary Figure 5 Sample assignment based on ClickTags from four-sample multiplexing experiment shown in **Supplementary Fig. 3a-e**. Facile thresholding can be achieved by taking the gradient of the rank-UMI plot. Examples are shown for two ClickTags from this experiment (**a** and **b**). Of n=3,800 cells analyzed, 3,627 cells had two ClickTags above threshold, which in all cases corresponded to a pair of sample-specific ClickTags. A doublet rate of ~4.5% is comparable with the doublet rate of

~3% estimated by 10x Genomics. (**d**) The Scrublet algorithm was used to computationally identify 168 cell multiplets based on ClickTags. The doublets identified by scrublet were extracted and grouped into six clustions by Louvain community detection. Violin plots for all six clusters are shown in (**e**), with each panel showing the distribution of ClickTags for a given cluster of cell doublets.

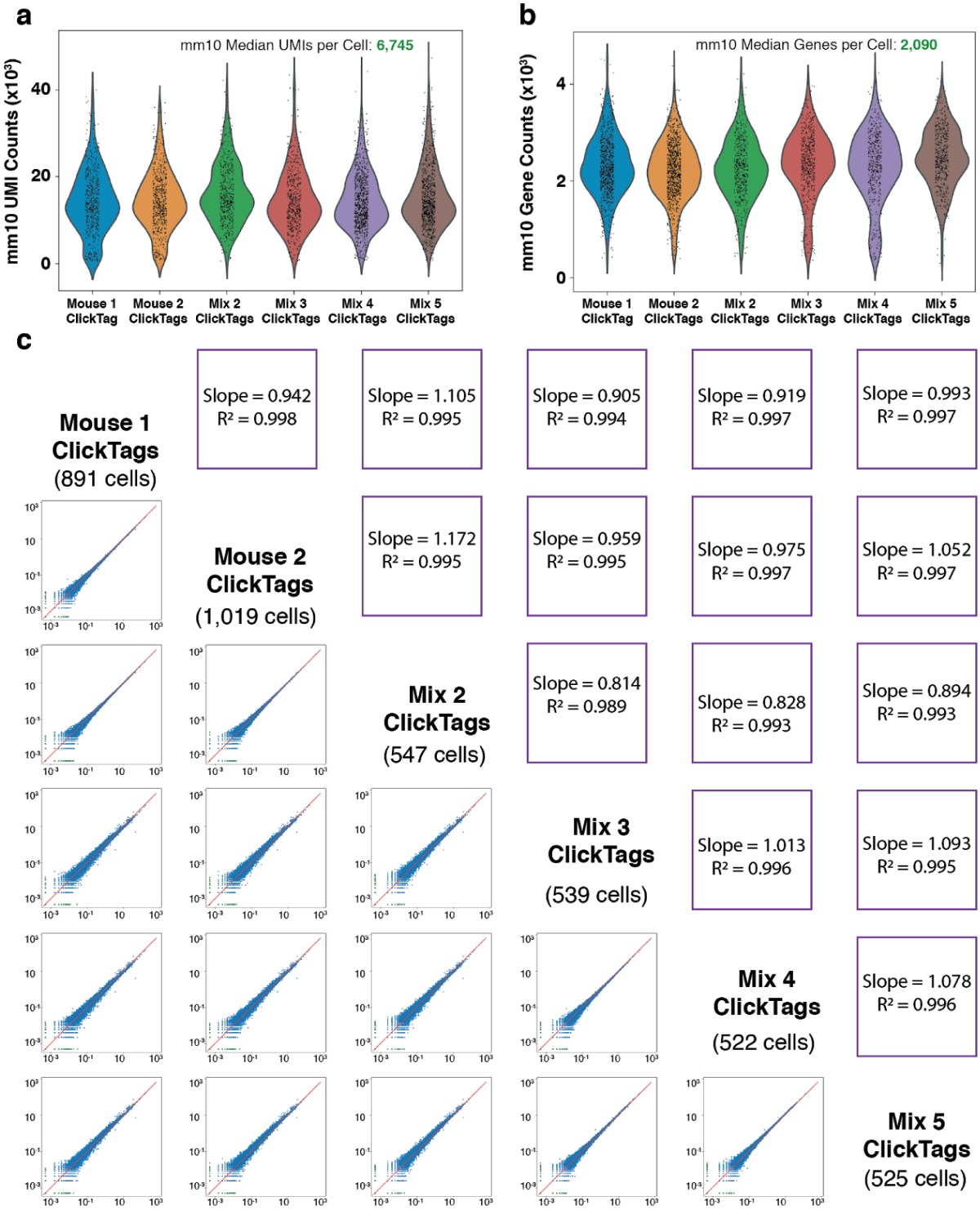


Supplementary Figure 6 Species-mixing experiment fidelity analysis. (**a**) Barnyard plot depicting cDNA counts for n=11,264 filtered cells colored by species as determined by CellRanger. (**b**) Pearson correlation of gene expression across n=27,998 mouse genes for methanol-fixed mouse NSCs treated with ClickTags versus untreated, methanol-fixed cells. cDNA from live NSCs was also used for comparison against methanol-fixed, ClickTagged cells (**c**) or untagged, methanol-fixed cells (**d**). Gene expression is shown as average counts per cell for each mouse gene.



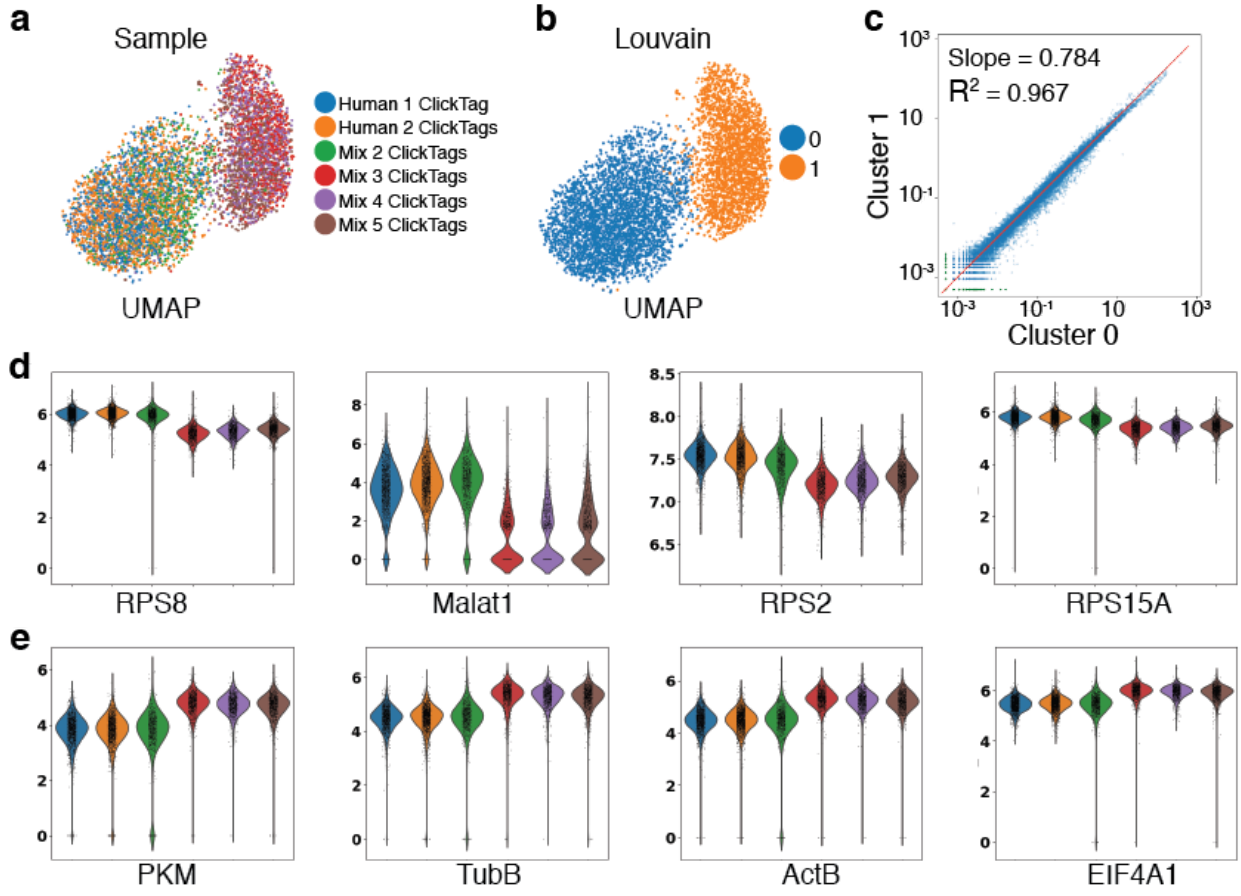
Supplementary Figure 7 Comparison of human cDNA libraries across samples from the species-mixing experiment. (**a**) Violin plot of UMIs per cell for all n=6,087 captured human cells identified by

CellRanger and grouped according to sample identification shown in **Supplementary Figure 12a**. (**b**) Violin plot of genes per cell for all n=6,087 captured human cells identified by CellRanger, grouped according to sample identification shown in **Supplementary Figure 12a**. (**c**) Pearson correlation of gene expression from all samples containing human cells. Gene expression is shown as average counts per cell for each of n=32,738 human genes. The number of human cells captured for each sample is shown also shown.



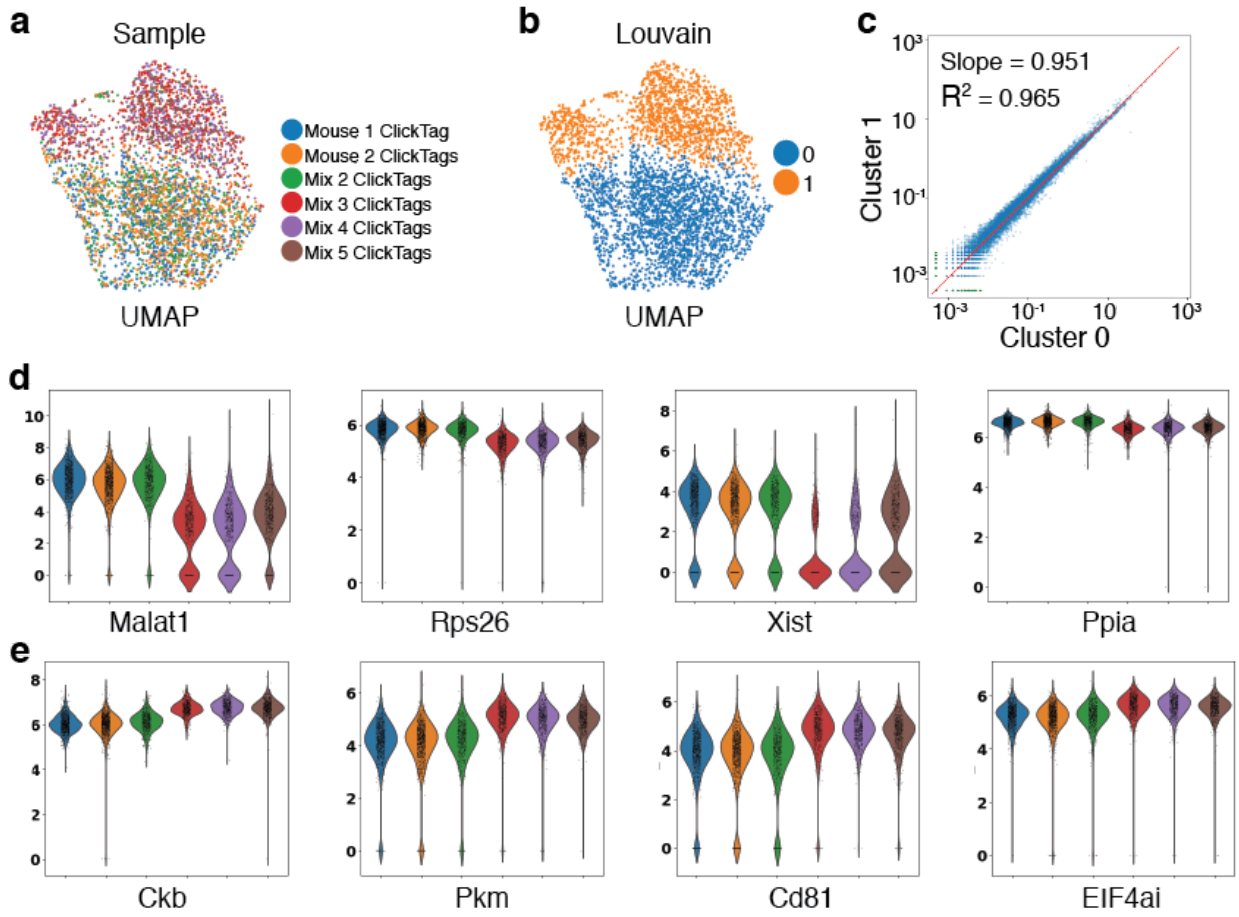
Supplementary Figure 8 Comparison of mouse cDNA libraries across samples from species-mixing experiment. (**a**) Violin plot of UMIs per cell for all n=6,087 captured mouse cells identified by

CellRanger and grouped according to sample identification shown in **Supplementary Figure 12a**. (**b**) Violin plot of genes per cell for all n=4,048 captured mouse cells identified by CellRanger, grouped according to sample identification shown in **Supplementary Figure 12a**. (**c**) Pearson correlation of gene expression from all samples containing mouse cells. Gene expression is shown as average counts per cell for each of n=27,998 mouse genes. The number of mouse cells captured for each sample is shown also shown.



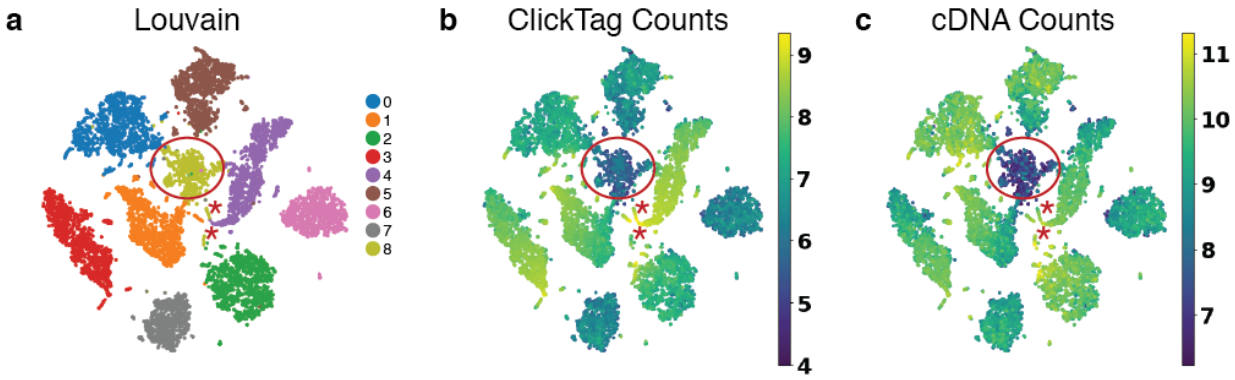
Supplementary Figure 9 Effect of ClickTag concentration on human gene expression quantification. (**a**) Embedding of cDNA from all n=5,880 "human" singlet cells showing mixing of cells from samples with

one/two ClickTags or three/four/five ClickTags, but separation of the two groups. (**b**) Clustering of the cells shown in panel **a**. (**c**) Pearson orrelation of gene expression for clusters shown in panel **b**, with each gene shown as the average counts per cell in each group (n=32,738 human genes). (**d**) Top differentially expressed genes for cluster 0, n=3,445 cells labeled with 1/2 ClickTags, colored as in panel **a**. (**e**) Top differentially expressed genes for cluster 1, n=2,364 cells labeled with three/four/five ClickTags, colored as in panel **a** showing gene expression across experimental sampels.

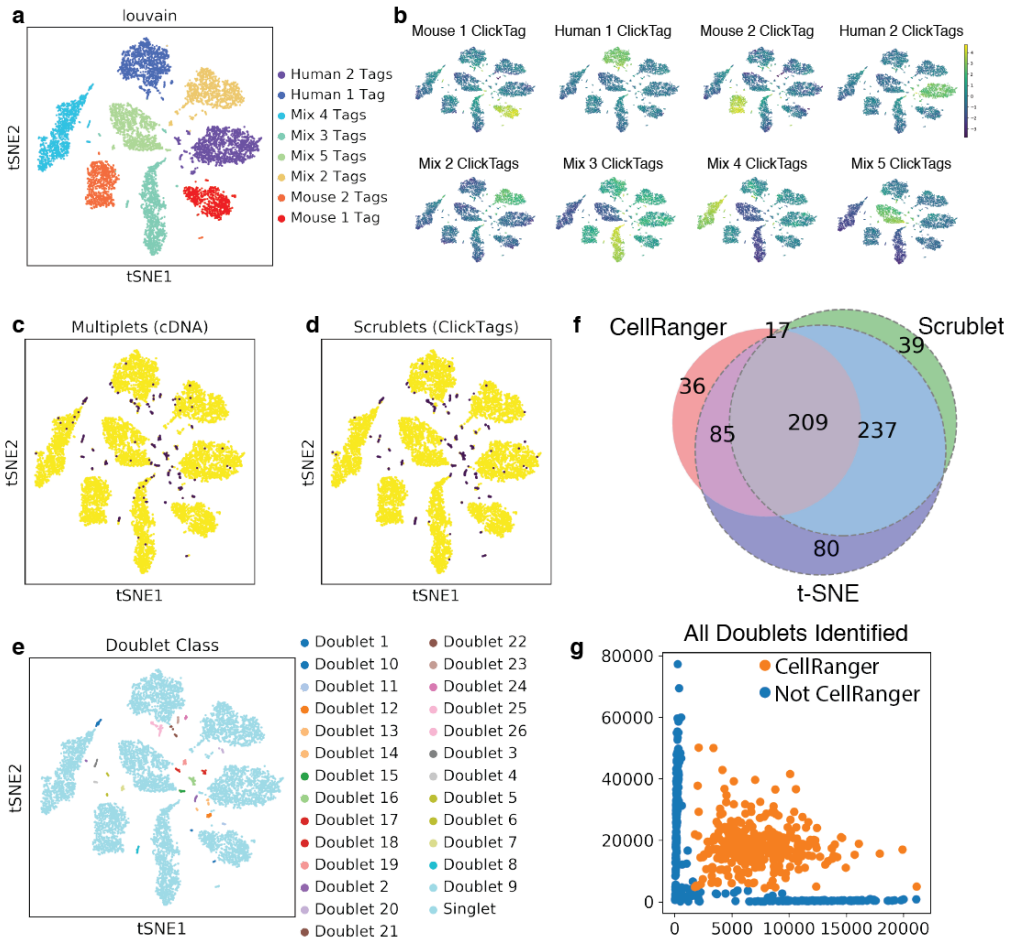


Supplementary Figure 10 Effect of ClickTag concentration on mouse gene expression quantification. (**a**) Embedding of cDNA from all n=3,938 “mouse” singlet cells showing mixing of cells from samples with one/two ClickTags or three/four/five ClickTags, but separation of the two groups. (**b**) Clustering of the

cells shown in panel **a**. (**c**) Pearson correlation of gene expression for clusters shown in panel **b**, with each gene shown as the average counts per cell in each group (n=27,998 mouse genes). (**d**) Top differentially expressed genes for cluster 0, n=2,406 cells labeled with one/two ClickTags, colored as in panel **a** showing gene expression across experimental samples. (**e**) Top differentially expressed genes for cluster 1, n=1,532 cells labeled with three/four/five ClickTags, colored as in panel **a** showing gene expression across experimental sampels.

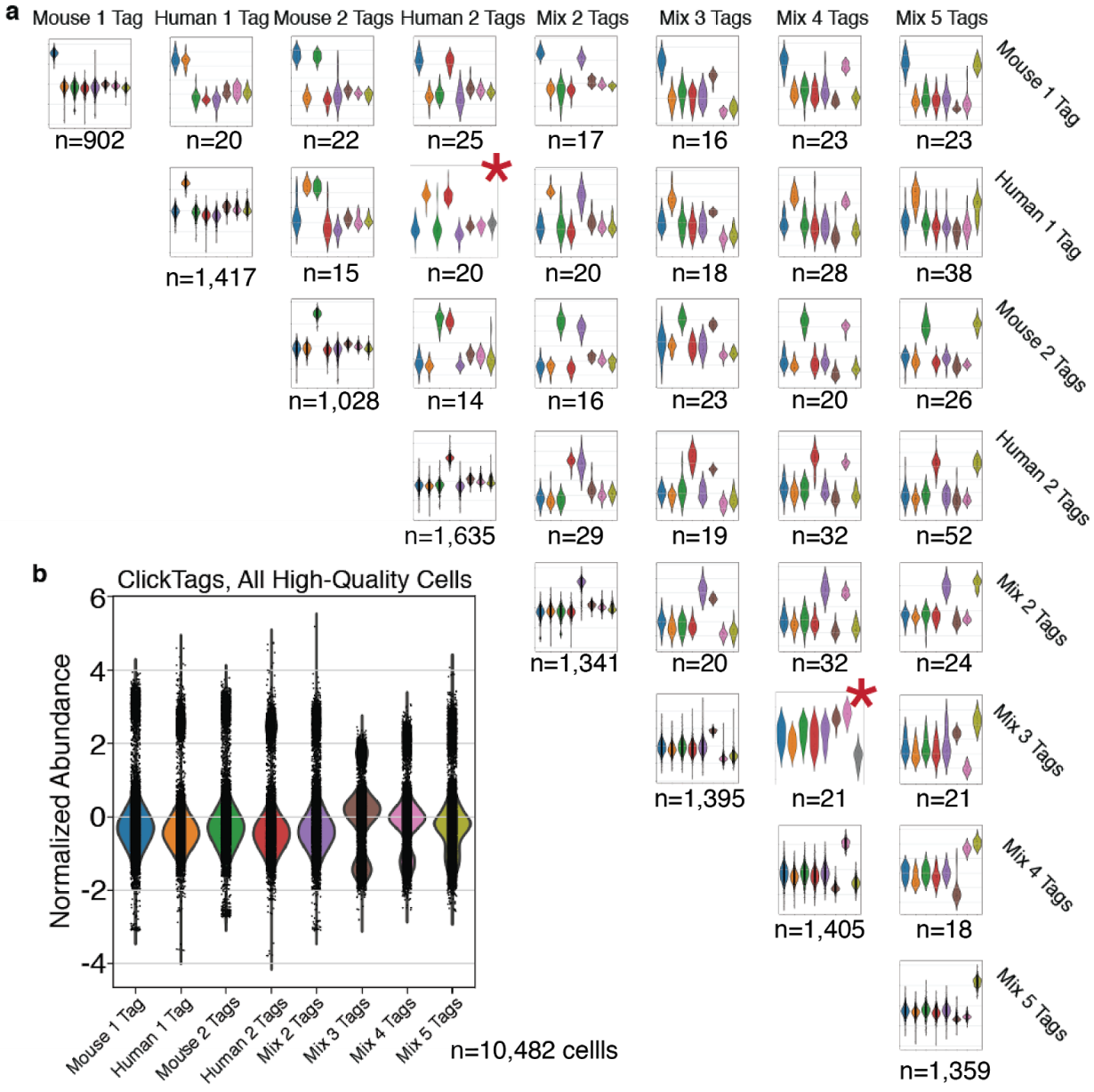


Supplementary Figure 11 Filtering for high-quality cells in the species-mixing experiment. ClickTag counts from all cells passing the Cell Ranger UMI filter were normalized, log-transformed, and embedded by t-SNE, generating nine distinct clusters. The t-SNE plots above display Louvain community detection (**a**), ClickTag counts (**b**), and cDNA counts (**c**) for n=11,264 cells. Cluster 8, circled, was found to have reduced UMI counts for both ClickTags and cDNAs and was removed from downstream analysis. Two sub-clusters (*) grouped with Cluster 8 were later found to correspond to two classes of inter-sample doublets and are similarly labeled in **Supplementary Figure 13** but were not included in the doublet detection comparison shown in **Supplementary Figure 12**.

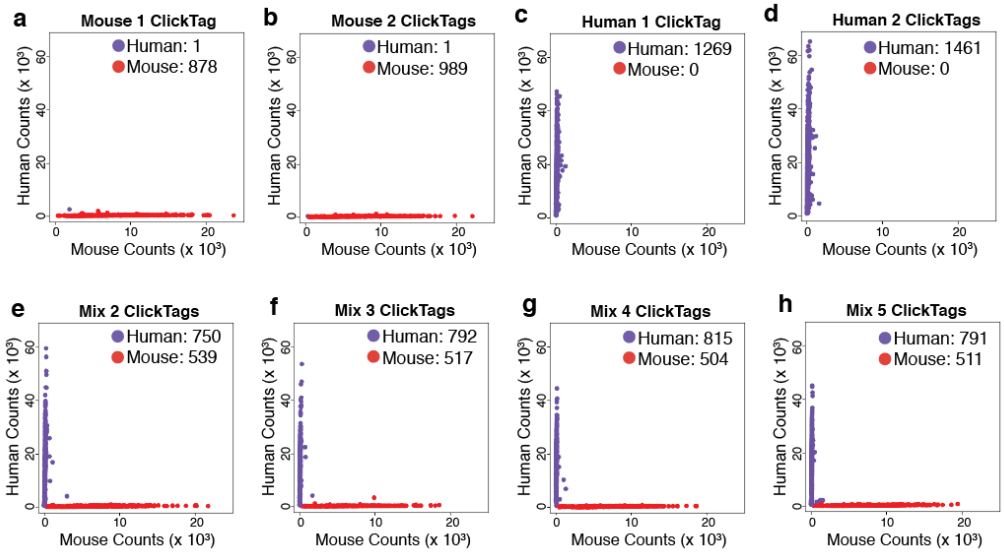


Supplementary Figure 12 The combination of ClickTags and species information presented a unique opportunity for comparison of doublet detection methods. Doublets identified by CellRanger are necessarily limited to cross-species events, while ClickTags are similarly only relevant for detection of doublets originating from different samples. (**a**) Clustering and embedding of ClickTag data after filtering low-quality cells (n=11,264). (**b**) ClickTag embedding colored according to summed, normalized, and log-transformed ClickTag counts from each sample. (**c**) Doublets as detected by Cell Ranger or Scrublet are found to predominantly label the same small sub-clusters on the ClickTag embedding. (**e**) Suspected cell doublet sub-clusters were manually selected from t-SNE embedding using FlowJo cytometry analysis software, identifying 26 sub-clusters that appeared to arise from inter-sample doublets. (**f**) Venn diagram showing agreement between all three doublet identification methods. (**g**) Human/mouse cDNA barnyard

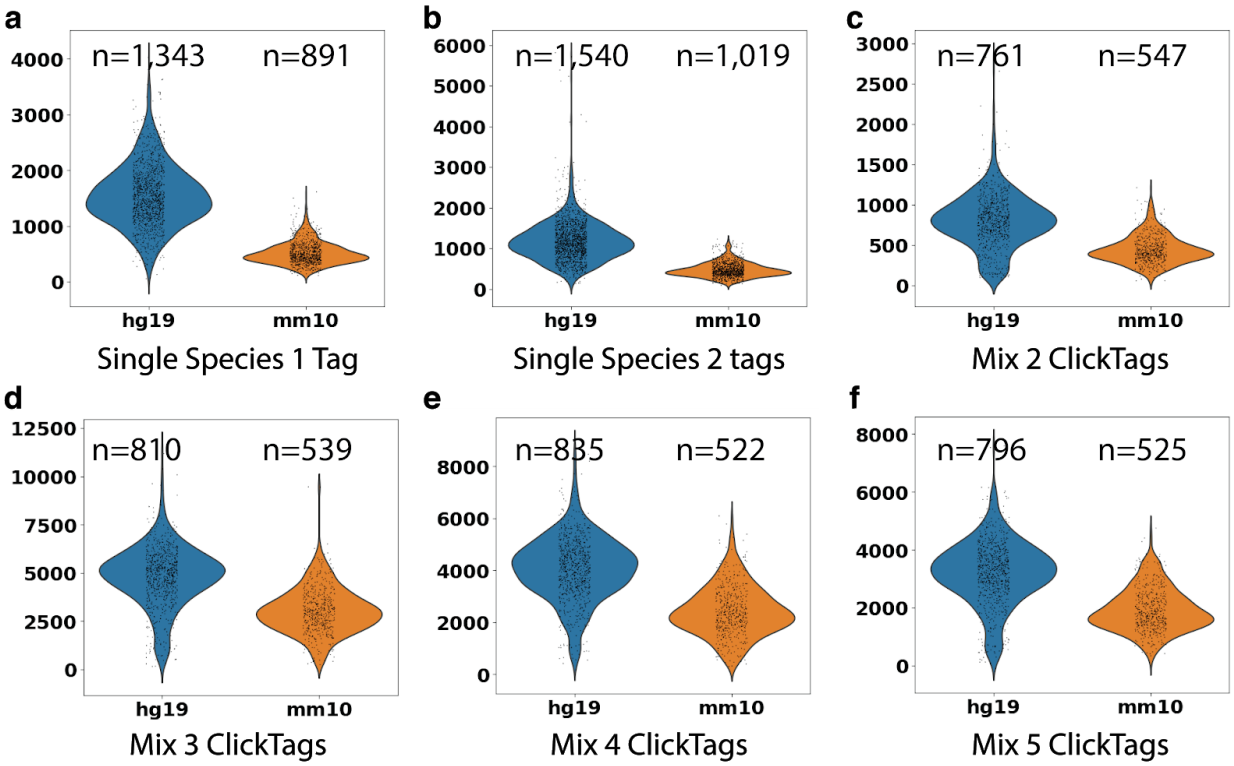
plot for all high-quality cells, colored according to Cell Ranger detection. Cell Ranger can only identify doublets between cells of different species, while ClickTag data can identify doublets between experimental samples regardless of species identity.



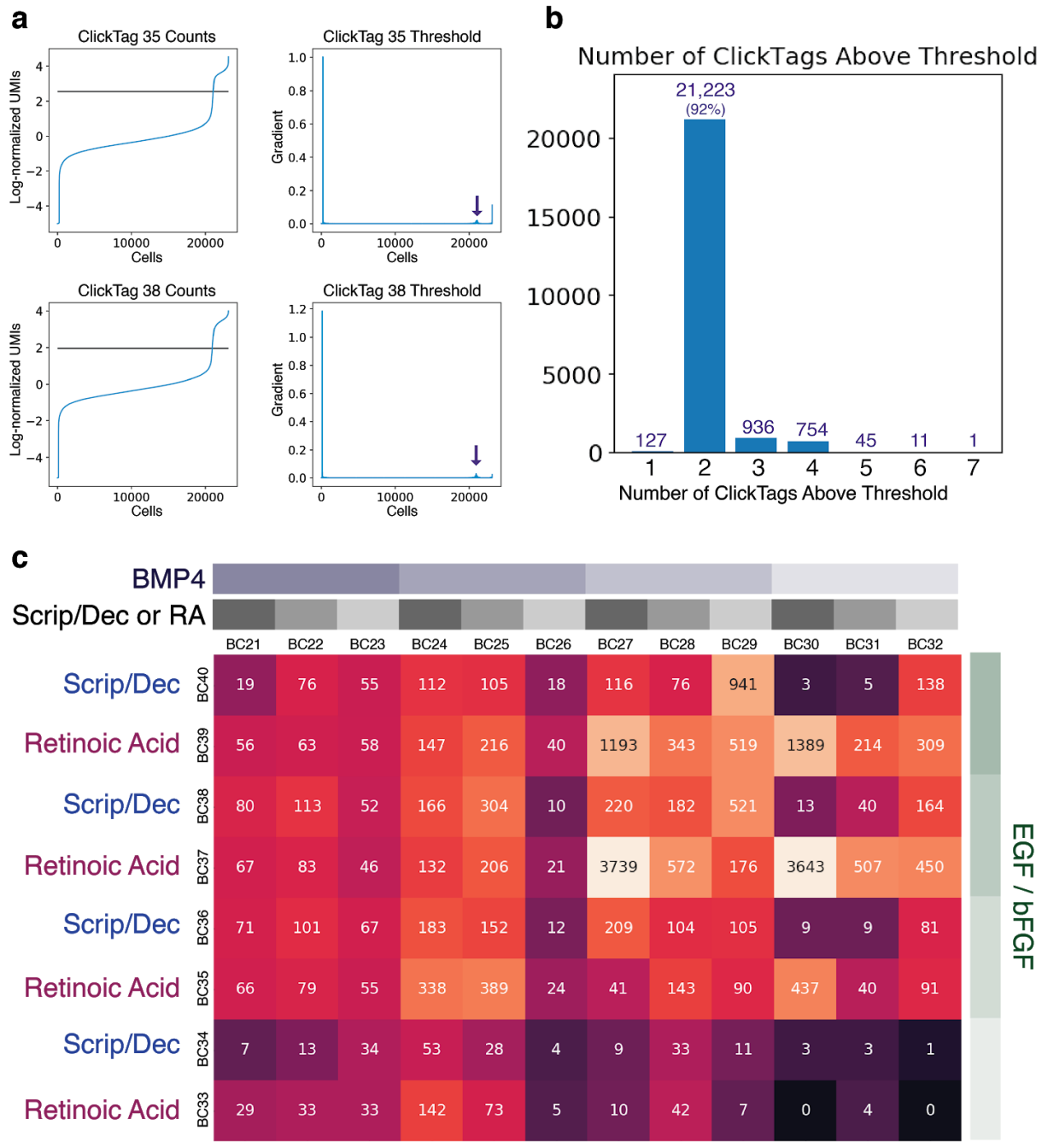
Supplementary Figure 13 Doublet classification for species-mixing experiment. (**a**) Violin plots generated from doublet sub-clusters manually isolated from a t-SNE embedding of ClickTag data shown in **Supplementary Figure 12e**. For comparison, ClickTag counts from single cells in each sample type are shown on the diagonal. Each of the 26 sub-clusters could be assigned to a specific doublet event between two well-defined samples, with the two remaining inter-sample doublet types (*) found to be filtered out due to low counts during the quality control step described in **Supplementary Fig. 11**. (**b**) Violin plots showing log-normalized ClickTag counts for all n=11,264 human and mouse cells across all 8 multiplexed samples. In each sample, a distinct group of positively labeled cells can be distinguished from negative cells originating from other samples.



Supplementary Figure 14 Barnyard plots for droplets identified as singlets following the manual selection procedure described in the legend of **Supplementary Figure 12**.

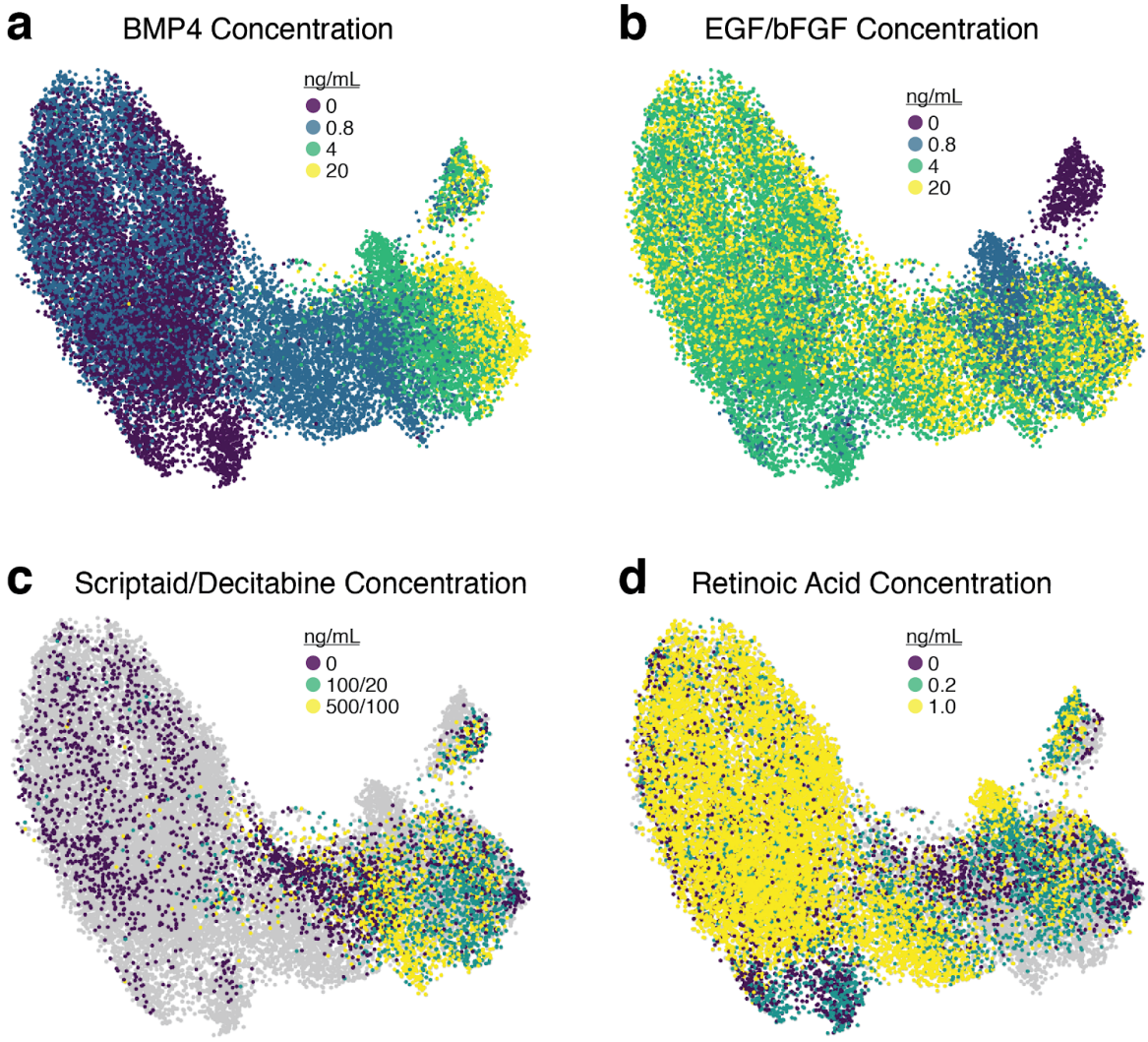


Supplementary Figure 15 Analysis of ClickTag counts from human HEK293T and mouse neural stem cells from the multiplexed species-mixing experiment. Human cells consistently yield more ClickTags than mouse cells from the same or similarly treated samples, consistent with the RNA yield as shown in **Supplementary Figure 14**.

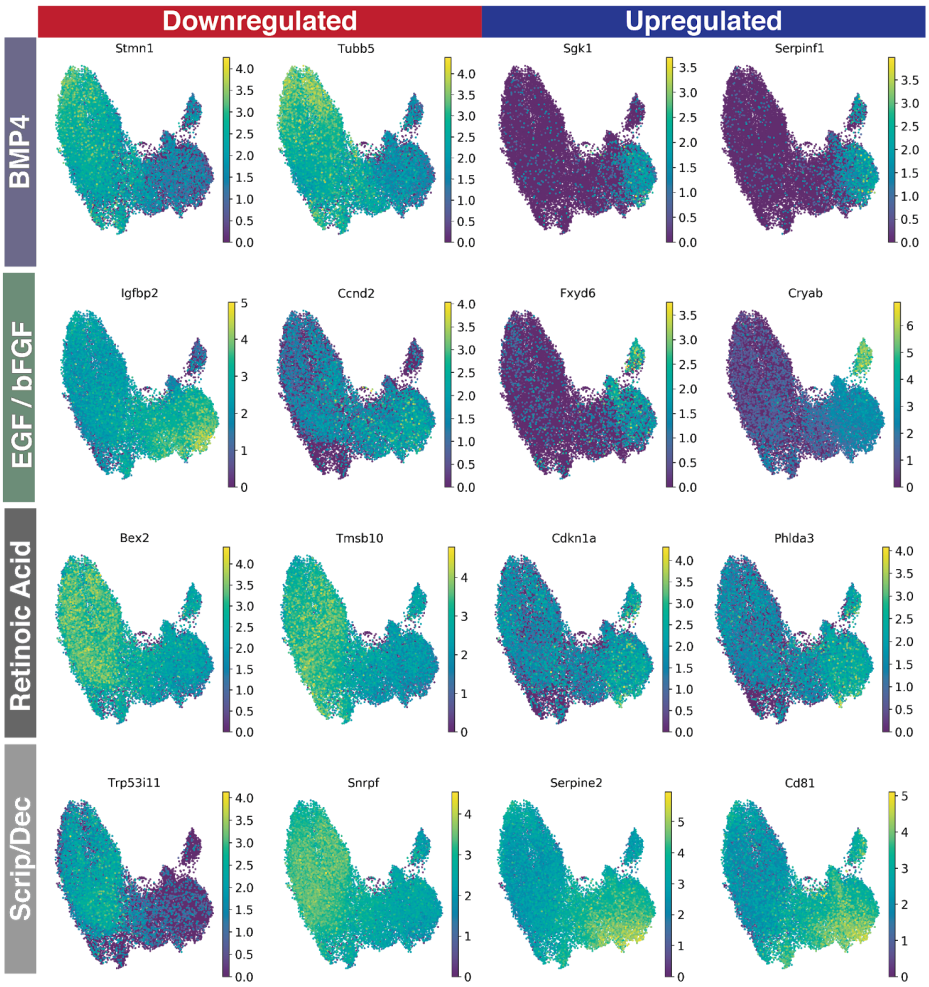


Supplementary Figure 16 Sample assignment for the 96-sample perturbation experiment. Thresholds were set using the maximum slope of the rank-UMI plot for each ClickTag across all cells as determined by the numpy gradient function. 21,223 cells (92%) were assigned to exactly two ClickTags (**b**), with 99.8% of those corresponding to a valid barcode combination from the experimental design. Only these

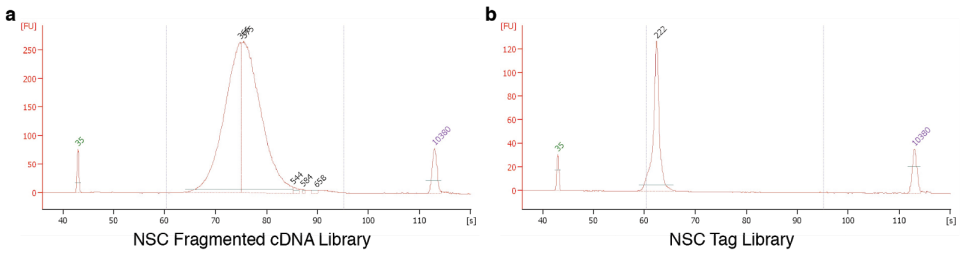
cells were used for downstream analysis. Distribution of cells recovered across the experimental conditions are shown in (**c**).



Supplementary Figure 17 UMAP embedding showing n=21,223 cells from the 96-sample perturbation experiment colored according to the experimental treatment for each cell. Global trends such as EGF/bFGF dependence, BMP4 response, and retinoic acid-driven proliferation are evident.



Supplementary Figure 18 A linear regression model was used to identify genes associated with individual perturbants. For each chemical, examples of upregulated and downregulated genes are shown as gene expression profiles across n=21,223 total cells from the 96-sample NSC perturbation experiment.



Supplementary Figure 19 BioAnalyzer traces for (**a**) fragmented cDNA libraries and (**b**) ClickTag amplicon libraries. Traces shown are representative of library preparation repeated three times from methanol-fixed, ClickTag barcoded cells.

Sample Number	Species	Tag(s)
1	Mouse 1 tag	BC41
2	Human 1 tag	BC42
3	Mouse 2 tags	BC43, BC44
4	Human 2 tags	BC45, BC46
5	Mouse and Human Mix 2 tags	BC47, BC48
6	Mouse and Human Mix 3 tags	BC49, BC50, BC51
7	Mouse and Human Mix 4 tags	BC52, BC53, BC54, BC55
8	Mouse and Human Mix 5 tags	BC56, BC57, BC58, BC59, BC60

Supplementary Table 1 Attached file. ClickTags used in multiplexed species-mixing experiment.

Metric	Multiplexed*	Fixed	Live
Number of Cells	4,611	3,808	9,719
Number of Reads	200 M	255 M	222 M
Mean Reads per Cell	26,022	67,059	22,803
Reads Mapped to Transcriptome	79.0%	77.0%	66.7%
Fraction Reads in Cells	90.6%	92.8%	92.6%
Sequencing Saturation	60.1%	79.8%	42.0%
Median Genes per Cell	2,090	2,241	2,169

Supplementary Table 2 Run statistics for multiplexed, fixed, and live mouse neural stem cells as described in **Supplementary Fig. 6**. The multiplexed sample contained both mouse and human cells, so the multiplexed statistics correspond only to "mouse" cells as identified by CellRanger. Note that the "Fixed" sample was sequenced much more deeply than the "Multiplexed" and "Live" samples.

Supplementary Table 3 Attached file. Top 100 marker genes from each cluster for n=21,223 cells in 96-sample multiplexed experiment. Clusters used for differential expression were determined by Louvain community detection with resolution=2.2. Differentially expressed genes were identified with the Wilcoxon test and the ScanPy `rank_genes_groups` function.

Supplementary Table 4 Attached file. Primer sequences used in this study.