








Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections

Yu-Chi Sun^{1,2}, Xiaoyin Chen^{1,2}  , Stephan Fischer¹, Shaina Lu^{1,2} , Huiqing Zhan¹, Jesse Gillis¹ and Anthony M. Zador¹  

Functional circuits consist of neurons with diverse axonal projections and gene expression. Understanding the molecular signature of projections requires high-throughput interrogation of both gene expression and projections to multiple targets in the same cells at cellular resolution, which is difficult to achieve using current technology. Here, we introduce BARseq2, a technique that simultaneously maps projections and detects multiplexed gene expression by in situ sequencing. We determined the expression of cadherins and cell-type markers in 29,933 cells and the projections of 3,164 cells in both the mouse motor cortex and auditory cortex. Associating gene expression and projections in 1,349 neurons revealed shared cadherin signatures of homologous projections across the two cortical areas. These cadherins were enriched across multiple branches of the transcriptomic taxonomy. By correlating multigene expression and projections to many targets in single neurons with high throughput, BARseq2 provides a potential path to uncovering the molecular logic underlying neuronal circuits.

Neural circuits are composed of neurons diverse in many properties, such as morphology^{1,2}, gene expression^{3,4} and projections^{5,6}. Although recent technological advances have made it possible to characterize the diversity in individual neuronal properties, associating multiple properties in single neurons with high throughput remains difficult to achieve. Investigating the relationship between multiple neuronal properties is essential for understanding the complex organization of neural circuits.

Of particular interest is the relationship between endogenous gene expression and long-range projections in the cortex. Cortical neurons have diverse patterns of long-range projections^{5,6} and diverse patterns of gene expression^{3,4}. The full diversity of neuronal projection patterns can often only be appreciated by assessing multiple projection targets simultaneously (Fig. 1a)^{2,6}. For example, Han et al.⁵ showed that neurons in mouse visual area V1 that project to area PM tend not to project to area AL and vice versa, a projection ‘motif’ that involves the relative probability that a single neuron projects to two targets and hence could not have been discovered by assessing projection targets one at a time. Gene expression patterns are also complex, and although the diversity in gene expression can be described by clustering neurons into transcriptomic types, these transcriptomic types have limited power in explaining the diversity of cortical projections beyond the major classes of projection neurons^{3,6–8}. Moreover, because the determination of a transcriptomic type relies on the expression of only a subset of genes, the inability of transcriptomic type to predict projection patterns raises the possibility that the expression of other genes—potentially in gene coexpression motifs—might be better correlated with projection patterns. Although transcriptomic methods can be combined with retrograde labeling^{3,9}, retrograde labeling is limited to one or at most a few brain areas at a time. Resolving the relationship between gene expression and projection patterns in the adult cortex thus requires high-throughput techniques that allow simultaneous multiplexed

gene detection with projection mapping to multiple target areas at single-neuron resolution, which remains difficult to achieve.

To achieve high-throughput mapping of projections to many brain areas, we recently introduced barcoded anatomy resolved by sequencing (BARseq), a projection mapping technique based on in situ sequencing of RNA barcodes⁶. In BARseq, each neuron is labeled with a unique virally encoded RNA barcode that is replicated in the somas and transported to the axon terminals. The barcodes at the axon terminals located at various target areas are sequenced and matched to somatic barcodes, which are sequenced in situ, to determine the projection patterns of each labeled neuron. Because BARseq preserves the location of somata with high spatial resolution, in principle it provides a platform to combine projection mapping with other neuronal properties also interrogated in situ, including gene expression. We have previously shown⁶ that BARseq can be combined with fluorescence in situ hybridization (FISH) and *Cre* labeling to uncover projections across neuronal subtypes defined by gene expression. However, these approaches can only interrogate one or a few genes at a time, which would be insufficient for unraveling the complex relationship between the expression of many genes to diverse cortical projections (Fig. 1a).

Here we aim to develop a technique to simultaneously map projections to multiple brain areas and detect the expression of dozens of genes in hundreds to thousands of neurons from a cortical area with high throughput, high spatial resolution and cellular resolution. To achieve this goal, we combine the high-throughput and multiplexed projection mapping capability of BARseq with state-of-the-art spatial transcriptomic techniques with high imaging throughput and multiplexing capacity^{10,11}. This second-generation BARseq (BARseq2) greatly improves the ability to correlate the expression of many genes to projections to many targets in the same neurons. As a proof of principle, we first demonstrate multiplexed gene detection using BARseq2 by mapping the spatial pattern of up

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. ²These authors contributed equally: Yu-Chi Sun, Xiaoyin Chen. ✉e-mail: xichen@cshl.edu; zador@cshl.edu

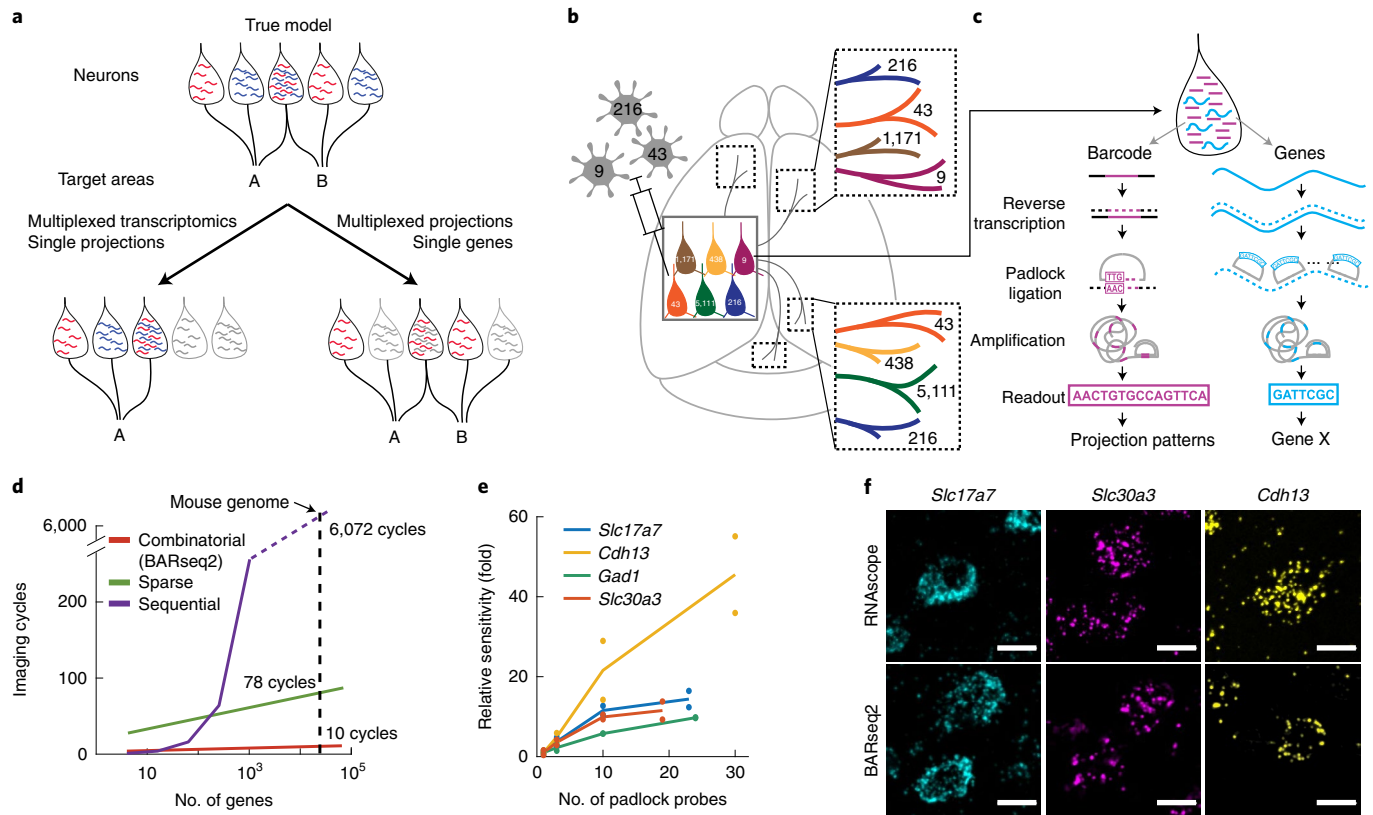


Fig. 1 | In situ sequencing of endogenous mRNAs using BARseq2. **a**, Cartoon of an example model in which the relationship between projections and gene expression can only be correctly inferred by multiplexed interrogation of both projections and gene expression. In this model, neurons that express both genes project to both targets A and B, whereas neurons that express only one of the two genes project randomly to either A or B, but not both. Methods that combine multiplexed single-neuron gene expression with data about only a single projection target will conclude that all three gene expression patterns project to target A, thus failing to detect the underlying ‘true’ relationship between gene expression and projections. Similarly, methods that combine multiplexed single-neuron projections with data about only a single gene will also fail to detect any relationship between gene expression and projections. **b,c**, BARseq2 correlates projections and gene expression at cellular resolution (**b**). In BARseq2, neurons are barcoded with random RNA sequences to allow projection mapping, and genes are also sequenced in the same barcoded neurons. RNA barcodes and genes are amplified and read out using different strategies (**c**). **d**, Theoretical imaging cycles using combinatorial coding (BARseq2), four-channel sequential coding or four-channel sparse coding as used by Eng et al.⁵⁰. Imaging cycles assumed three additional cycles for BARseq2, one additional round for sparse coding, and no extra cycle for sequential coding for error correction. **e**, Mean and individual data points of the relative sensitivity of BARseq2 in detecting the indicated genes using different numbers of padlock probes per gene. The sensitivity is normalized to that using one probe per gene. $n=2$ slices for each gene. **f**, Representative images of BARseq2 detection of the indicated genes using the maximum number of probes shown in **e** compared to RNAscope. Scale bars, 10 μ m.

to 65 cadherins and cell-type markers in 29,933 cells. We then correlate the expression of 20 cadherins to projections to 35 target areas in 1,349 neurons in mouse motor and auditory cortex. Our study reveals new sets of cadherins that correlate with homologous projections in both cortical areas. BARseq2 thus bridges transcriptomic signatures obtained through spatial transcriptional profiling with sequencing-based projection mapping to illuminate the molecular logic of long-range projections.

Results

To investigate how cadherin expression relates to diverse projections, we developed BARseq2 to combine high-throughput projection mapping with multiplexed detection of gene expression using in situ sequencing (Fig. 1b,c). BARseq2 is based on BARseq (Fig. 1c), which achieves high-throughput projection mapping by in situ sequencing of RNA barcodes⁶. Projection patterns observed using BARseq are consistent with those obtained using conventional neuroanatomical techniques in multiple circuits^{2,5}, but it can achieve throughput that is at least two to three orders of magnitude higher than the state-of-the-art single-cell tracing techniques². Possible technical concerns, including distinguishing fibers of passage from

axonal termini, sensitivity, double labeling of neurons and degenerate barcodes, have previously been addressed^{2,6,12,13} and will not be discussed in detail again here. Combining barcoded single-cell projection mapping with in situ detection of endogenous mRNAs exploits the unique advantage of BARseq in throughput to efficiently interrogate both neuronal gene expression and long-range projections simultaneously.

To detect gene expression using BARseq2, we used a non-gap-filled padlock probe-based approach to amplify target endogenous mRNAs^{10,11} (Fig. 1c). The elimination of gap filling, necessary for reading out extremely diverse sequences of barcodes, increases the sensitivity for endogenous gene detection. In this approach, the identity of the target is read out by sequencing a gene-identification index (GII) using Illumina sequencing chemistry in situ. Because the GII is a nucleotide barcode sequence that uniquely encodes the identity of a given gene, the multiplexing capacity increases exponentially as 4^N , where N is the number of sequencing cycles. This combinatorial coding by sequencing readout thereby allows simultaneous detection of a large number of genes using only a few cycles of imaging (Fig. 1d). Although sequencing readout offers many advantages, BARseq2 is also compatible with hybridization-based

readout when necessary. The combination of non-gap-filling in situ sequencing of endogenous genes and the gap-filling approach for sequencing barcodes allows many genes to be detected simultaneously with projections using BARseq2.

We first demonstrate that, by optimizing targeted in situ sequencing, BARseq2 could achieve sufficient sensitivity for detection of endogenous mRNAs. We next combined in situ sequencing of endogenous mRNAs with in situ sequencing of RNA barcodes to associate the expression of cadherins with projection patterns at cellular resolution. We then validated BARseq2 by demonstrating that it could be used to recapitulate projection patterns specific to transcriptomic neuronal subtypes and to identify cadherins that were differentially expressed across major projection classes. Finally, we identified a set of cadherins shared between the mouse auditory cortex and motor cortex that correlate with homologous projections of intratelencephalic (IT) neurons in both cortical areas.

BARseq2 robustly detects endogenous mRNAs. To adequately detect genes using BARseq2, we sought to improve the detection sensitivity. In most in situ hybridization (ISH) methods, high sensitivity is achieved by using many probes for each target mRNA^{14,15}. We reasoned that increasing the number of padlock probes for each gene might similarly improve the sensitivity of BARseq2. Indeed, we observed that tiling the whole gene with additional probes resulted in as much as a 46-fold increase in sensitivity compared to using a single probe (Fig. 1e and Methods). Combined with other technical optimizations (Extended Data Fig. 1a,b), we increased the sensitivity of BARseq2 to 60% of RNAscope, a sensitive and commercially available FISH method (Fig. 1f, Extended Data Fig. 1c,d and Methods). We further optimized in situ sequencing to robustly read out GIIs of single rolonies over many sequencing cycles (Extended Data Fig. 1e–j and Methods). These optimizations allowed BARseq2 to achieve sufficiently sensitive, fast and robust detection of mRNAs.

BARseq2 allows multiplexed detection of mRNAs in situ. To assess multiplexed detection of cadherins in situ using BARseq2, we examined the expression of 20 cadherins, along with either 3 (in auditory cortex) or 45 (in motor cortex) cell-type markers (Fig. 2a–c). We chose to focus on the cadherins because of their known roles in cortical development, including projection specification^{16,17}, and their differential expression among cardinal cell types defined by multiple properties¹⁸. These cadherins included most classical cadherins and nonclustered protocadherins expressed in auditory cortex and motor cortex. We successfully resolved and decoded 419,724 rolonies from two slices of mouse auditory cortex (1.7 mm² × 10 μm per slice) and 1,445,648 rolonies from four slices of primary motor cortex (2.8 mm² × 10 μm per slice). We recovered 20 rolonies in auditory cortex and 115 rolonies in motor cortex matching two GIIs that were not used in the experiment,

corresponding to an estimated error rate of 0.1% and 0.2%, respectively, for rolonie decoding.

Consistent with previous reports^{19,20}, many cadherins were enriched in specific layers and sublayers in the cortex (Fig. 2d). Interestingly, although most cadherins had similar laminar expression in both auditory cortex and motor cortex, some cadherins were differentially expressed across the two areas. For example, *Cdh9* and *Cdh13* were enriched in L2/3 in auditory cortex, but not in motor cortex (Fig. 2d and Extended Data Fig. 2). The laminar positions of peak cadherin expression were consistent with those obtained by other methods, including RNAscope (Fig. 2e) and the Allen Brain Atlas (ABA) database of ISH²¹ (Fig. 2f, Extended Data Fig. 3 and Methods). Thus, BARseq2 accurately resolved the laminar expression patterns of cadherins.

We then characterized gene expression obtained by BARseq2 at single-cell resolution (Methods). We assigned 228,371 rolonies to 3,377 excitatory or inhibitory neurons (67.6 ± 28.8 (mean ± s.d.) rolonies per neuron) in auditory cortex, and 752,687 rolonies to 11,492 excitatory or inhibitory neurons (65.5 ± 26.0 (mean ± s.d.) rolonies per neuron) in motor cortex. Most cadherins showed slight differences in single-cell expression levels in these two cortical areas (Extended Data Fig. 4). In auditory cortex, the total read counts per cell was higher in BARseq2 than in single-cell RNA sequencing (scRNA-seq) using 10x Genomics v3 (Fig. 2g; median read count was 64 for BARseq2 ($n = 3,337$ cells) compared to 57 for scRNA-seq ($n = 640$ cells); $P = 5.3 \times 10^{-5}$, rank-sum test). Thus, even using a limited number of probes, BARseq2 achieved sensitivity at least equal to scRNA-seq using 10x v3. For experiments requiring better quantification of genes present at low expression, the sensitivity could potentially be further improved by using more probes.

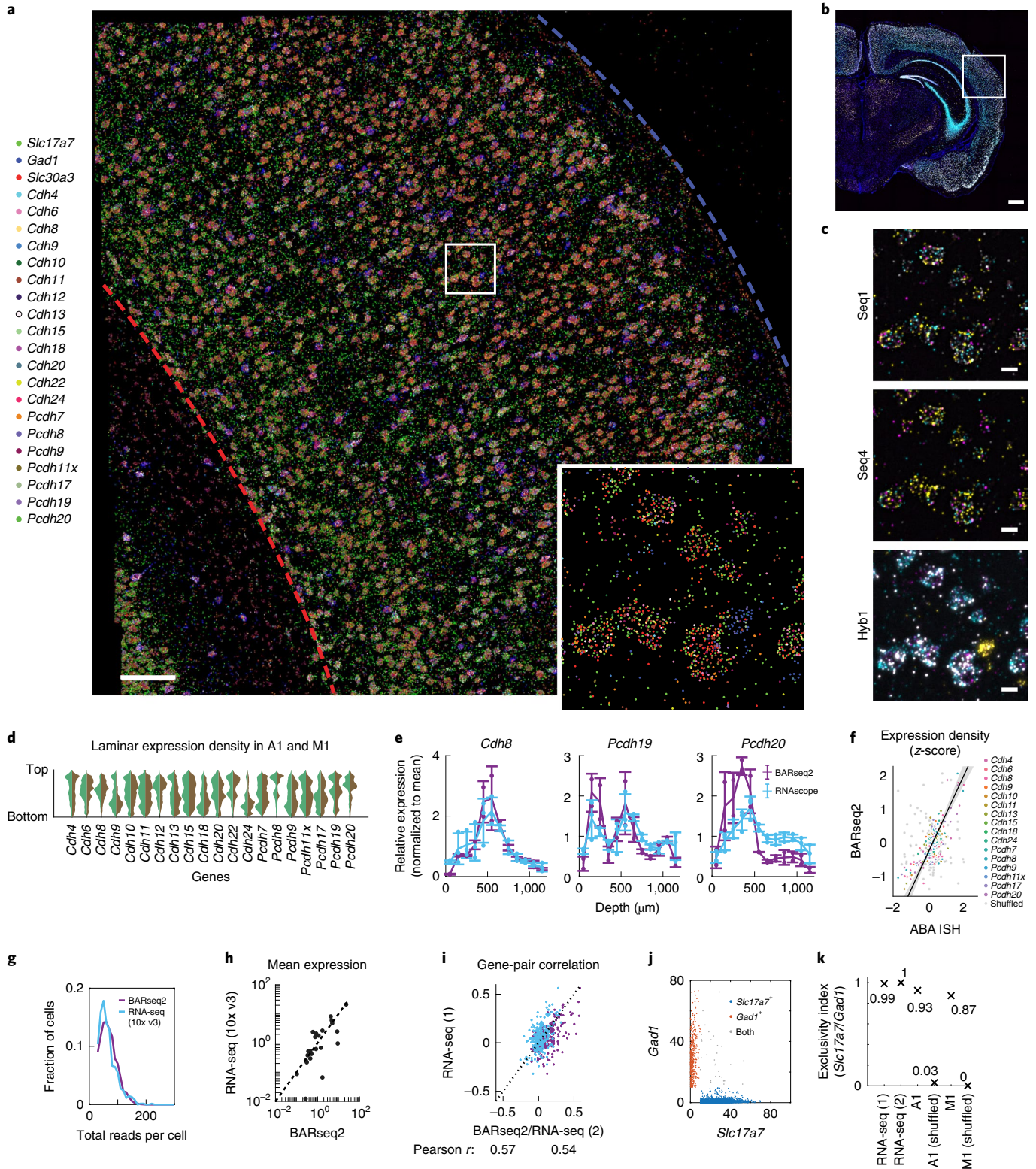
Further analyses showed that detection of mRNA by BARseq2 was specific. The mean expression of genes determined by BARseq2 was highly correlated with that determined by scRNA-seq using 10x v3 (Fig. 2h; Pearson correlation $r = 0.88$). A few outliers had substantially more counts in BARseq2 than in scRNA-seq, likely reflecting sampling differences across cell types, area-specific gene expression and differences in RNA accessibility in situ. For example, *Cdh6* expression observed by BARseq2 was 26 times that observed by scRNA-seq. This difference could be attributed to under-sampling of *Cdh6*-expressing pyramidal tract (PT) neurons in our scRNA-seq data⁶ and potentially variable sampling of neighboring cortical areas in which *Cdh6* is differentially expressed²². Furthermore, correlations between pairs of genes in single neurons determined by BARseq2 were consistent with scRNA-seq using 10x v3 to a similar extent as two independent 10x v3 experiments (Fig. 2i–k, Extended Data Fig. 5a,b and Methods). These results indicate that the single-cell gene expression patterns observed by BARseq2 were comparable to those of scRNA-seq.

We wondered if BARseq2 could detect more genes in parallel, and thus be potentially useful in associating projections with larger

Fig. 2 | Multiplexed detection of mRNAs using BARseq2. **a**, A representative image of rolonies in auditory cortex (from two slices sequenced). The top and the bottom of the cortex are indicated by the blue and red dashed lines, respectively. Scale bar, 100 μm. The inset shows a magnified view of the boxed area. **b**, Low-magnification image of the hybridization cycle, showing the location of the area imaged in **a**. Scale bar, 100 μm. **c**, Representative images of the indicated sequencing cycle and hybridization cycle of the boxed area in **a**. Scale bars, 10 μm. **d**, Violin plots showing the laminar distribution of cadherin expression in neuronal somata. Expression in auditory (green) and motor (brown) cortex is indicated. **e**, Laminar distribution of gene expression as detected by BARseq2 or FISH. Lines indicate means, error bars indicate s.d. values, and dots show individual data points. $n = 2$ slices for BARseq2 and $n = 3$ slices for FISH. **f**, Relative gene expression observed using BARseq2 and in Allen gene expression atlas. Each dot represents the expression of a gene in a 100-μm bin in laminar depth. Gray dots indicate the correlation between data randomized across laminar positions. A linear fit and 95% confidence intervals are shown by the diagonal line and the shaded area. $n = 2$ slices for BARseq2 and $n = 1$ slice for ABA ISH. **g**, Distribution of total read counts per cell in BARseq2 and scRNA-seq in auditory cortex. Only genes used in the panel detected by BARseq2 were included. **h**, Mean expression for each gene detected using BARseq2 or scRNA-seq. Each dot represents a gene. The dotted line indicates equal expression between BARseq2 and scRNA-seq. **i**, The correlations between pairs of genes observed in BARseq2 and scRNA-seq (purple dots), or in two scRNA-seq datasets (blue dots). **j**, Expression of *Slc17a7* and *Gad1* in single neurons. Neurons dominantly expressed *Slc17a7* (blue) or *Gad1* (red), or expressed both strongly (gray). **k**, Exclusivity indices (Methods) of *Slc17a7* and *Gad1* in neurons in two scRNA-seq datasets, BARseq2 in auditory or motor cortex, and shuffled BARseq2 data.

gene panels. Because BARseq2 imaging time scales logarithmically with the number of genes detected (Fig. 1d), the multiplexing capacity of BARseq2 is not limited by imaging time. Furthermore, targeting up to 65 genes did not significantly affect the detection sensitivity of each gene (Extended Data Fig. 5c and Methods). The detection of this 65-gene panel in motor cortex (Fig. 3a) allowed us to classify neurons to one of nine transcriptomic neuronal

types defined by scRNA-seq²³ (Fig. 3b, Methods and Extended Data Fig. 5d–h). Consistent with previous studies^{3,9}, these transcriptomic neuronal types displayed distinct laminar distributions (Fig. 3b,c and Methods) and cadherin expression (Fig. 3d). Most transcriptomic types were found in the expected layers with the notable exception of L5 PT and L6 IT Car3, which were seen in additional layers (for example, L2/3). These inaccuracies in cell typing likely



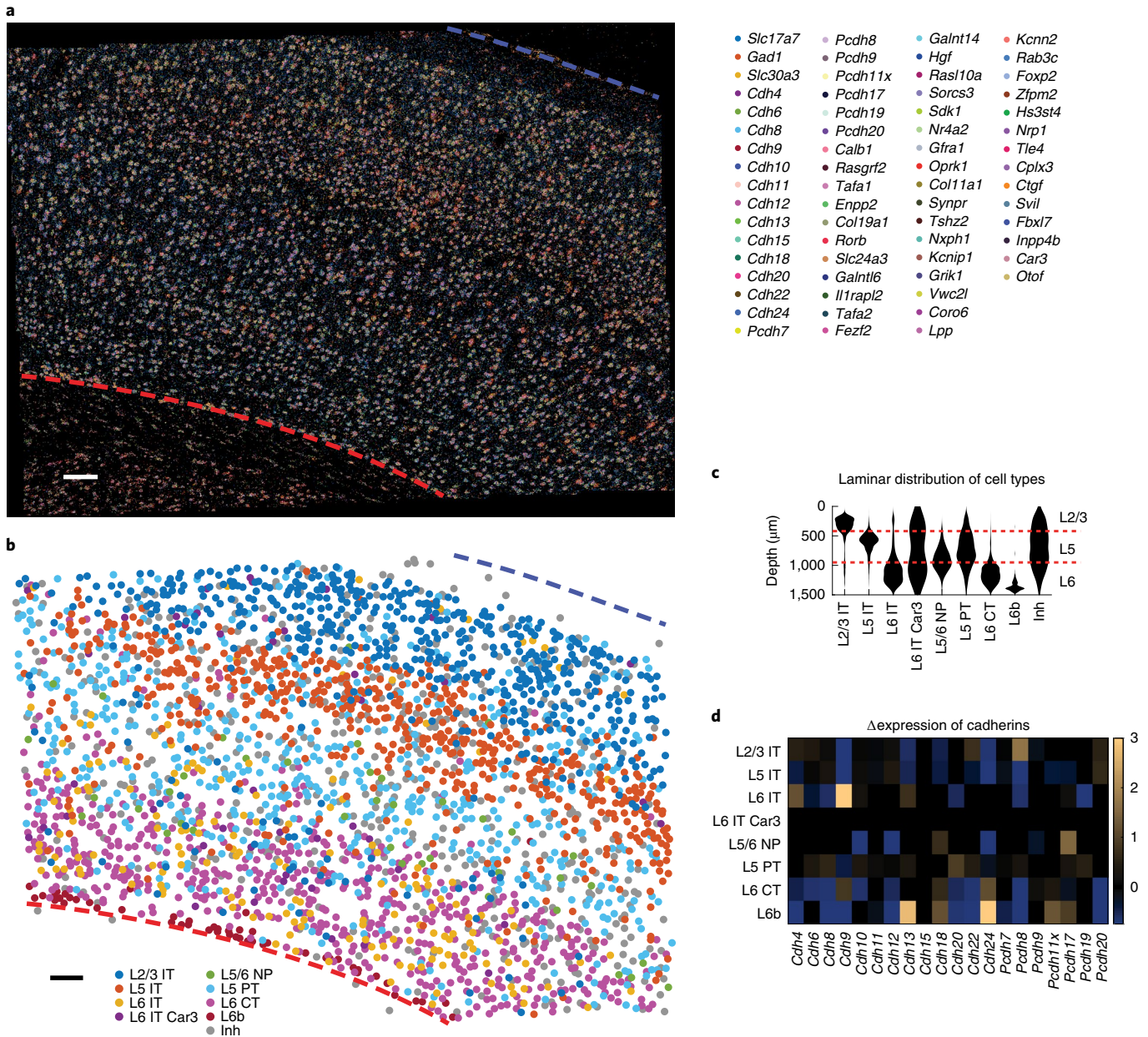


Fig. 3 | Cadherin expression across transcriptomic neuronal types in motor cortex. a, A representative image of rolonies in motor cortex (from four slices sequenced). mRNA identities are color-coded as indicated. The top and the bottom of the cortex are indicated by the blue and red dashed lines, respectively. Scale bar, 100 μm. **b**, Transcriptomic cell types called based on gene expression shown in **a**. **c**, Laminar distribution of transcriptomic neuronal types based on marker gene expression observed by BARseq2. Layer identities are shown on the right. NP, near-projecting. **d**, Differential expression of cadherins across transcriptomic neuronal types identified by BARseq2. Overexpression is indicated in yellow and underexpression is indicated in blue. Only differential expression that was statistically significant is shown. Statistical significance was determined using a two-tailed rank-sum test with Bonferroni correction for each gene between the indicated transcriptomic type and the expression of that gene across all other neuronal types.

resulted from suboptimal choice of marker genes (see Methods for a detailed discussion), and could potentially be improved in the future by optimizing the gene panels. These optimizations, however, were outside the scope of this study. These results demonstrate that BARseq2 can be applied to probe gene panels consisting of high dozens of genes, with minimal decrease in sensitivity and minimal increase in imaging time.

BARseq2 correlates gene expression to projections. Previous studies of the relationship between projection patterns and gene expression have largely focused on revealing the projection

patterns of transcriptomic neuronal types. Although this approach has identified some projection patterns biased in certain transcriptomic types^{6,8}, the diversity of projections in IT neurons remains largely unexplained by transcriptomic types^{3,6}. To further understand the relationship between gene expression and projections, we demonstrate an alternative approach that screens a targeted panel of genes for correlates of diverse projections. This approach relies on the ability of BARseq2 to interrogate both the expression of many genes and projections to many targets simultaneously, and thus would have been difficult to achieve using existing transcriptomic approaches that could only interrogate one or a small number of

projections (for example, Retro-seq^{3,9}) or barcoding-based projection mapping approaches that could only interrogate a small number of genes (for example, BARseq⁶).

As a proof-of-principle study, we examined long-range axonal projections and the expression of 20 cadherins, along with three marker genes, in motor cortex and auditory cortex in three mice. We optimized BARseq2 to detect both endogenous mRNAs and barcodes in the same barcoded neurons without compromising sensitivity (Extended Data Fig. 6a and Methods). In each barcoded cell, we segmented barcoded cell bodies using the barcode sequencing images (Fig. 4a). We then assigned rolonies amplified from endogenous genes that overlapped with these pixels to the barcoded cells (Fig. 4a). This allowed us to map both projection patterns and gene expression (Fig. 4b) in the same neurons. We matched barcodes in these target sites to 3,164 well-segmented barcoded neurons (1,283 from auditory cortex and 1,881 from motor cortex) from 15 slices of auditory cortex and 16 slices of motor cortex, each with 10- μ m thickness. Of the barcoded neurons, 624 and 791 neurons had projections above the noise floor in auditory cortex and motor cortex, respectively. Most neurons (53% (329/624) in auditory and 89% (703/791) in motor cortex) projected to multiple brain areas. We then focused on 598 neurons in auditory cortex and 751 neurons in motor cortex, which also had sufficient endogenous mRNAs detected in each cell, for further analysis (Fig. 4c). These observations were largely consistent with previous BARseq experiments in auditory and motor cortex performed without assessing gene expression^{3,6}, confirming that the modifications for BARseq2 did not compromise projection mapping.

BARseq2 recapitulates known projection biases. Although BARseq2 can read out gene expression and projections in the same neurons, one might be concerned that barcoding neurons using Sindbis virus could disrupt gene expression²⁴. To determine the relationship between genes and projections, one would require that the gene–gene relationship in Sindbis-infected single neurons reflects that in noninfected neurons, and that any change in absolute gene expression level would have little effect. Reassuringly, previous reports have shown that the relationship among genes in single neurons is indeed largely preserved despite a reduction in the absolute expression of genes in Sindbis-infected cells^{6,25}. Furthermore, correlations between transcriptomic types and projections revealed in Sindbis-infected neurons were corroborated by other methods that did not require Sindbis infection^{6,26}. In agreement with these previous reports, we observed that the correlations between pairs of genes in the barcoded neurons were consistent with those in non-barcoded neurons despite an overall reduction in gene expression (Extended Data Fig. 6b–f and Methods). Therefore, the relationship between gene expression and projections resolved by BARseq2 likely reflects that in non-barcoded neurons.

To further test whether BARseq2 can capture the relationship between gene expression and projections, we asked if we could identify differences in projection patterns across transcriptomic neuronal types that could also be validated by previous studies and/or other experimental techniques. We performed these validation

analyses at three different levels of granularity. First, BARseq2 confirmed that most barcoded neurons with long-range projections were excitatory, not inhibitory; whereas about 8–9% of all barcoded neurons were inhibitory (100 of 1,047 in auditory cortex and 140 of 1,689 in motor cortex; Fig. 4d), only 7 of 240 (3%) inhibitory neurons (5 in auditory cortex and 2 in motor cortex) had detectable projections (Fig. 4e, Methods and Extended Data Fig. 6g,h). Second, BARseq2 identified many cadherins (8 for auditory cortex and 12 for motor cortex) that were differentially expressed across IT, PT and corticothalamic (CT) neurons²⁷ (Fig. 5a–d); the differential expression of these genes was consistent with the expression observed by scRNA-seq³ (Extended Data Fig. 7a and Methods). Finally, BARseq2 confirmed known biases in projection patterns across transcriptionally defined IT subtypes in auditory cortex (Extended Data Fig. 7b,c and Methods). Thus, BARseq2 recapitulated known projection differences across transcriptomic subtypes of IT neurons.

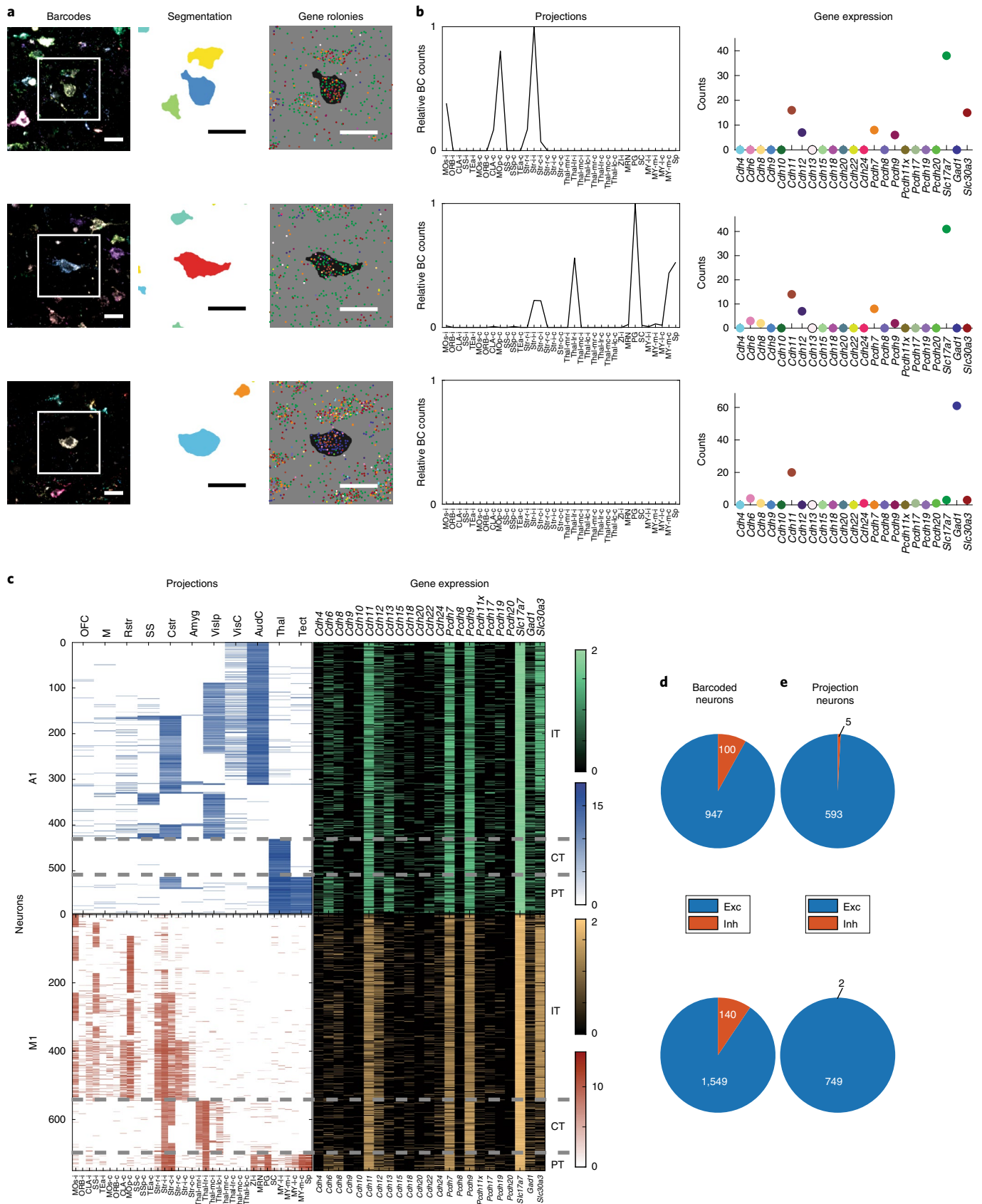
BARseq2 identifies cadherin correlates of IT projections. Having established that BARseq2 identified gene correlates of projections that were consistent with previous studies, we then asked whether cadherin expression correlates with projection patterns within the IT class of neurons. Although cadherins and other cell adhesion molecules are involved in projection specification and axonal growth during development^{16,28}, many take on other functions unrelated to projection specification in later developmental stages^{29,30}. In addition, other mechanisms such as axonal pruning could further shape the projection patterns of neurons independent of initial genetic programs. Therefore, any correlation between cadherins and projections is likely a remnant, or ‘echo,’ of the developmental program that initially specified projections, and may thus be weak and further obscured by gene expression associated with later developmental stages. To overcome the challenges of identifying potentially weak relationships between gene expression and projections, we used BARseq2 to identify correlations between projections and cadherins using a module-based strategy inspired by similar approaches in transcriptomics³¹. Projection modules and gene modules average over the noise in the measurement of individual projections and genes, respectively, and are thus easier to detect when there is considerable biological and/or technical noise in the measurements. This approach requires knowing the projections to many brain areas from individual neurons, a unique advantage of barcoding-based projection mapping techniques (that is, BARseq and BARseq2) compared to retrograde labeling-based approaches^{3,9}. Next, we identify modest associations between cadherin expression and projections in IT neurons, including several associated pairs of cadherins/projections that were shared across cortical areas.

The projections of an IT neuron to its targets are not random. Rather, in both auditory cortex and motor cortex, these projections are organized and show statistical regularities that can be uncovered within the large datasets obtained by BARseq^{2,6} (Fig. 6a). For example, neurons in the auditory cortex that projected to the somatosensory cortex were also more likely to project to the ipsilateral visual cortex, but not the contralateral auditory cortex. To

Fig. 4 | Correlating gene expression to projections using BARseq2. **a**, False-colored barcode sequencing images, soma segmentations and gene rolonies of three representative neurons from the motor cortex. The segmentation and gene rolonie images correspond to the white squared area in the barcode images. In the gene rolonie images, the areas corresponding to the soma segmentations of the target neurons are in black. Scale bars, 20 μ m. **b**, Projections and gene expression of the target neurons shown in **a**. The dots indicating gene expression use the same color coding as in **a**. The neurons shown in the first two rows are excitatory projection neurons, whereas the neuron shown in the bottom row is an inhibitory neuron without projections. See Supplementary Table 2 for the brain areas corresponding to each abbreviated target area. BC, barcode. **c**, Projections and gene expression of neurons in auditory cortex (A1) and motor cortex (M1). Each row represents a barcoded projection neuron. Both projections and gene expression are shown in log scale. Major projection neuron classes determined by projection patterns are indicated on the right. **d,e**, The number of excitatory (exc) or inhibitory (inh) neurons in all barcoded neurons (**d**) or barcoded projection neurons (**e**). Neurons in auditory cortex are shown in the top row and those in motor cortex are shown in the bottom row.

exploit these correlations, we used nonnegative matrix factorization (NMF)³² to represent the projection pattern of each neuron as the sum of several ‘projection modules.’ (NMF is an algorithm related

to principal-component analysis, but imposes the added constraint that projections are nonnegative). Each of these modules (six modules for the motor cortex and three for the auditory cortex; Fig. 6b)



consisted of subsets of projections that were likely to co-occur. We named these modules by the main projections (cortex (CTX) or striatum (STR)) followed by the side of the projection (ipsilateral (I) or contralateral (C)). For some modules, we further indicated that the projections were to the caudal part of the structure by prefixing with 'C' (for example, CSTR-I or CCTX-I). A small number of projection modules could explain most of the variance in projections (three modules and six modules explained 84% and 87% of the variance in projections to nine areas in auditory cortex and 18 areas in motor cortex that IT neurons project to, respectively; Fig. 6c).

Because both the projection patterns of neurons^{2,27} and their transcriptomic types^{3,9} are well correlated with laminae, we first asked how well cadherins explained the diversity of projections in IT neurons compared to the laminar positions of neurons (Methods). Although most cadherins had no predictive power on the projection modules, some individual cadherins could explain a substantial fraction of the variance in projections compared to that explained by the laminar positions of neurons (Extended Data Fig. 8). For example, *Cdh13* and *Pcdh7* explained $6.0\% \pm 0.3\%$ and $7.0\% \pm 0.3\%$ (mean \pm s.d.) of the variations in CTX-C in auditory cortex, compared to $19.4 \pm 0.3\%$ (mean \pm s.d.) explained by the laminar positions of neurons. Strikingly, *Pcdh19* and *Pcdh7* were predictive of CSTR-I in auditory cortex, whereas the laminar positions were not. These results indicate that some but not all cadherins were modestly predictive of projections, and that the predictive power of these cadherins could be comparable in magnitude to the laminar positions of neurons, one of the strongest known predictors of projection patterns.

To further understand how cadherin expression relates to projections, we examined how it covaried with projection modules (Supplementary Fig. 1). Interestingly, the expression of several cadherins covaried with similar projection modules in both cortical areas. For example, auditory cortex neurons expressing *Pcdh19* were stronger in the CSTR-I projection module than those not expressing *Pcdh19* (Fig. 6d; $P=5 \times 10^{-4}$ comparing the CSTR-I module in neurons with ($n=83$) or without ($n=346$) *Pcdh19* expression using rank-sum test); the same association between *Pcdh19* and the CSTR-I projection module was also seen in motor cortex (Fig. 6d; $P=4 \times 10^{-6}$ using rank-sum test, $n=31$ for *Pcdh19*⁺ neurons and $n=512$ for *Pcdh19*⁻ neurons). Similarly, *Cdh8* was correlated with the CTX-I module and *Cdh12* was correlated with the CTX-C module (Fig. 6e; false discovery rate (FDR) < 0.1) in both auditory and motor cortex. These correlations were independently validated by retrograde tracing using cholera toxin subunit B (CTB) and FISH (Extended Data Fig. 9a–e and Methods). *Pcdh19*, together with *Cdh8* or *Cdh11*, correlated with CTX-I or CSTR-I modules, respectively, in motor cortex (Fig. 6e and Extended Data Fig. 8), consistent with a potential combinatorial nature of cadherin correlates of projections. Although the correlations between individual cadherins and projections were relatively modest, our observations that the same cadherins correlated with similar projection modules in both areas suggest that a common molecular logic might underscore the

organization of projections across cortical areas beyond class-level divisions.

Analyses based on the expression of single genes suffer from biological and technical noise of gene expression in single neurons. We reasoned that the correlations among genes might allow us to identify additional relationships between gene expression and projections that were missed by analyzing each gene separately. This ability to leverage the relationship among genes represents an advantage of BARseq2 over the original BARseq because of the improved capacity of BARseq2 for multiplexed gene detection. To exploit the correlations among genes, we grouped 16 cadherins into three meta-analytic coexpression modules based on seven scRNA-seq datasets of IT neurons in motor cortex (Fig. 7a and Extended Data Fig. 10a,b)²³. To obtain the modules, we followed the rank-based network aggregation procedure defined by Ballouz et al.³³ and Crow et al.³⁴ to combine the seven dataset-specific gene–gene coexpression networks into an aggregated network, and then grouped together genes showing consistent excess correlation using the dynamic tree-cutting algorithm³¹. Two coexpressed modules were associated with projections: module 1 was associated with contralateral striatal projections (STR-C projection module), and module 2 was associated with ipsilateral caudal striatal projections (CSTR-I; Fig. 7b,c and Extended Data Fig. 10c,d). These associations between the coexpression modules and projections were consistent with, but stronger than, associations between individual genes contained in each module and the same projections (Extended Data Fig. 10e). Interestingly, these coexpression modules were enriched in multiple transcriptomic subtypes of IT neurons, but these transcriptomic subtypes were found in multiple branches of the transcriptomic taxonomy (Fig. 7d and Extended Data Fig. 10f). For example, module 1 was associated with transcriptomic subtypes of IT neurons in L2/3, L5 and L6. This result is consistent with previous observations^{3,6} that first-tier transcriptomic subtypes of IT neurons (that is, subtypes of the highest level in the transcriptomic taxonomy within the IT class) appeared to share projection patterns, and further raises the possibility that transcriptomic taxonomy does not necessarily capture differences in projections. Taken together, our finding that projections correlate with cadherin coexpression modules independent of transcriptomic subtypes demonstrates that BARseq2 can reveal intricate relationships between gene expression and projection patterns.

Discussion

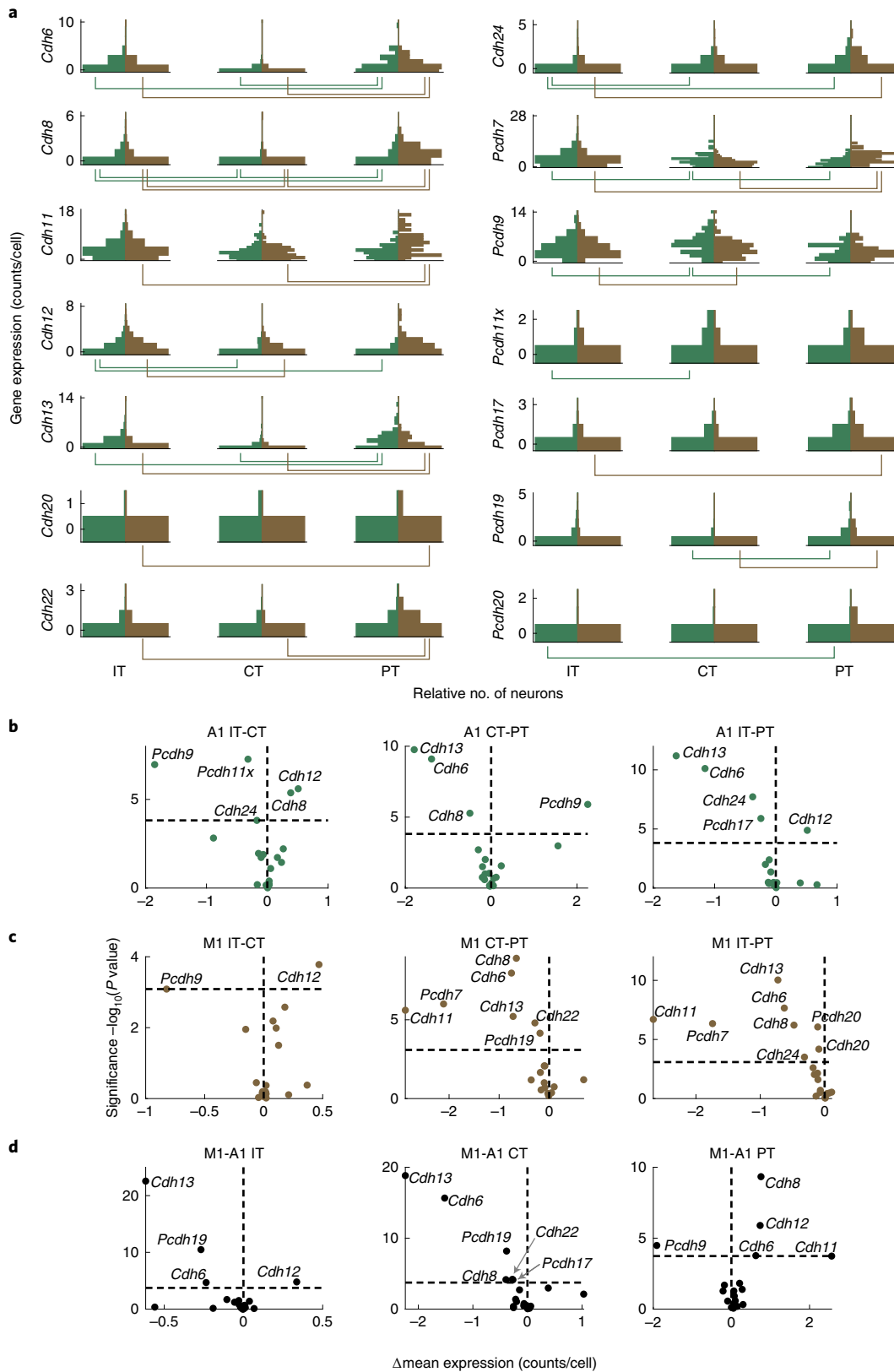
BARseq2 combines high-throughput mapping of projections to many brain areas with multiplexed detection of gene expression at single-cell resolution. Because BARseq2 is high throughput, we are able to correlate gene expression and projection patterns of thousands of individual neurons in a single experiment, and thereby achieve statistical power that would be challenging to obtain using other single-cell techniques. By applying BARseq2 to two distant cortical areas—primary motor and auditory cortex—in the adult mouse, we identified cadherin correlates of diverse projections. Our results suggest that BARseq2 provides a path to discovering

Fig. 5 | Differential cadherin expression across major classes and cortical areas. **a**, Vertical histograms of the expression (raw counts per cell) of cadherins that were differentially expressed across major classes in either auditory or motor cortex. *y* axes indicate gene expression level (counts per cell) and *x* axes indicate number of neurons at that expression level. The numbers of neurons are normalized across plots so that the bins with the maximum number of neurons have equal bar lengths. In each plot, gene expression in auditory cortex (green) is shown on the left, and gene expression in motor cortex (brown) is shown on the right. Lines beneath each plot indicate pairs of major classes with different expression of the gene (FDR < 0.05). **b,c**, Volcano plots of cadherins that were differentially expressed across pairs of major classes in auditory cortex (**b**) or motor cortex (**c**). *y* axes indicate significance and *x* axes indicate effect size. The horizontal dashed lines indicate significance level for FDR < 0.05 , and the vertical dashed lines indicate equal expression. **d**, Volcano plots of cadherins that were differentially expressed across auditory and motor cortex in the indicated major classes. *y* axes indicate significance and *x* axes indicate effect size. Gene identities for points close together are noted with gray arrows for clarity. The horizontal dashed lines indicate significance level for FDR < 0.05 , and the vertical dashed lines indicate equal expression. For all panels, *P* values were calculated using two-tailed rank-sum tests.

the general organization of gene expression and projections that are shared across the cortex.

High-throughput and multiplexed gene detection by BARseq2.
To correlate panels of genes to projections, we designed BARseq2

to detect gene expression with high throughput, for multiplexing to dozens of genes, to have sufficient sensitivity, and be compatible with barcoding-based projection mapping. To satisfy these needs, we based BARseq2 on padlock probe-based approaches^{10,11}. With additional optimizations for sensitivity, sequencing readout



and compatibility with barcode sequencing, we successfully used BARseq2 to identify gene correlates of projections.

One of the critical requirements for BARseq2 is high throughput when reading out many genes. Through strong amplification of mRNAs, combinatorial coding and robust readout using Illumina sequencing chemistry^{6,35}, BARseq2 achieves fast imaging at low optical resolution compared to many other imaging-based spatial transcriptomic methods^{14,36}. Further optimizations, including computational approaches for resolving spatially mixed colonies³⁷, have the potential to increase imaging throughput even further. Although the gene multiplexing capacity of BARseq2 may ultimately be limited by other physical constraints, such as crowding of colonies and reduced detection sensitivity, these factors are unlikely to be limiting when multiplexing to dozens to hundreds of genes¹¹.

Another critical optimization was increasing the low sensitivity that early versions of the padlock probe-based technique was susceptible to, unless special and expensive primers were used¹⁰. Inspired by other spatial transcriptomic methods, we and others¹¹ have found that tiling target genes with multiple probes could greatly improve the sensitivity. This design allowed variable sensitivity for different experimental purposes. Although in the present work we identified cadherin correlates of projections using only a modest number of probes per gene to achieve sensitivity similar to scRNA-seq using 10x Genomics v3, the sensitivity of BARseq2 can be considerably higher when more probes are used (Fig. 1e). This high and tunable sensitivity, combined with the fact that the gene multiplexing capacity of BARseq2 is not limited by imaging time, opens potential application of BARseq2 to a wide range of questions that require high-throughput interrogation of gene expression in situ.

BARseq2 reveals gene correlates of projections. BARseq2 exploits the high-throughput axonal projection mapping that BARseq offers to identify gene correlates of diverse projections. BARseq has sensitivity comparable to single-neuron tracing⁵. Although the spatial resolution of BARseq for projections is lower than that of conventional single-neuron tracing, it offers throughput that is several orders of magnitude higher than the state-of-the-art single-cell tracing techniques^{1,2}. This high throughput allows BARseq to reveal higher-order statistical structure in projection patterns that would have been difficult to observe using existing techniques, such as single-cell tracing^{5,6}. The increased statistical power of BARseq, obtained at the cost of some spatial resolution, is reminiscent of different clustering power across scRNA-seq techniques of varying throughput and read depth^{23,38}. The high throughput of BARseq thereby provides a powerful asset for investigating the organization of projection patterns and their relationship to gene expression.

BARseq2 enables simultaneous measurement of multiplexed gene expression and axonal projections to many brain areas, at single-neuron resolution and at a scale that would be difficult to achieve with other approaches. For example, *Cre*-dependent

labeling allows interrogation of the gene expression and projection patterns of a genetically defined subpopulation of neurons⁶. However, this approach lacks cellular resolution, is limited by the availability of *Cre* lines, and requires that a neuronal population of interest be specifically distinguished by the expression of one or two genes. The combination of single-cell transcriptomic techniques with retrograde labeling does provide cellular resolution, but can only interrogate projections to one or at most a small number of brain areas at a time^{3,9}. The inability to interrogate projections to many brain areas from the same neuron would miss higher-order statistical structures in projections, which are nonrandom⁵ and provide additional information regarding other properties of the neurons, such as laminar position and gene expression^{2,6}. The projections of individual neurons to multiple brain areas can be obtained using multiplexed single-cell tracing¹, but the throughput of these methods remains relatively low. Moreover, many advanced single-cell tracing techniques require special sample processing that hinders multiplexed interrogation of gene expression in the same sample. The throughput of single-cell projection mapping was addressed by the original BARseq⁶, but the small number of genes (up to three) that could be co-interrogated with projections limited its use in identifying the general relationship between gene expression and projections. BARseq2 thus addresses limitations of existing techniques and provides a powerful approach for probing the relationships between gene expression and projection patterns.

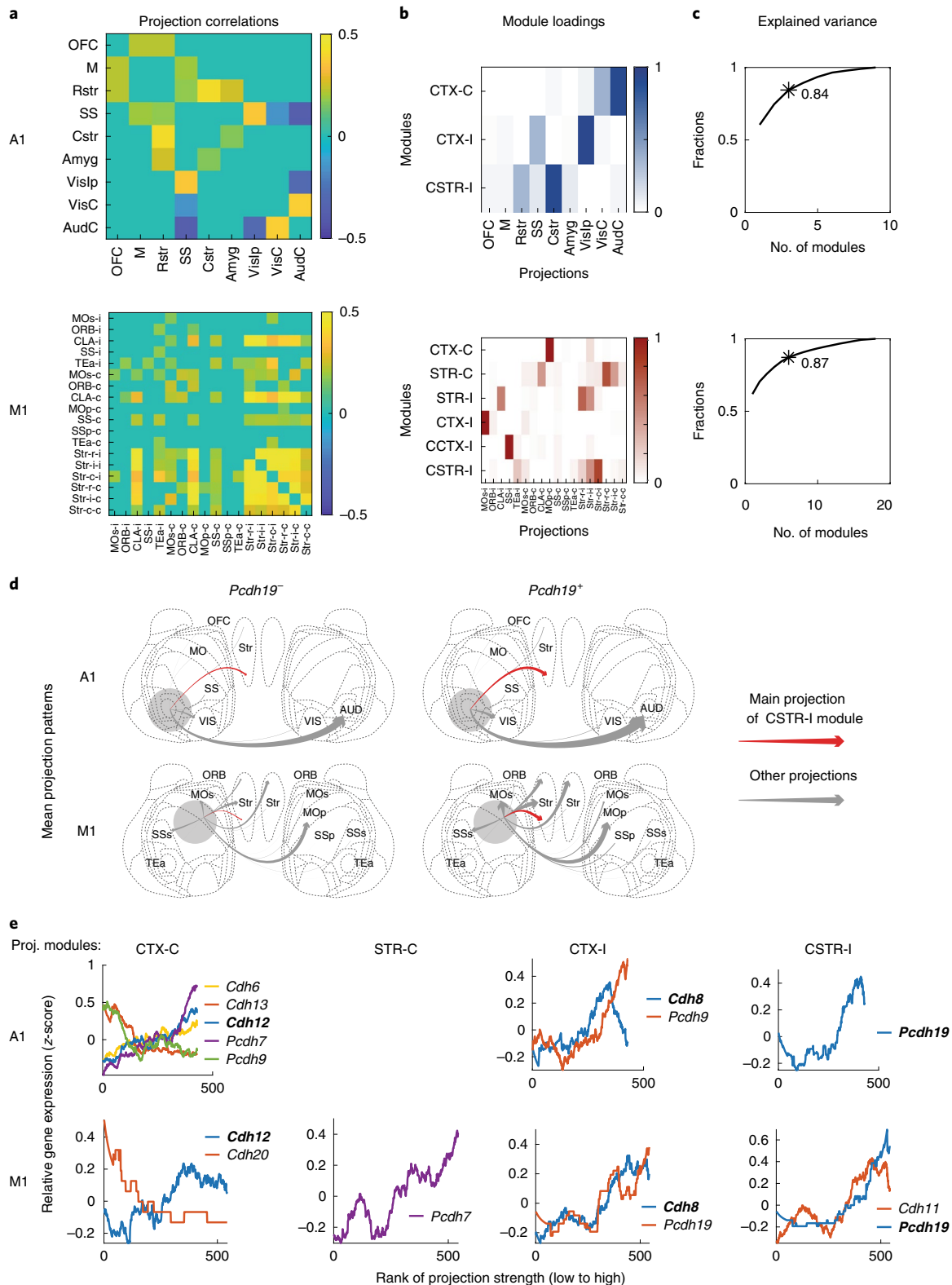
Cadherins correlate with diverse projections of IT neurons. As a proof-of-principle study, we used BARseq2 to identify several cadherins that correlate with homologous IT projections in both auditory and motor cortex, two spatially and transcriptomically distant areas with distinct cortical and subcortical projection targets. In addition, cadherin coexpression modules that correlated with projections were associated with multiple branches of the transcriptomic taxonomy. This type of correlation between neuronal connectivity and variations in gene expression independent of transcriptomic types is not unique to the cortex and has previously been observed in other brain areas, such as the hippocampus³⁹. Therefore, our findings are consistent with the hypothesis that a shared cell adhesion molecule code might underlie the diversity of cortical projections independent of transcriptomic types^{18,39}.

Even though the power of some cadherins to predict projections was comparable in magnitude to that of laminar position, a strong predictor of projection patterns, these cadherins could only explain a small fraction of the overall variance in projections. This noisy association between cadherin expression and projection patterns contrasts with the known roles of cadherins in specifying neuronal connectivity in the cortex and other circuits^{20,40}, but the relatively small magnitude of these associations is not surprising for a few reasons. First, gene expression programs and signaling cues needed for specifying projections are usually transient in development⁴¹, so it is likely that these cadherins only represent the remnants of

Fig. 6 | Cadherins correlate with diverse projections of IT neurons. **a**, Pearson correlation of projections to different brain areas in IT neurons of auditory cortex (A1) or motor cortex (M1). Only significant correlations are shown. OFC, orbitofrontal cortex; M, motor cortex; Rstr, rostral striatum; SS, somatosensory cortex; Cstr, caudal striatum; Amyg, amygdala; Vislp, ipsilateral visual cortex; VisC, contralateral visual cortex; AudC, contralateral auditory cortex. **b**, Projection modules of IT neurons in auditory cortex (top) or motor cortex (bottom). Each row represents a projection module. Columns indicate projections to different brain areas. **c**, The fractions of variance explained by different numbers of projection modules in auditory cortex (top) and motor cortex (bottom). The numbers of projection modules that correspond to those in **b** are labeled with an asterisk with the fraction of variance explained indicated. **d**, Mean projection patterns of neurons in A1 and M1 with or without *Pcdh19* expression. The thickness of arrows indicates projection strength (barcode counts). Red arrows indicate projections that correspond to the strongest projection in the CSTR-I projection modules. ORB, orbitofrontal cortex; MOs, secondary motor cortex; MOp, primary motor cortex; SSp, primary somatosensory cortex; SSs, secondary somatosensory cortex; TEa, temporal association cortex. **e**, The expression of cadherins (*y* axes) that were rank correlated with the indicated projection modules in auditory cortex and motor cortex. Neurons (*x* axes) were sorted by the strengths of the indicated projection modules. Only genes that were significantly correlated with projection modules are shown (FDR < 0.1 using two-tailed rank-sum tests). Genes that were correlated with the same projection modules in both areas are shown in bold.

a common developmental program that establish projections⁴², or may be needed for ongoing functions or maintenance of projections. Second, non-cadherin cell adhesion molecules (for example, IgCAMs^{43,44}) and other cell-surface molecules (for example, plexins, semaphorins⁴⁵ and teneurins⁴⁶) are also involved in specifying projections, so cadherins likely only represent a fraction of the molecular programs that specify projections. Finally, cortical projections

undergo extensive activity-dependent modifications after the initial specification, so the overall diversity in cortical projections is likely much higher than that produced by the initial molecular program. These possibilities can be better resolved by applying BARseq2 to reveal gene expression in both the projection neurons and the areas they project to during development, in combination with perturbation experiments. BARseq2 thus provides a path to discovering



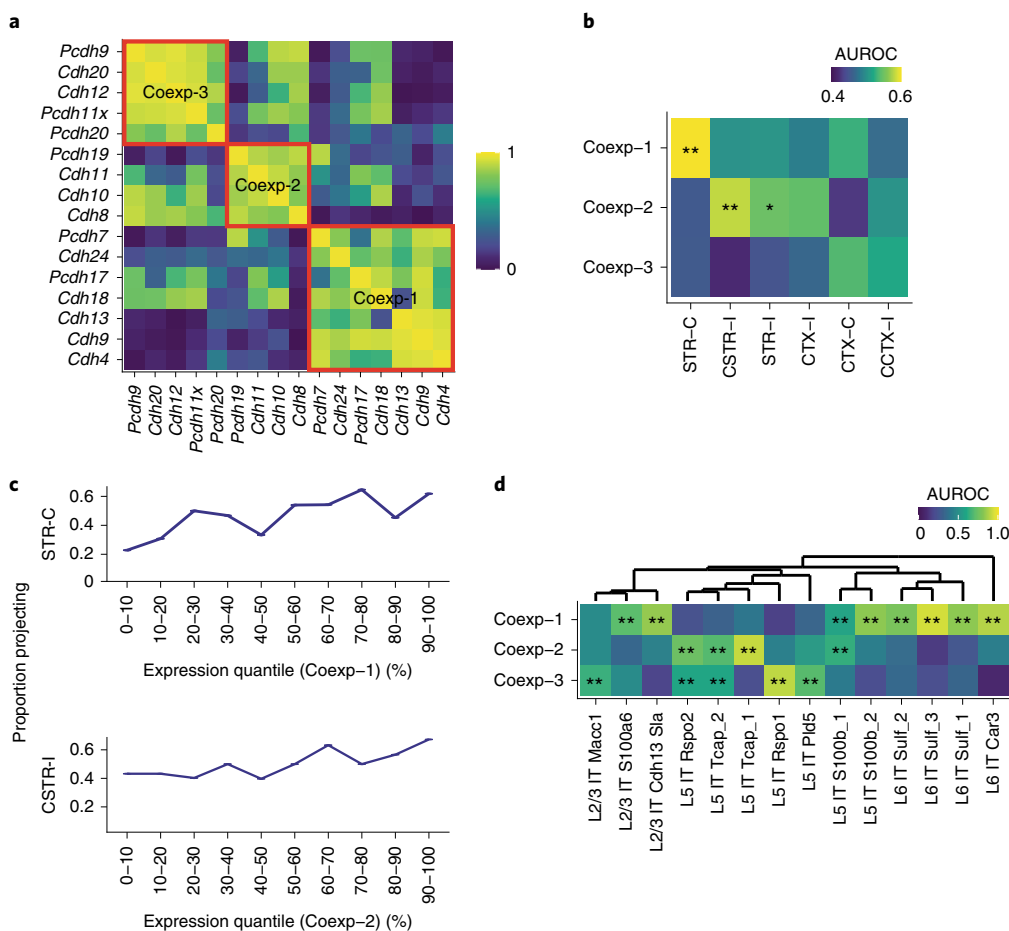


Fig. 7 | Gene coexpression modules correlate with diverse projections of IT neurons. **a**, Correlation among cadherins as identified using scRNA-seq in IT neurons in motor cortex²³. Three coexpression modules are marked by red squares. Cadherins that did not belong to any module are not shown. **b**, Association between cadherin coexpression modules and projection modules (AUROC, area under the receiver-operator characteristic curve). Significant associations are marked by asterisks (*FDR < 0.1, **FDR < 0.05). **c**, Fractions of neurons with the indicated projection modules as a function of coexpression module expression. Neurons are binned by gene module quantiles as indicated. **d**, Association of the three coexpression modules in transcriptomic IT neurons in the single-cell SmartSeq dataset (AUROC, significance shown as in **b**).

the myriad of genetic programs that specify and/or correlate with long-range projections in both developing and mature animals.

BARseq2 builds a unified description of neuronal diversity. Neuronal barcoding was originally proposed as a method for untangling circuit connectivity at synaptic resolution^{47,48}. Solving neuronal connectivity with barcode sequencing not only has the potential to achieve high-throughput and single-cell resolution by exploiting advances in sequencing technology, but also provides a path to integrate measurements of multiple neuronal properties in single neurons—toward the ‘Rosetta brain’⁴⁹. BARseq2 is a step toward this goal. Although BARseq2 currently only resolves projections at relatively low spatial resolution (brain areas, that is hundreds of microns), this limitation can be addressed in the future by using in situ sequencing to read out axonal barcodes (Yuan et al., unpublished data), which would resolve axonal projections at sub-cellular spatial resolution. Further combining in situ sequencing of axonal barcodes with synaptic labeling, expansion microscopy and/or transsynaptic viral labeling could yield information regarding the synaptic connectivity of neurons. Because BARseq2 integrates neuronal properties using spatial information, it is potentially compatible with other in situ assays, such as immunohistochemistry, two-photon calcium imaging and dendritic morphological reconstruction. By spatially correlating various neuronal properties in

single neurons, BARseq2 represents a feasible path toward achieving a comprehensive description of neuronal circuits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-021-00842-4>.

Received: 25 August 2020; Accepted: 19 March 2021; Published online: 10 May 2021

References

1. Winnubst, J. et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell* **179**, 268–281 (2019).
2. Muñoz-Castañeda, R. et al. Cellular anatomy of the mouse primary motor cortex. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.10.02.323154> (2020).
3. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
4. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
5. Han, Y. et al. The logic of single-cell projections from visual cortex. *Nature* **556**, 51–56 (2018).

6. Chen, X. et al. High-throughput mapping of long-range neuronal projection using in situ sequencing. *Cell* **179**, 772–786 (2019).
7. Kim, D. W. et al. Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell* **179**, 713–728 (2019).
8. Economo, M. N. et al. Distinct descending motor cortex pathways and their roles in movement. *Nature* **563**, 79–84 (2018).
9. Zhang, M. et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.06.04.105700> (2020).
10. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
11. Qian, X. et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat. Methods* **17**, 101–106 (2020).
12. Kecsichull, J. M. et al. High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron* **91**, 975–987 (2016).
13. Huang, L. et al. BRICseq bridges brain-wide interregional connectivity to neural activity and gene expression in single animals. *Cell* **182**, 177–188 (2020).
14. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
15. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
16. Hayano, Y. et al. The role of T-cadherin in axonal pathway formation in neocortical circuits. *Development* **141**, 4784–4793 (2014).
17. Friedman, L. G. et al. Cadherin-8 expression, synaptic localization, and molecular control of neuronal form in prefrontal corticostriatal circuits. *J. Comp. Neurol.* **523**, 75–92 (2015).
18. Paul, A. et al. Transcriptional architecture of synaptic communication delineates GABAergic neuron identity. *Cell* **171**, 522–539 (2017).
19. Matsunaga, E., Nambu, S., Oka, M. & Iriki, A. Complex and dynamic expression of cadherins in the embryonic marmoset cerebral cortex. *Dev. Growth Differ.* **57**, 474–483 (2015).
20. Redies, C. Cadherins and the formation of neural circuitry in the vertebrate CNS. *Cell Tissue Res.* **290**, 405–413 (1997).
21. Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
22. Terakawa, Y. W., Inoue, Y. U., Asami, J., Hoshino, M. & Inoue, T. A sharp cadherin-6 gene expression boundary in the developing mouse cortical plate demarcates the future functional areal border. *Cereb. Cortex* **23**, 2293–2308 (2013).
23. Yao, Z. et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.02.29.970558> (2020).
24. Fros, J. J. & Pijlman, G. P. Alphavirus infection: Host cell shut-off and inhibition of antiviral responses. *Viruses* <https://doi.org/10.3390/v8060166> (2016).
25. Klingler, E. et al. Single-cell molecular connectomics of intracortically projecting neurons. Preprint at *bioRxiv* <https://doi.org/10.1101/378760> (2018).
26. Wang, Y. et al. Complete single-neuron reconstruction reveals morphological diversity in molecularly defined claustral and cortical neuron types. Preprint at *bioRxiv* <https://doi.org/10.1101/675280> (2019).
27. Harris, K. D. & Shepherd, G. M. The neocortical circuit: themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
28. Duan, X., Krishnaswamy, A., De la Huerta, I. & Sanes, J. R. Type II cadherins guide assembly of a direction-selective retinal circuit. *Cell* **158**, 793–807 (2014).
29. Friedman, L. G., Benson, D. L. & Huntley, G. W. Cadherin-based transsynaptic networks in establishing and modifying neural connectivity. *Curr. Top. Dev. Biol.* **112**, 415–465 (2015).
30. Jontes, J. D. The cadherin superfamily in neural circuit assembly. *Cold Spring Harb. Perspect. Biol.* **10**, a029306 (2018).
31. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
32. Lee, D. D. & Seung, H. S. Learning the parts of objects by nonnegative matrix factorization. *Nature* **401**, 788–791 (1999).
33. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq coexpression network construction and analysis: safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).
34. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Exploiting single-cell expression to characterize coexpression replicability. *Genome Biol.* **17**, 101 (2016).
35. Chen, X., Sun, Y. C., Church, G. M., Lee, J. H. & Zador, A. M. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* **46**, e22 (2018).
36. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
37. Chen, S. et al. BARcode DEMixing through Non-negative Spatial Regression (BarDensr). *PLoS Comput. Biol.* **17**, e1008256 (2021).
38. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
39. Harris, K. D. et al. Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS Biol.* **16**, e2006387 (2018).
40. Duan, X. et al. Cadherin combinations recruit dendrites of distinct retinal neurons to a shared interneuronal scaffold. *Neuron* **99**, 1145–1154 (2018).
41. Li, H. et al. Classifying *Drosophila* olfactory projection neuron subtypes by single-cell RNA sequencing. *Cell* **171**, 1206–1220 (2017).
42. Custo Greig, L. F., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat. Rev. Neurosci.* **14**, 755–769 (2013).
43. Bagri, A. et al. Slit proteins prevent midline crossing and determine the dorsoventral position of major axonal pathways in the mammalian forebrain. *Neuron* **33**, 233–248 (2002).
44. Shu, T., Sundaresan, V., McCarthy, M. M. & Richards, L. J. Slit2 guides both precrossing and postcrossing callosal axons at the midline in vivo. *J. Neurosci.* **23**, 8176–8184 (2003).
45. Yoshida, Y. Semaphorin signaling in vertebrate neural circuit assembly. *Front. Mol. Neurosci.* **5**, 71 (2012).
46. Berns, D. S., DeNardo, L. A., Pederick, D. T. & Luo, L. Teneurin-3 controls topographic circuit assembly in the hippocampus. *Nature* **554**, 328–333 (2018).
47. Zador, A. M. et al. Sequencing the connectome. *PLoS Biol.* **10**, e1001411 (2012).
48. Peikon, I. D. et al. Using high-throughput barcode sequencing to efficiently map connectomes. *Nucleic Acids Res.* **45**, e115 (2017).
49. Marblestone, A. H., et al. Rosetta brains: a strategy for molecularly-annotated connectomics. Preprint at *arXiv* <https://arxiv.org/abs/1404.5103> (2014).
50. Eng, C. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Animal processing and tissue preparation. All animal procedures were carried out in accordance with the Institutional Animal Care and Use Committee (protocol no. 19-16-10-07-03-00-4) at Cold Spring Harbor Laboratory. The animals were housed at maximum of five in a cage on a 12-h on/12-h off light cycle. The temperature in the facility was kept at 22°C with a range not exceeding 20.5°C to 26°C. Humidity was maintained at around 45–55%, not exceeding a range of 30–70%. A list of animals used is provided in Supplementary Table 1.

For samples used for only endogenous mRNA detection, 8- to 10-week-old male C57BL/6 mice were anesthetized and decapitated. We immediately embedded the brain in optimal cutting temperature (OCT) compound in a 22-mm² cryomold and snap froze the tissue in an isopentane bath submerged in liquid nitrogen. Sections were cut into 10- μ m-thick slices on Superfrost Plus Gold Slides (Electron Microscopy Sciences). Unlike in the original BARseq, the sections were directly melted onto slides without the use of a tape-transfer system. This change in mounting methods allowed increased efficiency in gene detection. The slides were stored at –80°C until use.

For BARseq2 samples, 8- to 10-week-old male C57BL/6 mice were injected as indicated in Supplementary Table 1. After 24 h, we anesthetized and decapitated the animal, punched out the injection site and snap froze the rest of the brain on a razor blade on dry ice for conventional MAPseq⁶. The injection site was embedded, cryosectioned and stored as described above.

To prepare samples for BARseq2 experiments, we immersed slides from –80°C instantly into freshly made 4% paraformaldehyde (10-ml vials of 20% PFA; Electron Microscopy Sciences) in PBS for 30 min at room temperature. We washed the samples in PBS for 5 min before installing HybriWell-FL chambers (22 mm \times 22 mm \times 0.25 mm; Grace Bio-Labs) for subsequent reactions on the samples. We then dehydrated the samples in 70%, 85% and 100% ethanol for 5 min each, followed by washing in 100% ethanol for at least 1 h at 4°C. Finally, we rehydrated the samples in PBST (0.5% Tween-20 in PBS).

For retrograde labeling experiments, we prepared 1.0 mg ml⁻¹ of CTB in PBS from 100 μ g for injections (see Supplementary Table 1 for a list of animals and coordinates used). We perfused the animals with fresh 4% PFA 96 h after injection, post-fixed for 24 h in 4% PFA, and cryoprotected in 10% sucrose in PBS for 12 h, 20% sucrose in PBS for 12 h and 30% sucrose in PBS for 12 h. The brain was then frozen in OCT and cryosectioned into 20- μ m slices using a tape-transfer system.

BARseq2 detection of endogenous genes. We prepared a master mix of reverse transcription primers at 0.5 μ M each for all target mRNAs. For volumes exceeding the amount required for reverse transcription, we speed-vacuum concentrated the primer mix into a smaller volume. We then prepared the reaction (0.5 μ M per gene of RT primer (IDT), 1 U μ l⁻¹ RiboLock RNase inhibitor (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 500 μ M dNTPs (Thermo Fisher Scientific), 20 U μ l⁻¹ RevertAid H-Minus M-MuLV reverse transcriptase (Thermo Fisher Scientific) in 1 \times RT buffer). We incubated the samples in the reverse transcription mixture at 37°C overnight. After reverse transcription, we cross-linked the cDNAs in 50 mM BS(PEG)₉ (Thermo Fisher Scientific) for 1 h and neutralized excess cross-linker with 1 M Tris-HCl at pH 8.0 for 30 min, and then washed the sample with PBST twice to eliminate excess Tris buffer. We then prepared a master padlock mix with 200 nM per padlock probe for each target mRNA and speed-vacuum concentrated the mixture for a higher concentration at a smaller volume, if necessary. We ligated the gene padlock probes on the cDNA (200 nM per gene padlock (IDT), 1 U μ l⁻¹ RiboLock RNase Inhibitor, 20% formamide (Thermo Fisher Scientific), 50 mM KCl, 0.4 U μ l⁻¹ RNase H (Qiagen) and 0.5 U μ l⁻¹ Ampligase (Epicentre) in 1 \times Ampligase buffer) for 30 min at 37°C and 45 min at 45°C. Finally, we performed rolling circle amplification (RCA; 125 μ M amino-allyl dUTP (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 250 μ M dNTPs, 5% glycerol and 1 U μ l⁻¹ ϕ 29 DNA polymerase (Thermo Fisher Scientific) in 1 \times ϕ 29 DNA polymerase buffer) overnight at room temperature. After RCA, we again cross-linked the colonies in 50 mM BS(PEG)₉ for 1 h, neutralized with 1 M Tris-HCl at pH 8.0 for 30 min and washed with PBST. We washed the sample in hybridization buffer (10% formamide in 2 \times SSC) and then either added probe detection hybridization solution (0.25 μ M fluorescent probe in hybridization buffer) or gene sequencing primer hybridization solution (1 μ M of sequencing primer in hybridization buffer) for 10 min at room temperature. We then washed the sample with hybridization buffer three times at 2 min each, rinsed the sample in PBST twice, and proceeded to imaging or continued with Illumina sequencing.

BARseq2 simultaneous detection of endogenous genes and barcodes. We prepared a master mix of reverse transcription primers at 0.5 μ M each for all target mRNAs. For volumes exceeding the amount required for reverse transcription, we speed-vacuum concentrated the primer mix into a smaller volume. We then prepared the reaction (0.5 μ M per gene RT primer (IDT), 1 μ M barcode LNA RT primer (Qiagen), 1 U μ l⁻¹ RiboLock RNase inhibitor (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 500 μ M dNTPs (Thermo Fisher Scientific), 20 U μ l⁻¹ RevertAid H-Minus M-MuLV reverse transcriptase (Thermo Fisher Scientific) in 1 \times RT buffer), adding the barcode LNA primer last into the reaction mix to reduce cross-hybridization due to the LNA strong binding affinity. We incubated the samples in the reverse transcription mixture at 37°C overnight. After reverse

transcription, we cross-linked the cDNAs in 50 mM BS(PEG)₉ (Thermo Fisher Scientific) for 1 h and neutralized excess cross-linker with 1 M Tris-HCl at pH 8.0 for 30 min, and then washed the sample with PBST twice to eliminate excess Tris buffer. We then prepared a master padlock mix with 200 nM per padlock probe for each target mRNA and speed-vacuum concentrated the mixture for a higher concentration at a smaller volume, if necessary. We ligated the gene padlock probes on the cDNA (200 nM per gene padlock (IDT), 1 U μ l⁻¹ RiboLock RNase Inhibitor, 20% formamide (Thermo Fisher Scientific), 50 mM KCl, 0.4 U μ l⁻¹ RNase H (Qiagen) and 0.5 U μ l⁻¹ Ampligase (Epicentre) in 1 \times Ampligase buffer) for 30 min at 37°C and 45 min at 45°C. After ligating padlock probes for our target genes, we ligated the padlock probe for the barcode cDNA (100 nM barcode padlock (IDT), 50 μ M dNTPs, 5% glycerol, 1 U μ l⁻¹ RiboLock RNase Inhibitor, 20% formamide (Thermo Fisher Scientific), 50 mM KCl, 0.4 U μ l⁻¹ RNase H (Qiagen), 0.001 U μ l⁻¹ Phusion DNA polymerase (NEB) and 0.5 U μ l⁻¹ Ampligase (Epicentre) in 1 \times Ampligase buffer) without any wash in between, and incubated the reaction for 5 min at 37°C and 40 min at 45°C. We then washed the sample twice with PBST and once with hybridization buffer (10% formamide in 2 \times SSC), before hybridizing 1 μ M of RCA primer in hybridization buffer for 15 min at room temperature. We washed the sample with hybridization buffer three times at 2 min each. Finally, we performed RCA (125 μ M aadUTP (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 250 μ M dNTPs, 5% glycerol and 1 U μ l⁻¹ ϕ 29 DNA polymerase (Thermo Fisher Scientific) in 1 \times ϕ 29 DNA polymerase buffer) overnight at room temperature. After RCA, we again cross-linked the colonies in 50 mM BS(PEG)₉ for 1 h, neutralized with 1 M Tris-HCl at pH 8.0 for 30 min, and washed with PBST. We washed the sample in hybridization buffer (10% formamide in 2 \times SSC) and then added gene sequencing primer hybridization solution (1 μ M of sequencing primer in hybridization buffer) for 10 min at room temperature. We then washed the sample with hybridization buffer three times at 2 min each, rinsed the sample in PBST twice and proceeded to Illumina sequencing.

In situ sequencing of endogenous genes. To sequence the endogenous genes using Illumina sequencing chemistry, we used the HiSeq Rapid SBS Kit v2 reagents to reduce cost from the original sequencing protocol⁶. For the first cycle, we incubated samples in universal sequencing buffer (USB) at 60°C for 3 min, then washed in PBST, followed by incubation in iodoacetamide (9.3 mg in 2 ml PBST) at 60°C for 3 min. We washed the sample in PBST again, rinsed with USB twice more, and then incubated in incorporation mix (IRM) at 60°C for 3 min. We repeated the IRM step again to ensure the reaction was as close to 100% complete as possible. We then washed the sample in PBST once and then continued to wash in PBST four more times at 60°C for 3 min each time. To reduce bleaching during imaging, we imaged the sample in universal scan mix (USM).

For subsequent cycles, we first washed samples in USB, then incubated in cleavage reagent master mix (CRM) at 60°C for 3 min. We repeated the CRM step to ensure complete reaction and washed out residual CRM twice with cleavage wash mix (CWM). We then washed the sample with USB, and then with PBST, before incubating in iodoacetamide at 60°C for 3 min. We repeated this step again to ensure we blocked as many of the free thiol groups as possible to reduce background. We then continued with IRM and PBST washes as described for the first cycle and imaged after each cycle. We performed four sequencing cycles and seven sequencing cycles in total for our cadherins panel of 23 genes and our motor cell-type markers and cadherins panel of 65 genes, respectively.

To visualize high expressors, we cleaved the fluorophores in the last sequencing cycle and washed the sample with CWM and PBST. We then washed our sample in hybridization buffer and added probe detection solution (0.5 μ M for each probe in hybridization buffer) for four different fluorescent probes detecting *Slc17a7*, *Gad1*, *Slc30a3* and all previously sequenced genes, respectively, for 10 min at room temperature. We washed the sample in the same hybridization buffer three times for 2 min each, washed in PBST, before adding DAPI stain (ACDBio) for 2 min at room temperature. We rinsed in PBST again and finally in USM for imaging.

In situ sequencing of barcodes. After sequencing and hybridizing for endogenous genes as described above, we stripped the sample of all hybridized oligonucleotides and sequenced bases by incubating twice in strip buffer (40% formamide in 2 \times SSC with 0.01% Triton-X) at 60°C for 10 min. We washed with PBST, then washed with hybridization buffer, and then incubated samples in barcode sequencing primer hybridization solution (1 μ M sequencing primer in hybridization buffer) for 10 min at room temperature. We washed with hybridization solution three times for 2 min each, before rinsing twice in PBST. We sequenced barcodes with the same sequencing procedure as described for endogenous genes but for 15 cycles in total. At around cycle 4 or 5, we eliminate the iodoacetamide blocker incubation for the rest of sequencing because iodoacetamide blockage is irreversible, so further incubation in this blocker becomes unnecessary after several cycles.

Target area barcode sequencing. Barcode sequencing in target brain areas was performed by the Cold Spring Harbor Laboratory MAPseq Core following procedures used in a previous study⁶. The target areas were dissected to match two other studies in A1 (ref. ⁹) and in M1 (ref. ³) resulting in 11 and 35 projection targets for neurons in auditory cortex and motor cortex, respectively; these projection targets corresponded to most of the major projection targets based on

bulk tracing⁵¹. A detailed description of each dissected area and correspondence to the Allen reference atlas are shown in Supplementary Table 2.

Fluorescence in situ hybridization. FISH experiments were performed using RNAscope Fluorescent Multiplex Kit v1 according to the manufacturer's protocols with minor modifications to sample preprocessing. For FISH experiments in comparison to BARseq2 endogenous mRNA detection (Figs. 1f and 2e), the samples were fresh frozen in an isopentane bath as described above. From -80°C storage, the samples were immediately submerged in freshly prepared 4% PFA (Electron Microscopy Sciences) for 15 min at 4°C , then dehydrated in 75%, 85% and 100% ethanol twice for 5 min each. After air-drying, we assembled HybriWell-FL chambers (22 mm \times 22 mm \times 0.25 mm; Grace Bio-Labs) and digested the samples in Protease IV for 30 min at room temperature. We washed the samples in PBST, and then proceeded with probe hybridization and subsequent amplification and visualization steps following the manufacturer's protocol, and mounted the samples with coverslips finally for imaging.

For FISH experiments in samples labeled in retrograde, we first imaged the samples before performing FISH. The samples were then dehydrated in 50%, 75% and 100% ethanol twice for 5 min each. After air-drying the samples, we either assembled HybriWell-FL chambers (22 mm \times 22 mm \times 0.25 mm; Grace Bio-Labs) or drew a barrier around the samples using a ImmEdge hydrophobic barrier pen. The samples were then digested in Protease III for 30 min at 40°C , and washed in nuclease-free H_2O twice. We then proceeded to probe hybridization and subsequent amplification and visualization steps following the manufacturer's protocol, and mounted the samples with coverslips for imaging.

For Fig. 1f, the FISH probes used were Mm-Slc17a7-C1, Mm-Slc30a3-C2 and Mm-Cdh13-C3, visualized with Amp4 A It A. For Fig. 2e, the FISH probes used were Mm-Pcdh19-C1, Mm-Cdh8-C2 and Mm-Pcdh20-C3, visualized with Amp4 A It A. For retrograde labeling experiments in Extended Data Fig. 9a–e, the FISH probes used for the cadherins were Mm-Cdh12-C1 (custom ordered, no. 842531), Mm-Cdh8-C1 or Mm-Pcdh19-C1, in addition to Mm-Slc30a3-C2 and Mm-Slc17a7-C3, visualized with Amp4 A It C.

Imaging. All sequencing experiments were performed on an Olympus IX81 microscope with Crest X-light V2 spinning disk confocal, a Photometrics BSI prime camera and an 89North LDI seven-channel laser bank. Retrograde labeling experiments were imaged on either the same microscope or an LSM 710 laser scanning confocal microscope. Filters and lasers used for imaging are listed in Supplementary Table 3. Images were acquired using Micro-Manager (v1.4.23)⁵² on the spinning disk confocal and Zeiss Zen 2012 SP5 FP2 (v14.0.0.0) on the laser scanning confocal.

For all BARseq2 experiments, we imaged endogenous genes using an Olympus UPLFLN $\times 40$ 0.75-NA air objective and tiled 5×5 or 7×5 with 15% overlap between tiles for all sequencing cycles and the hybridization cycles. For each sequencing cycle, the four sequencing channels (G, T, A and C) and the DIC channel were captured. For hybridization cycles, GFP, RFP, Texas Red, Cy5 and DIC channels were captured. At the last cycle (usually the hybridization cycle for high expressors), we also imaged the DAPI channel.

For barcode sequencing, we imaged the first three cycles using the same imaging settings described above at $\times 40$ magnification. The third sequencing cycle was additionally reimaged at $\times 10$ magnification using an Olympus UPLANAPO $\times 10$ 0.45-NA air objective without tiling. All subsequent barcode sequencing cycles were imaged at $\times 10$ magnification.

On the spinning disk confocal, all $\times 40$ BARseq2 and FISH images were acquired as z-stacks with 1- μm step size and 0.16- μm xy pixel size, and all $\times 10$ images were acquired as z-stacks with 5- μm step size.

On the LSM 710, CTB-labeled samples were first imaged using a Plan-Apochromat $\times 10$ 0.45-NA objective without a coverslip as a z-stack with 7- μm z-step size and 0.7- μm xy pixel size. After FISH, the same samples were imaged using a Plan-Apochromat $\times 20$ 0.8-NA objective as a z-stack with 2- μm step size and 0.35- μm xy pixel size.

Probe design. A detailed description of probe sets used for each experiment and their sequences is provided in Supplementary Table 4.

To design reverse transcription primers and padlock probes, we tried to design as many probe sets as possible on each transcript while avoiding the end (~ 20 nucleotides) of the mRNA transcripts and ensuring at least a 3-nucleotide-long gap between two adjacent probe sets. Specific reverse transcription primers were designed to be 25 to 26 nucleotides long with amino modifier C6 at the 5' end and purified by high-performance liquid chromatography. In addition, we avoided sequences that contained G/C quadruplexes and/or had a low melting temperature, T_m (below 55°C). Padlock probes were designed to have two arms of 21 to 23 nucleotides long with a minimum T_m of 58°C , GC content between 40% and 60% and high complexity. The two arms were connected by a backbone consisting of a 32-nucleotide-long sequencing primer or detection probe target site, a 7-nucleotide gene-specific index, and a 3-nucleotide-long 3' linker. For padlock probes designed for hybridization readout, different backbone sequences were used for different genes. We further filtered out padlock probe sequences with potential nonspecific binding. To find potential nonspecific binding targets, we blasted the ligated

padlock arm sequences against the mouse genome and identified all targets with (1) 3 nucleotides of perfect match on either side of the ligation junction, (2) no gap and/or insertion within 7 nucleotides of the ligation junction and (3) melting temperatures of at least 37°C for nonspecific binding of each arm.

We maximized the number of padlock probe sets for *Slc17a7* (23 probes), *Slc30a3* (19 probes), *Gad1* (24 probes) and *Cdh13* (30 probes). These probe sets were used to evaluate the relationship between detection sensitivity and probe numbers. For the cadherin panels and the cell-type marker panels, we selected a subset of probes for each gene so that we had at most 12 probe sets per gene. Some shorter genes had fewer than 12 probes. These panels resulted in sensitivity that was sufficient for the present experiments, albeit somewhat below the maximum achievable with more probes. All but three genes (*Slc17a7*, *Slc30a3* and *Gad1*) were visualized using combinatorial GII codes (4 nucleotides in auditory cortex and 7 nucleotides in motor cortex; Supplementary Table 4); only a small subset of all possible GIIs were used, ensuring a Hamming distance of at least two bases between all pairs of GIIs in auditory cortex (from 4 nucleotides) and three bases in motor cortex (from 7 nucleotides) for error correction. The three remaining genes with high expression (*Slc17a7*, *Gad1* and *Slc30a3*) were detected by hybridization.

Optimization of endogenous mRNA detection. We optimized padlock probes, tissue pretreatment and reverse transcription to maximize detection sensitivity. We found that using multiple padlocks per mRNA transcript, with each padlock targeting a different site on the mRNA coding sequence, increased detection efficiency substantially (Fig. 1e). The increase in sensitivity varied across genes, but this was likely caused by differences in sensitivity of the single probe to which we normalized the sensitivities. For tissue pretreatment, we found that thin fresh-frozen tissue cryosections fixed with 4% PFA for 30 min to 1 h (Extended Data Fig. 1a) yielded higher mRNA sensitivity than shorter fixation or other pretreatments, such as PFA-perfused tissue slices with or without post-fixation. For reverse transcription, we found that reverse transcription primers specific to the targets at a concentration of 0.5–5 μM each yielded higher sensitivity than using random primers at concentrations up to 50 μM (Extended Data Fig. 1b). Altogether, these optimizations were crucial for increased mRNA detection sensitivity comparable to hybridization-based techniques.

To quantify the sensitivity of BARseq2 compared to conventional FISH methods, we detected two genes, *Slc30a3* and *Cdh13*, using both BARseq2 and RNAscope (Fig. 1f). We also probed for a third gene, *Slc17a7*, but at the resolution we imaged at, we were unable to fully resolve the signals from both BARseq2 and RNAscope; therefore, we only used *Slc30a3* and *Cdh13*, not *Slc17a7*, to evaluate the sensitivity of BARseq2. Linear regression between BARseq2 and RNAscope counts of *Slc30a3* and *Cdh13* genes in these two genes resulted in a slope of 1.65 (Extended Data Fig. 1c,d; $R^2 = 0.73$), indicating that BARseq2 achieved a sensitivity of about $1/1.65 \approx 60\%$, compared to RNAscope.

To multiplex gene detection with high imaging throughput, we optimized in situ sequencing to robustly read out GIIs of single ronolies over many sequencing cycles. We had previously adapted Illumina sequencing chemistry to sequence neuronal somata filled abundantly with RNA barcode ronolies, that is, DNA nanoballs generated by RCA^{6,35}. However, directly applying this method to sequence single ronolies generated from individual mRNAs proved difficult due to heating cycles and harsh stripping treatments that led to loss and/or jittering of ronolies (Extended Data Fig. 1e). To allow robust sequencing of single ronolies, we optimized cryosectioning and amino-allyl dUTP concentration³⁵ to cross-link ronolies more extensively, achieving less spatial jitter of single ronolies between imaging cycles (Extended Data Fig. 1e–h) and stronger signals (Extended Data Fig. 1i,j) retained over cycles. This robust in situ sequencing of combinatorial GII codes allowed BARseq2 to achieve fast imaging critical for high-throughput correlation of gene expression with projections.

Simultaneous detection of endogenous mRNAs and barcodes using BARseq2.

To assess multiplex gene expression and long-range projections in the same cells, we optimized for simultaneous detection and amplification of both endogenous mRNAs and barcodes. Although both endogenous mRNAs and barcodes are amplified using padlock probe-based approaches, amplifying barcodes required the addition of a DNA polymerase to copy barcode sequences into padlock probes to allow direct sequencing of diverse barcodes (up to $\sim 10^{18}$ diversity; Fig. 1c). Directly combining the two processes reduced the detection sensitivity of target mRNAs due to the addition of the DNA polymerase (Extended Data Fig. 6a; $37\% \pm 3\%$ (mean \pm s.d.); comparing the control condition to the zero-polymerase concentration). To preserve detection sensitivity for endogenous mRNAs while allowing the sequencing of diverse barcodes, we adjusted the concentration of the DNA polymerase to 0.001 U μl^{-1} (1/200 of the amount in the original BARseq), which doubled the sensitivity for endogenous mRNAs while also maintaining the sensitivity for barcodes (Extended Data Fig. 6a). This optimization allowed BARseq2 to detect both endogenous mRNAs and RNA barcodes together in the same neurons without compromising sensitivity.

Single-cell RNA-seq of auditory cortex. To dissociate neurons for scRNA-seq, we anesthetized animals with isoflurane and decapitated the animals. We then used a 2-mm biopsy punch to remove the auditory cortex. The tissue was then

dissected in ice-cold HABG medium (40 ml Hibernate A (Brainbits), 0.8 ml B27 (Thermo Fisher Scientific) and 0.1 ml Glutamax (Thermo Fisher Scientific)) into small pieces and digested in 3 ml prewarmed papain solution (3 ml Hibernate A-Ca (Brainbits), 6 mg papain (Brainbits) and 7.5 μ l Glutamax) at 30°C for 40 min. The digested tissues were then triturated in 2 ml prewarmed HABG for ten times using a salinized pipette with a 500- μ m opening. The undissociated tissues were transferred to a new tube with 2 ml HABG and triturated another ten times. The undissociated tissues were transferred again to a new tube with 2 ml HABG and triturated five times. The three tubes of HABG were combined and laid on top of a density gradient of 17.3%, 12.4%, 9.9% and 7.4% (vol/vol) Optiprep (Sigma) in HABG and centrifuged at 750g for 15 min. After removing the top two fractions, we collected the next two and half fractions and diluted in 5 ml HABG and centrifuged at 300g for 5 min. The pellet was washed in 5 ml HABG, pelleted again and resuspended in 100 μ l HABG. The cell suspension was then processed for library preparation using 10x Genomics Chromium Single Cell 3' Kits v3 according to the manufacturer's protocol. One of the scRNA-seq datasets was previously published⁶, and a new dataset was obtained in this study.

BARseq2 data processing. Sequencing data for projection target areas were acquired through the MAPseq core facility at Cold Spring Harbor Laboratory. We first demultiplexed raw sequencing reads and applied a threshold by read counts per molecule to remove PCR errors. This produced a list of unique barcode sequences with molecule counts in each target area. We then corrected for sequencing and amplification errors, allowing up to three mismatches. The resulting error-corrected barcode molecule counts were used to generate the projection matrix.

To process in situ sequencing data for genes, we first performed maximum projection of the image stacks along the *z* axis. Each maximum projection image was then corrected for sequencing channel bleed-through and lateral shift across channels. The images were then filtered with a median filter and background subtracted using a rolling ball with a radius of ten pixels. The sequencing cycle images were then registered to the first sequencing cycle using the sum of all four sequencing channels, and the hybridization images were registered to the first sequencing cycle using the channel that labeled all sequenced colonies. Registrations were performed by maximizing enhanced cross-correlation⁵⁴. After all images were registered, putative colonies were then picked from the first sequencing cycle by finding all peaks that were at least brighter than all surrounding valleys by a certain threshold determined empirically. This was achieved by first performing morphological reconstruction using the original image as the mask and the image minus the threshold as the marker, followed by identification of all local maxima. We then deconvolved all registered images and found the signal intensities for all colonies across all sequencing cycles and channels.

At this point, the signal for each colony is represented by an $m \times 1$ vector, in which m equals four (sequencing channels) times the number of cycles. To identify the gene that each colony corresponds to, we project the signal vector onto the signal vector of all genes and find the two genes with the highest projections, I_1 and I_2 . For colonies whose $(I_1 - I_2)/I_1$ is above a threshold, we assign the genes with the highest projections to these colonies. The remaining colonies are filtered out. For hybridization cycles, the channel in which the colonies are found is used directly to identify the genes.

For experiments in which genes were detected without barcodes for projection mapping, we segmented somas based on the rolon signals, background fluorescence from somas and nuclear staining using Cellpose⁵⁵, and assigned the rolonies to the segmented cells.

For experiments in which genes were detected in conjugation with barcodes, we further registered barcode sequencing cycles to the first sequencing cycle for genes using the DIC channel. The barcode sequencing images were then filtered with a median filter and background subtracted using a rolling ball with a radius of 50 pixels. The high-resolution images of the second and third cycles were then registered to the first sequencing cycle of barcodes using the sum of all four sequencing channels. The low-resolution images of the third sequencing cycle were then registered to the high-resolution image of the same cycle.

To segment the barcoded cells from the high-resolution images, we first determined 'seed' pixels by identifying local maxima in the first sequencing cycle image as described above. These seed pixels are positions of the strongest signal within putative cell bodies. Then for each seed pixel, we calculated the projection of signal vectors for all other pixels within a local area on the signal vector of the seed pixel and the rejection of signal vectors for these pixels from the signal vector of the seed pixel. We then segmented the cell bodies by finding all pixels that fulfill the following criteria: (1) the projections of their signal vectors are above a threshold; (2) the ratios between the rejections and projections are below a threshold; and (3) they are connected to the seed pixel. In parallel, we performed a second segmentation using only the DAPI signals and gene sequencing images with a marker-based watershed without using the barcode sequencing images, and found the segmented cells that overlapped with the barcode segmented cells. We then visually inspected the sequencing images and segmentations for each cell to determine which segmentation produced better results and to eliminate badly segmented cells. We then assign gene rolonies to the filtered segmented cells to produce the expression matrix.

To find the barcode sequences of the segmented cell, we integrated signals over the whole segmented cells and called the channel with the strongest signal as the base in both the high-resolution images and the low-resolution images. We then concatenated the sequences from the high-resolution images and the low-resolution images to produce the full barcode sequences. To find the projection patterns, these in situ sequenced barcodes were then matched to the barcodes identified in the projection areas allowing one mismatch but not ambiguous matches (that is, one in situ barcode matching to multiple barcodes found in projection sites).

Analysis of BARseq2 gene expression data. All analyses were carried out in MATLAB. For analysis of gene-only datasets, neurons were first filtered by requiring at least ten counts of *Slc17a7* or *Gad1* and were positioned within the cortex. To make the data comparable to previous studies⁶, the cortical depths of neurons were normalized to a total thickness of 1,200 μ m for auditory cortex and 1,500 μ m for motor cortex. To find cadherins that were differentially expressed in cell types, the expression of cadherins in each cell type was compared to the expression of cadherins in all other cell types using rank-sum tests.

Laminar distribution of cadherins. Because many genes, especially cell adhesion molecules, are differentially expressed across cortical layers, we evaluated how well BARseq2 can capture spatial organization of cadherins compared to existing methods, such as FISH. To compare laminar distribution observed by BARseq2, FISH and Allen Brain Atlas, we quantified gene expression signal densities across 100- μ m bins in laminar depth. For BARseq2 and FISH, the quantification was performed by counting dots. For Allen Brain Atlas, the quantifications were done by integrating signal intensities over all pixels in each bin. Because each bin had a different number of pixels sampled in our data, we then divided the gene expression signals by the area observed in the images to calculate the density. We then *z*-scored the densities within each gene to produce the laminar profiles for each gene.

RNAscope against *Cdh8*, *Pcdh19* and *Pcdh20* revealed laminar expression profiles that were qualitatively similar to those obtained by BARseq2 (Fig. 2e). For *Pcdh20*, the dynamic range of gene expression (that is, the differences between peaks and valleys in expression) was more pronounced in the BARseq2 data than that observed by RNAscope. Because low sensitivity and/or low specificity would likely result in a reduction, not an increase, in the dynamic range of expression, it is unlikely that such quantitative differences in the laminar profiles of gene expression were caused by sensitivity and/or specificity issues with BARseq2. We suspect that the reduced dynamic range in RNAscope is caused by nonspecific signals inherent to amplified FISH methods. We therefore sought to compare BARseq2 to other FISH datasets to confirm its accuracy.

We then compared the distributions of genes obtained by BARseq2 to those in the Allen gene expression atlas⁵¹ (Fig. 2f and Extended Data Fig. 3). The laminar distribution of gene expression revealed by BARseq2 was highly correlated with that in the Allen gene expression atlas (Spearman correlation $\rho = 0.696$, $P = 3.8 \times 10^{-29}$). Specifically, the laminar distribution of *Pcdh20* obtained by BARseq2 matched very well with *Pcdh20* in the Allen gene expression atlas (Extended Data Fig. 3). These results indicate that BARseq2 accurately captured the laminar distribution of cadherin expression.

Gene-pair expression in single neurons. To test whether BARseq2 accurately captures gene expression, we compared the expression of two pairs of genes in single neurons. First, we compared the expression of *Slc17a7* and *Gad1*, two genes that are expressed in two distinct classes of neurons. Second, we compared the expression of *Slc30a3* and *Cdh24*, two genes that are anti-correlated at the subtype level based on scRNA-seq³.

Slc17a7 and *Gad1* are expressed in excitatory and inhibitory neurons, respectively. They are thus almost never expressed in the same neuron in the cortex. To quantify the mutual exclusivity of *Slc17a7* and *Gad1* in neurons, we defined the exclusivity index $E = P(Gad1|Slc17a7)/P(Gad1)$, where $P(Gad1|Slc17a7)$ indicates the probability of a cell expressing at least ten counts of *Gad1* conditioned on the expression of at least ten counts of *Slc17a7*, and $P(Gad1)$ indicates the probability of a cell expressing at least ten counts of *Gad1* in all filtered neurons.

BARseq2 recapitulated the mutual exclusivity between these two genes (Fig. 2j,k), but a small number of neurons did express both *Slc17a7* and *Gad1* (gray cells in Fig. 2j). This could be caused by overlapping cells (that is, an inhibitory neuron and an excitatory neuron at the same *x/y* position, but in different *z* planes were merged together in the maximum projection images) or cell segmentation errors (two adjacent cells incorrectly segmented as a single cell). Because the sections we used were 10- μ m thick, comparable to the diameter of an average neuron, the latter source of error was likely to be more common.

This type of error was similar to doublets in droplet-based scRNA-seq techniques. Assuming that the mutual exclusions of *Slc17a7* and *Gad1* were absolute, then we could estimate the 'doublet' rate as the ratio between the probability of neurons expressing both genes and the product of the probabilities of neurons expressing either gene. Using this formula, we estimated the doublet rate of BARseq2 to be 7.5%, which is in a similar range as droplet-based scRNA-seq

techniques (usually <5%). Improvement in cell segmentation algorithms may further reduce the doublet rate.

In addition to cells that express both *Gad1* and *Slc17a7* at substantial levels, most cells that expressed one of the two genes dominantly also had nonzero expression of the other gene, albeit at much lower levels. This noise floor could be caused by mRNAs in dendrites that were incorrectly assigned to other neurons. Because the expression levels of these genes in the somata were much higher than those in the dendrites, this type of error was unlikely to substantially affect the determination of excitatory and inhibitory neurons.

Similarly, consistent with a previous scRNA-seq study³, BARseq2 also confirmed the observation that *Slc30a3* was more highly expressed in subtypes of excitatory neurons that did not express *Cdh24* compared to projection neurons that did express *Cdh24* (Extended Data 5a,b; $P = 5 \times 10^{-26}$ using two-tailed rank-sum test on scRNA-seq data using Smart-Seq2 ($n = 10,044$ neurons)³, and $P = 4 \times 10^{-65}$ on BARseq2 data ($n = 2,947$ neurons)).

Cell typing in BARseq2 and single-cell data. To select a panel of marker genes, we chose meta-analytic markers from seven scRNA-seq datasets in the motor cortex²³, accessed from the NeMO archive. In each dataset and for each cell type, we extracted differentially expressed genes (DEGs) among excitatory neurons ('glutamatergic' class, one-versus-all DEGs, fold change > 2, Mann-Whitney FDR < 0.05). We filtered out genes with low expression (average counts per million (CPM) < 100), then ranked genes primarily by the number of datasets where they were DEGs and secondarily by average fold change, and selected the top five markers.

To examine if multiplexing affects detection sensitivity, we probed for *Slc17a7*, *Slc30a3* and *Gad1* either as a separate three-gene panel or as part of the 65-gene panel (20 cadherins and 45 marker genes). The mean expression densities across laminar positions for the three genes were similar between the three-gene panel and the 65-gene panel (Extended Data Fig. 5c; $P = 0.22$ for *Slc17a7*, $P = 0.49$ for *Slc30a3* and $P = 0.66$ for *Gad1* using two-tailed rank-sum tests), suggesting that targeting more genes did not affect detection sensitivity of each gene.

To call cell types in BARseq2 and single-cell data, we used the following procedure. First, we normalized counts to $\log(1 + \text{CPM})$, then we computed the average marker expression for each cell type and assigned the cell type with the highest average expression. If two marker sets were tied for highest expression, the cell was left unassigned. This method of cell typing achieved good precision and recall for most cell types when applied to scRNA-seq data (Extended Data Fig. 5d). We applied the procedure across nine datasets to check whether it is robust across technologies and sequencing depth (Extended Data Fig. 5e,f). Overall, we observed extremely high performance for NP and CT subtypes in all cases, while L6b was slightly better predicted in high-depth datasets. The cell-typing method always predicted IT cells correctly, but not always the correct layer (L2/3, L5, L6 and Car3; Extended Data Fig. 5g). This is consistent with the observation that IT types form a continuum in single-cell datasets, making it difficult to fully separate subtypes by layer. Finally, the PT type proved to be the most difficult cell type to predict. While all PT cells were correctly annotated as PT (Extended Data Fig. 5h), numerous L2/3 IT and L5 IT cells were wrongly annotated as PT, in particular in high-depth datasets (Extended Data Fig. 5f,g). We believe that this was due to an imbalance in the marker panel, with PT markers showing higher expression than markers from other cell types. We tested various normalization procedures to overcome this effect but found that results were insensitive to normalization overall (Extended Data Fig. 5f).

Using this panel and cell-typing method, we determined the transcriptomic types of excitatory neurons in motor cortex using BARseq2 (Fig. 3b). Most transcriptomic types were found enriched in the correct layers. One exception to this was the L6 Car3 IT type. In general, few L6 Car3 IT neurons were identified by BARseq2. Furthermore, even though L6 Car3 IT neurons were predominantly in L6, some were identified in L2/3 by BARseq2 (Fig. 3c). This result was surprising, given that L6 Car3 IT neurons, when present, were only rarely mistyped as L2/3 in our preliminary analyses (Extended Data Fig. 5g). L6 Car3 IT neurons were only rarely detected in the datasets used to select markers, so we expect that using additional data will lead to a more robust marker selection and better cell-typing performance with BARseq2. These optimizations, however, are beyond the scope of this paper.

Gene expression in barcoded neurons. Gene expression in Sindbis-infected barcoded neurons largely reflect the gene expression in non-barcoded neurons. For example, the expression of the excitatory marker *Slc17a7* and the inhibitory marker *Gad1* remained mutually exclusive in barcoded neurons in both auditory cortex and motor cortex (Extended Data Fig. 6c,d). This mutual exclusivity was preserved despite an overall reduction in mRNA expression (Extended Data Fig. 6e; median read of 38 in barcoded cells in both auditory and motor cortex, compared to 64 and 48 in non-barcoded cells in the two cortical areas, respectively). Similarly, *Slc30a3* remained differentially expressed across barcoded excitatory neurons with or without *Cdh24* expression as it was in non-barcoded excitatory neurons (Extended Data Fig. 6f; $P = 1 \times 10^{-6}$ using rank-sum test, $n = 810$ neurons). Although our observations cannot rule out the possibility that a small subset of genes (for example, viral response genes) may be disrupted by Sindbis infection, these results

suggest that the coexpression relationships of most genes in Sindbis-infected neurons reflect those in noninfected cells.

Analysis of BARseq2 gene expression and projection dataset. For analysis of BARseq2 datasets with both gene expression and projections, we first evaluated the mutual exclusivity of *Slc17a7* and *Gad1* expression (see below). For this purpose, the neurons were filtered with the same thresholds as in the gene-only dataset. For all other analyses, we used a more relaxed filtering to compensate for the reduced gene expression in barcoded cells, requiring neurons to have at least five counts of *Slc17a7* or *Gad1*. In this filtered set, neurons were considered excitatory if the counts of *Slc17a7* were larger than the counts of *Gad1*, and were considered inhibitory if the counts of *Gad1* were larger than the counts of *Slc17a7*. Projection data were log normalized as in previous studies⁴. We further normalized the projection strengths of each area to two previous clustered BARseq datasets⁶ and used a random forest classifier to assign neurons to projection clusters.

To find cadherins that were differentially expressed across major projection classes and between auditory and motor cortex, we performed rank-sum tests for pairwise comparisons among major classes or the two areas for each cadherin and calculated the FDRs.

Projection modules were identified using NMF³². To find the variance in projections explained by cadherins and/or laminar positions (Extended Data Fig. 8), we used Gaussian process regression to predict projection modules using the laminar position of neurons as a predictor and linear regression to predict projection modules using the expression of individual cadherins. The variance explained by each predictor was reported after 100 iterations of ten-fold cross validation. To find cadherins that were associated with projection modules, we calculated the Spearman correlation between the coefficients for projection modules and gene counts. To generate the plots of differential gene expression in Fig. 6e, we sorted the neurons by the coefficients for projection modules and smoothed gene expression using a window of 101 neurons.

Projections of excitatory and inhibitory neurons. BARseq2 accurately observed the fact that projection neurons in the cortex are predominantly excitatory and express the excitatory marker *Slc17a7*, not the inhibitory marker *Gad1*. To distinguish between excitatory and inhibitory neurons, we categorized a neuron as excitatory or inhibitory if (1) the neuron had higher expression of the excitatory marker *Slc17a7* or the inhibitory marker *Gad1*, respectively, and (2) the marker was expressed at greater than five reads in the cell. This threshold resulted in 2,496 excitatory neurons (947 in auditory cortex and 1,549 in motor cortex) and 240 inhibitory neurons (100 in auditory cortex and 140 in motor cortex; Fig. 4d). Consistent with previous observations, most cortical projection neurons identified by BARseq2 were excitatory (Fig. 4e). However, we also identified a small fraction of inhibitory projection neurons. Some of these neurons could be caused by 'doublets' as discussed above. Consistent with this hypothesis, the inhibitory projection neurons (and some excitatory projection neurons) in motor cortex expressed both *Gad1* and *Slc17a7* at similar levels (Extended Data Fig. 6g). However, inhibitory projection neurons in auditory cortex expressed only *Gad1*, not *Slc17a7* (Extended Data Fig. 6h), suggesting that these were real inhibitory projection neurons. This observation was consistent with previous reports of rare inhibitory projection neurons in the cortex^{6,56}. We did not further analyze these inhibitory projection neurons.

We also observed many excitatory neurons without projections (Fig. 4d,e), similar to those observed in previous BARseq experiments⁶. These neurons were likely non-projecting excitatory neurons and neurons that project only locally or to neighboring cortical areas³ that we did not sample.

Differential expression of cadherins across IT, PT and CT neurons. BARseq2 revealed differential gene expression across major classes of neurons defined by projections. We found that many cadherins (8 for auditory cortex and 12 for motor cortex) were differentially expressed across IT, PT and CT neurons that were defined by projections as in previous studies^{2,6} (Fig. 5a-c). Several cadherins were consistently differentially expressed in both cortical areas. For example, *Cdh6* and *Cdh13* were overexpressed in PT neurons compared to the other two classes, whereas *Cdh8* was underexpressed in CT neurons compared to the other two classes (FDR < 0.05, rank-sum test). In addition, we also found nine cadherins that were differentially expressed across the two cortical areas in at least one class (Fig. 5d; FDR < 0.05, rank-sum test).

Major classes of projection neurons (IT, PT and CT) differ in both gene expression and projection patterns. Therefore, the differential expression of cadherins observed across these three major classes defined by projection patterns should be consistent with the differential expression across the classes defined by transcriptomic methods. To test this, we compared the differences in mean expression of cadherins in the three classes in motor cortex and auditory cortex observed by BARseq2 to those observed using scRNA-seq in neighboring cortical areas (V1 and ALM)³. Generally, differentially expressed cadherins identified by BARseq were also differentially expressed in scRNA-seq (Extended Data Fig. 7a; the rank correlation of the differences in cadherin expression across major neuronal types was 0.61 between BARseq and scRNA-seq, compared to 0.39 between auditory cortex and motor cortex in BARseq). Importantly, all cadherins that were consistently differentially expressed in both A1 and M1 were also differentially

expressed across the same pairs of major classes in V1 and ALM as shown by scRNA-seq (Extended Data Fig. 7a). Several cadherins, including *Pcdh7* and *Cdh11*, were differentially expressed with the opposite signs in scRNA-seq and in BARseq2 (Extended Data Fig. 7a). However, these cadherins were not consistently expressed across motor and auditory cortex. For example, *Pcdh7* was expressed at a significantly higher level in PT neurons than CT neurons in motor cortex ($P < 10^{-8}$; Fig. 5c), but at a lower level in PT neurons than CT neurons in auditory cortex ($P = 0.0011$, not statistically significant at FDR < 0.05). It is thus likely that these differences between observations by BARseq2 and by scRNA-seq reflect area-to-area differences, not methodological differences. These results confirm the differential expression of cadherins across major classes identified by BARseq2.

Projection differences across transcriptionally defined IT subtypes. BARseq2 confirmed known biases in projection patterns across transcriptomic IT subtypes in auditory cortex (Extended Data Fig. 7b,c). Previous studies using both barcoding-based strategies and single-cell tracing have identified distinctive projection patterns for two transcriptomic subtypes of IT neurons, IT3 (L6 IT) and IT4 (L6 Car3 IT)^{6,26}. To test if we could capture the same projection specificity of transcriptomic subtypes, we mapped projection patterns to projection clusters identified in a previous study in auditory cortex, and used a combination of gene expression and laminar position to distinguish four transcriptomic subtypes of IT neurons⁶. These subtypes were defined consistently with a previous study⁶ for ease of comparison. Specifically, we defined IT1 as neurons with depths of less than 590 μm , IT2 as neurons with depths between 590 and 830 μm and did not express *Cdh13*; IT3 as neurons between 590 and 830 μm that expressed *Cdh13* or neurons deeper than 830 μm that expressed *Slc30a3*; and IT4 as neurons deeper than 830 μm that did not express *Slc30a3*.

As expected, the two transcriptomic subtypes (IT3 and IT4) predominantly found in L5 and L6 were indeed more likely to project only to the ipsilateral cortex, without projections to the contralateral cortex or the striatum ($P = 4 \times 10^{-7}$ comparing the fraction of neurons with only ipsilateral cortical projections in IT3/IT4 to the fraction of them in IT1/IT2 using Fisher's test; Extended Data Fig. 7b,c). Between IT3 and IT4, IT4 neurons were more likely to project ipsilaterally (58% of IT3 neurons compared to 92% of IT4 neurons; $P = 1 \times 10^{-4}$ using Fisher's test), whereas IT3 neurons were more likely to project contralaterally (66% of IT3 neurons compared to 14% of IT4 neurons, $P = 5 \times 10^{-8}$ using Fisher's test). Thus, BARseq2 recapitulated known projection differences across transcriptomic subtypes of IT neurons.

Cadherin coexpression module analysis. To extract robust modules of coexpressed cadherins, we used a previously developed approach to combine multiple datasets by meta-analysis, a crucial step to attenuate technical and biological noise^{33,34}. Briefly, we built coexpression networks using the Spearman correlation for seven scRNA-seq datasets in the motor cortex²³, accessed from the NeMO archive and subset to the following subclasses: 'L2/3 IT', 'L4/5IT', 'L5 IT', 'L6 IT' and 'L6 IT Car3'. We ranked each network, then averaged the networks to obtain our final meta-analytic network. We then applied hierarchical clustering with average linkage and extracted modules using the dynamic tree-cutting tree algorithm³¹.

To compute the association between coexpression modules and projection patterns, we framed the association as a classification task: can we predict projection patterns from module expression? First, we generated labels by binarizing each projection pattern—cells with a projection strictly greater than the median projection strength were marked as positives. Next, we generated predictors by computing gene module expression as the average $\log(\text{CPM} + 1)$ across all genes in the module. We reported the association strength (classification results) as an AUROC. To compute the association between coexpression modules and cell types, we used a similar approach, using clusters defined by the BRAIN Initiative Cell Census Network²³ as labels. For visualization, cell types were organized according to the following procedure: cell types were reduced to a centroid by taking the median expression for each gene, then cell types were clustered according to hierarchical clustering with average linkage with correlation-based distance.

Validation of cadherin correlates of IT projections using in situ hybridization and retrograde labeling. To confirm that *Cdh8*, *Cdh12* and *Pcdh19* correlated with ipsilateral, contralateral and striatal projections, respectively, we performed CTB retrograde labeling from the projection targets and performed FISH against *Slc17a7*, *Slc30a3* and the cadherins in both A1 and M1 (Extended Data Fig. 9a; see Supplementary Table 1 for injection coordinates). We then quantified cadherin expression and CTB labeling in IT neurons that had sufficient DAPI signals and expressed both *Slc17a7*, an excitatory cell marker, and *Slc30a3*, which labeled the majority of IT neurons (Extended Data Fig. 9b). Neurons that had weak and/or ambiguous CTB signals were excluded from the analyses. Indeed, we observed that the three cadherins were expressed at higher levels in CTB⁺ neurons in both areas despite notable overlap in expression between CTB⁺ and CTB⁻ neurons (Extended Data Fig. 9c–e). This overlap was expected because CTB was unlikely to have labeled all neurons that projected to the areas that we sampled with BARseq2. For example, in a previous study, we found that less than half of neurons with

projections detected by BARseq were also labeled by injection of CTB into the same target area⁶. These results thus provide further support for the finding that cadherins correlate with similar projections in both A1 and M1.

Statistics and reproducibility. No statistical method was used to predetermine sample size, but our sample sizes are similar to those reported in previous publications^{6,14}. No data were excluded from the analyses. Because only wild-type animals were used and the findings did not rely on comparison across animals, the experiments were not randomized and the investigators were not blinded to allocation of animals during experiments and outcome assessment. All statistical tests performed are indicated in the text. Two-tailed tests and Bonferroni correction were used for all P values reported unless noted otherwise. Wherever indicated, FDRs were computed according to the Benjamini–Hochberg procedure⁵⁷. All statistical tests used were non-parametric except when statistical significance was estimated for the Pearson correlation (Fig. 6a). When estimating statistical significance for the Pearson correlation, normal distribution was assumed, but this was not formally tested.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw target area sequencing data (Fig. 4c; SRR12247894, SRR12245390 and SRR12245389) and scRNA-seq data (Fig. 2g–i) are deposited at the Sequence Read Archive (SRR13716225). Raw in situ sequencing images (Figs. 2–4) are deposited at the Brain Image Library (<https://download.brainimaginglibrary.org/06/35/0635a0b3b0954c7e/>). Example annotated images from the dissected brain slices and other data and intermediate processed sequencing data are deposited at Mendeley Data (<https://doi.org/10.17632/jnx89bmv4s.2>).

Code availability

Processing scripts are deposited at Mendeley Data (<https://doi.org/10.17632/jnx89bmv4s.2>).

References

- Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
- Edelstein, A. D. et al. Advanced methods of microscope control using μ Manager software. *J. Biol. Methods* **1**, e10 (2014).
- Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
- Evangelidis, G. D. & Psarakis, E. Z. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1858–1865 (2008).
- Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
- Rock, C., Zurita, H., Wilson, C. & Apicella, A. J. An inhibitory corticostriatal pathway. *Life* **5**, e15890 (2016).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

Acknowledgements

The authors thank members of the MAPseq core facility, H. Zhan, Y. Li and N. Gemmill, for MAPseq data production; K. Matho and Z. J. Huang for dissection coordinates in motor cortex; H. Zhan, L. Yuan, H. L. Gilbert, K. Matho, J. Keschull and D. Fürth for useful discussions; and W. Wadolowski, B. Burbach, K. Lucere and E. Fong for technical support. This work was supported by the National Institutes of Health (5RO1NS073129, 5RO1DA036913, RF1MH114132 and U01MH109113 to A.M.Z.; R01MH113005 and R01LM012736 to J.G.; and U19MH114821 to A.M.Z. and A.G.), the Brain Research Foundation (BRF-SIA-2014-03 to A.M.Z.), IARPA MICrONS (D16PC0008 to A.M.Z.), Paul Allen Distinguished Investigator Award (to A.M.Z.), Simons Foundation (350789 to X.C.), Chan Zuckerberg Initiative (2017-0530 ZADOR/ALLEN INST (SVCF) SUB to A.M.Z.) and Robert Lourie award (to A.M.Z.). This work was additionally supported by the Assistant Secretary of Defense for Health Affairs endorsed by the Department of Defense, through the FY18 PRMRP Discovery Award Program (W81XWH1910083 to X.C.) Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the U.S. Army. In conducting research using animals, the investigators adhered to the laws of the United States and regulations of the Department of Agriculture.

Author contributions

Y.-C.S., X.C. and A.M.Z. conceived the study. Y.-C.S. and X.C. optimized and performed BARseq2. Y.-C.S., X.C. and H.Z. collected BARseq2 data. X.C., S.F. and J.G. analyzed data. Y.-C.S., X.C. and S.F. selected gene panels. X.C. and S.L. compared gene expression between BARseq2 and Allen ISH. Y.-C.S. and X.C. performed retrograde tracing combined with FISH validations. Y.-C.S., X.C., S.F. and A.M.Z. wrote the paper.

Competing interests

A.M.Z. is a founder and equity owner of Cajal Neuroscience and a member of its scientific advisory board. The remaining authors declare no competing interests.

Additional information

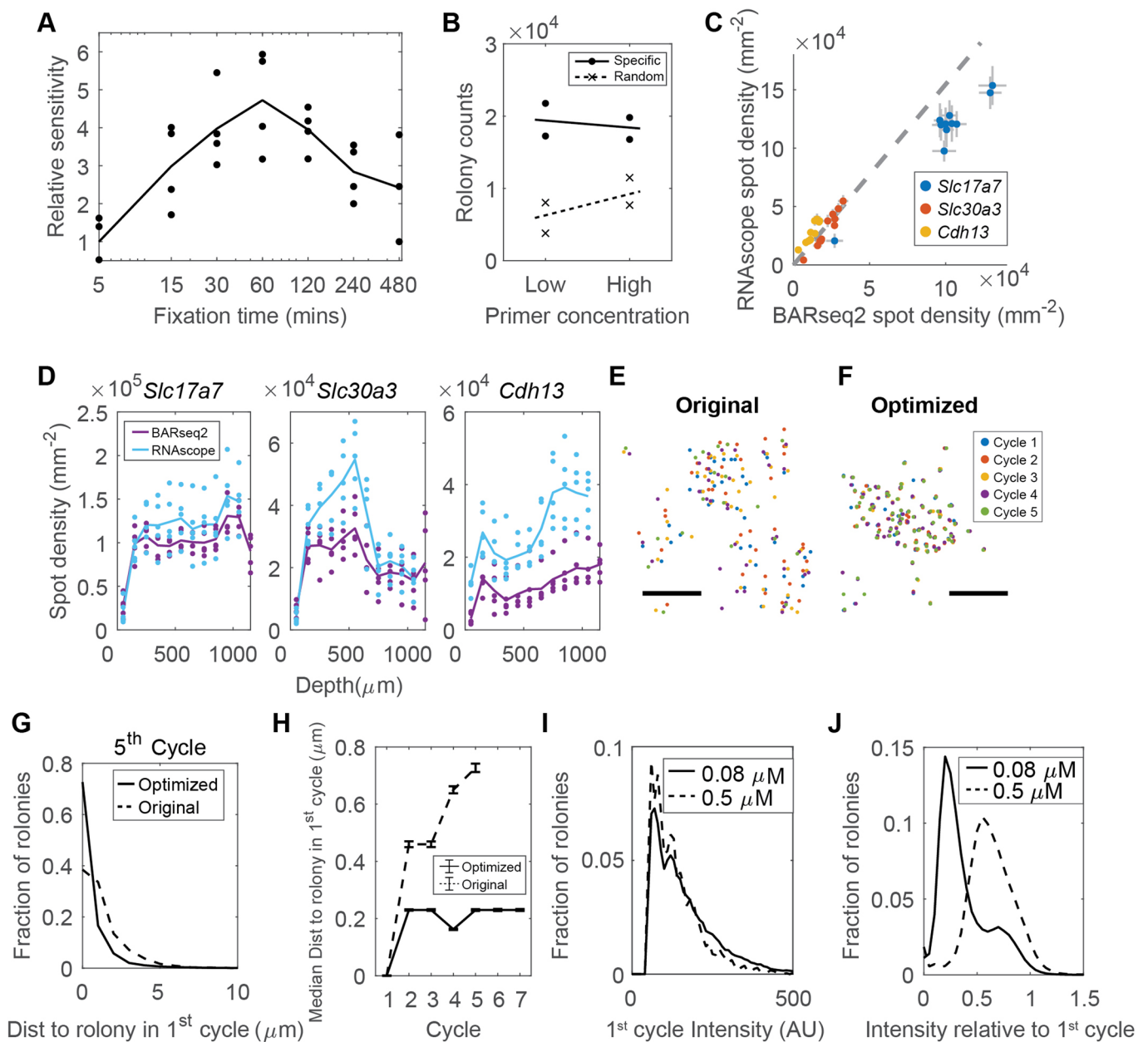
Extended data is available for this paper at <https://doi.org/10.1038/s41593-021-00842-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41593-021-00842-4>.

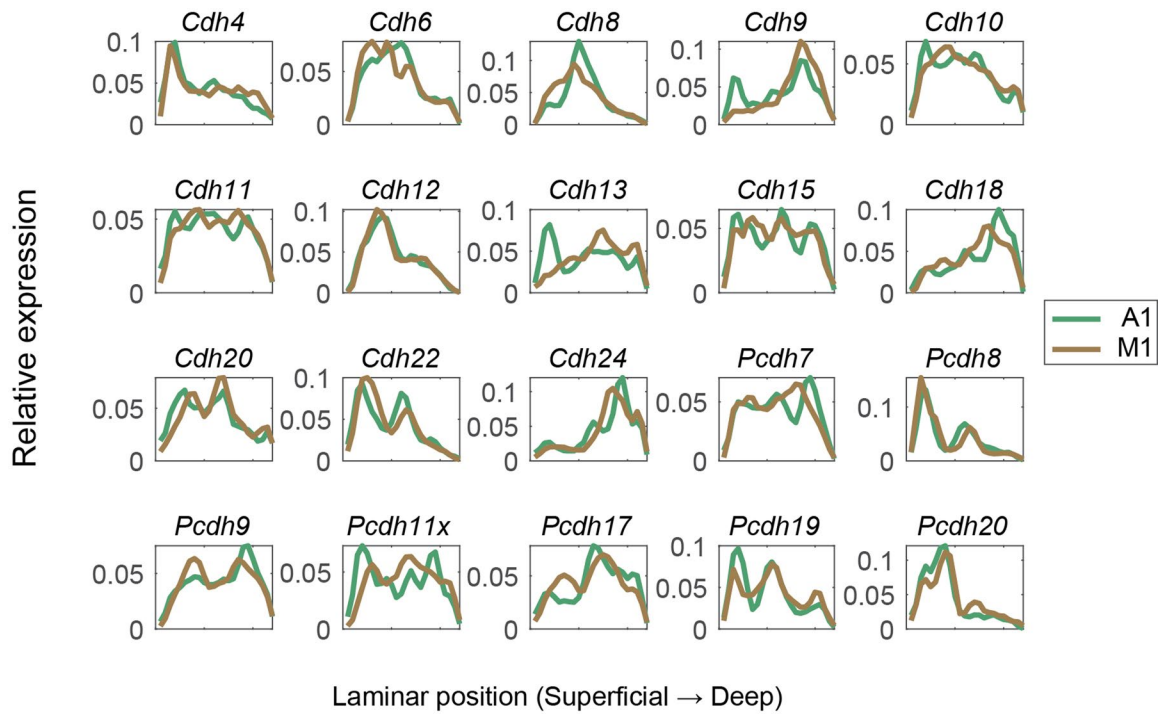
Correspondence and requests for materials should be addressed to X.C. or A.M.Z.

Peer review information *Nature Neuroscience* thanks Kenneth Harris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

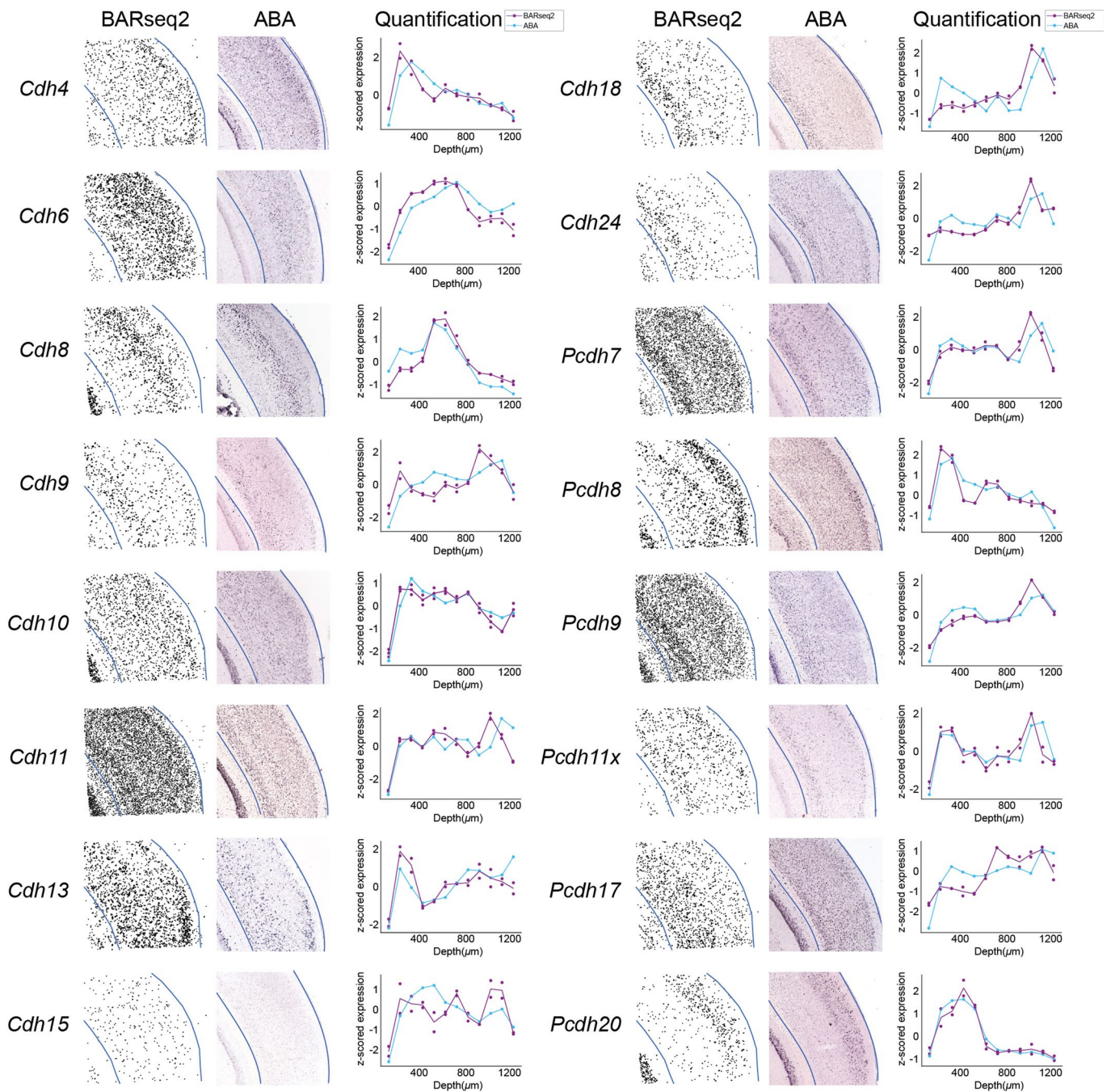
Reprints and permissions information is available at www.nature.com/reprints.



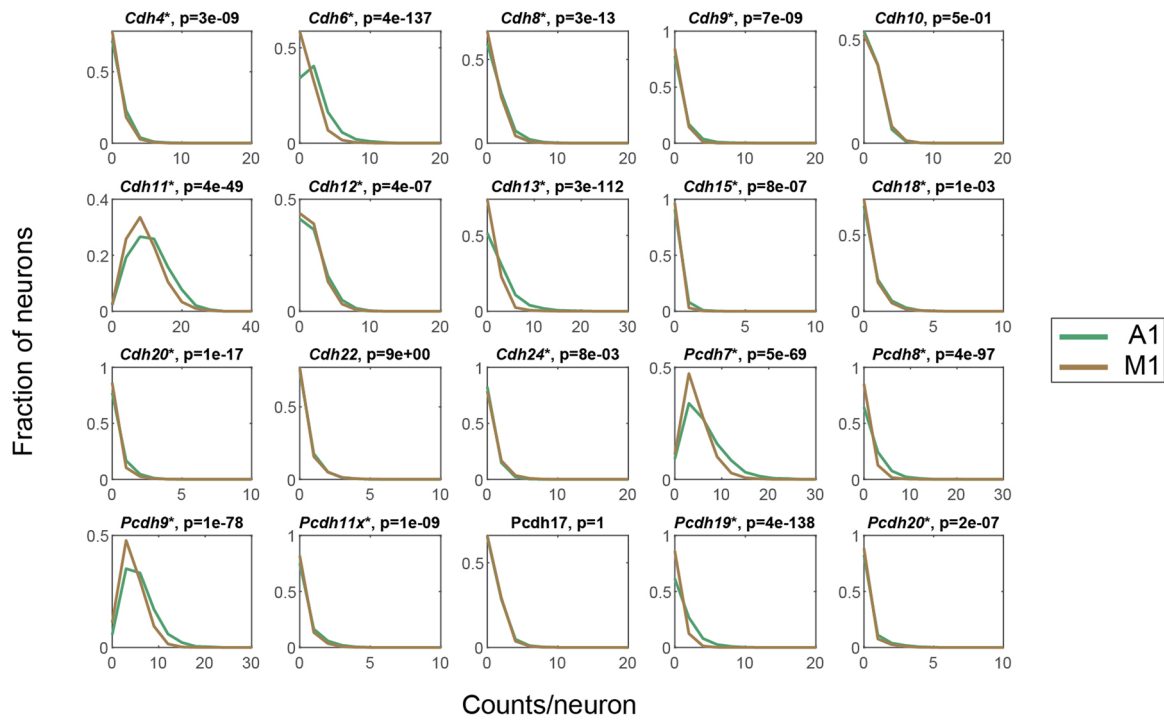
Extended Data Fig. 1 | Optimization of BARseq2 for detecting endogenous mRNAs. **a**, Relative sensitivity (means and individual data points) of BARseq2 in detecting *Slc17a7* using the indicated fixation times, normalized to that achieved with 5 mins of fixation. $n=3$ for 480 mins and $n=4$ for other conditions. **b**, Rolony counts for *Slc17a7* using either random primers or specific primers at two different concentrations. The two concentrations used were 5 μM (low) and 50 μM (high) for random primers, and 0.5 μM (low) and 5 μM (high) for specific primers. Lines indicate means and dots/crosses represent individual samples. $n=2$ slices for each condition. **c,d**, BARseq2 sensitivity compared to RNAscope. **c**, Spot density detected by BARseq2 or RNAscope in each 100 μm bin along the laminar axis in auditory cortex. Error bars indicate standard errors. The dashed line indicates linear fit for *Slc30a3* and *Cdh13*. Slope = 1.65 and $R^2=0.73$. $n=5$ slices for both BARseq2 and RNAscope. **d**, shows the means and individual samples for each gene. **e,f**, Positions of rolony across five sequencing cycles using the original (**e**) or the optimized (**f**) sequencing protocol. Scale bars = 10 μm. **g**, The distribution of minimum distance between rolony imaged in the first cycle and in the fifth cycle using the original or the optimized protocol. **h**, Median distance between rolony imaged in the indicated cycles and the closest rolony imaged in the first cycle using the original or the optimized protocol. Error bars indicate standard errors. For both (**g**) and (**h**), $n=148,708$ rolony for optimized condition and $n=12,114$ for original condition. **i,j**, The distribution of absolute rolony intensities for the first sequencing cycle (**i**) and relative rolony intensities after 6 sequencing cycles and one stripping step, normalized to the intensities in the first sequencing cycle (**j**). Amino-allyl dUTP concentrations used are indicated. In (**i**), $n=63,852$ rolony for 0.08 μM and $n=4,286$ rolony for 0.5 μM; in (**j**), $n=128,976$ rolony for 0.08 μM and $n=113,235$ rolony for 0.5 μM.



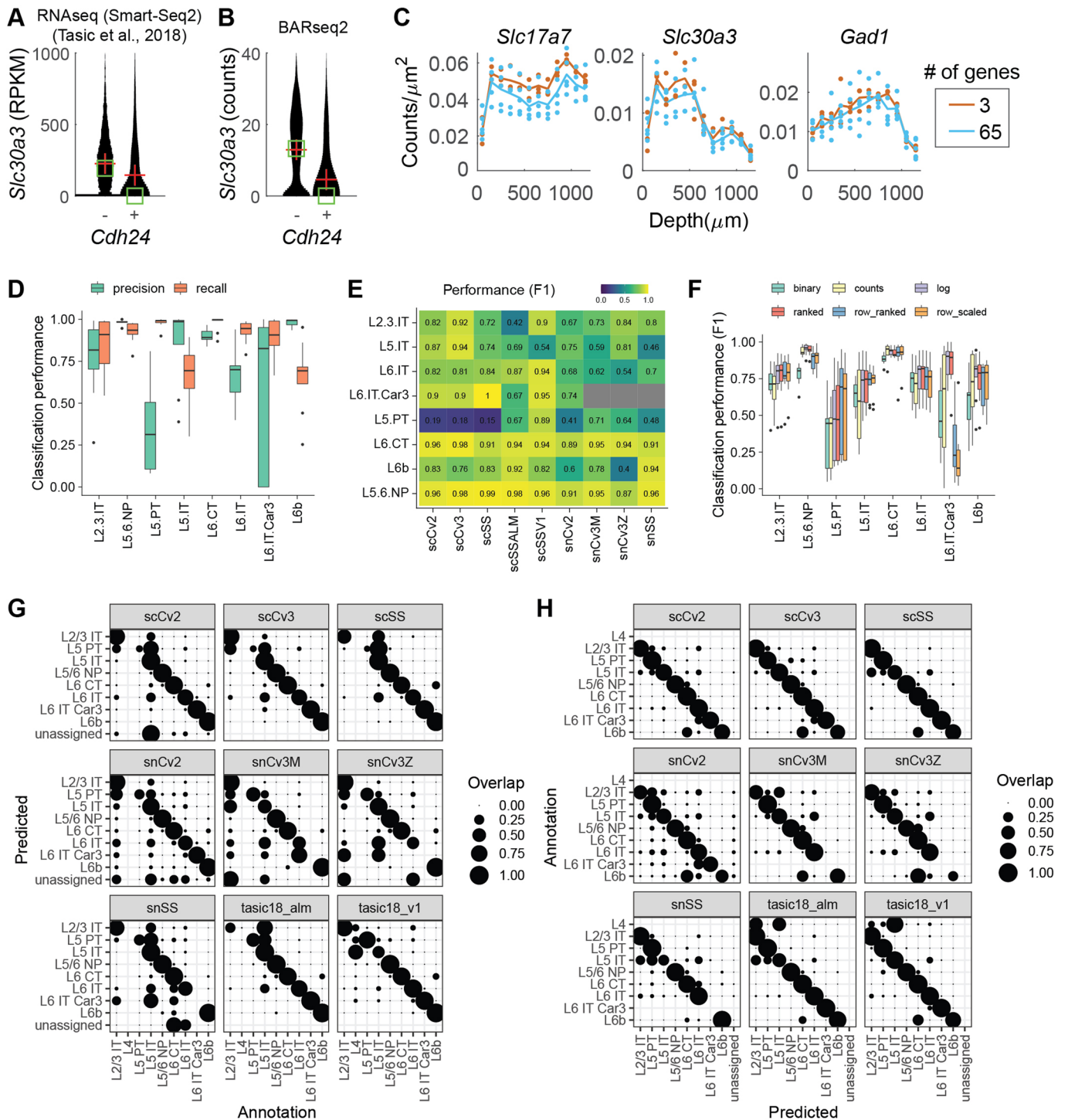
Extended Data Fig. 2 | Laminar distribution of cadherins in auditory cortex (green) and motor cortex (brown). In both cortical areas, cortical depth is normalized so that the bottom and the top of the cortex match between M1 and A1.



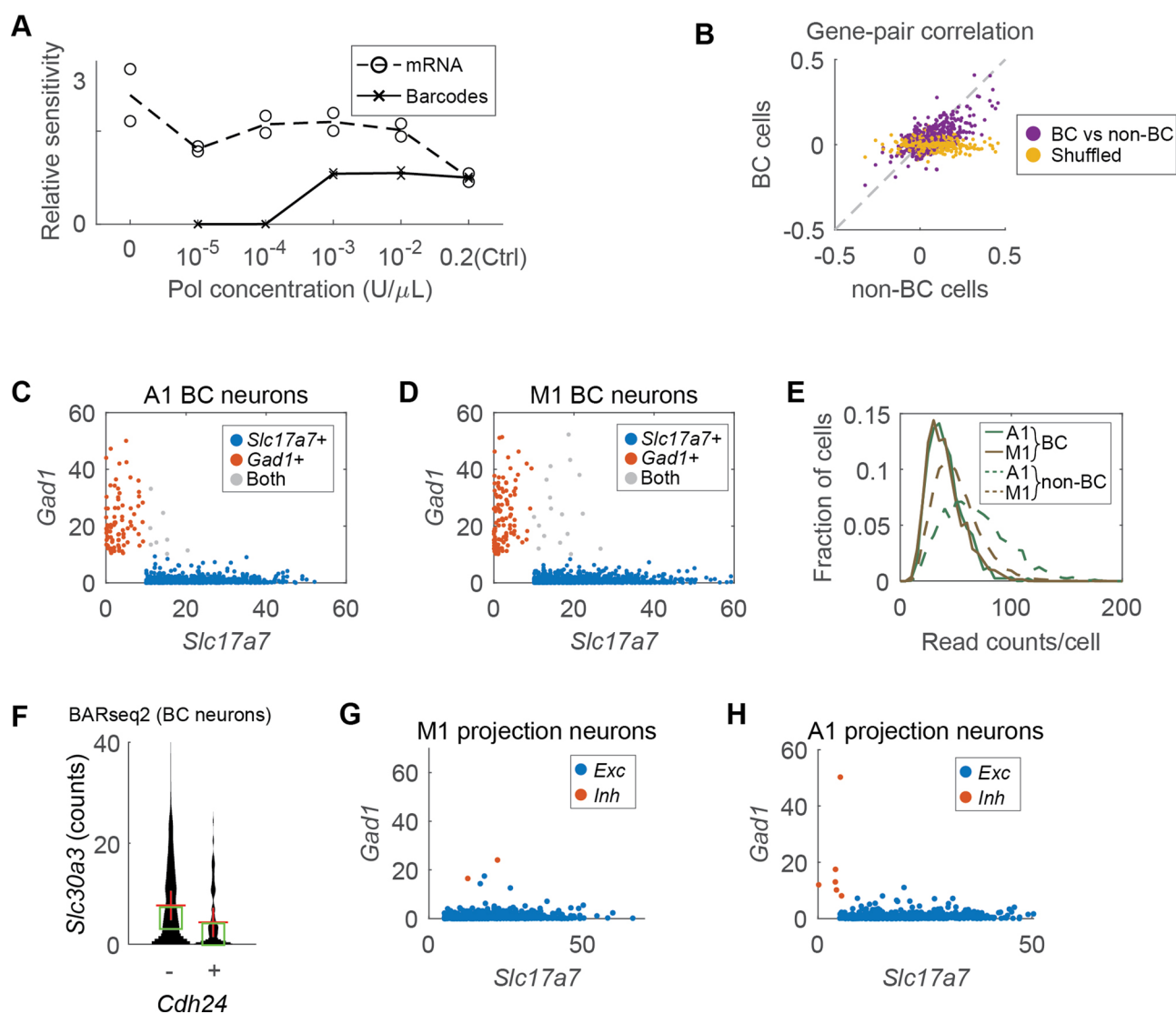
Extended Data Fig. 3 | Comparison between BARseq2 and Allen gene expression atlas. Gene expression patterns in auditory cortex identified by BARseq2 are plotted next to in situ hybridization images of the same genes in Allen gene expression atlas (ABA) and the quantified laminar distribution of the gene in both datasets. Only genes that had coronal images in the Allen gene expression atlas are shown. Blue lines indicate the boundaries of the cortex in both BARseq2 and ABA images. In the laminar distribution plots, dots represent values from two BARseq2 samples (purple) and one ABA sample (blue) per gene. Lines indicate means across samples.



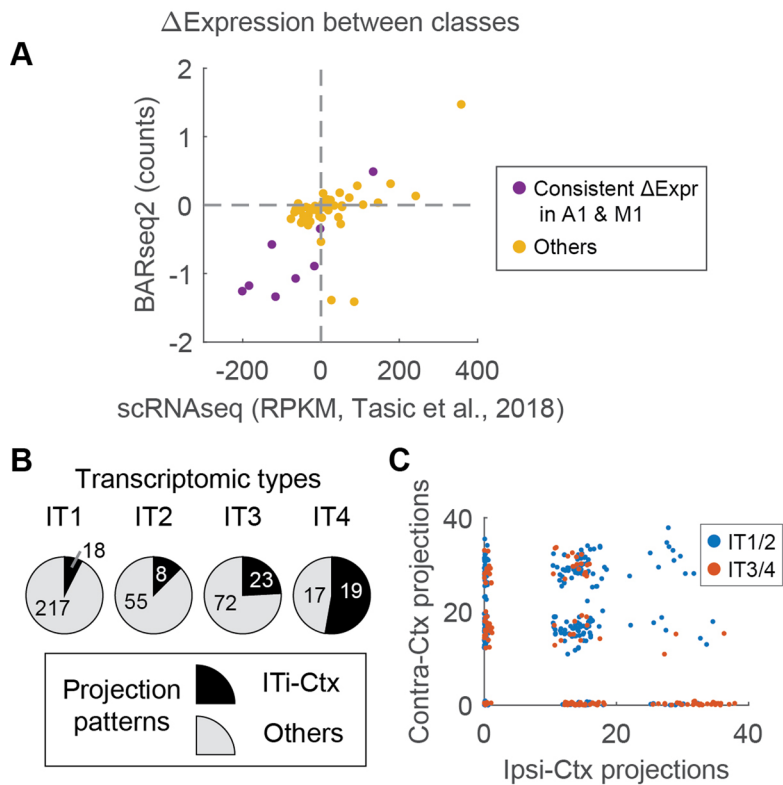
Extended Data Fig. 4 | The distribution of read counts per cell for the indicated genes in auditory cortex (green) and motor cortex (brown). Asterisks indicate genes with significant difference in expression between the two areas ($p < 0.05$ using two-tailed rank sum test after Bonferroni correction). p values after Bonferroni correction are indicated on top.



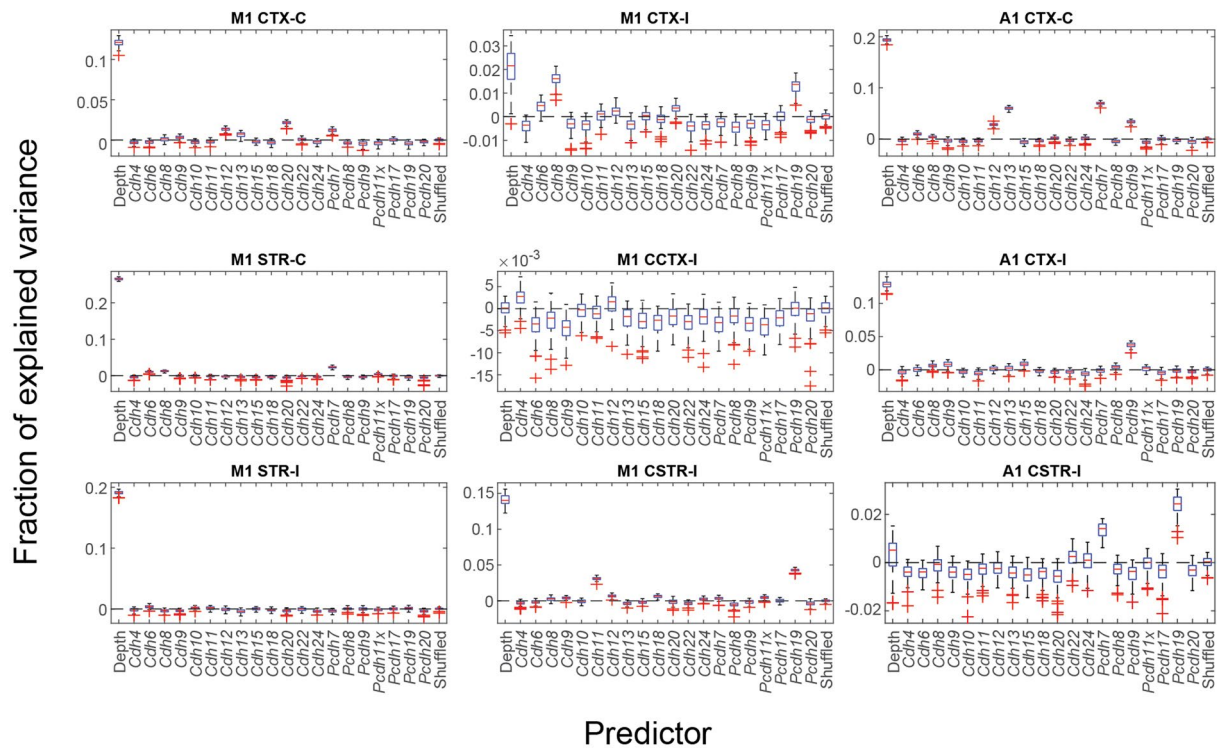
Extended Data Fig. 5 | Transcriptomic typing using BARseq2. a, b, *Slc30a3* expression in excitatory neurons with or without *Cdh24* expression in single-cell RNAseq (a) from Tasic, et al.³ or in BARseq2 (b). A cell is considered expressing *Cdh24* if the expression is higher than 10 RPKM in RNAseq or 1 count in BARseq2. Red crosses indicate means and green squares indicate medians. c, Expression density (means and individual data points) across laminar positions for the indicated genes. n = 3 slices for the three-gene panel and n = 5 slices for the 65-gene panel. d, Precision and recall of cell typing using the marker gene panel across nine single cell datasets. N = 9 independent datasets shown in (e). In each box, the center shows the median, the bounds of the box show the 1st and 3rd quartiles, the whiskers show the range of the data, and points further than 1.5 IQR (Inter-Quartile Range) from the box are shown as outliers. e, Breakdown of average performance for each cell type in each dataset. The datasets are: scSSALM and scSSV1 are single cell SmartSeq datasets from ALM and V1 respectively³. All other datasets are BICCN M1 datasets²³ and the name indicates the technology used (sc = single cell, sn = single nuclei, Cv2/3 = Chromium v2/3, SS = SmartSeq). f, Average cell typing performance for six normalization strategies. N = 9 independent datasets shown in (e). The box plots are generated in the same way as (d). g, Confusion matrix showing overlap between prediction and annotations, normalized by predictions. This plot emphasizes precision; it indicates the probability that a given prediction was correct. h, Confusion matrix showing overlap between prediction and annotations, normalized by annotations. This plot emphasizes recall; it indicates the probability that a given annotation was recovered.



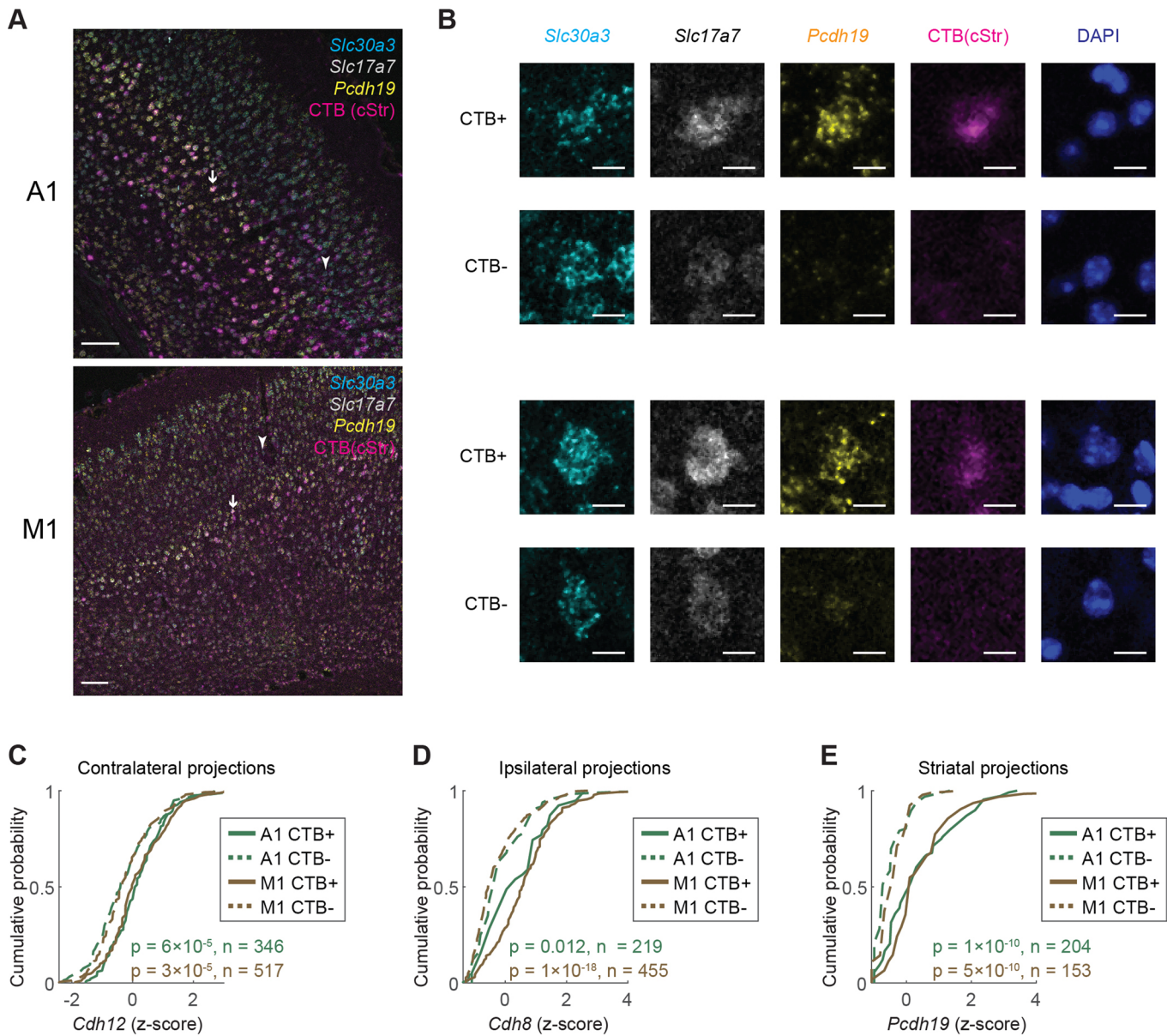
Extended Data Fig. 6 | Correlating gene expression to projections using BARseq2. **a**, Relative sensitivity of BARseq2 to barcodes (solid line) and endogenous mRNAs (dashed line) using the indicated concentration of Phusion DNA polymerase. Sensitivities are normalized to the original BARseq condition (*Ctrl*). Circles and crosses show individual data points across $n=2$ slices. **b**, Correlation between pairs of genes in barcoded cells (y-axis) and in non-barcoded cells (x-axis) as determined by BARseq2. Shuffled data (yellow) are also plotted for comparison. **c,d**, $Slc17a7$ (x-axes) and $Gad1$ (y-axes) expression in barcoded neurons in auditory (**c**) or motor cortex (**d**). Only neurons with more than 10 counts in either gene are shown. **e**, The distributions of read counts per barcoded neuron (solid lines) or non-barcoded neuron (dashed lines) in auditory (green) and motor (brown) cortex. **f**, $Slc30a3$ expression in barcoded excitatory neurons with or without $Cdh24$ expression in BARseq2. A cell is considered expressing $Cdh24$ if the expression is higher than 1 count. Red crosses indicate means and green squares indicate median. **g,h**, $Slc17a7$ (x-axes) and $Gad1$ (y-axes) expression in barcoded projection neurons in motor (**g**) or auditory cortex (**h**). Excitatory and inhibitory neurons are color-coded as indicated.



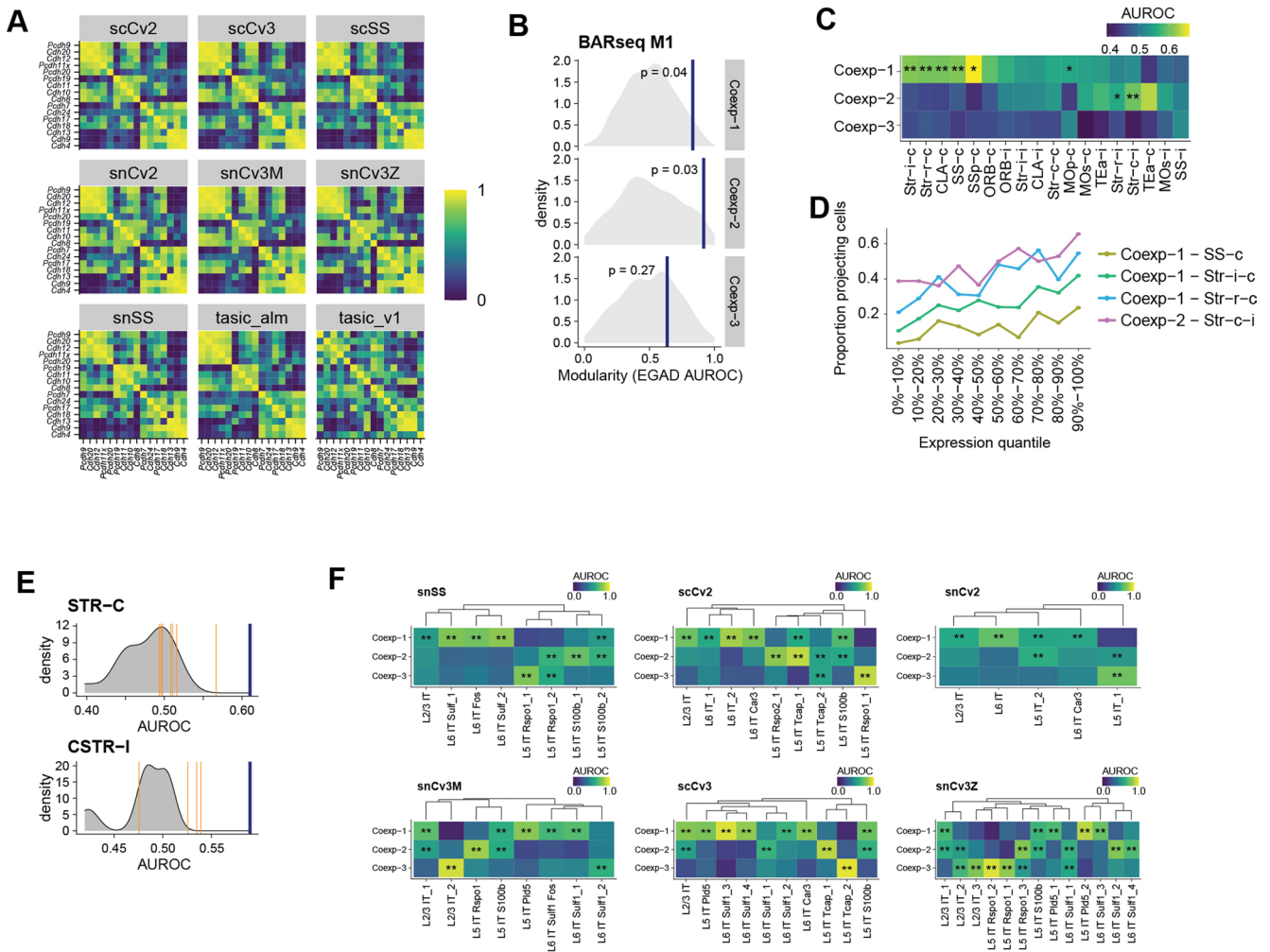
Extended Data Fig. 7 | BARseq2 reveals projection and gene expression differences across major classes and IT subtypes. **a**, Differential gene expression across major classes (IT, PT, and CT) observed using BARseq2 and single-cell RNAseq. Each dot shows the difference in mean expression of a gene across a pair of major classes observed using BARseq2 (y-axis) or single-cell RNAseq (x-axis). Differences in expression that were statistically significant (FDR < 0.05 using two-tailed rank sum tests) in both A1 and M1 as shown by BARseq2 are labeled purple; otherwise they are labeled yellow. The single-cell RNAseq data used were collected in the visual cortex and anterior-lateral motor cortex³. **b**, The fraction of ITi-Ctx neurons in four transcriptomic types of IT neurons in auditory cortex. ITi-Ctx neurons have only ipsilateral cortical projections and no striatal projections or contralateral projections⁶. The number of ITi-Ctx neurons and neurons with other projection patterns for each transcriptomic type are labeled on top of the pie charts. **c**, The projection strengths for contralateral (y-axis) and ipsilateral (x-axis) cortical projections for each IT neuron in auditory cortex. IT1/IT2 neurons are labeled blue and IT3/IT4 neurons are labeled red.



Extended Data Fig. 8 | Variance in projections explained by cadherins and laminar positions. Box plots of variance in each projection modules explained by the indicated predictors after 100 iterations of 10-fold cross validation. Boxes indicate second and third quartiles and whiskers indicate minimum and maximum values excluding outliers. Outliers are shown in red.



Extended Data Fig. 9 | Validation of correlation between cadherins and IT projections. **a**, Representative images of in situ hybridization in A1 (top) and M1 (bottom) slices with CTB labeling in the caudal striatum. Three marker genes and CTB labeling are shown in the indicated colors. Scale bars = 100 μ m. Arrows and arrowheads indicate example CTB+ and CTB- neurons, respectively. Experiments for each combination of targeted gene and CTB labeling condition (*Cdh12* with contralateral labeling, *Cdh8* with ipsilateral labeling, and *Pcdh19* with striatal labeling) were performed in slices from two animals. **b**, Crops of the indicated individual channels of example neurons from (a). Scale bars = 10 μ m. **c,d,e**, Cumulative probability distribution of the expression of *Cdh12* (c), *Cdh8* (d), and *Pcdh19* (e) in neurons with or without retrograde labeling of contralateral (c), ipsilateral (d), or caudal striatal (e) projections. p values from two-tailed rank sum tests after Bonferroni correction and numbers of neurons used for each experiment are indicated. N=2 animals for each experiment.



Extended Data Fig. 10 | Cadherin co-expression modules correlate with IT projections. **a**, Correlation among cadherins in IT neurons in motor cortex identified in the indicated single-cell RNAseq datasets^{3,23}. The datasets included are: tasic_alm and tasic_v1 are single cell SmartSeq datasets from ALM and V1 respectively²; all other datasets are BICCN M1 datasets²³; the name indicates the technology used (sc = single cell, sn = single nuclei, Cv2/3 = Chromium v2/3, SS = SmartSeq). **b**, Modularity (EGAD AUROC) of co-expression modules in BARseq2 M1 against null distribution of modularity (node permutation). BARseq2 modularity is shown by the blue lines with the corresponding p-values. P values are calculated using a one-sided non-parametric node permutation test without multiple comparison correction. **c**, Association (AUROC) between cadherin co-expression modules and the indicated projections. Significant associations are marked by asterisks (* FDR < 0.1, ** FDR < 0.05). **d**, Fractions of neurons with the indicated projections as a function of co-expression module expression. **e**, Distribution of associations of the indicated projection modules with gene expression. Association with significant gene module is shown by a blue line; association with single genes from that module is shown by orange lines; association with all other genes is shown by a gray density. **f**, Association of the three co-expression modules in transcriptomic IT neurons in the indicated datasets (AUROC, significance shown as in **c**).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection micro-manager (Version 1.4.23, Build 20190321) and Zeiss Zen 2012 SP5 FP2 (Version 14.0.0.0) were used to capture microscopy data.

Data analysis Custom MATLAB codes are provided at Mendeley Data (<http://dx.doi.org/10.17632/jnx89bmv4s.1>) as indicated in the data availability section in Methods. The codes were run on MATLAB 2018b. The gene expression module analysis uses the R library "dynamicTreeCut" (v1.63.1) ran in R v4.0.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Target area sequencing data (Fig. 4) are deposited at SRA (SRR12247894, SRR12245390, and SRR12245389). Single-cell RNAseq data are deposited at SRA (SRR13716225). Raw in situ sequencing images (Fig. 2-4) are deposited at Brain Image Library (<https://download.brainimaginglibrary.org/06/35/0635a0b3b0954c7e/>). Other data and processed sequencing data are deposited at Mendeley Data (<http://dx.doi.org/10.17632/jnx89bmv4s.1>).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. Because this manuscript focuses on reporting the development of the BARseq2 technique, we performed BARseq2 on three biological replicates to ensure the reproducibility of the technique.
Data exclusions	No data was excluded
Replication	The main BARseq2 experiments were performed on three animals and multiple findings from BARseq2 are validated using other techniques. The main new findings were validated using FISH and retrograde labeling. This validation experiments was replicated successfully across two animals for each experiment.
Randomization	Randomization was not applicable because the findings did not rely on comparison across different biological samples, but among different neurons. The scoring of neuronal properties (i.e. gene expression and projections) was automated and did not involve human scoring.
Blinding	Blinding was not relevant to the main dataset because the findings did not rely on comparison across different biological samples, but among different neurons. The scoring of neuronal properties (i.e. gene expression and projections) was automated and did not involve human scoring. Other experiments (i.e. comparison across experimental conditions) were also quantified without human input.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Male C57BL/6J mice of age 8-10 wks were used. The animals were housed at maximum of 5 in a cage on a 12 hrs on/12 hrs off light cycle. The temperature in the facility was kept at 22 C with a range not exceeding 20.5 C to 26 C. Humidity was maintained at around 45-55% not exceeding a range of 30-70%. A list of animals used is provided in Supp. Table S1.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected in the field.
Ethics oversight	All animal procedures were carried out in accordance with the Institutional Animal Care and Use Committee protocol 19-16-10-07-03-00-4 at Cold Spring Harbor Laboratory.

Note that full information on the approval of the study protocol must also be provided in the manuscript.