Supplemental Notes

Supplemental Note 1: Sindbis virus

MAPseq requires that a propagation-incompetent virus be used for barcode delivery, i.e. after a neuron is infected with a particular barcode, the virus carrying this barcode should not propagate and spread to other cells. If the virus did propagate, barcodes would spread from cell to cell, and unique labeling of neurons by barcodes would break down as many neurons would now share the same barcode.

Initial Sindbis virus libraries prepared with the conventional helper construct DH(26S)5'SIN induced GFP labeling not only at the injection site, but also occasionally at sites far away from the primary site of injection (Supplementary Fig. 1). Such distal labeling has previously been interpreted as retrograde infection (Furuta et al., 2001), but we recently showed that it does not arise from retrograde spread, but is due instead to secondary infection (Kebschull et al., 2015). Consistent with secondary spread, a subset of barcodes showed unexpectedly high expression levels in single target areas ("spikes"; Supplementary Fig. 1e) when we performed MAPseq with barcoded virus packaged using the conventional helper construct. Indeed, we find that the barcode expression level in these spikes is comparable to the expression level of the same barcodes at the primary injection site (Supplementary Fig. 1f). These observations strongly suggested that the observed spikes originate from ectopically infected cell bodies that are labeled with the same barcode as a neuron at the injection site. Given the high diversity of the viral library, such double labeling is exceedingly unlikely to occur by chance if labeling were due to retrograde infection, but would be expected if the virus used propagated inside the brain.

We therefore designed a new helper construct, (DH-BB(5'SIN;TE12ORF); (Kebschull et al., 2015)), to eliminate secondary spread. When we used this modified helper construct, which minimized co-packaging of Sindbis virus Defective Helper RNA, we almost completely eliminated secondary infection, and were unable to detect any more spikes by sequencing (Supplementary Fig. 1g). Sindbis virus packaged by DH-BB(5'SIN;TE12ORF) thus fulfills the requirements for use in MAPseq. In all subsequent MAPseq experiments we used viral libraries prepared with the modified helper virus.

Supplemental Note 2: Labeling neurons with barcodes

In MAPseq we randomly label neurons with barcodes from a viral library to provide them with a unique identity. Ideally, every infected neuron would have a single, unique barcode. There are two deviations from this ideal scenario: (i) multiple neurons per barcode; and (ii) multiple barcodes per neuron. We consider the implications of the former in more detail below.

Multiple neurons per barcode

Multiple neurons per barcode, i.e. degenerate labeling, is problematic as it leads to incorrect results. Consider for example two neurons, A and B, that project to distinct cortical target areas. If by chance A and B are labeled with the same barcode (e.g. barcode 11), then MAPseq will return the merged projection pattern of A and B as the projection pattern of barcode 11 (Fig. 3c). While this is indeed the projection pattern of barcode 11, it cannot be interpreted as the projection of a single neuron. Errors of this type can be avoided by using sufficiently diverse viral libraries, thereby minimizing the probability that the same barcode will label two different neurons. This implies that the requisite diversity of the viral library depends on the number of neurons infected.

Here we formulate the mathematical problem: Given a population of k neurons, labeled randomly from a pool of N barcodes, what is the probability that a given neuron will be uniquely labeled? This is closely related to the problem: What is the probability that a given barcode will appear in more than one neuron? These problems are related to the classical problem of drawing balls with replacement from an urn, where every ball corresponds to a barcode sequence, and the probability of drawing each ball is determined by the abundance of this barcode in the library.

We first consider a simplified case, in which we assume that every barcode is equally abundant in the virus library, i.e. that the barcode probability distribution is uniform. What then is the expected number of neurons that share a barcode with at least one other labeled cell? If there are only two neurons A and B, then the probability of neuron B having the same barcode as neuron A is $P(A)=1/N$, so the probability that A's barcode is unique is $1-P(A)$. Generalizing to k infected neurons, the probability that A's barcode is unique is $(1-P(A))^{(k-1)}$, and the probability that

it is not unique is $1-(1-P(A))^{(k-1)}$. As the expected value of a sum is the sum of its expected values, the expected number of non-uniquely labeled neurons is

$$E(X) = k(1-(1-P(A))^{(k-1)}).$$

The fraction of uniquely labeled neurons F is then

$$F = 1-E(X)/k = (1-P(A))^{(k-1)} = (1-1/N)^{(k-1)}.$$

Similarly, the expected number D of barcodes used more than once is

$$D=(k^2)/(2N),$$

where N is the number of barcodes and k is the number of infected cells, and we have assumed $N>>k$.

Now, let us consider the more realistic case, in which the distribution of barcode abundance is not uniform, so neurons are more likely to be labeled with some barcodes than others. To calculate the expected value of non-uniquely labeled neurons in this case, we generalize the reasoning above by including a sum over all barcodes, weighted by their probability. E(X) is then given by

$$E(X) = k * \sum_{i=1}^{N} p_i(1 - (1 - p_i)^{k-1}) ,$$

where p_i is the probability of barcode i=1..N, k is the number of infected neurons and N is the total number of barcodes in the virus library.

To determine the empirical distribution of barcodes in the virus library, we directly sequenced the genomic RNA of an aliquot of our Sindbis virus. Sequencing was performed at sufficient depth to overcome Poisson sampling introduced by Illumina sequencing. After error correction, the absolute abundance of different barcode sequences is a direct measure of the barcode probability distribution (Fig. 3e). Despite error correction, there is a chance of including erroneous barcode sequences when counting barcodes that have a very low molecule count. For all calculations, we therefore chose a conservative threshold, and required at least 3 counts for barcodes to be included in the virus library. Based on this empirically determined distribution of barcode abundances, we then calculated the fraction of uniquely labeled cells as a function of the number of infected cells based on the above derivations (Fig. 3f). Simulations indicate that removing the most abundant barcodes have little effect on the capacity of the library to label neurons uniquely (Supplementary Fig. 2a). These results indicate that the observed non-uniformity in the abundance of barcodes in the library does not substantially interfere with the capacity of the library to uniquely label large numbers of cells.

Supplemental Note 3: False positive and negative rates of MAPseq

MAPseq false negative rate

Like every experimental method, MAPseq is susceptible to both false negatives and false positives. First, we sought to relate the efficiency of MAPseq and thus its false negative rate to established neuroanatomical methods. MAPseq is conceptually closest to GFP-based methods (Oh et al., 2014; Zingg et al., 2014), in which a genetically-encoded fluorophore is expressed in a neuronal population, and fluorescence is detected in targets. The sensitivity and selectivity of such fluorophore-based methods depend on many factors, including expression level, imaging conditions, background fluorescence, etc. To our knowledge there has not been a rigorous and precise quantification of the sensitivity and selectivity of such methods, which would allow us to compute e.g. the probability of detecting a small axon for e.g. a given fluorophore expression level, etc; nor indeed is it clear how one would ground-truth such a quantification. Moreover, direct comparison of MAPseq and fluorophore-based methods on a section-by-section basis would be challenging because the optimal conditions for imaging and RNA extraction differ. We therefore did not attempt a quantitative comparison of the efficiency of MAPseq with that of fluorophore-based methods.

Instead, we compared the efficiency of MAPseq to that of another well-established method, Lumafluor retrobeads which allows us to directly compare the efficiency of bead labeling and MAPseq within the same animal. Briefly, we injected red retrobeads into the olfactory bulb, and MAPseq Sindbis virus into LC (Supplementary Fig. 6). Retrobeads taken up by axons in the olfactory bulb are actively transported back to cell bodies and label bulb-projecting cells. Barcodes from infected LC cells that are labeled with retrobeads should therefore be present in the

bulb. The fraction of barcodes recovered from retrobead-labeled LC neurons that are also detected in the olfactory bulb by MAPseq thus provides a neuron-by-neuron estimate of the MAPseq false-negative rate.

To calculate this measure of efficiency, we performed MAPseq on the olfactory bulb and sequenced the barcode complement of individual bead and Sindbis labeled LC cells by dissociating LC, and picking individual red and green cells using glass pipets (Sugino et al., 2006). Producing a single cell suspension from tissue slices involves digestion of the extracellular matrix and trituration of the tissue, which inevitably leads to breaking of processes and release of barcode mRNA into the bath. Given the very high expression levels of Sindbis virus, it was critical to determine the contribution of barcodes present freely floating in the bath or in cell debris, as these barcodes will be collected alongside the labeled cells and sequenced, and will later be indistinguishable from cell resident barcodes except for their abundance. We measured this background noise distribution by collecting cells that were GFP-negative, but were bead labeled. Since GFP-negative neurons do not express barcodes, any barcodes recovered from such cells represent contamination. We used the level of such contamination to establish the threshold for true barcode expression in intact isolated neurons.

We collected 45 neurons that were labeled with both GFP/barcodes and with red retrobeads from the olfactory bulb, and 9 neurons labeled only with red retrobeads to determine the background noise level of barcode expression. We found that MAPseq efficiency is high: 91.4+/-6% (mean +/- std error) of all barcodes from cells that project to the bulb as determined by bead labeling also appear to project to the bulb by sequencing (across 3 animals; Supplementary Fig. 6d). This estimate is robust over a large range of reasonable estimates for the level of background barcode contamination (Supplementary Fig. 6f). We therefore conclude that the false negative rate of MAPseq is 8.6+/-6%.

MAPseq false positive rate

A false positive event in MAPseq is the detection of a barcode in a target area to which the neuron expressing the barcode does not project. There are two potential sources of false positives. First, we might correctly detect a barcode that does indeed target this particular area, but we might mistakenly identify it as a different barcode (due e.g. to sequencing errors). Alternatively, barcodes that arise from other samples (slices), or from outside sources, might contaminate the target sample. (A third kind of error, those arising from insufficient barcode diversity, might also be considered a special case of false positives, but are considered separately above in “Unique labeling of neurons with barcodes”).

Due to the large combinatorial space of barcodes, it is exceedingly unlikely to mistake one barcode for another because of PCR or sequencing errors (see Supplementary Note 4). Contamination, however, is a concern and needs to be quantified.

LC neurons project primarily to the ipsilateral hemisphere (Waterhouse et al., 1983), and only a small fraction of LC neurons project to both ipsilateral and contralateral cortex (Room et al., 1981). Quantifying the projection strength of neurons to the contralateral hemisphere relative to their projection to the ipsilateral hemisphere therefore provides an upper bound on the rate of contamination, and thus on the false positive rate of MAPseq. Note that samples from the ipsi- and contralateral hemisphere were processed intermixed and out of order. Cross-contamination between samples from the ipsi and contralateral side should therefore be comparable to contamination between samples from the ipsilateral side only, and should be a good measure of overall contamination levels.

We used the MAPseq dataset of the bilaterally injected animal (Fig. 7) to calculate this upper bound to the false positive rate. Briefly, we calculated the ratio of the total number of barcode molecules detected in the contralateral hemisphere to the total number of barcode molecules detected in the ipsilateral hemisphere for all barcodes that projected more strongly to the ipsi- than contralateral side (n=115). The mean ratio, and thus upper bound to the MAPseq false positive rate is 1.4+/-0.8% (mean+/-std error). Note that our assumption that LC neurons project only ipsilaterally is conservative; violations of this assumption would increase the estimated false positive rate. Thus we conclude that MAPseq has a low false positive rate. These results indicate that MAPseq provides both sensitive and reliable mapping of long-range projection targets of a large number of neurons.

Supplemental Note 4: Bioinformatics

Raw MAPseq data consist of two .fastq files containing Illumina sequencing results, where paired end 1 covers the barcode sequence, and paired end 2 covers the 12-nt UMI and the 6-nt SSI (Supplementary Fig. 1h and 3). To convert these sequencing data into projection maps, we first preprocessed the data in bash, before analyzing them in Matlab (Mathworks).

Preprocessing of sequencing data.

Briefly, we stripped the fastq files of their quality information and trimmed the reads to the relevant length, then merged paired end 1 and 2 into a single file. Each line of this file corresponded to a single read containing the 30-nt barcode, the 2-nt pyrimidine anchor (YY), the 12-nt UMI and the 6-nt SSI. We de-multiplexed the reads based on the SSI using the *fastx_barcode_splitter* tool and filtered the reads to remove any ambiguous bases. We then collapsed the reads to unique sequences and sorted them.

Next, we selected a threshold of how many reads a sequence has to have to be considered for analysis. We were guided by earlier work on the effect of PCR amplification during Illumina library generation on next generation sequencing data (Kebschull and Zador, 2015). In this previous work, we found that when amplifying a pool of unique barcode sequences by PCR, the sequence rank profile of the Illumina results consists of a plateau of sequences with roughly equal read counts, followed by a shoulder and a long tail. The tail of this distribution is formed almost exclusively by PCR errors. In the MAPseq datasets, we therefore manually selected a minimum read threshold to remove the tail of the sequence rank profile from the analysis. This avoids contamination of our dataset with large numbers of PCR and sequencing errors and simplifies subsequent error correction and analysis steps.

We then collapsed the remaining reads (30-nt barcode+YY+12-nt UMI) after removal of the 12-nt UMI to convert reads into molecule counts. Note that we here ignored any potential PCR or sequencing errors in the 12-nt UMI, which will lead to a slight, but uniform, overestimation of molecule counts as two copies of the same cDNA with an error in the UMI only will be counted as two distinct molecules rather than one.

Split of barcodes from spike-ins.

Spike-in molecules are barcodes of length 24 followed by the constant sequence ATCAGTCA, and are therefore easily distinguished from barcodes expressed from the virus (Supplementary Fig. 1h). As they carry different information, we split the uncorrected barcode data into spike-ins (perfect match to N_{24}ATCAGTCA) and virally expressed barcodes (no N_{24}ATCAGTCA sequence, but N_{30}YY) and processed them separately.

Error correction.

A random barcode of 30nt length has a potential diversity of $4^{30} \approx 10^{18}$ different sequences. If we sample a relatively small number of barcodes from this enormous diversity, the chosen barcodes are likely very different from each other. Therefore many mutations to any given barcode are necessary to convert it into any other barcode of the chosen set.

We exploited this fact to correct errors in the sequenced barcodes. Using the short read aligner *bowtie* (Langmead et al., 2009), we performed an all-against-all mapping of all barcode sequences with >1 counts, allowing up to 3 mismatches, and forcing bowtie to output all possible alignments. We then constructed a connectivity matrix of all barcode sequences, where bowtie alignments are the connections between sequences. We used Matlab to find all connected graph components, that is all barcodes that mapped to each other, and collapsed the molecule counts of each of the members of such a connected component to the sequence of the most abundant member. We then removed low complexity sequences—a common artifact of Illumina sequencing—by filtering barcodes with stretches of more than 6 identical nucleotides. Finally, we compared all error corrected barcode sequences to the error corrected barcode sequences found in the original virus library and kept only those barcodes for analysis that had a perfect match in the virus library.

Code for preprocessing of all MAPseq libraries can be found in *preprocessing.sh* and *matlab_preprocessing.m*. The viral library was processed using *viruslibrary_preprocessing.sh* and *viruslibrary_matlabcode.m*.

Analyzing the projection pattern.

The described workflow results in a list of barcode sequences and their molecule counts in each target area and the injection site. Using Matlab, we then matched the barcode sequences in the injection site (reference barcodes) with the barcode sequences in the target sites, constructing a barcode matrix of size [# of reference barcodes]x[# of target

sites + # of injection sites] which then acts as the basis of all further analysis. Note that barcodes that appear in target areas and not the injection site ('orphans') are very rare and have low abundances, consistent with an interpretation of orphan barcodes as contaminants.

To exclude low confidence projection patterns from analysis, we required each barcode to have more than 100 counts in the injection site and at least one target area with more than 30 counts.

Code can be found in *producebarcodematrix_unilateral.m* and *producebarcodematrix_bilateral.m*.

Barcode matrix normalization.

Raw barcode counts are very useful to survey the data available and to form intuitions about the mapped projection patterns. However, to compute summary statistics, we normalized the raw barcode matrix. We first normalized each target area by the number of unique spike-in molecules detected in each, to normalize for varying reverse transcription, PCR or library making efficiencies. We then normalized each area by the amount of β-actin per µl of total RNA (as measured by qPCR) to correct for varying tissue input and RNA extraction efficiencies. Lastly, we normalized all barcodes to sum to 1 across all target areas to correct for different expression levels of different barcodes.

Code for all analysis of the barcode matrix can be found in *analyse_unilateralinjections.m* and *analyse_bilateralinjections.m*. The raw and normalized projection patterns of all 995 traced neurons can be found in *bigmatrixcounts.mat* and *bigmatrix.mat*, respectively.

Peak finding.

To summarize LC projection patterns, we set a number of criteria to define peaks for each barcode. First, peaks need be at least half as high as the maximal barcode count across all target sites. Second, peaks need to be separated by at least 3 slices, and third, peaks need to rise at least their half maximal height from their surroundings ('prominence'). Code used to find peaks can the found in *detectpeaks.m*.

Identification of double-infected cells.

In order to identify pairs of barcodes that originated from double-infected cells, we looked for projection profiles from individual mice that are more similar than expected for barcodes from different cells. Briefly, we calculated the minimum pairwise Euclidean distance of every barcode profile to any other barcode profile of a particular mouse in z-scored space ("within mouse"). We then constructed a null-distribution by repeatedly calculating the minimum pairwise distances for every barcode profile from that mouse to a random sample of the same size of barcodes profiles obtained by sampling the other three mice in this MAPseq dataset ("between mice"). Distances that appear in the "between mice" null-distribution result from the similarity of the projection profiles of different cells. Therefore, distances in the "within mouse" set lower than those explained by this null distribution suggest that the two barcode profiles are more similar than would be expected for two separate cells. The two barcodes that correspond to this low distance probably arise from a single double-infection cell. Accordingly, we defined those barcode pairs as originating from double-infected cells that had distances in the left tail of the null distribution subject to Bonferroni correction for multiple hypothesis testing.

To estimate the overall number of barcodes from double-infected cells in every animal, we calculated area between the probability density function of the "within mouse" distances and the "between mouse" distances and multiplied it by the total number of barcode pairs in the dataset. We then took the number of barcode pairs as our estimate of the number of barcodes in double infected cells, which corresponds roughly to 2x the number of double infected cells.

Code for this analysis can be found in *finddoubles.m*.

Single cell analysis.

Code for analysis of single cells sequencing data can be found in *preprocessing_singlecells.sh*, *matlab_preprocessing_singlecells.m* and *analyse_singlecells.m*.

False positive rate.

Code for the calculation of the false positive rate can be found in *analyse_bilateralinjections.m*.

Dimensionality reduction and clustering.

Code for t-SNE dimensionality reduction and hierarchical clustering of cortico-cerulear neurons can be found in *doclustering.m*.

Supplemental Note 5: Spike-in recovery

To assess the efficiency of barcode recovery in MAPseq, we added a known amount of spike-in RNA (Supplementary Fig. 1h) into every sample (Fig. 4a, Supplementary Fig. 3) and quantified the number of distinct spike in molecules in the sequencing results. The ratio of the number of recovered spike-in molecules to the number of input molecules then is the probability of detection of any given barcode molecule.

Detection efficiencies are relatively constant across areas and animals (Supplementary Fig. 7c,d) and average to P(detection)=0.024 for target areas. This implies that when we do not detect a barcode in an area, there are less than 123 barcode mRNA molecules present in that sample with a confidence >95%, as dictated by the negative binomial distribution.

Note here, that this measure of barcode detection probability is based on the efficiency of going from total RNA to sequencing results. It is blind to losses incurred during extraction of total RNA from tissue, such that the overall MAPseq detection efficiency is likely somewhat lower than what we estimate.

Supplemental Figures

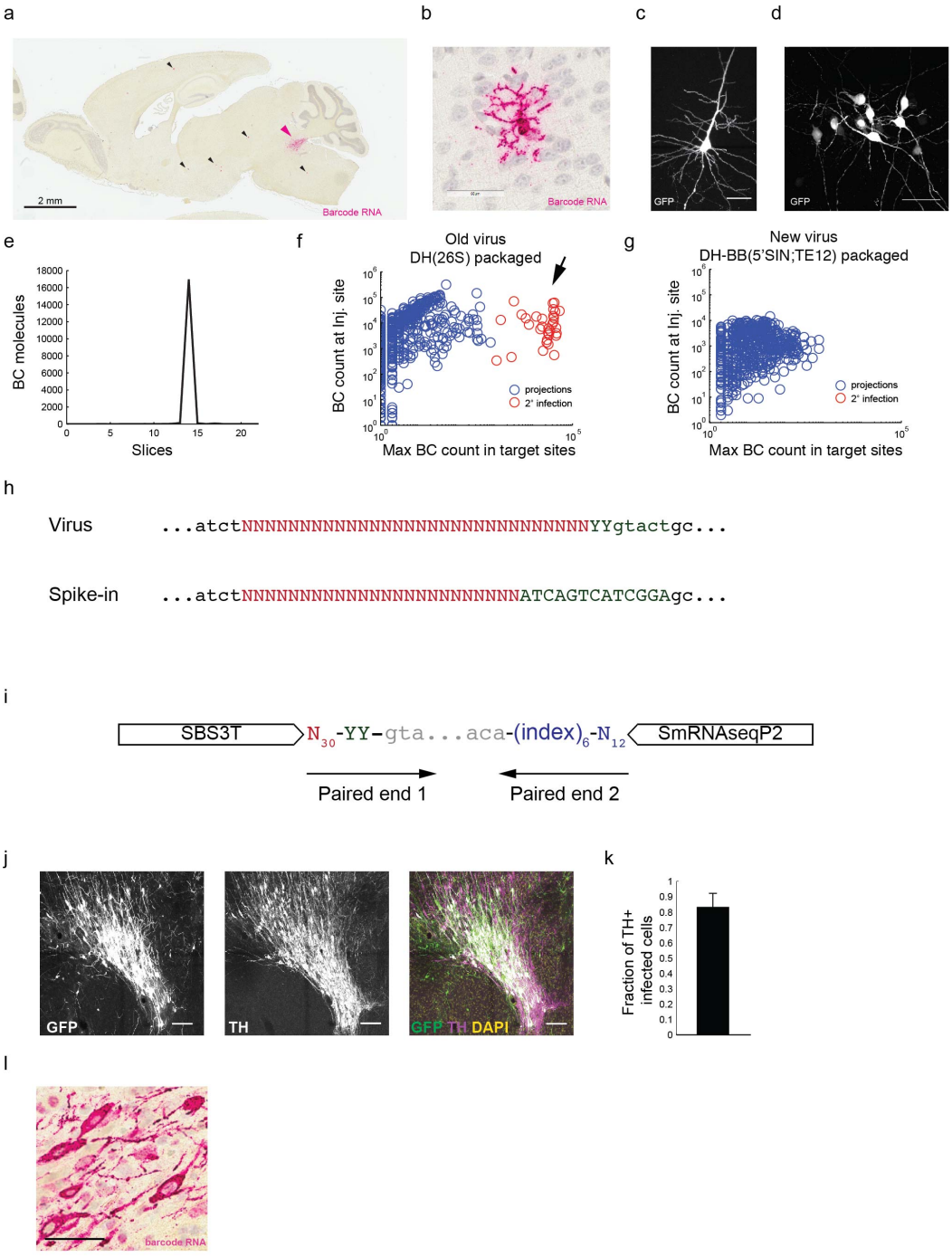

Figure S1; related to Fig. 1 and Fig 2

Barcodes are delivered to LC neurons using recombinant Sindbis virus. (a-g) The replacement of the conventional packaging system, DH(26S)5'SIN, with a modified packaging system we developed, DH-BB(5'SIN;TE12ORF), largely eliminates infection of cells distal to the injection site. After injection of conventionally packaged virus, (a,b) *in situ* hybridization for barcode mRNA labels cells far away from the injection site. Pink arrow = primary injection site; black arrows = secondary infection. (c,d) Similarly we can detect GFP positive cells or clusters of cells far away from the injection site after injection of conventionally produced virus. Scale bar = 50μm. (e) MAPseq data

produced using DH(26S)5'SIN packaged virus shows spurious barcodes with extremely high abundance in a single target site only ("spikes"), which arise from barcodes expressed in cortical somata secondarily infected by propagation of viral particles from the axons of infected LC neurons. (f) Expression levels of these high abundance barcodes are comparable to that of barcodes in the injection site. (g) Changing the packaging system to the new DH-BB(5'SIN;TE12ORF) produces a propagation incompetent Sindbis virus and eliminates these high abundance barcodes. All MAPseq results described in this manuscript made use of this new virus. (h) Differences in the sequence of viral barcodes and spike-in RNA allow easy discrimination of the two. (i) Structure of the final sequencing amplicon. (j,k,l) Stereotaxic injection of Sindbis virus reliably infects LC and fills cell bodies and axons with barcode mRNA. (j) Maximum z-projection of a representative Sindbis injection shows excellent overlap with the TH-stained LC, confirming successful stereotactic targeting of the nucleus. Scale bar=100μm. (k) Quantification of the fraction of infected cells that are also TH+ confirms reliable targeting of LC by streotactic injection. N=6. Mean +/- s.d. is shown. (l) RNA *in situ* of barcode mRNA showing good fills of cell bodies at the injection site. Scale bar=50μm.

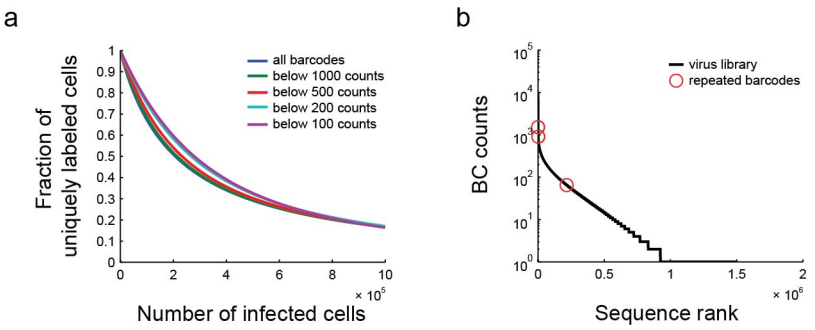

Figure S2; related to Fig.3

The diversity of the MAPseq virus library is sufficient to uniquely label many cells. (a) The number of cells that can be uniquely labeled using our virus library does not change dramatically when we bioinformatically remove overrepresented barcodes from the library. The legend indicates which barcodes are still considered for labeling. (b) Position of the three barcodes that were traced in more than one of four animals. Two of the three are highly abundant in the virus library.

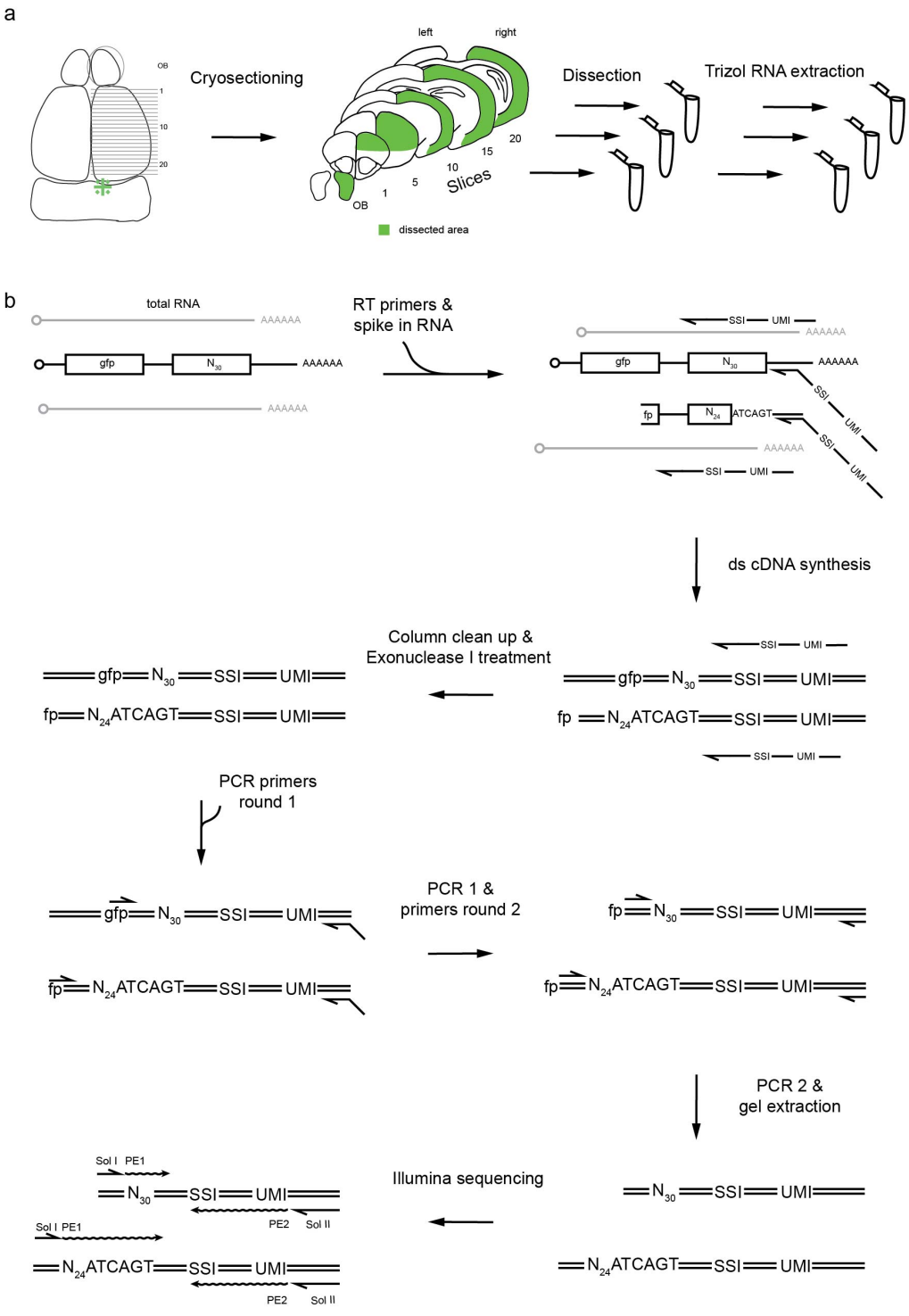

Figure S3; related to Fig. 4

MAPseq workflow. (a) We cryosection a flash frozen brain and dissect out areas of interest. We then extract total RNA from every area individually. (b) To the total RNA from every area, we add a known amount of spike-in RNA and reverse transcription primers containing unique SSIs and UMIs. We produce double stranded cDNA, and digest leftover reverse transcription primers using Exonuclease I to avoid UMI containing primers to participate in

subsequent PCR reactions. We then perform two rounds of nested PCR, bringing in the PE2 sequencing primer binding site and P7 sequence as 5' overhangs of the reverse primer. After gel extraction, the amplicons are ready for Illumina sequencing.

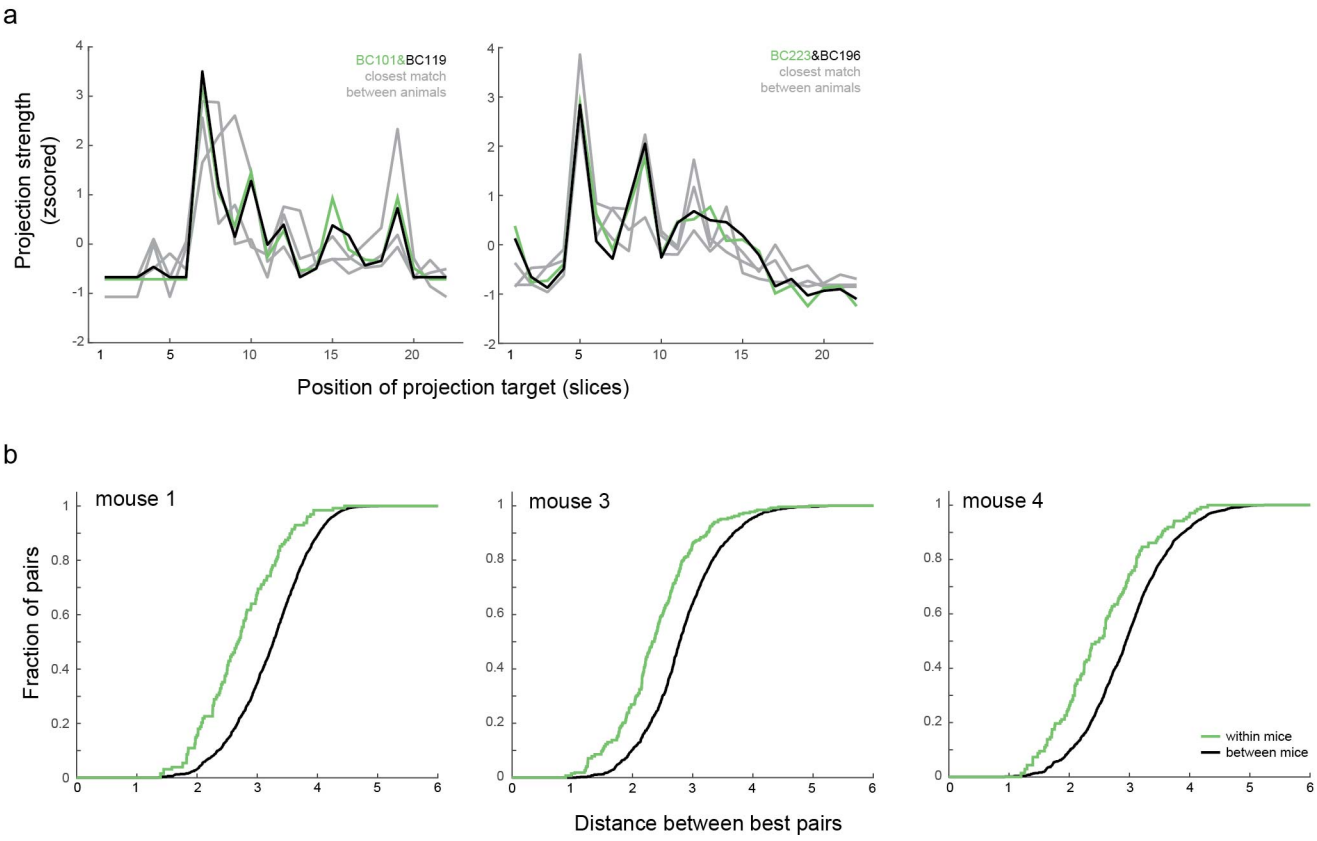

Figure S4; related to Fig. 5

MAPseq provides a robust readout of single neuron projection patterns. (a) The same example pairs of barcodes profiles that are more similar than expected by chance as shown in Fig 5a. In grey we indicated best matches of the barcode profiles across animals from 5 independent samplings of the comparison animal. (b) Cumulative distribution function of distances of best barcode pairs within and across animals for animals 1, 3 and 4.

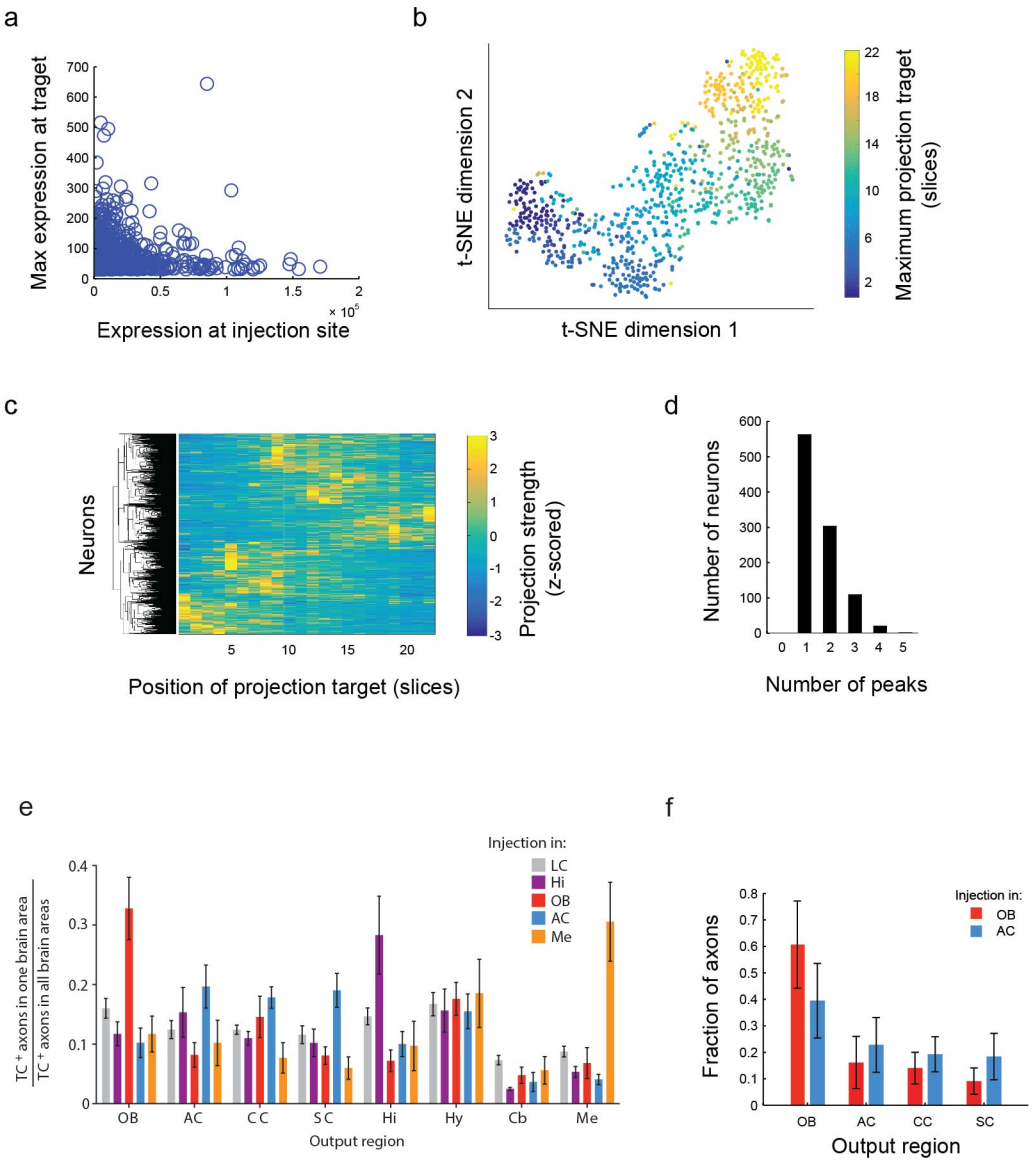

Figure S5; related to Fig. 6

Aggregate projection of MAPseq traced neurons reproduces homogeneous bulk projection, but individual projection patterns are non-homogeneous. (a) There is no correlation between the expression level of a barcode at the injection site and its maximum projection strength to a target area. (b) t-SNE dimensionality reduction of cortically projection LC neurons reveals an orderly separation of neurons according to their maximum projection location. (c) Hierarchical clustering of z-scored projection profiles of cortical projection neurons however reveals no striking clustering. (d) Histogram of the number of detected peaks for all MAPseq traced neurons. For peak definitions see Supplementary Note 4. (e,f) Simulation of CAV-cre injection and axon tracing from MAPseq data reproduces the non-specific output pattern of LC neurons reported by Schwarz *et al*. (Schwarz et al., 2015). (e) Reproduction of Figure 4d of ref (Schwarz et al., 2015). Briefly, Schwarz *et al*. injected retrograde CAV-cre virus into a number of areas including olfactory bulb and auditory cortex, and cre dependent TVA-mCherry-AAV into LC. They then counted the number of mCherry labeled LC axons in a number of output areas and normalized the number of axons across all output areas. They could thereby quantify the projection strength of groups of LC neurons defined by their projection to the injection site and found that most groups of LC neurons project equally to all output areas. (f) Results of our MAPseq data based simulation of the experiment preformed by Schwarz *et al*, plotted in the same way. Briefly, we simulated CAV-cre injections into olfactory bulb or auditory cortex by labeling barcodes that are

present at more than 50 counts in either olfactory bulb or auditory cortex. We then summed up the normalized counts of the labeled barcodes in slices containing the output regions and normalized the resulting projection strength across all output regions, thus mimicking the counting of labeled axons in output regions. In contrast to the idiosyncratic single cell projection pattern reported by MAPseq, this simulation recapitulates the findings of Schwarz *et al*., highlighting the importance of single neuron resolution in connectivity mapping. OB = olfactory bulb; AC = auditory cortex; CC = cingulate cortex; SC = somatosensory cortex.

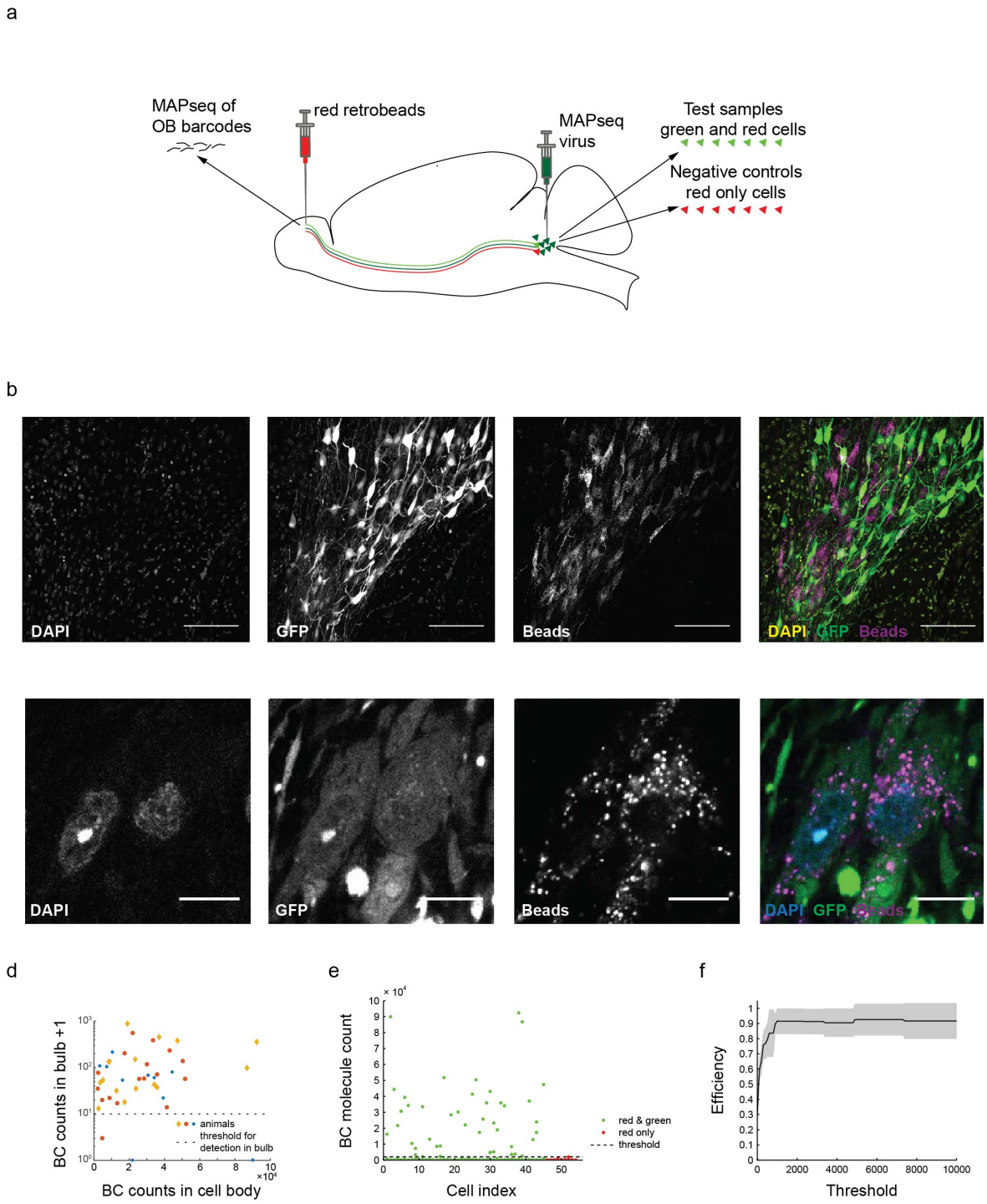

Figure S6; related to Fig. 6

Sequencing of single LC cells reveals low MOI and high MAPseq efficiency. (a) Overview of the experimental design. Red Lumafluor retrobeads label bulb projecting cells in LC. Barcodes present in these cells should also be present in the bulb. (b) Overview image of LC, showing bead and GFP labeling of cells. Scale bar = 100μm. (c) Detailed image of retrobeads and GFP labeled cells. Scale bar = 10μm. (d) Scatter plot showing the relationship of barcode abundance in the olfactory bulb to barcode abundance in individual cells. The dashed line indicates the minimum barcode abundance in the bulb chosen as detection threshold. (e) Scatter plot of abundance of all barcodes

found in the sequenced single cells for both bead and Sindbis labeled cells (n=45 from 3 animals, green) and negative control cells (bead labeled only; n=9 from 3 animals; red). Dotted line indicates the height of the most abundant barcode from red only cells, the threshold chosen to distinguish real from artefactual barcodes. (f) MAPseq efficiency as a function of an increasingly stringent noise threshold. The MAPseq efficiency estimate is not very sensitive to changes in the threshold value. Shaded area indicates s.d. across animals.

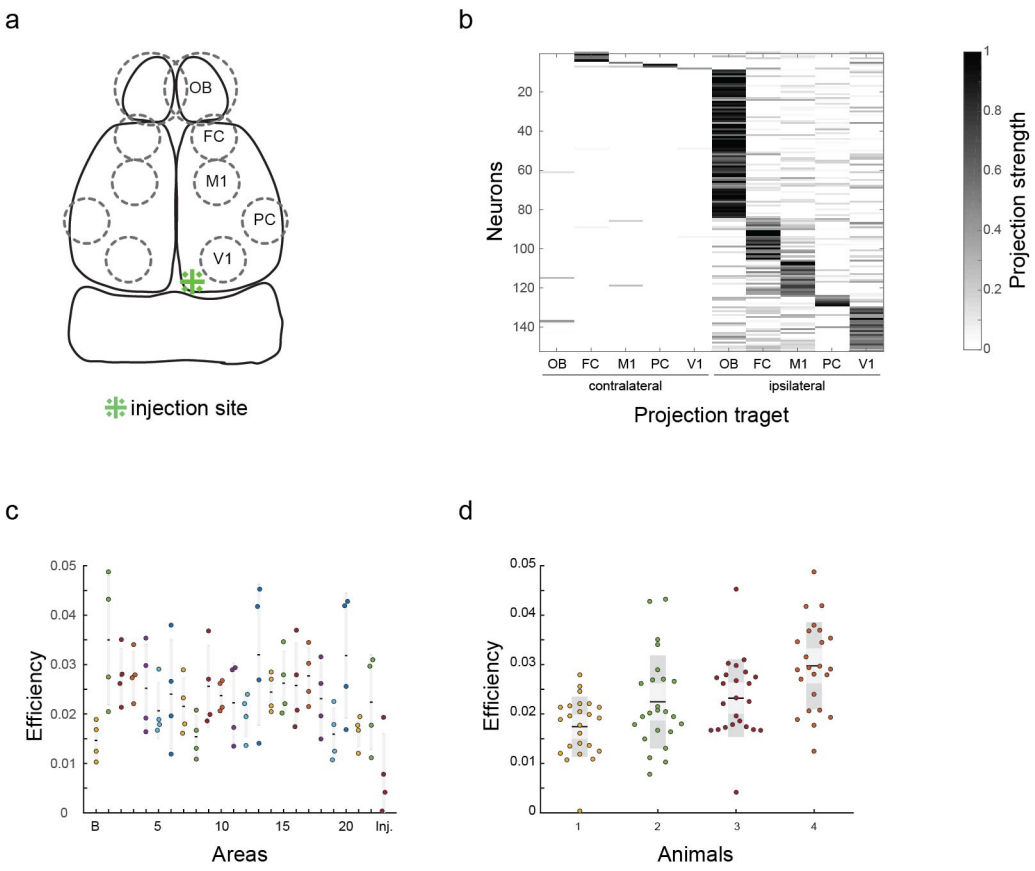

Figure S7; related to Fig. 6

MAPseq can be performed on small target areas. (a) Schematic of dissected areas. FC = frontal cortex; M1 = primary motor cortex; PC = piriform cortex; V1 = primary visual cortex. (b) A heatmap of all ~140 neurons traced across 3 independent animals using DH(26S)5'SIN packaged MAPseq virus. We removed all ectopically infected cells (see Supplementary Fig. 1f) that could have confused tracing results by a maximum abundance cutoff of 1000. Preferential targeting of different ipsilateral areas is clearly evident. (c,d) Efficiency of barcode recovery from total RNA samples in MAPseq is relatively constant across areas (c; n=4 animals) and animals (d) as measured by spike-in RNA recovery.

Extended Experimental Procedures

MAPP-nλ

MAPP-nλ is a modified version of pre-mGRASP (Kim et al., 2011). We stripped the pre-mGRASP protein of the 2A-cerulean fusion and added four repeats of the nλ RNA binding domain (Daigle and Ellenberg, 2007) in the cytoplasmic tail after amino acid 287 of the original pre-mGRASP sequence. We also added a Myc epitope tag followed by the CLIP-tag domain (Gautier et al., 2008) after amino acid 59 of the original pre-mGRASP protein.

Sindbis virus barcode library

The virus used in this study is based on a dual promoter pSinEGdsp construct (Kawamura et al., 2003). We inserted MAPP-nλ after the first subgenomic promoter. Downstream of the second subgenomic promoter, we inserted the GFP coding region followed by closely spaced NotI and MluI restriction sites and four repeats of the boxB motif (Daigle and Ellenberg, 2007). Using this construct, we produced a high diversity plasmid library by inserting a diverse pool of double stranded ultramers (Integrated DNA Technologies) with sequence 5'-AAG TAA ACG CGT AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNN NNN NNN NNN NNN NNN NNN NNN NYY GTA CTG CGG CCG CTA CCT A-3' between the NotI and MluI sites. We then produced Sindbis virus as previously described (Kebschull et al., 2015) using either the conventional DH(26S)5'SIN helper (Bredenbeek et al., 1993) or the new DH-BB(5'SIN;TE12) (Kebschull et al., 2015) helper. We determined the titer of the resulting virus by qPCR as previously described (Kebschull et al., 2015) and determined the viral library diversity by Illumina sequencing of the RNaseI protected genomic virus RNA. Both the genomic construct and the helper construct are available from Addgene under accessions 73074 and 72309, respectively.

Injections

Animal procedures were approved by the Cold Spring Harbor Laboratory Animal Care and Use Committee and carried out in accordance with National Institutes of Health standards.

We pressure injected 180nl of $2x10^{10}$ GC/ml barcoded Sindbis virus uni- or bilaterally into LC of 8-10 week old C57BL/6 males (Jackson Labs) as described (Cetin et al., 2007). We leveled the animal skulls on two axes using lambda and bregma for the AP axis and 2mm laterally from the midpoint between lambda and bregma for the lateral axis. We used coordinates AP=-5.4mm, ML=0.8mm, DV=2.9mm and 3.1mm for LC and measured depth from the surface of the brain. We injected each DV coordinate with 90nl of virus, waiting ten minutes in between each depth. We sacrificed animals 44 hours post injection. For immunofluorescence, RNA *in situ* and histology, we transcardially perfused animals with ice cold saline (9g/l) followed by 4% paraformaldehyde (Electron Microscopy Sciences) in 0.1M Phosphate buffer. For RNA work we extracted the fresh brain and flash froze it on dry ice.

For measurements of MAPseq efficiency, we injected red retrobeads (Lumafluor) into the right olfactory bulb of 8-12 week old C57BL/6 males (Jackson Labs). Briefly, we roughly determined the center of the right olfactory bulb, and measured +/-1mm from the center in the AP axis and performed two craniotomies 2mm apart. We sonicated the beads for 20 minutes prior to injection to homogenize the solution and injected 210nl of stock concentration of beads at three different depths (0.3mm, 0.6mm and 0.9mm DV from the surface of the olfactory bulb) as described (Cetin et al., 2007). Twenty-four hours later, we injected barcoded Sindbis virus into right LC as above and sacrificed the animals 44-48 hours after Sindbis injection.

Immunofluorescence and ISH

We performed anti-GFP staining and RNA *in situ* hybridization on 6μm thick paraffin sections. For immunofluorescence, we used a rabbit anti-GFP antibody ab290 (Abcam; RRID:AB_303395) after heat induced antigen retrieval. We performed RNA *in situ* hybridization using the Panomics ViewRNA ISH Tissue kit (Affymetrix) using anti-GFP probe VF1-10141 according to the manufacturer's protocol (10 minutes boiling and 10 minutes protease treatment). We performed anti-TH staining on floating 70μm vibratome sections using rabbit anti-TH antibody SAB4300675 (Sigma-Aldrich; RRID:AB_11130236).

Spike-in RNA

To produce spike-in RNA, we double stranded an ultramer (Integrated DNA Technologies) with sequence 5'-GTC ATG ATC ATA ATA CGA CTC ACT ATA GGG GAC GAG CTG TAC AAG TAA ACG CGT AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNN NNN NNN

NNN NNN NNN NNN NAT CAG TCA TCG GAG CGG CCG CTA CCT AAT TGC CGT CGT GAG GTA CGA CCA CCG CTA GCT GTA CA-3' and then *in vitro* transcribed the resulting dsDNA using the mMessage mMachine T7 *in vitro* transcription kit (Thermo Fisher) according to the manufacturer's instructions.

qPCR

We reverse transcribed total RNA using oligodT primers and Superscript III reverse transcriptase (Thermo Fisher) according to the manufacturer's instructions. We then quantified the amount of barcode and β-actin cDNA by qPCR in SYBR green power master mix (Thermo Fisher) according to the manufacturer's instructions using primers 5'-GAC GAC GGC AAC TAC AAG AC-3' and 5'-TAG TTG TAC TCC AGC TTG TGC-3' for barcode cDNA and 5'-CGG TTC CGA TGC CCT GAG GCT CTT-3' and 5'-CGT CAC ACT TCA TGA TGG AAT TGA-3' for β-actin cDNA.

MAPseq

We cut 300μm thick coronal sections of fresh frozen brains using a Leica CM 3050S cryostat at -12ºC chamber temperature and -10ºC object temperature. To avoid cross-contamination between samples, we took care to cut each section with a fresh, unused part of the blade. We melted each section onto a clean microscope slide and rapidly froze the section again on dry ice before dissecting out the cortex on dry ice using a cold scalpel blade. During dissection, we aimed to avoid known fiber tracts to minimize the contamination of our dataset with fibers of passage. After sample collection, we processed all samples out of order to avoid potential sample cross-contamination from impacting interpretation of MAPseq results.

We extracted total RNA from tissue samples using Trizol reagent (Thermo Fisher) according to the manufacturer's instructions. We mixed the total RNA from the tissue samples with spike-in RNA. We then produced ds cDNA as previously described (Morris et al., 2011) using a gene specific primer of from 5'-CTT GGC ACC CGA GAA TTC CAN NNN NNN NNN NNX XXX XXT GTA CAG CTA GCG GTG GTC G-3', where XXXXXX is one of 65 trueseq like SSI and N_{12} is the UMI. We then cleaned the reaction using the Qiagen MinElute PCR purification kit according to the manufacturer's instructions and treated the eluted ds cDNA with ExonucleaseI (New England Biolabs) to remove remaining primers. We amplified the barcode amplicons by nested PCR using primers 5'-CTC GGC ATG GAC GAG CTG TA-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGT GAT GTG ACT GGA GTT CCT TGG CAC CC GAG AAT TCC A-3' for the first PCR and primers 5'-AAT GAT ACG GCG ACC ACC GA-3' and 5'- CAA GCA GAA GAC GGC ATA CGA-3' for the second PCR in Accuprime Pfx Supermix (Thermo Fisher). We then gel extracted the amplicons using the Qiagen MinElute Gel extraction kit according to the manufacturer's instructions and pooled the individual sequencing libraries based on qPCR quantification using primers 5'-AAT GAT ACG GCG ACC ACC GA-3' and 5'- CAA GCA GAA GAC GGC ATA CGA-3'. We then sequenced the pooled libraries on an Illumina NextSeq500 high output run at paired end 36 using the SBS3T sequencing primer for paired end 1 and the Illumina small RNA sequencing primer 2 for paired end 2.

Efficiency measurements and single cell isolation

After transcardial perfusion with ice-cold artificial cerebrospinal fluid (127mM NaCl, 25mM $NaHCO_3$, 1.25mM $NaPO_4$, 2.5mM KCl, 2mM $CaCl_2$, 1mM $MgCl_2$, and 25mM D-glucose), we extracted the unfixed brain and flash froze the bead-injected olfactory bulb on dry ice before processing it for sequencing as described above. We cut 400μm thick acute sagittal slices of the remaining right hemisphere in dissection solution (110mM choline chloride, 11.6mM ascorbic acid, 3.1mM Na pyruvic acid, 25mM $NaHCO_3$, 1.25mM $NaPO_4$, 2.5mM KCl, 0.5mM $CaCl_2$, 7mM $MgCl_2$, and 25mM D-glucose) using a Microm HM650V vibratome. We incubated sections containing LC in artificial cerebrospinal fluid (126mM NaCl, 20mM $NaHCO_3$, 3mM KCl, 1.25mM NaH_2PO_4, 2mM $CaCl_2$, 2mM $MgSO_4$, and 20mM D-glucose) containing synaptic blockers (0.05mM APV, 0.02mM DNQX and 0.1μM TTX) for 20 minutes at room temperature. We then digested the slices in artificial cerebrospinal fluid with streptomyces griseus protease (Sigma P5147) at 1mg/ml at room temperature for 30 min. After washing in artificial cerebrospinal fluid with synaptic blockers, we dissected LC from the digested section and triturated the tissue to produce a single cell suspension. Using an inverted fluorescent microscope (Zeiss Observer), we picked individual cells by hand, deposited the cells directly into lysis buffer (2.4μl 0.2% triton, 1μl 10mM dNTPs, 1μl 10mM RT primer per cell) and proceeded to preparing sequencing libraries from the cells as described above for tissue samples.

Animals used

Number of animals	Manipulation	Figures based on these animals
4	Right LC injection with MAPseq virus; dissection of right cortex and olfactory bulb; qPCR and sequencing of barcode RNA	Fig. 2e, 4, 5, 6 Supp. Fig. 1, 2, 4, 5, 7c,d
2	Right LC injection with MAPseq virus; dissection of right and left cortex; qPCR of barcode RNA	Fig. 2f
1	Bilateral LC injection with MAPseq virus; dissection of right and left cortex and olfactory bulb; qPCR and sequencing of barcode mRNA	Fig. 7
3	Right LC injection with MAPseq virus and retrobeads injection into right olfactory bulb; single cell isolation from LC	Fig. 3b; Supp. Fig. 6
3	Right LC injection with DH(26S)5'SIN packaged MAPseq virus and dissection of select cortical targets and the olfactory bulb and sequencing of barcode mRNA	Supp. Fig. 7a,b
6	Right LC injection with DH(26S)5'SIN packaged MAPseq virus. TH staining of LC and quantification of overlap and count of TH+ neurons	Supp. Fig. 1j,k,l
3	Right LC injection with DH(26S)5'SIN packaged MAPseq virus; ISH for barcode mRNA and IF for GFP protein	Fig. 3b Supp. Fig. 1
2	Right LC injection with DH(26S)5'SIN packaged MAPseq virus and dissection of the olfactory bulb and ipsilateral cortex and sequencing of barcode mRNA	Supp. Fig. 1e,f