Methods S1. Workflows and detection rates. Related to STAR Methods and Figure 1.

Experimental workflow

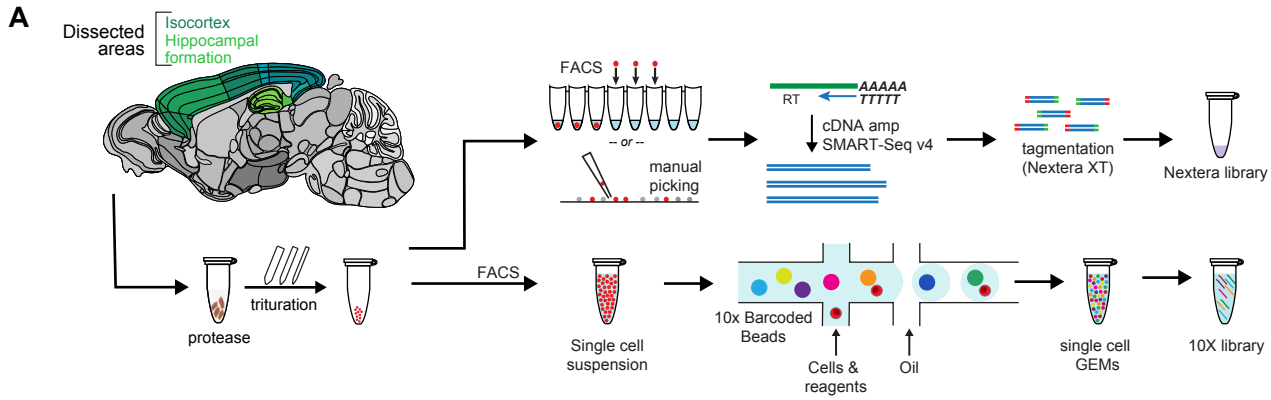
(A) Schematic diagram illustrating cell isolation process for single cell RNA-sequencing. Brain regions were dissected according to the Allen CCFv3. Each sample was digested and triturated to obtain single cell suspensions. For SMART-Seq v4 (SSv4) processing, individual cells were sorted into 8-well strip PCR tubes by FACS or by manual picking. Cells were lysed, and SSv4 was used to reverse-transcribe and amplify full-length cDNAs from each cell. cDNAs were then tagmented by Nextera XT, PCR-amplified, and processed for Illumina sequencing. For 10x processing, debris was removed from single cell suspension by FACS, suspensions were loaded on the 10x Genomics Chromium™ Controller to create single-cell libraries, and libraries were processed for Illumina sequencing.

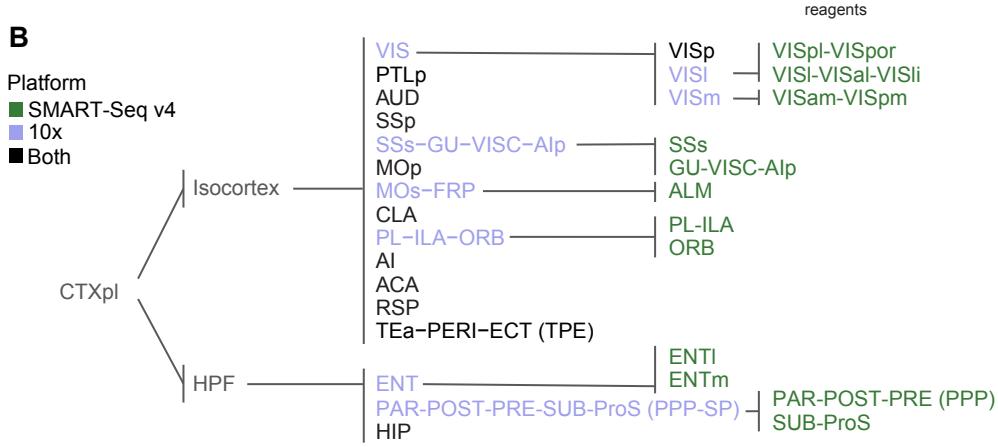
(B) Ontology of dissected brain regions according to CCFv3. See **Table S1** for each region's full name and sampling.

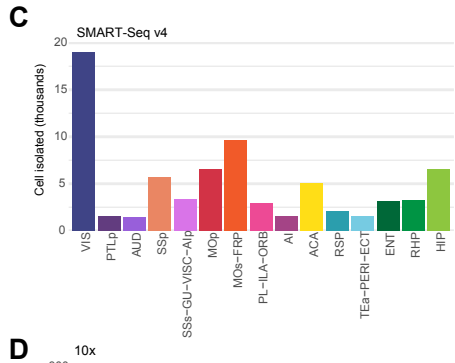
(C-D) The number of cells sampled from each dissection region for SSv4 (**C**) or 10xv2 (**D**). Bars are colored by region.

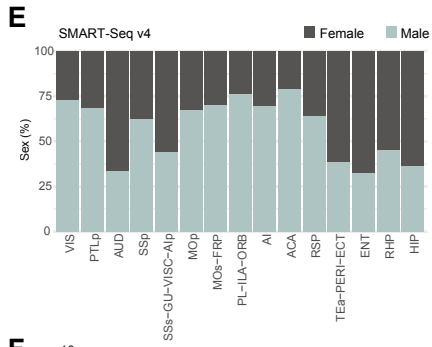
(E-F) Sex sampling proportion per joint region for SSv4 (**E**) or 10xv2 (**F**).

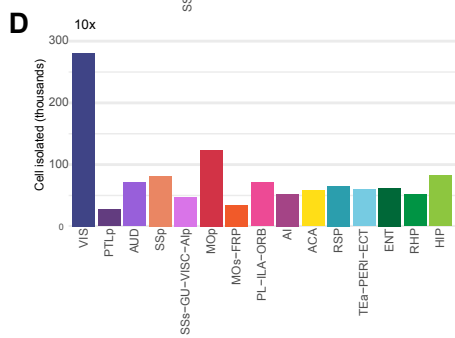
(G-H) Genotype sampling proportion per joint region for SSv4 (**G**) or 10xv2 (**H**), colored by region. Columns add up to 100%.

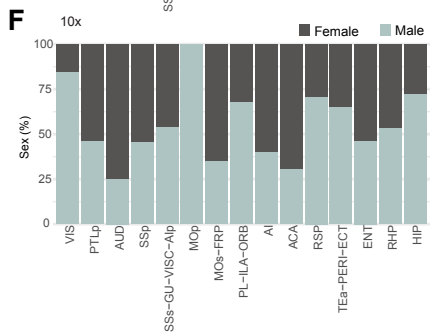

A
Dissected areas
Isocortex
Hippocampal formation
FACS
-- or --
manual picking
RT
AAAAA
TTTTT
cDNA amp
SMART-Seq v4
tagmentation
(Nextera XT)
Nextera library
protease
trituration
FACS
Single cell suspension
10x Barcoded Beads
Cells & reagents
Oil
single cell GEMs
10X library

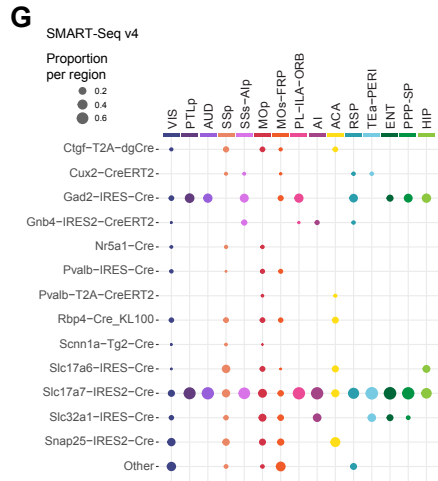

B
Platform
SMART-Seq v4
10x
Both
CTXpl
Isocortex
HPF
VIS
PTLp
AUD
SSp
SSs-GU-VISC-AIp
MOp
MOs-FRP
CLA
PL-ILA-ORB
AI
ACA
RSP
TEa-PERI-ECT (TPE)
ENT
PAR-POST-PRE-SUB-ProS (PPP-SP)
HIP
VISp
VISl
VISm
VISpl-VISpor
VISl-VISal-VISli
VISam-VISpm
SSs
GU-VISC-AIp
ALM
PL-ILA
ORB
ENTl
ENTm
PAR-POST-PRE (PPP)
SUB-ProS

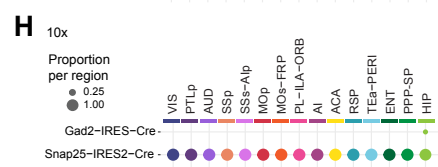

C
SMART-Seq v4
Cell isolated (thousands)


E
SMART-Seq v4
Female
Male
Sex (%)


D
10x
Cell isolated (thousands)

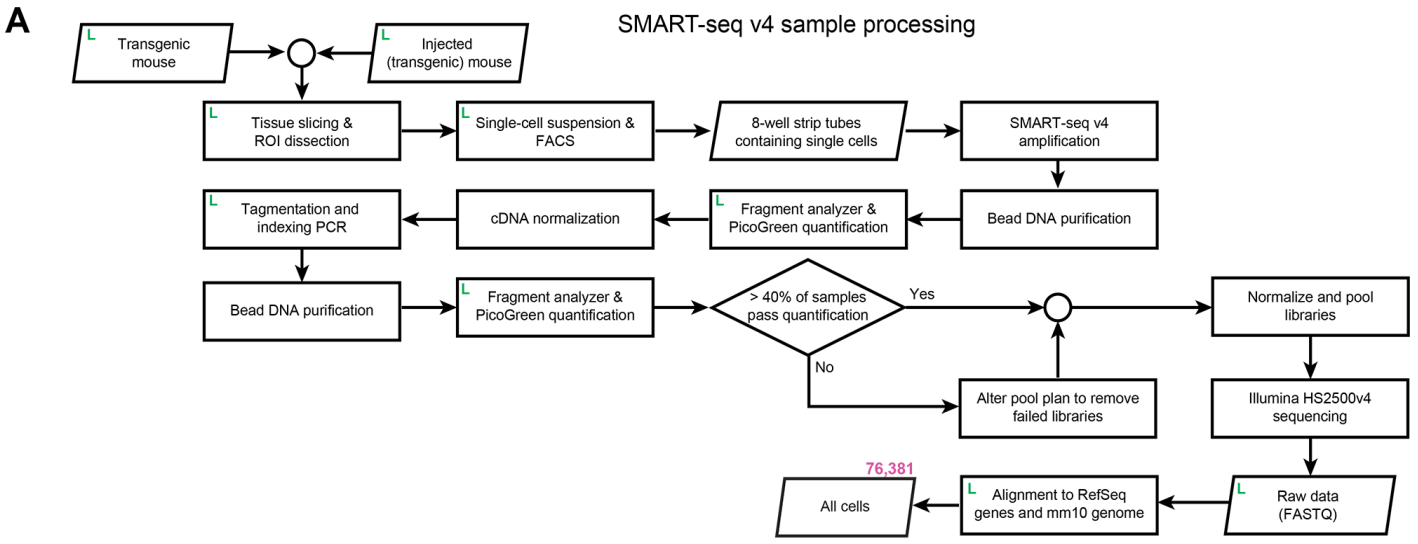

F
10x
Female
Male
Sex (%)

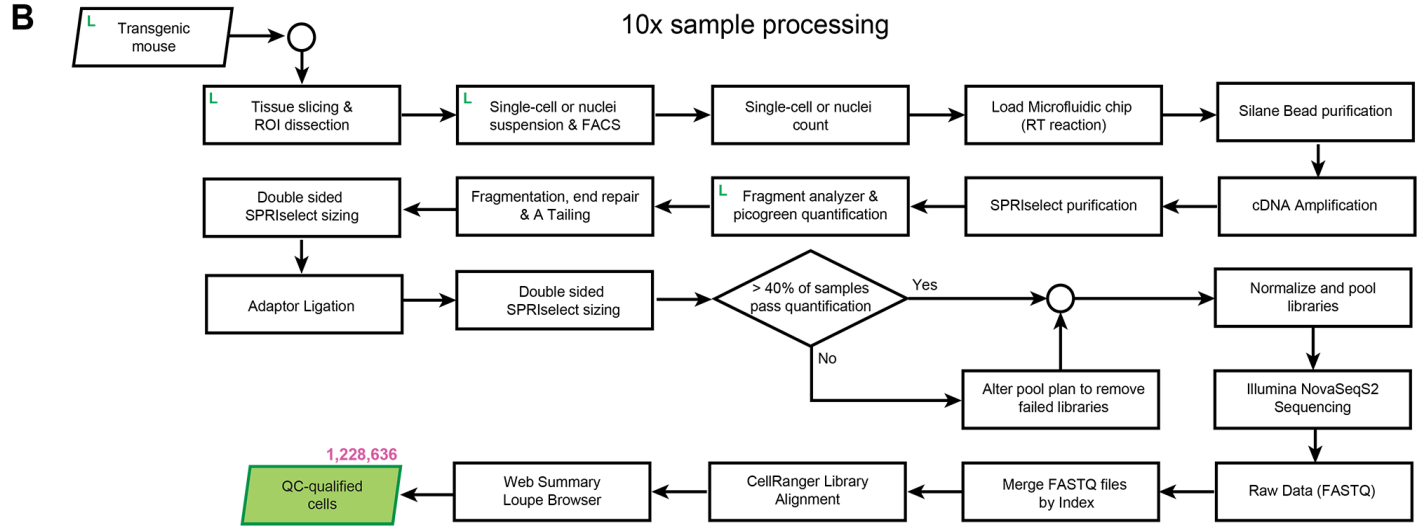

G
SMART-Seq v4
Proportion per region
Ctgf-T2A-dgCre
Cux2-CreERT2
Gad2-IRES-Cre
Gnb4-IRES2-CreERT2
Nr5a1-Cre
Pvalb-IRES-Cre
Pvalb-T2A-CreERT2
Rbp4-Cre_KL100
Scnn1a-Tg2-Cre
Slc17a6-IRES-Cre
Slc17a7-IRES2-Cre
Slc32a1-IRES-Cre
Snap25-IRES2-Cre
Other


H
10x
Proportion per region
Gad2-IRES-Cre
Snap25-IRES2-Cre

Data workflow

(A-B) scRNA-seq pipeline and data preprocessing workflow outlining the path from individual experimental animals to quality control (QC)-qualified scRNA-seq data for SSv4 (**A**) or 10xv2 (**B**). At multiple points throughout sample processing, cell and sample metadata are recorded in a laboratory information management system (LIMS, labeled as L), which informs QC processes. Samples must pass QC benchmarks to continue through sample processing.



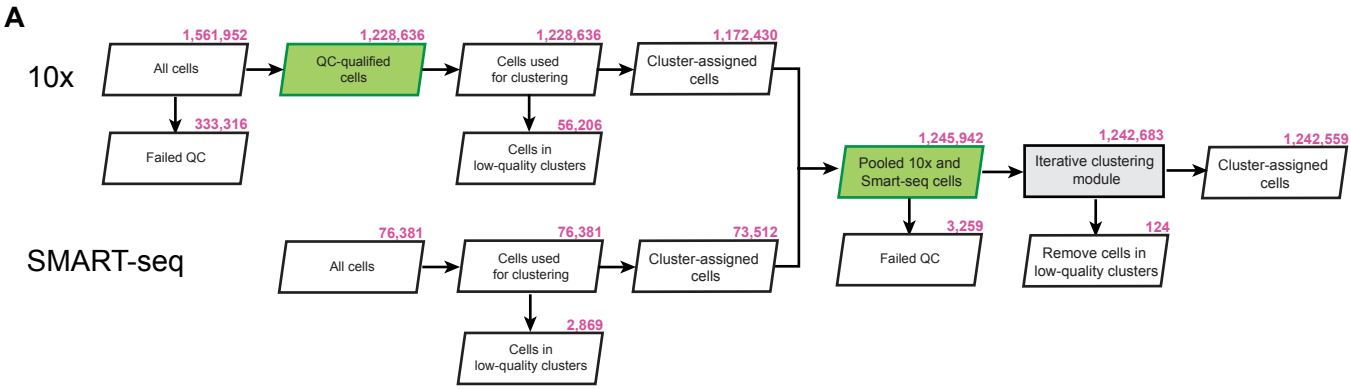


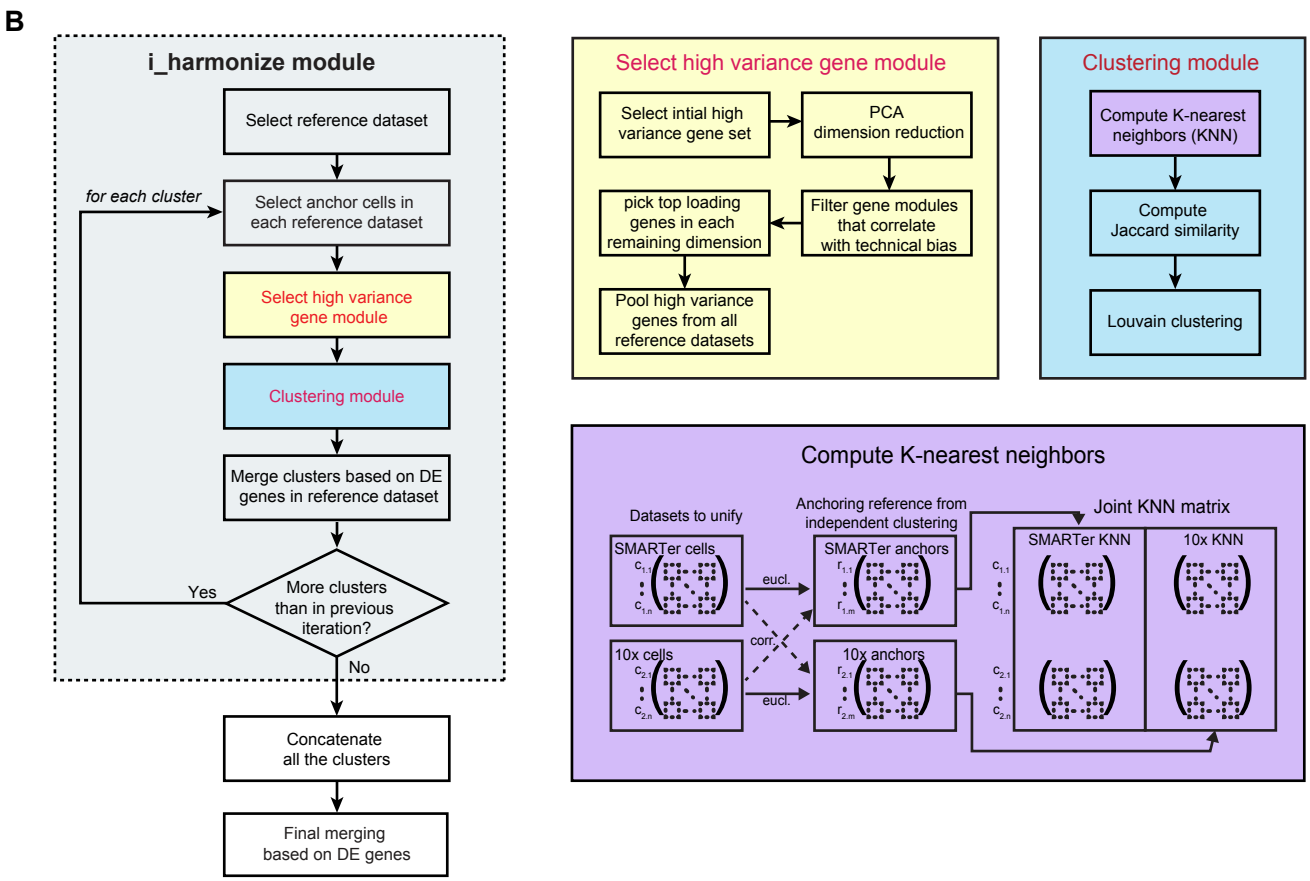
Analysis workflow

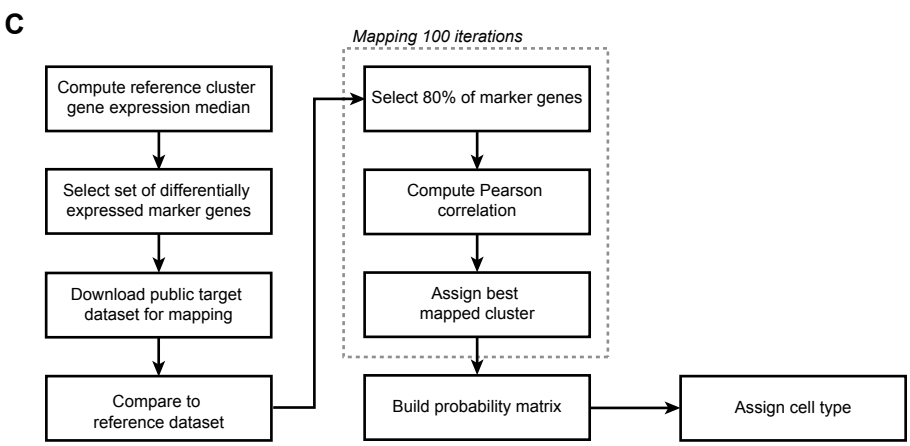
(A) The number of cells at each step in the scRNA-seq data analysis pipeline. The identification of doublets and low-quality clusters is described in more detail in **Methods**. The 10xv2 and SSv4 data was first QC-ed and analyzed separately. After initial clustering the datasets were combined and QC-ed again before and after joint clustering.

(B) Joint clustering procedure using the new i_harmonize function from the scrattch.hicat package. For this study, as the 10xv2 dataset includes more cells while the SSv4 dataset provides more sensitive gene detection, both datasets were used as reference datasets. For each reference dataset anchor cells were selected to achieve uniform coverage of all cell types. Based on these anchor cells high variance genes were selected (select high variance gene module, yellow box), and high variance genes from each reference dataset were pooled. Next, a common adjacency graph using all cells from all datasets was built (purple box) and the standard Jaccard-Louvain clustering algorithm was applied (clustering module, blue box). Resulting clusters were merged to ensure that all pairs of clusters, even at the finest level, were separable by conserved differentially expressed genes across platforms. This i_harmonize function applies the integrative clustering across datasets iteratively, while ensuring that all clusters at each iteration are separable by conserved differentially expressed genes.

(C) To assess correspondence of cell types identified in this study to previously published datasets, cells from published datasets were mapped to our clusters using the nearest centroid classifier based on a set of shared markers that were detected in both datasets (expression > 0). To estimate the robustness of mapping, classification was repeated 100 times, each time using 80% of randomly sampled markers, and the probability for each cell to map to every reference cluster was computed.


A
10x
1,561,952
All cells
1,228,636
QC-qualified cells
1,228,636
Cells used for clustering
1,172,430
Cluster-assigned cells
333,316
Failed QC
56,206
Cells in low-quality clusters
SMART-seq
76,381
All cells
76,381
Cells used for clustering
73,512
Cluster-assigned cells
2,869
Cells in low-quality clusters
1,245,942
Pooled 10x and Smart-seq cells
3,259
Failed QC
1,242,683
Iterative clustering module
124
Remove cells in low-quality clusters
1,242,559
Cluster-assigned cells


B
i_harmonize module
Select reference dataset
for each cluster
Select anchor cells in each reference dataset
Select high variance gene module
Clustering module
Merge clusters based on DE genes in reference dataset
Yes
More clusters than in previous iteration?
No
Concatenate all the clusters
Final merging based on DE genes
Select high variance gene module
Select intial high variance gene set
PCA dimension reduction
pick top loading genes in each remaining dimension
Filter gene modules that correlate with technical bias
Pool high variance genes from all reference datasets
Clustering module
Compute K-nearest neighbors (KNN)
Compute Jaccard similarity
Louvain clustering
Compute K-nearest neighbors
Datasets to unify
Anchoring reference from independent clustering
Joint KNN matrix
SMARTer cells
SMARTer anchors
SMARTer KNN
10x KNN
eucl.
corr.
10x cells
10x anchors
eucl.


C
Mapping 100 iterations
Compute reference cluster gene expression median
Select set of differentially expressed marker genes
Download public target dataset for mapping
Compare to reference dataset
Select 80% of marker genes
Compute Pearson correlation
Assign best mapped cluster
Build probability matrix
Assign cell type

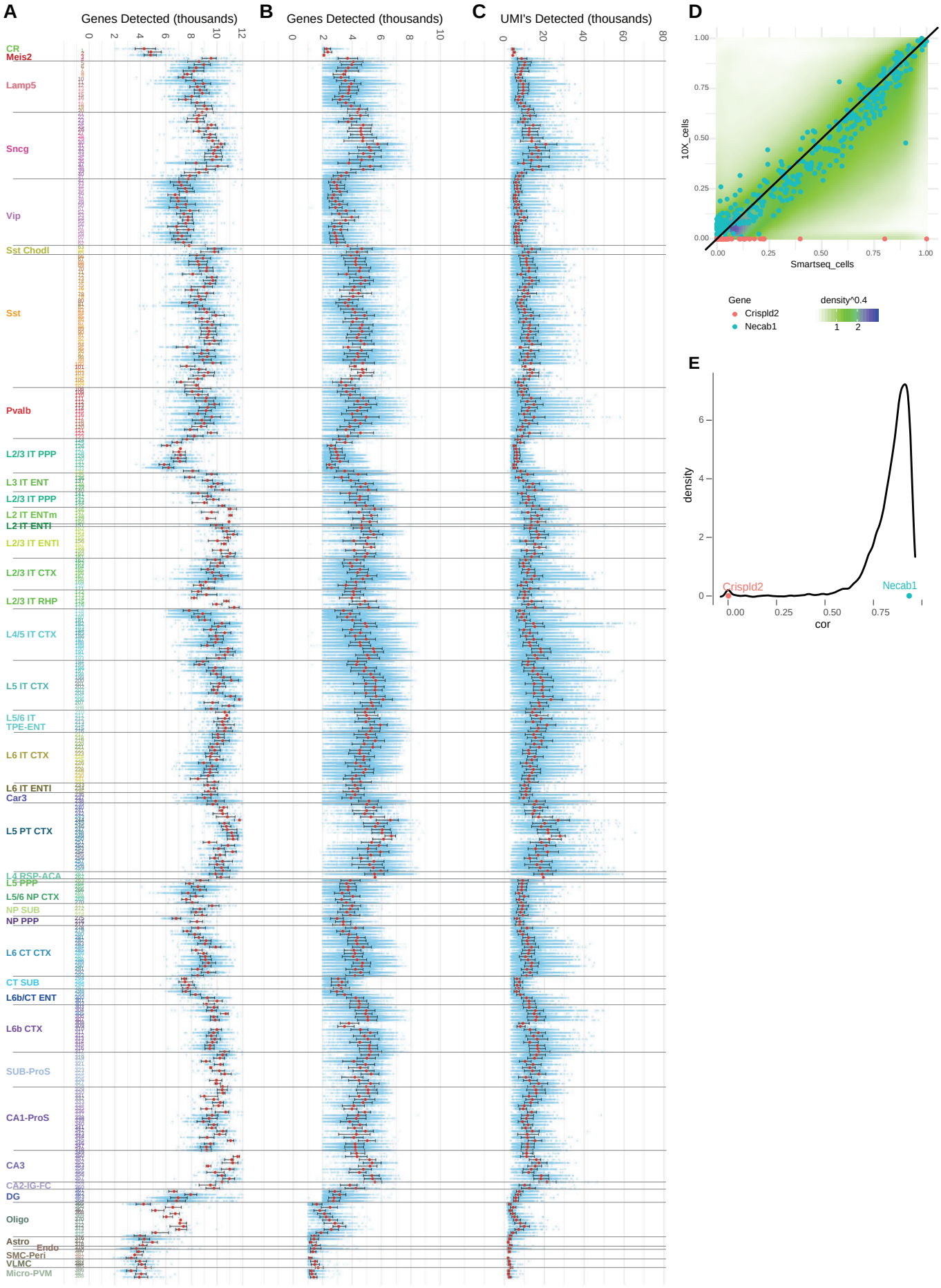
Detection rates

(A-B) Number of genes detected per SSv4 (**A**) or 10xv2 (**B**) cell for each cluster.

(C) Number of UMI's detected per 10xv2 cell for each cluster. The average number of UMIs detected per cell in 10xv2 data was 10,576. The numbers of genes or UMIs detected were largely consistent within each class and subclass of cells; non-neuronal cells and the CR and Meis2 neurons had substantially lower numbers of genes detected than other neurons.

(D) Comparison of the relative expression level of marker genes across all clusters between the SSv4 and 10xv2 datasets. Since the two datasets differ in experimental platform, gene expression quantification software and gene annotation reference, for each gene, we normalized the average log2(CPM+1) values at the cluster level in the range [0,1] by subtracting the minimum value and then dividing them by the maximum value for that gene. The smooth scatter plot represents normalized gene expression for all marker genes across all clusters in the two datasets. The areas with the highest density of data points are colored blue, and the lowest density white. Two example genes are shown, each dot representing the average expression level for each gene in each cluster.

(E) Distribution of gene expression conservation between the two platforms. For each of 5,981 marker genes, we computed the correlation of its average expression across all overlapping cell types between 10xv2 and SSv4, and distribution of such correlation values ('cor') is shown in the density plot. Two example genes shown in panel **D** are highlighted: *Necab1* represents a gene with high correspondence between the two datasets, while *Crispld2* is detectable in SSv4 cells but not in 10xv2 cells.


A
Genes Detected (thousands)
B
Genes Detected (thousands)
C
UMI's Detected (thousands)
CR
Meis2
Lamp5
Sncg
Vip
Sst Chodl
Sst
Pvalb
L2/3 IT PPP
L3 IT ENT
L2/3 IT PPP
L2 IT ENTm
L2 IT ENTl
L2/3 IT ENTl
L2/3 IT CTX
L2/3 IT RHP
L4/5 IT CTX
L5 IT CTX
L5/6 IT TPE-ENT
L6 IT CTX
L6 IT ENTl
Car3
L5 PT CTX
L4 RSP-ACA
L5 PPP
L5/6 NP CTX
NP SUB
NP PPP
L6 CT CTX
CT SUB
L6b/CT ENT
L6b CTX
SUB-ProS
CA1-ProS
CA3
CA2-IG-FC
DG
Oligo
Astro
Endo
SMC-Peri
VLMC
Micro-PVM
D
10X_cells
Smartseq_cells
Gene
Crispld2
Necab1
density^0.4
E
density
Crispld2
Necab1
cor