Neuron, Volume *92*

Supplemental Information

In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus

Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai

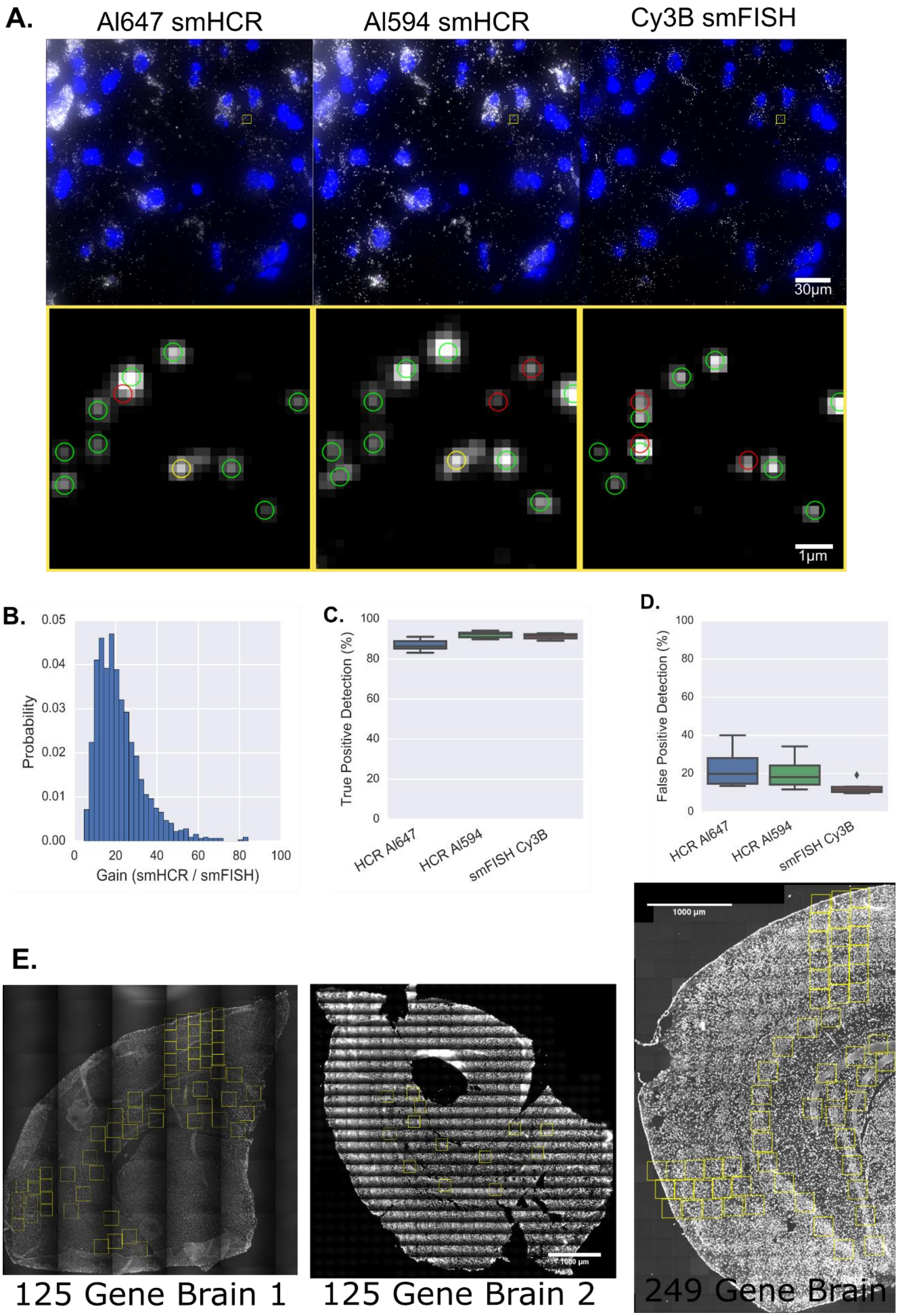
A.
Al647 smHCR
Al594 smHCR
Cy3B smFISH
30μm
1μm
B.
Probability
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
Gain (smHCR / smFISH)
C.
True Positive Detection (%)
100
80
60
40
20
0
HCR Al647
HCR Al594
smFISH Cy3B
D.
False Positive Detection (%)
100
80
60
40
20
0
HCR Al647
HCR Al594
smFISH Cy3B
E.
1000 μm
1000 μm
125 Gene Brain 1
125 Gene Brain 2
249 Gene Brain

Fig S1. SmHCR performance metrics as compared to smFISH, Related to Figure 1. A. Raw data of Pgk1 transcripts imaged in a brain slice. The transcript was targeted with 2 hcr probes sets and 1 smFISH probe set, each consisted of 24 oligonucleotide probes. The probe sets were hybridized together and were imaged in 3 different channels. Green circles are transcripts detected in all channels, yellow circles signify transcripts detected in 2 out of 3 channels, and red circles represent signal found in only 1 channel (false positives due to nonspecific binding). These images show that smHCR and smFISH have similar sensitivity, specificity, and spot size. **B.** Gain of smHCR vs smFISH. The mean gain of smHCR is 22.1 ± 11.55 vs smFISH (n=1338). **C.** True positive detection rate of smHCR and smFISH per channel. The percent of true positives (transcripts detected with at least 2 out of 3 probe sets) detected with each probe set (n=1338). **D.** False positive rate of smHCR and smFISH. Percent of total dots in a channel not detected in any other channel for 3 color Pgk1 (n=1338). **E.** All the regions imaged in the coronal section are boxed. Each box represents a field of 216 um x 216 um. The brain section used for figure 4 and 5 is shown on the left. The middle section is used for figure 6 and the right section is used for figure 7.

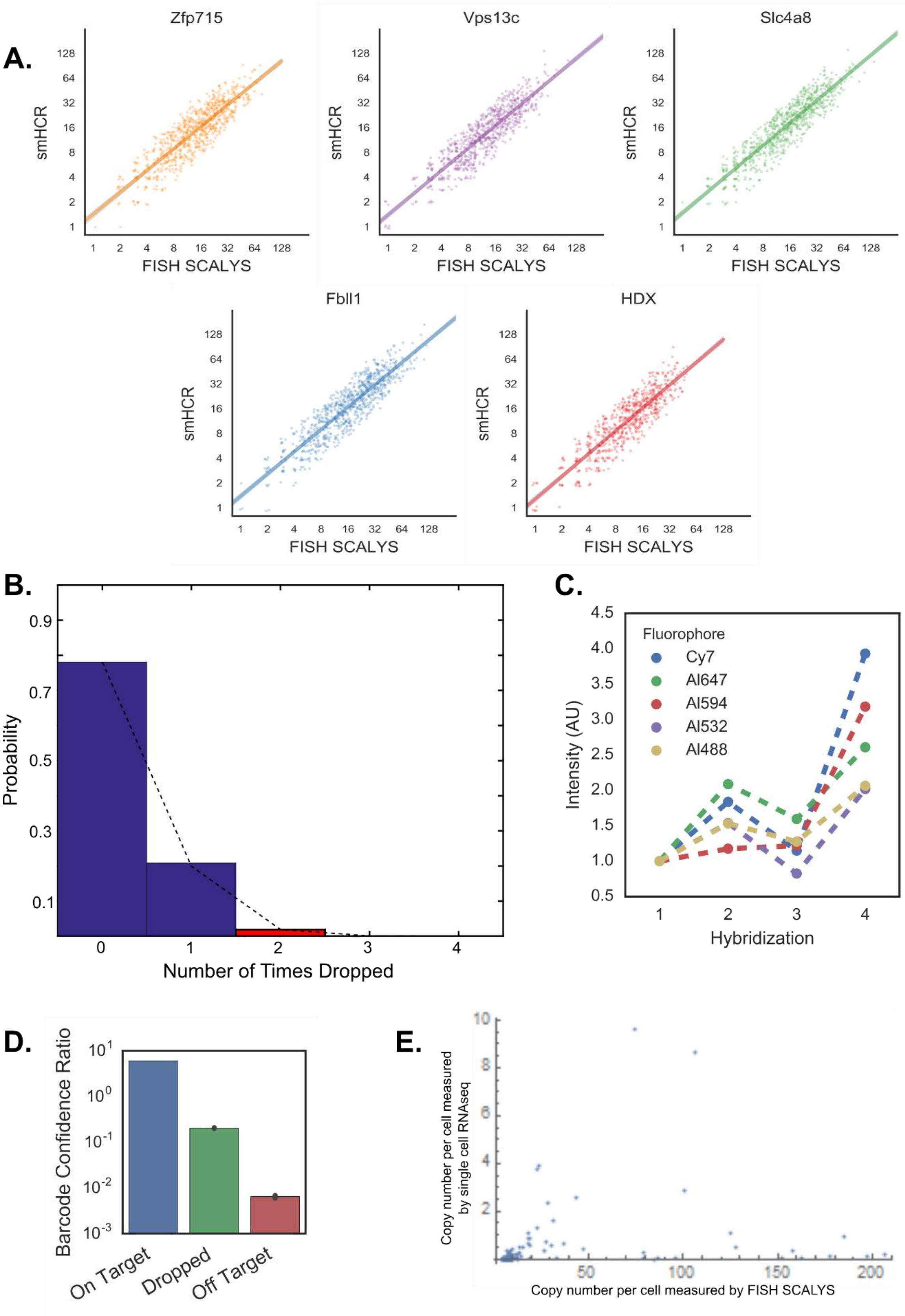
A.
Zfp715
Vps13c
Slc4a8
Fbll1
HDX
smHCR
FISH SCALYS
1
2
4
8
16
32
64
128
B.
Probability
0.1
0.3
0.5
0.7
0.9
0
1
2
3
4
Number of Times Dropped
C.
Fluorophore
Cy7
Al647
Al594
Al532
Al488
Intensity (AU)
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
1
2
3
4
Hybridization
D.
Barcode Confidence Ratio
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
On Target
Dropped
Off Target
E.
Copy number per cell measured by single cell RNAseq
2
4
6
8
10
50
100
150
200
Copy number per cell measured by FISH SCALYS

Fig S2. Quantitation of seqFISH, Related to Figure 2. A. All control genes show high correlations between seqFISH and smHCR. **B.** Number of dropped hybridizations from the barcode. Blue bars represent measured probability and the red bars represent inferred values from binomial distribution fitting of measured probability. The ratio of the full barcodes (4 hybridizations) vs 3 hybridization barcodes indicate that transcripts that are mis-hybridized in 2 rounds are rare. Transcripts missed in 2 or more hybridizations (red bars) could not be recovered from the error-correction algorithm and would be dropped from our quantifications (N=2,115,477 total barcodes). **C.** Intensity of barcode hybridizations overtime. All dots belonging to barcodes are quantified in each hybridization and their mean intensity is plotted over time normalized to the first hybridization. 99% CI ratio of mean is plotted as a bar over points, but is not visible due to its small size (n=60143 to 111284 points per channel). **D.** Barcoding confidence ratio. Barcode classes in D are compared to a null model of barcode observations where random chance observation should give a ratio of 1. Off target barcodes are observed 0.005 times less than expected, suggesting that seqFISH has high accuracy in correctly counting barcoded transcripts (n=3493 cells). Dark bars on top of bar plots correspond to 99.999% confidence interval determined by bootstrap resampling. **E.** Comparison of average copy numbers per gene as measured by Zeisel *et al.*[4] and seqFISH. Single cell RNA-seq underestimates copy numbers compared to seqFISH.

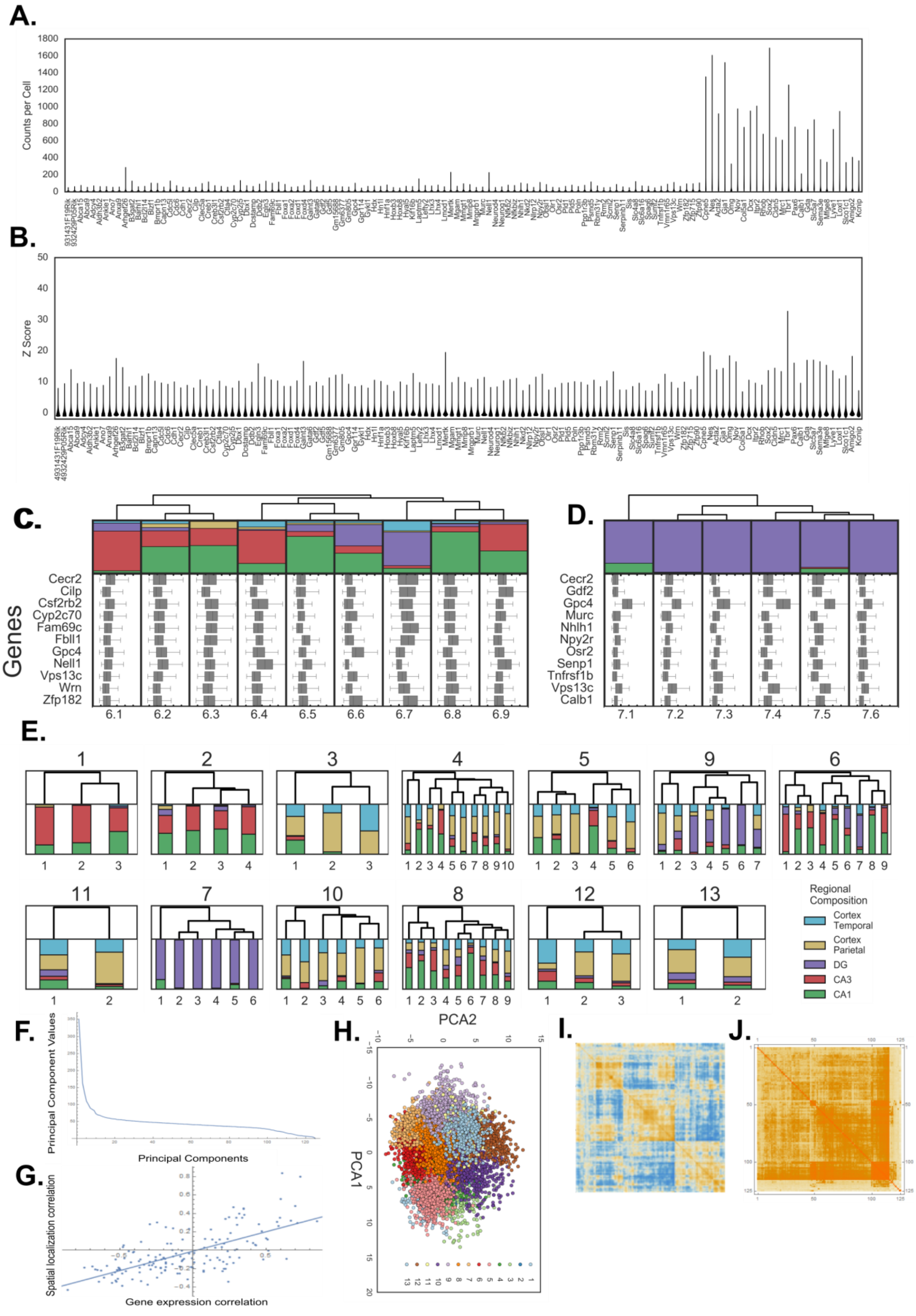
A.
Counts per Cell
1800
1600
1400
1200
1000
800
600
400
200
0
931431F19Rik
932429P05Rik
Abca15
Abca9
Adcy4
Aldh3b2
Ankle1
Ano7
Anxa9
Arhgef26
B3gat2
Barhl1
Bcl2l14
Blzf1
Bmpr1b
Capn13
Cdc5l
Cdc6
Cdh1
Cecr2
Cilp
Clec5a
Creb1
Creb3l1
Csf2rb2
Ctla4
Cyp2c70
Cyp2j5
Dbx1
Dcstamp
Ddb2
Egln3
Fam69c
Fbll1
Foxa1
Foxa2
Foxd1
Foxd4
Galnt3
Gata6
Gdf2
Gdf5
Gm15688
Gm6377
Gm805
Gpc4
Gpr114
Gykl1
Hdx
Hn1l
Hnf1a
Hoxb3
Hoxb8
Hyal5
Kif16b
Laptm5
Lefty2
Lhx3
Lhx4
Lmod1
Mertk
Mgam
Mmgt1
Mmp8
Mrgprb1
Murc
Nell1
Neurod4
Neurog1
Nfkb2
Nfkbiz
Nhlh1
Nkd2
Nlrp12
Npy2r
Obsl1
Olr1
Osr2
Pld1
Pld5
Poln
Ppp1r3b
Psmd5
Rbm31y
Rrm2
Scml2
Senp1
Serpinb11
Sis
Slc4a8
Slc6a16
Spag6
Sumf2
Tnfrsf1b
Vmn1r65
Vps13c
Wrn
Zfp182
Zfp715
Zfp90
Cpne5
Nes
Acta2
Gja1
Omg
Nov
Col5a1
Dcx
Itpr2
Rhob
Sox2
Cldn5
Mrc1
Tbr1
Pax6
Calb1
Gda
Slc5a7
Sema3e
Mfge8
Lyve1
Loxl1
Slco1c1
Amigo2
Kcnip
B.
Z Score
50
40
30
20
10
0
4931431F19Rik
4932429P05Rik
Abca15
Abca9
Adcy4
Aldh3b2
Ankle1
Ano7
Anxa9
Arhgef26
B3gat2
Barhl1
Bcl2l14
Blzf1
Bmpr1b
Capn13
Cdc5l
Cdc6
Cdh1
Cecr2
Cilp
Clec5a
Creb1
Creb3l1
Csf2rb2
Ctla4
Cyp2c70
Cyp2j5
Dbx1
Dcstamp
Ddb2
Egln3
Fam69c
Fbll1
Foxa1
Foxa2
Foxd1
Foxd4
Galnt3
Gata6
Gdf2
Gdf5
Gm15688
Gm6377
Gm805
Gpc4
Gpr114
Gykl1
Hdx
Hn1l
Hnf1a
Hoxb3
Hoxb8
Hyal5
Kif16b
Laptm5
Lefty2
Lhx3
Lhx4
Lmod1
Mertk
Mgam
Mmgt1
Mmp8
Mrgprb1
Murc
Nell1
Neurod4
Neurog1
Nfkb2
Nfkbiz
Nhlh1
Nkd2
Nlrp12
Npy2r
Obsl1
Olr1
Osr2
Pld1
Pld5
Poln
Ppp1r3b
Psmd5
Rbm31y
Rrm2
Scml2
Senp1
Serpinb11
Sis
Slc4a8
Slc6a16
Spag6
Sumf2
Tnfrsf1b
Vmn1r65
Vps13c
Wrn
Zfp182
Zfp715
Zfp90
Cpne5
Nes
Acta2
Gja1
Omg
Nov
Col5a1
Dcx
Itpr2
Rhob
Sox2
Cldn5
Mrc1
Tbr1
Pax6
Calb1
Gda
Slc5a7
Sema3e
Mfge8
Lyve1
Loxl1
Slco1c1
Amigo2
Kcnip
C.
Genes
Cecr2
Cilp
Csf2rb2
Cyp2c70
Fam69c
Fbll1
Gpc4
Nell1
Vps13c
Wrn
Zfp182
6.1
6.2
6.3
6.4
6.5
6.6
6.7
6.8
6.9
D.
Cecr2
Gdf2
Gpc4
Murc
Nhlh1
Npy2r
Osr2
Senp1
Tnfrsf1b
Vps13c
Calb1
7.1
7.2
7.3
7.4
7.5
7.6
E.
1
1 2 3
2
1 2 3 4
3
1 2 3
4
1 2 3 4 5 6 7 8 9 10
5
1 2 3 4 5 6
9
1 2 3 4 5 6 7
6
1 2 3 4 5 6 7 8 9
11
1 2
7
1 2 3 4 5 6
10
1 2 3 4 5 6
8
1 2 3 4 5 6 7 8 9
12
1 2 3
13
1 2
Regional Composition
Cortex Temporal
Cortex Parietal
DG
CA3
CA1
F.
Principal Component Values
350
300
250
200
150
100
50
20
40
60
80
100
120
Principal Components
G.
Spatial localization correlation
0.8
0.6
0.4
0.2
-0.2
-0.4
-0.5
0.5
Gene expression correlation
H.
PCA2
-10
-5
0
5
10
15
PCA1
-15
-10
-5
0
5
10
15
20
13
12
11
10
9
8
7
6
5
4
3
2
1
I.
J.
1
50
100
125
1
50
100
125

Fig S3. Gene expression patterns and clustering of the 125-gene dataset, Related to Figure 3. A. Overview of 125 gene expression. Plots show the distribution of each transcript in all 14,908 imaged cells. Note the last 25 genes have higher expression and were imaged with serial hybridization. **B.** Violin plots of Z-score distribution for 125 genes. **C.** Subclusters of cluster 6 cells and their regional localization and gene expression profile displayed under the dendrogram. Subcluster 6.1 is enriched in the CA3, while 6.7 is enriched in the DG. **D.** Subclusters of cluster 7 cells are shown. Almost all cells are localized in the GCL but have different combinatorial expression profiles. Note Calb1 expression, which marks out granule cell maturation, differs amongst subclusters. **E.** Sub-cluster hierarchy of each of the 13 clusters identified in Figure 3B. **F.** PCA eigenvalue analysis of the cell-to-cell correlation matrix. First 125 PC and their eigenvalues are shown. As observed in Fig 3, the first 10 PCs explain 59.5% of the variation in the data, while the remaining 115 PCs are needed to explain remaining data. Reflecting this, the eigenvalues of the first 10 components are high, while the remaining eigenvalues are uniform. **G.** Correlation between gene expression and spatial localization. Each dot represents a pair of cell classes and their correlations in gene expression expression space (x) and spatial localization patterns (y) (N=153 pairwise correlations between classes, R=0.67). Classes that are similar in expression have similar localization patterns. **H.** PCA decomposition separates cells into coherent clusters corresponding to cell classes. Cells are colored according to the clusters displayed in the dendrogram. **I.** Cell to cell correlation map for all 14,908 cells images in the 125 gene experiment. **J.** Gene to gene correlation map for all 125 genes measured in the 125 gene experiment.

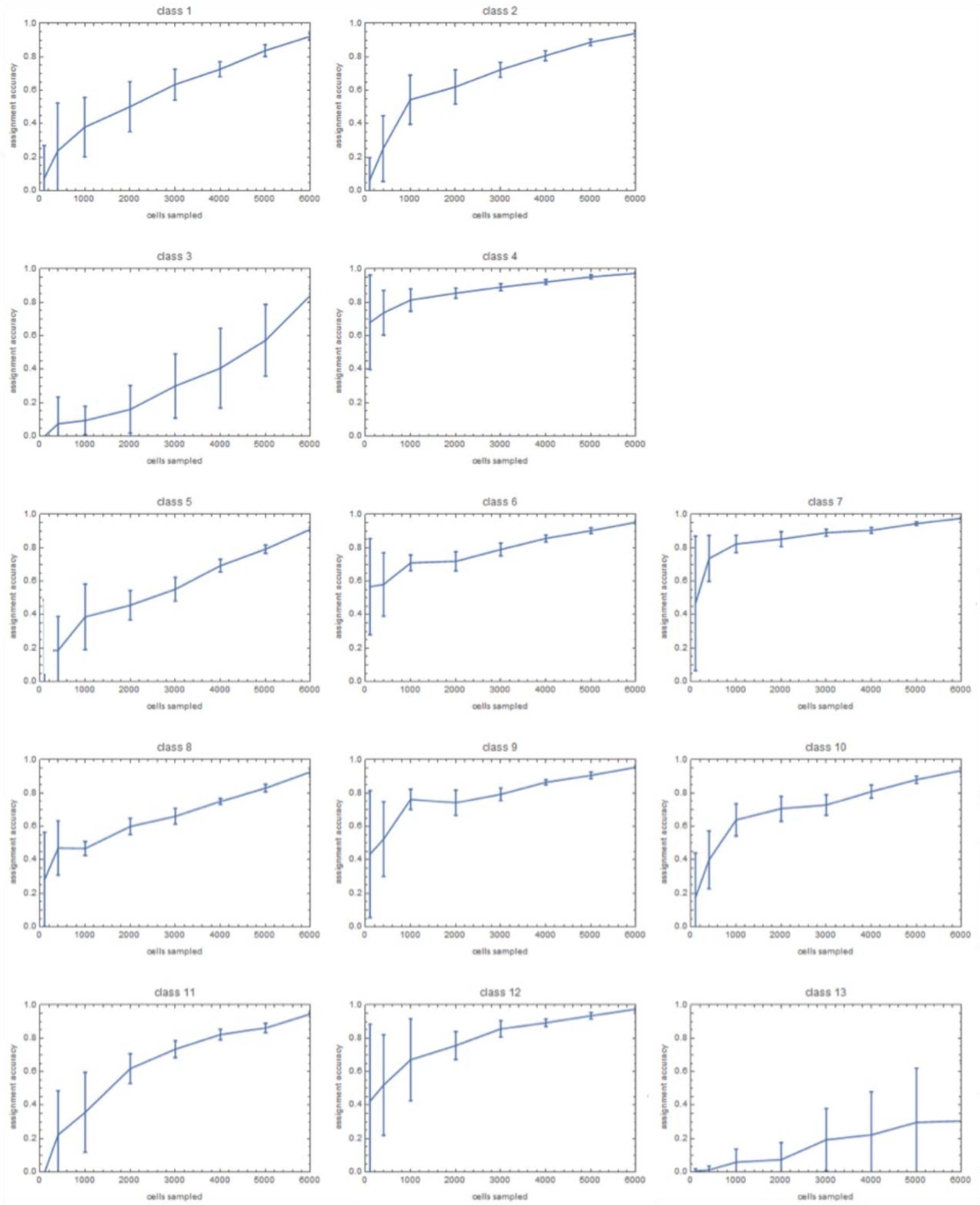

Fig S4. Robustness of cell classes to downsampling of cells, Related to Figure 3. To measure how well cluster assignments perform with a limited number of cells, a random forest model was trained on the cell-to-cell correlation matrix of subsets of 14,908 cells. The robustness of the clusters was calculated by applying this model to classify the remaining cells and determining the percent accuracy of correct assignment to the clusters presented in 3b. While some classes can be assigned accurately even with a small number of cells as the initial training set, several classes require large number of cells to accurately assign (n=10 bootstrap replicates, S.E.)

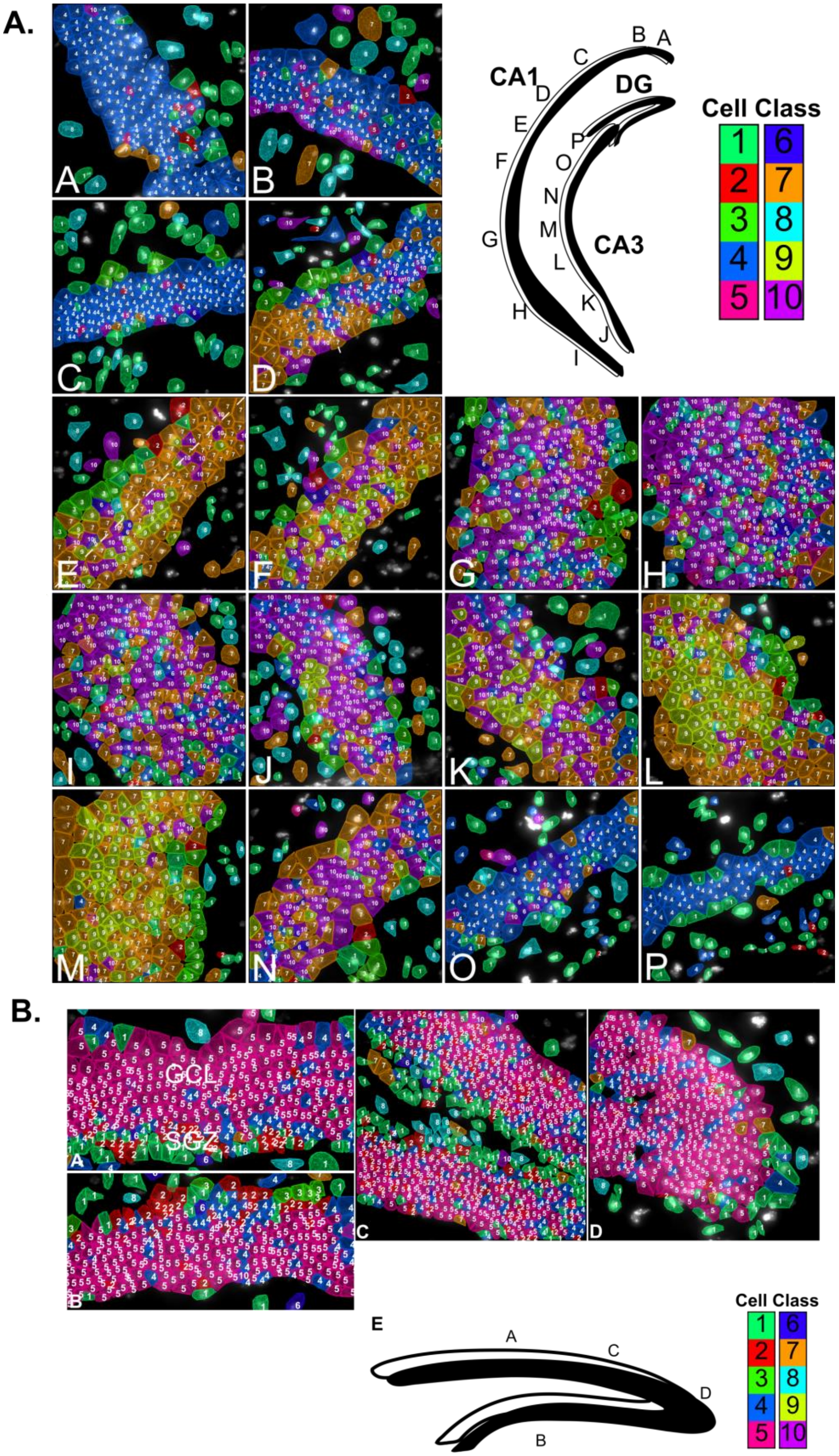
A.
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
CA1
DG
CA3
Cell Class
1
2
3
4
5
6
7
8
9
10
B.
GCL
SGZ
A
B
C
D
E
Cell Class
1
2
3
4
5
6
7
8
9
10

Figure S5. The same pattern of hippocampal subregions are observed when only hippocampal cells are clustered, related to Figures 3-5. **A.** In Figure 5 and 6, both cortex and hippocampal cells were used in the clustering. When only cells from the hippocampus are used for clustering, the same patterns are observed with homogenous cell populations in CA1d and CA3d. The intermediate and ventral subregions contain heterogeneous cell clusters. **B.** The laminar patterning in the dentate gyrus is also observed similar to Figure 4.

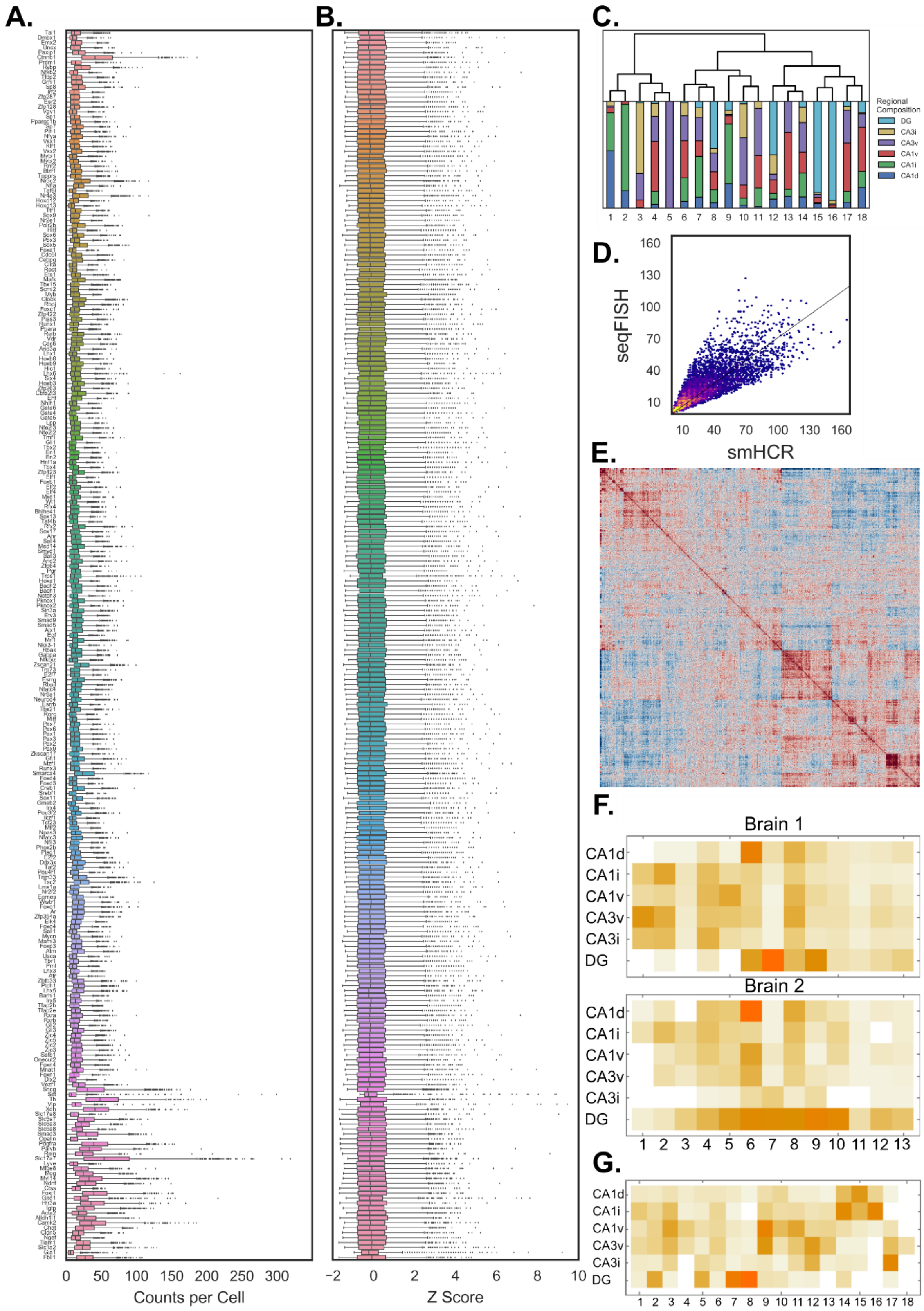

Fig S6. Gene expression patterns and clustering of the 249-gene dataset, Related to Figure 7. A. Overview of 249-gene expression. Plots show the distribution of each transcript in all

2050 imaged cells in the hippocampus. Note the last 35 genes have higher expression and were imaged with serial hybridization. **B.** Violin plots of Z-score distribution for 249 genes. **C.** Dendrogram with regional localization of the 18 cell clusters for the 249-gene experiment. **D.** Correlation of seqFISH counts to smHCR counts for the 249-gene experiment. The 2D density histogram shows a high density of points around the regression line that fall off towards the edges of the distribution. **E.** Cell-to-cell correlation for all 2050 cells in the 249-gene dataset. **F.** Heat map of the percentage of each cell class in each region of the hippocampus for both the 125-gene experiments. These heat maps show that in both 125-gene experiments the same cell classes are used in roughly the same proportions. **G.** Heat map of the percentage of each cell class in each region of the hippocampus for the 249-gene experiment. The same patterns are seen as the 125 gene experiment (i.e. different regions use different cell classes in varying amounts).

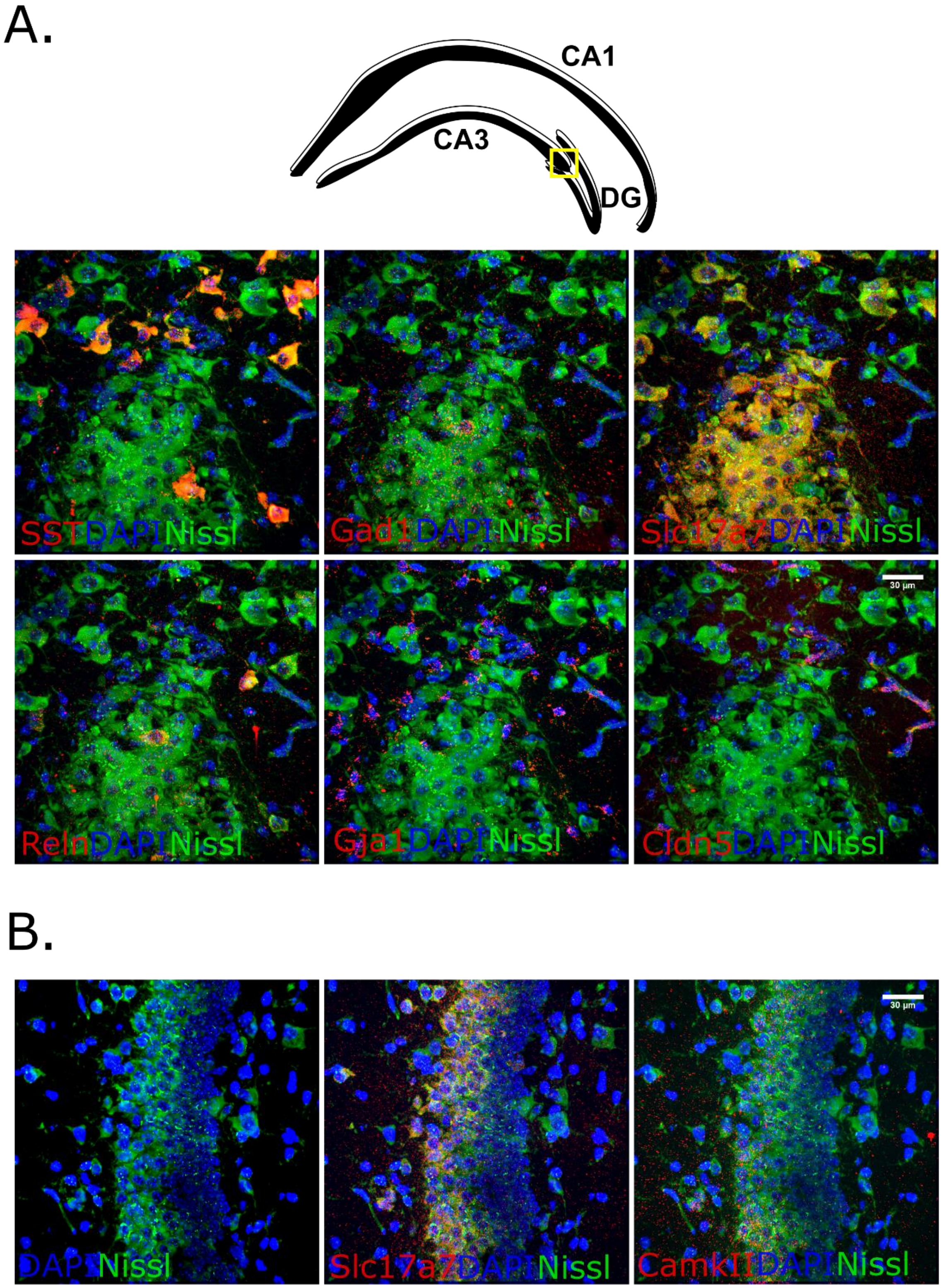

Fig S7. Marker genes robustly identify cell types, Related to Figure 7. A. The top panel outlines the region of the hippocampus being shown in a yellow box. The images show the raw

gene expression patterns seen using smHCR in our data at the dorsal most tip of the CA3 for a representative set of cell identity markers used in the 249 gene experiment. The transcript expression profile is shown in red, Nissl staining is shown in green, and DAPI staining is shown in blue. Each image shown is the full field of view and a maximum intensity projection over 15 um. **B.** Set of images showing the distinction between the GCL and SGZ. The GCL shows a high level of Nissl staining and expression of neuronal genes such as *slc17a7* and *camkII*. The SGZ shows an absence of Nissl staining and terminal neuron marker genes. The transcript expression profile is shown in red, Nissl staining is shown in green, and DAPI staining is shown in blue. Each image shown is the full field of view (216 um x 216 um) and a maximum intensity projection over 15 um.

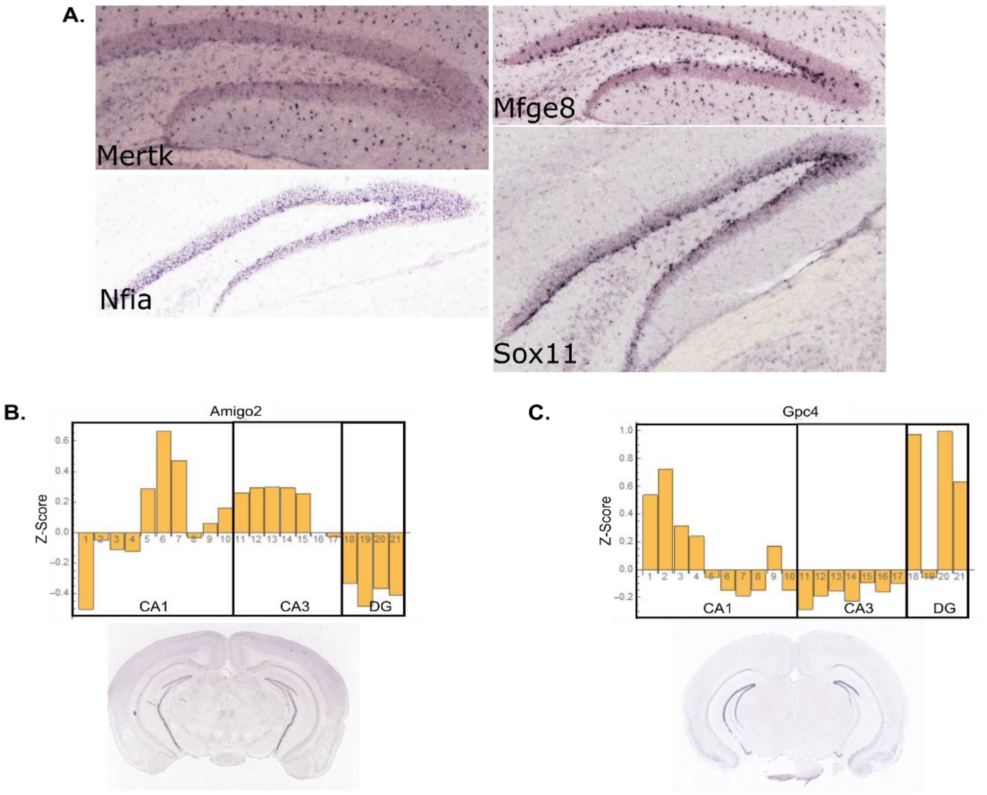

Fig S8. Comparison of SeqFISH expression data to Allen Brain Atlas expression data, Related to Figure 8. A. ISH data from the Allen Brain Atlas for genes seen to be enriched in the SGZ in the 125 and 249 gene seqFISH experiments. In the 125 gene experiment, *mertk* and *mfge8* were found to be enriched in the SGZ. In the 249 gene experiment, *nfia* and *sox11* were seen to be enriched in the SGZ. ABA ISH data shows similar patterns to those observed with seqFISH for the SGZ. **B-C.** Comparison of averaged z-score values per cell from seqFISH to ABA data across hippocampus.   **B.** Amigo2 Z-score profile found across the different fields of the hippocampus using seqFISH is shown on top and the ABA ISH image for Amigo2 is shown on the bottom. **C.** Gpc4 Z-score profile found across the different fields of the hippocampus using seqFISH is shown on top and ABA ISH image for Gpc4 is shown on the bottom.

	Barcode Hyb 1	Barcode Hyb 2	Barcode Hyb 3	Barcode Hyb 4	Control Hyb 1	Control Hyb 2	Serial Hyb 1	Serial Hyb 2	Serial Hyb 3	Serial Hyb 4	Serial Hyb 5
4931431F19Rik	1	2	1	3							
4932429P05Rik	1	3	1	4							
Abca15	1	4	1	5							
Abca9	1	3	2	3							
Acta2							3				
Adcy4	1	4	2	4							
Aldh3b2	1	5	2	5							
Amigo2											4
Ankle1	1	2	3	1							
Ano7	1	3	3	2							
Anxa9	1	4	3	3							
Arhgef26	1	5	3	4							
B3gat2	1	1	3	5							
Barhl1	1	3	4	1							
Bcl2l14	1	4	4	2							
Blzf1	1	5	4	3							
Bmpr1b	1	1	4	4							
Calb1										1	
Capn13	1	2	4	5							
Cdc5l	1	4	5	1							
Cdc6	1	5	5	2							
Cdh1	1	1	5	3							
Cecr2	1	2	5	4							
Cilp	1	3	5	5							
Cldn5									2		
Clec5a	2	4	1	1							
Col5a1								2			
Cpne5							1				
Creb1	2	5	1	2							
Creb3l1	2	1	1	3							
Csf2rb2	2	2	1	4							
Ctla4	2	3	1	5							
Cyp2c70	2	5	2	1							
Cyp2j5	2	3	2	4							
Dbx1	2	4	2	5							
Dcstamp	2	1	3	1							
Dcx								3			
Ddb2	2	4	3	4							
Egln3	2	5	3	5							
Fam69c	2	2	4	1							
Fbll1	2	3	4	2	5	5					
Foxa1	2	4	4	3							
Foxa2	2	5	4	4							
Foxd1	2	1	4	5							
Foxd4	2	3	5	1							
Galnt3	2	4	5	2							
Gata6	2	5	5	3							
Gda										2	
Gdf2	2	1	5	4							
Gdf5	2	2	5	5							
Gja1							4				
Gm15688	3	3	1	1							
Gm6377	3	4	1	2							
Gm805	3	5	1	3							
Gpc4	3	1	1	4							
Gpr114	3	2	1	5							
Gykl1	3	4	2	1							
Hdx	3	5	2	2	1	1					
Hn1l	3	1	2	3							
Hnf1a	3	2	2	4							
Hoxb3	3	3	2	5							
Hoxb8	3	5	3	1							

	Barcode Hyb 1	Barcode Hyb 2	Barcode Hyb 3	Barcode Hyb 4	Control Hyb 1	Control Hyb 2	Serial Hyb 1	Serial Hyb 2	Serial Hyb 3	Serial Hyb 4	Serial Hyb 5
Hyal5	3	1	3	2							
Itpr2								4			
Kcnip											5
Kif16b	3	4	3	5							
Laptm5	3	1	4	1							
Lefty2	3	2	4	2							
Lhx3	3	5	4	5							
Lhx4	3	2	5	1							
Lmod1	3	3	5	2							
Loxl1											2
Lvye1											1
Mertk	3	4	5	3							
Mfge8										5	
Mgam	3	5	5	4							
Mmgt1	3	1	5	5							
Mmp8	4	2	1	1							
Mrc1									3		
Mrgprb1	4	3	1	2							
Murc	4	4	1	3							
Nell1	4	5	1	4							
Nes							2				
Neurod4	4	1	1	5							
Neurog1	4	3	2	1							
Nfkb2	4	4	2	2							
Nfkbiz	4	5	2	3							
Nhlh1	4	1	2	4							
Nkd2	4	2	2	5							
Nlrp12	4	4	3	1							
Nov								1			
Npy2r	4	5	3	2							
Obsl1	4	1	3	3							
Olr1	4	2	3	4							
Omg							5				
Osr2	4	3	3	5							
Pax6									5		
Pld1	4	5	4	1							
Pld5	4	1	4	2							
Poln	4	2	4	3							
Ppp1r3b	4	1	5	1							
Psmd5	4	2	5	2							
Rbm31y	4	3	5	3							
Rhob								5			
Rrm2	5	2	1	2							
Scml2	5	3	1	3							
Sema3e										4	
Senp1	5	4	1	4							
Serpinb11	5	2	2	1							
Sis	5	3	2	2							
Slc4a8	5	4	2	3	4	4					
Slc5a7										3	
Slc6a16	5	5	2	4							
Slco1c1											3
Sox2									1		
Spag6	5	1	2	5							
Sumf2	5	3	3	1							
Tbr1									4		
Tnfrsf1b	5	4	3	2							
Vmn1r65	5	5	3	3							
Vps13c	5	1	3	4	2	2					
Wrn	5	2	3	5							
Zfp182	5	4	4	1							
Zfp715	5	5	4	2	3	3					
Zfp90	5	1	4	3							

Table S1, Related to Figure 1. Barcode assignments in the 125-gene seqFISH and serial experiment. 125 genes are profiled, 100 of which are barcoded and 25 are identified by serial smHCR hybridizations. Five control genes (Hdx, Vps13c, Zfp715, Fbll1, Slc4a8) were quantified by both techniques. The smHCR round of hybridization of control genes were performed twice to colocalize signal to obtain an absolute count.

Table S2, Related to Figure 3-6. (Provided as a separate Excel file)
Cluster group data for both 125-gene experiments. The "Major Cluster" column A defines the large cluster number. The "sub-cluster index" column B gives the subcluster number of the cluster within the major cluster. The number of cells in each subcluster and the location of those cells are tabulated in the columns C-I. Column J lists the top 4 enriched genes in the subcluster.

Table S3, related to Figure 1 and Figure S1. (Provided as a separate Excel file)
Sequences for RNA integrity test. 48 probes targeting the PGK1 transcript was used. 24 probes were amplified with initiator B1 and the remaining 24 probes were amplified with initiator B3.

Table S4, Related to Figures 1-6. (Provided as a separate Excel file)
Probe list for 125-gene barcoding experiment. Sheet 1 to sheet 4 gives the full oligoarray synthesized sequence for hybridizations 1 to 4 respectively. Sheet 5 and 6 give the sequences of the 25 high copy number genes that were targeted. For sheets 1-4, Column A gives the gene name. Columns B and J give the forward and reverse amplification primer, respectively. Columns C and I give the restriction site sequences. Column D and H give the restriction site spacer sequences. The final probe sequence is can be made by concatenating columns E-G. For sheets 5-6, column A is the gene name and concatenating Column B-D gives the final probe. The second sheet gives the readout probes. Column A is the gene name and concatenating Column B-D gives the final readout probe for HCR.

Table S5, related to Figure 4, 5, and 6. (Provided as a separate Excel file)
Raw expression data for 125 genes in brain 1 and brain 2 cells. For sheet 1, each row represents a single cell and each column represents the mRNA count within a cell for a specific gene in brain 1. For sheets 2-4, each column represents a single cell and each row represents the mRNA count within a cell for a specific gene.

Table S6, Related to Figure 7. (Provided as a separate Excel file)
Raw expression data for 249 genes in Brain 3 cells. Each column represents a single cell and each row represents the mRNA count within a cell for a specific gene.

Table S7, Related to Figure 7. (Provided as a separate Excel file)
Cluster group data for 249-gene experiment. The "Major Cluster" column A defines the cluster number. The number of cells in each sub-cluster and the location of those cells are tabulated in the columns C-I. Column J lists the top 4 enriched genes in the cluster.

Table S8, Related to Figure S7. (Provided as a separate Excel file)
Barcode assignments in the 249-gene seqFISH and serial experiment. 249 genes are profiled, 214 of which are barcoded and 35 are identified by serial smHCR hybridizations. Four control genes (*Smarca4*, *Sin3a*, *Npas3*, and *Neurod4*) were quantified by both techniques.

Supplementary Experimental Procedure

Probe Design. Genes were selected from the Allen Brain Atlas database. We identified genes that are heterogeneously expressed in coronal sections containing the hippocampus at Bregma coordinates -2.68 mm anterior. Using the ABA region definitions, we break down the voxels representing the ABA data in those brain sections into 160 distinct regions and average the expression values within each region. We selected 100 genes that had high variances across these distinct regions and that also had low-medium expression levels. These genes included transcription factors and signaling pathways components as well as ion channels and other functional genes. Lastly, we chose 25 genes from single cell RNA-seq data that were enriched in certain cell types. Briefly, the design criteria used were 1) constant regions of all spliced isoforms were identified, 2) Masked regions of UCSC genome were removed from possible probe design, 3) 35mer sequences were tiled 4nt apart, 4) sets of non-overlapping probes with tightest GC range around 55% were found, 5) probes were blasted for off-target hits. Any probe with an expected total off-target copy number of more than 5000 was dropped. Once all possible probes for every target gene was acquired, the probe set oligo-pool was optimized using the following criteria: 1) Expected # of off-target hits for entire probe pool was calculated, 2) probes were sequentially dropped from genes until any off-target gene was hit by no more than 6 probes from entire pool, 3) HCR adapters were added to designed probes and 10nt in either direction of the adapter junction was blasted and screened for off-target hits, 4) probe pools were searched for regions of 18mer complementary, 5) the probe sets for a given transcript was refined down to 24 probes by dropping probes in order of the expected number of off-target hits, 6) Cutting sites and hybridization specific primers were added to probes.

Probe Generation. All oligoarray pools were purchased as 92k synthesis from Customarray Inc. Probes were amplified from array-synthesized oligo pool as previously described *(36)*, with the following modifications: (i) a 35nt RNA-targeting sequence for in situ hybridization, (ii) a 35nt HCR initiator sequence designed to initiate one color of 5 possible HCR polymers, (iii) two hybridization specific flanking primer sequences to allow PCR amplification of the probe set and (iv) EcoRI (5'-GAATTC—3') and KpnI (5'-GGTACC-3') sites for cutting out flanking primers to reduce probe size. Ethanol precipitation was used to purify the final digested probes.

Brain extraction and sample mounting. C57BL/6 with Ai6Cre-reporter (uncrossed) (Jackson Labs, SN: 007906) female mice aged 50-80 days were anesthetized with isoflurane according to institute protocols (protocol #1701-14) (38). No randomization of mice was used and blinding was not necessary as the study was exploratory. Mice were perfused for 8 minutes with perfusion buffer (10U/ml heparin, 0.5% $NaNO_2$(w/v) in 0.1M PBS at 4C). Mice were then perfused with fresh 4% PFA\0.1M PBS buffer at 4C for 8 minutes. The mouse brain was dissected out of the skull and immediately placed in a 4% PFA buffer for 2 hours at room temperature under gentle mixing. The brain was then immersed in 4C 30% RNAse-free Sucrose (Amresco 0335-2.5KG)\1x PBS until the brain sank. After the brain sank, the brain was frozen in an dry ice\isopropanol bath in OCT media and stored at -80C. Fifteen micron sections were cut using a cryotome and immediately placed on an aminosilane modified coverslip.

Sample permeabilization, hybridization, and Imaging. Brain sections mounted to coverslips were permeabilized in 4C 70% EtOH for 12-18 hours. Brains were further permeabilized by the addition of rnase-free 8% SDS (Ambion AM9822) for 10 minutes. Samples were rinsed to remove SDS, desiccated and a hybridization chamber (Grace Bio-Labs 621505) was adhered around the brain section. Samples were hybridized overnight at 37C with Split Color PGK1 Probes (Table S3) in Hybridization Buffer (2X SSC (Invitrogen 15557-036), 10% Formaldehyde (v/v) (Ambion AM9344), 10% Dextran Sulfate (Sigma D8906), 2mM Vanadyl Ribonucleoside Complex (VRC; NEB S1402S) in Ultrapure water (Invitrogen 10977-015)). Samples were washed in 30% Wash Buffer (WBT: 2X SSC, 30% Formaldehyde (v/v)] 10% Dextran Sulfate, 0.1% Triton-X 100 (Sigma X-100), 2mM VRC in Ultrapure water) for 30 minutes. While washing aliquoted HCR hairpins (Molecular Instruments Inc) were heated to 95C for 1.5 minutes and allowed to cool to RT for 30 minutes. HCR hairpins were diluted to a concentration of 120nM per hairpin in amplification buffer (2X SSC, 10% Dextran Sulfate) and added to washed tissue for 45 minutes. Following amplification, samples were washed in the same 30% WBT for at least 10 minutes to remove excess hairpins. Samples were stained with DAPI and submerged in pyranose oxidase antibleaching buffer *(12).* Sample port covers were closed with a glass coverslip or a transparent polycarbonate sheet to exclude oxygen.

Samples were imaged using a standard epifluorescence microscope (Nikon Ti Eclipse with custom built laser assembly) for the 125-gene experiment. Exposures times were 200 ms for cy7 and alexa 488 channels and 100 ms for alexa 647, alexa 594, and cy3b channels. For the 249-gene experiment, a Yokogawa CSU-W1 spinning disk confocal unit attached to an Olympus IX-81 base was used for imaging. The exposure times were 500 ms for each channel. At this stage, intact and accessible mRNA should always appear in two channels. If the RNA was deemed to be intact, DAPI data was collected in this hybridization. Samples were digested with DNAse I (Roche 04716728001) for 4 hours at room temperature on the scope. Following DNAse I the sample was washed several times with 30% WBT and hybridized overnight with 70% Formamide HB and the experiment probes at 1 nM concentration per probe sequence at room temperature (Table S4 and S5). Samples were again washed and amplified as before. Barcode digits were developed by repeating this cycle with the appropriate probes for each hybridization. Fluorescent Nissl stain (ThermoFisher N-21480) was collected at the end of the experiment along with images of multispectral beads to aid chromatic aberation corrections.

Image Processing. To remove the effects of chromatic aberration, the multispectral beads were first used to create geometric transforms to align all fluorescence channels. Next, the background illumination profile of every fluorescence channel was mapped using a morphological image opening with a large structuring element. These illumination profile maps were used to flatten the illumination in post-processing resulting in relatively uniform background intensity and preservation of the intensity profile of fluorescent points. The background signal was then subtracted using the imagej rolling ball background subtraction algorithm with a radius of 3 pixels. Finally, the calculated geometric transforms were applied to each channel respectively.

Image Registration. The processed images were then registered by first taking a maximum intensity projection along the z direction in each channel. All of the maximum projections of the channels of a single hybridization were then collapsed resulting in 4 composite images containing all the points in a particular round of hybridization. Each of these composite images of hybridization 1-3 were then cross-correlated individually with the composite image of hybridization 4 and the position of the maxima of the cross-correlation was used as the translation factor to align hybridizations 1-3 to hybridization 4.

Cell Segmentation. For cells in the cortex, the cells were segmented manually using the DAPI images taken in the first round of hybridization and the fluorescent nissl stain taken at the end of the experiment. Furthermore, the density of the point cloud surrounding a cell was taken into account when forming cell boundaries, especially in cells that did not stain with the nissl stain. For the hippocampus, the cells were segmented by first manually selecting the centroid in 3D of each DAPI signal of every cell. Transcripts were first assigned based on nearest centroids. These point clouds were then used to refine the centroid estimate and create a 3D voronoi tessellation with a 10% boundary-shrinking factor to eliminate ambiguous mRNA assignments from neighboring cells.

Barcode calling. The potential mRNA signals were then found by LOG filtering the registered images and finding points of local maxima above a specified threshold value[8]. Once all potential points in all channels of all hybridizations were obtained, dots were matched to potential barcode partners in all other channels of all other hybridizations using a 1 pixel search radius to find symmetric nearest neighbors. Point combinations that constructed only a single barcode were immediately matched to the on-target barcode set. For points that matched to construct multiple barcodes, first the point sets were filtered by calculating the residual spatial distance of each potential barcode point set and only the point sets giving the minimum residuals were used to match to a barcode. If multiple barcodes were still possible, the point was matched to its closest on-target barcode with a hamming distance of 1. If multiple on target barcodes were still possible, then the point was dropped from the analysis as an ambiguous barcode. This procedure was repeated using each hybridization as a seed for barcode finding and only barcodes that were called similarly in at least 3 out of 4 rounds were used in the analysis. The number of each barcode was then counted in each of the assigned cell volumes and transcript numbers were assigned based on the number of on-target barcodes present in the cell volume.

Clustering. To cluster the dataset with 14,908 cells and 125 genes profiled, we first z-score normalized the data based on gene expression (Table S6). Once the single cell gene expression data is converted into z-scores, we compute a matrix of cell-to-cell correlations using Pearson correlation coefficients. Then hierarchical clustering with Ward linkage is performed on the cell-to-cell correlation data with cells in the center field of view. The cluster definitions are then propagated to the remaining cells using a random forest machine learning algorithm. To analyze the robustness of individual clusters, a random forest model was trained using varying subsets of the data and used to predict the cluster assignment of the remaining cells *(22)*. A bootstrap analysis by dropping different sets of cells was performed in increments (Fig S5). To determine the effect of dropping out genes on the accuracy of the clustering analysis, we used a random forest decision tree to learn the cluster definition based on the 125 gene data. Then we ask the decision tree to re-compute the cluster assignment on cell-to-cell correlation matrices with fewer and fewer genes (Fig 3F, green line). Bootstrap resampling was also performed with this analysis (Fig 3F, bluelines). The PCA and tSNE analysis were performed using the same cell-to-cell z-scored Pearson correlation matrix. The cell-to-cell correlation in Fig 3E was calculated with increasing number of principal components dropped (have their eigenvalues set to zero). The cluster assignment accuracy is again computed through the random forest decision tree.

Supplementary Text

Error correction barcode design

Designing an error correction code to correct for k number of errors in a message of n length is analogous to packing as many spheres of radius k in a n dimensional cube. There are examples of "perfect codes" such as Golay and Hamming codes that can be as efficient as possible in this packing design. These perfect codes are important in digital communication because the word lengths are long, up to billions of letters for gigabytes of data, and many forms of errors can occur, including deletion and insertions. However, in the seqFISH experiments, as the code lengths are short, a perfect code correction system is not necessary, especially as the "correct" codes are already defined. One of the major source of error is deletions due to loss of a hybridization. Thus, it is possible to design simple correction schemes that are not completely efficient (i.e. obtain the tightest packing density for the n-spheres) but can achieve good error correction with just a few extra rounds of hybridization.

To design a barcode scheme that can tolerate loss of a single round of hybridization is akin to a problem where any n-dimensional hypercube is collapsed by 1 dimension to a n-1 dimensional hypercube without having any two points on the n-dimensional hypercube mapping to the same point. In order for this to be true, no two barcodes can be connected by a 1D line running parallel to any of the axes. There are many solutions to generate this 1 round loss tolerant code. A barcode generator (i, (i+j+k) mod 5,j,k) is used to generate the barcodes used in our experiment. This design can correct for loss of 1 hybridization for an arbitrarily long barcode sequence with minimal extra effort. For example, 7 rounds of hybridization with 5 colors can cover 5^7= 78,125 transcripts, more than the transcriptome, with 8 hybridizations the entire transcriptome can be coded with error correction using the barcoding system proposed.

Another consideration in designing error-tolerant barcodes is that the mechanism of re-hybridzation should guide the robustness of error correction. In the merFISH implementation of seqFISH, null signal, or "0", along with "1" which is cy5 fluoroscence, is used to form a binary barcode. However, it is difficult to determine whether no signal is due to mis-hybridization or actual null signal. In our seqFISH implementation using positive signals as readouts during each round of hybridization reduces the need for error correction because false positive signal is unlikely to re-occur in the same position during another hybridization due to DNAse stripping between hybridizations. Thus implementation of seqFISH with 5 colors and 1 extra round of hybridization to error correct is both efficient and accurate, and allows imaging of a large tissue sections since imaging time is ultimately limiting in multiplexing experiments.

Optical Space for Barcodes in Cells

The theoretical upper limit for the number of barcodes that can be identified accurately within a cells primarily depends on the volume of the cell. As mRNA spots are diffraction limited, if a microscope is configured to have sub-diffraction limited pixel size, the ability to identify smFISH signal without any super-resolution would require no two mRNA signals to be immediately adjacent to each other in x, y or z dimension. We will call these minimum required voxels "coding voxels." The absolute upper limit of the number of transcripts that can be coded unambiguously without any super-resolution methods is solely a function of the number of coding voxels present in a cell. Assuming a diffraction limit of λ um and a resolution of z um in the z direction, there exists $\frac{V}{(3\lambda)^2 z}$ coding voxels per cell, where V is the volume of the cell in microns. In our seqFISH method, we use 5 or more channels to hold mRNA spots which would increase the total number of coding voxels by a multiplicative factor equal to the number of channels used for barcoding. Therefore,

$$\#B = \frac{F\,V}{(3\lambda)^2 z}$$

where #B is the maximum number of unambiguous barcodes a cell can hold, and F is the number of channels used. As mammalian cells range from about 500 – 4000 microns in volume, these cells can accommodate roughly between 6100 – 49,000 barcodes assuming 5 fluorescence channels are being used, the diffraction limit is 0.3 um, and the z resolution is 0.5 um. In principle, this calculation would provide the total number of perfectly discernible spots a cell can accommodate. In our actual experimental data, we have some amount of dropped barcodes due to ambiguity in barcode assignment due to spot overlaps. This is one of the main factors that reduces the efficiency of seqFISH as compared to single transcript detection (i.e. smFISH or smHCR). Expansion microscopy could further increase the number of coding voxels in a cell by the expansion factor leading to fewer drops and imaging of denser transcripts.

seqFISH clustering analysis vs single cell RNAseq analysis

As we do not sample the entire transcriptome with seqFISH, the dendrogram structure and cell clusters will be influenced by the genes chosen. This may be best illustrated by an analogy: at an international conference, you are surveying the attendees. You ask 100 questions, 80 of them about their profession and 20 on their country of origin. You might find that many people are biologists, and others are engineers. Similarly, there are many people from the US and others from elsewhere. If you cluster the data, you will find that because 80 out of the 100 questions are about their profession, the biologist vs engineer split will be the first split along the dendrogram, and the citizenship will split next. This is because two biologists from different countries will look more similar to each other (according the 100 questions asked), than an engineer. If the questions were 80 about citizenship, and 20 about profession, then the citizenship would be first split in the clustering. Now if you throw in more questions about gender in the survey, then you can get clusters that have both men and women engineers in the US. But if there is only 1 question about gender and the remaining 99 questions are about other things, then it is expected that gender would not factor importantly into the unsupervised clustering of the data.

The clustering dendrogram is determined by the distance between different clusters, and this distance is affected by how many marker genes are in the target list. If we had 2 genes in our target list that differ in their expression between neuron vs non-neurons, and 10 genes that differ between DG vs CA1 neurons all with similar expression distributions, then the dendrogram is going to split the DG and CA1 neurons first because the distance between those clusters are larger than the neurons vs non-neurons distances (10 vs 2). Thus the dendrogram for the 100-200 genes experiment should not match completely the dendrogram for the RNAseq data, because the composition of marker genes in the 100 or 200 gene list is not the same proportions as the transcriptome.

In our analysis, all genes are weighted equally regardless of expression levels or “canonical importance.” We believe this is the most direct and unbiased way to perform the analysis on our data. Because our data is 100-200 genes, but with higher accuracy per gene than single cell RNAseq for those 100-200 genes, it is going to be fundamentally different than single cell RNAseq data and our analysis method was selected to best match the nature of our data.

Most single cell RNAseq analysis methods either select subsets of genes for clustering at each level or iteratively select genes with the highest variations to define cell types. Our analysis uses the full set of genes probed to detect combinatorial expression differences amongst cells. We have tried to implemented several analysis methods similar to ones used in single cell RNAseq, but obtained poor separation between cells clusters as compared to the analysis method herein presented. This is because we lose significant combinatorial information by ignoring genes when defining clusters. The accurate measurement of 100 genes can provide a great deal of explanatory power because the combinatorial expression pattern contains more information than just individual genes.

On the other hand, the main limitation of the 100-200 gene experiment is that it does not measure all of the genes. So the data will not cluster the same as RNAseq data, and does not identify all the big “cell types” in the top branches of the dendrogram. However, we can detect fine differences between cells. single cell RNAseq can be used for cataloguing all the major cell types, while seqFISH can be used to focus on a specific region or “cell type” and investigate the spatial mRNA expression patterns between cells or fine differences within a “cell type”.