Materials and Methods

Animals. 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. For muMAPseq, experimental subjects were 8-week-old male C57BL/6J mice or BTBR T^+ Itpr3^{tf}/J mice from the Jackson Laboratory. For functional imaging, triple transgenic mice Emx-Cre; Ai93; LSL-tTA were generated. A small fraction of mice used for functional imaging also harbored a CamKII-tTA allele to enhance the expression of GCaMP6f.

Sindbis virus barcode libraries. The Sindbis virus used in muMAPseq was made as described previously (*1*, *2*). Briefly, based on a dual promoter pSinEGdsp construct, we inserted MAPP-nλ after the first subgenomic promoter, and GFP-BC(barcode)-4×boxB after the second subgenomic promoter. Sequences (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 NNN NNN GTA CTG CGG CCG CTA CCT A(3') were inserted between MluI and NotI sites which were between GFP and 4×boxB. In barcode library 1, the 32-nt BC ended with 2 purines, while in barcode library 2, the 32-nt BC ended with 2 pyrimidines. Sindbis virus was produced using the DH-BB(5'SIN;TE12ORF) helper plasmid (*3*). The viral barcode library diversity was determined by Illumina sequencing. $\sim 2 \times 10^6$ barcodes were sequenced in the viral library 1, and $\sim 8 \times 10^6$ barcodes were sequenced in the viral library 2.

Injections. For muMAPseq, Sindbis virus of barcode library 2 was injected into the right cortical hemispheres of experimental animals. Anesthesia was initially induced with isoflurane (4% mixed with oxygen, 0.5 L/min). Meloxican (2 mg/kg), dexamethasone (1 mg/kg) and baytril (10 mg/kg) were then administered subcutanesouly. For Sindbis injections, the whole skull above the right cortical hemisphere was removed. More than 100 injection pipette penetrations were made to cover the entire exposed brain, each spaced by 0.5 mm, both in the AP axis and ML axis. Nanoject III (Drummond Scientific) was used to inject Sindbis virus (2×10^{10} GC/mL), at 3-4 depths per penetration site (Supplementary Table 1). At each penetration site and depth, 23 nL virus was injected. The full injection surgery required about 8 hours, and constant isoflurane (1% mixed with oxygen, 0.5 L/min) was administered to maintain anesthesia. After injection, sterile Kwik-Cast (World Precision Instruments) was gently applied to cover the exposed brain region, and the skin was closed with sutures. Meloxican (2 mg/kg), dexamethasone (1 mg/kg) and baytril (10 mg/kg) were then routinely administered to animals subcutaneously every 12 hours post surgery, and animal condition was inspected every 6 – 12 hours. Similarly, we injected Sindbis virus of barcode library 1 into control animals. In control animals, instead of injecting the virus into the whole right cortex, we only made ~6 penetrations covering a small cortical area.

For control experiments testing the soma calling strategy (Fig. 1E), the same muMAPseq protocol was followed, but Sindbis virus of barcode library 1 was injected into the secondary motor areas, and Sindbis virus of barcode library 2 into the primary motor areas.

For control experiments testing template switches (Fig. S3B,C), we followed the muMAPseq protocol above, but injected Sindbis virus of barcode library 2 into two separate animals.

For AAV CAG-tdTomato tracing experiments (Fig. S6), we used coordinates AP = -4 mm, ML = 0.5 mm, 1 mm and 1.5 mm, DV = 0.25 mm and 0.5 mm for retrosplenial cortex in C57BL/6J and coordinates AP = -4 mm, ML = 0.75 mm, 1 mm and 1.5 mm, DV = 0.25 mm and 0.5 mm for retrosplenial in BTBR. In BTBR, as two hemispheres began to separate at AP = -4 mm and there was no cerebral cortex at ML = 0.5 mm, we used ML = 0.75 mm instead. In each coordinate, 20 nL of AAV1 CAG-tdTomato AAV (2×10^{13} GC/mL Penn Vector Core) was injected.

Cryosectioning and laser microdissection (LMD). In muMAPseq, 44 hours after Sindbis viral injection, the brain was harvested and fresh frozen at -80 °C. Olfactory bulbs and rostral spinal cord/caudal medulla were cut from the brain and collected separately. We then cut 300 μm coronal sections using a Leica CM 3050S cryostat at -12 °C chamber temperature and -10 °C object temperature. Each slice was cut with a fresh part of a blade, and the platform and brushes were carefully cleaned between slices. Each slice was immediately mounted onto a steel-framed PEN (polyethylene naphthalate)-membrane slide (Leica). After mounting on the slide, the slice was fixed in 75% ethanol at 4 °C for 3 min, washed in Milli-Q water (Millipore) briefly, stained in 0.5% toluidine blue (Sigma-Aldrich, MO) Milli-Q solution at room temperature for 30 sec, washed in Milli-Q water at room temperature for 3 times (15 sec each time), and fixed again in 75% ethanol at room temperature twice (2 min each time). The slide was then left in a vacuum desiccator for 30 min. Next, another fresh frame slide was used to sandwich the brain slice, and the two slides tightly taped to prevent the slice from falling. The sandwiched slice was stored in the vacuum desiccator at room temperature until LMD. If LMD was performed more than 1 week after cryosectioning, the sandwiched slices were stored at -80 °C in a desiccated container.

Cubelet dissection was performed with Leica LMD 7000. During LMD, cortical cubelets with ~1 mm arc length were dissected from each coronal slice, from the surface to the deepest layer above the white matter. Orbitofrontal cortical cubelets (in rostral slices), anterior cingulate cortical cubelets, and retrosplenial cortical cubelets were also collected separately. For subcortical areas including striatum, thalamus, amygdala, tectum and pons/medulla, tissue belonging to each brain area was pooled every 1-3 consecutive slices. About 12~21 cubelets were also collected from injection sites and contralateral homotopic areas of the injection sites in the barcode library 1 control animal, and 2 cortical cubelets in the uninjected control animal. Pictures were taken before and after every cubelet was dissected. After dissecting every 4 cubelets, we transferred them into homogenizing tubes with homogenizing beads, and added 100 μL lysis solution (RNAqueous-Micro Total RNA Isolation Kit, Thermo Fisher) into each cubelet. The collected tissues were stored temporally on dry ice and then at -80 °C.

Sequencing library preparation. After LMD, each cubelet was homogenized in lysis solution with a tissue lyser (Qiagen) at 20 Hz for 6 min. Then we extracted RNA molecules from each cubelet with RNAqueous-Micro Total RNA Isolation Kit (Thermo

Fisher). We did not treat products with DNase I as DNA did not influence following experiments. The final product was eluted in 20 µL elution solution.

After RNA extraction, we performed reverse transcription (RT) with barcoded RT primers using SuperScript IV (Thermo Fisher). Barcoded RT primers were in the form of (5')CTT GGC ACC CGA GAA TTC CAX XXX XXX XXX XXZ ZZZ ZZZ ZTG TAC AGC TAG CGG TGG TCG(3'), where Z_8 is one of 288 CSIs (cubelet-specific identifiers) and X_{12} is the UMI (unique molecular identifier). 1 µL of 1 × 10^{-9} µg/µL spike-in RNAs were also added. The sequence of spike-in RNAs were (5')GUC AUG AUC AUA AUA CGA CUC ACU AUA GGG GAC GAG CUG UAC AAG UAA ACG CGU AAU GAU ACG GCG ACC ACC GAG AUC UAC ACU CUU UCC CUA CAC GAC GCU CUU CCG AUC UNN NNN NNN NNN NNN NNN NNN NNN NAU CAG UCA UCG GAG CGG CCG CUA CCU AAU UGC CGU CGU GAG GUA CGA CCA CCG CUA GCU GUA CA(3').

We then cleaned up RT products with 1.8×SPRI select beads (Beckman Coulter), synthesized double-stranded cDNA with previously described methods (*3*), cleaned up 2nd strand synthesis products again with 1.8× SPRI select beads, and treated the eluted ds cDNA with Exonuclease I (New England Biolabs) (incubated the mix at 37°C for 1 hr and inactivated the enzyme at 80°C for 20 min). As cDNA molecules from different cubelets were already CSI-barcoded after RT, we pooled every 12 RT products for 1st bead purification and 2nd strand synthesis, and pooled all the products for 2nd bead purification and Exonuclease I treatment.

We next amplified the cDNA library by nested PCR using primers (5')GGA CGA GCT G(3') and (5') CAA GCA GAA GAC GGC ATA CGA GAT CGT GAT GTG ACT GGA GTT CCT TGG CAC CCG AGA ATT CCA(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). First PCR was performed for 5 cycles in 720 µL; after Exonuclease I treatment (incubated the mix at 37°C for 30 min and inactivated the enzyme at 80°C for 20 min), ¼ of the first PCR products were used for second PCR. Second PCR was performed for 5-10 cycles in 12 mL. Standard Accuprime protocol was used for PCR except that the extension time in each cycle was set to 2 min to reduce incomplete elongation and template switches.

Nested PCR products were then purified and eluted in 600 µL with a Wizard SV Gel and PCR Clean-Up System (Promega), and further concentrated with Ampure XP beads (Beckman Coulter) in 25 µL Milli-Q H2O. After running in a 2% agarose gel, the 230 bp band was cut out and cleaned up with the Qiagen MinElute Gel Extraction Kit (Qiagen). We sequenced the library 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.

Most of the molecular experiments were performed according to the reagent manufacturer's protocol unless otherwise stated.

Sequencing. We sequenced the pooled libraries prepared as above 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.

Confocal imaging. In AAV tracing experiments, brains were harvested 14 days after viral injection, fixed in 4% paraformaldehyde, washed in phosphate-buffered saline, and cut into 100 µm slices with a vibrotome (LeicaVT1000S, Leica). Slices were then mounted onto slides in Fluoroshield (Sigma-Aldrich), and imaged in a Laser Scanning Microscope 710 system (Leica).

Wide-field calcium imaging and behavior. For Fig. 3 and Fig. S5, imaging and behavior are as described in ref (*4*). To preprocess widefield data, we used SVD to compute the 500 highest dimensions accounting for more than 88 % of the variance in the data. The original data matrix M (of size pixels × frames) was decomposed as

$$M = USV$$

, which returns 'spatial components' U (of size pixels × components), 'temporal components' V (of size components × frames) and singular values S (of size components × components) to scale components to match the original data. Further analysis methods are described in Supplementary Note 5.10**.**

Image processing. Wholebrain toolbox (http://www.wholebrainsoftware.org) was used to register Toluidine Blue-stained coronal slices into Allen Reference Atlas semi-automatically. Using Matlab, we determined the coordinates of each cubelet by processing pictures taken before and after each cubelet was dissected. Combining image registration results and cubelet coordinates, we mapped each cubelet into one or multiple brain areas.

MuMAPseq data analysis. The details on muMAPseq analysis, including bioinformatics, statistics and computational methods are in Supplementary Note 5.

MuMAPseq data visualization. MuMAPseq data were visualized in a 3D brain in Fig 2A. To reconstruct the cubelet-to-cubelet connection pathways, the position in stereotactic coordinates for each registered cubelet source node was used to query Allen Mouse Brain Connectivity Atlas (*5*) for injection sites within 500 µm from each source node. Out of all the injection sites the injection with largest injection volume was used to download projection density volumes with 200 µm voxel resolution. 92 out of 99 cubelet source nodes could be mapped to a unique projection density volume. Next, we used A* search algorithm (*6*) implemented in C/C++ to find the optimal path between muMAPseq source and target cubelet nodes using binary projection density volume to represent graph nodes and blocked obstacles. The optimal path for 1677 out of 3015 non-zero connection could be determined (56%). The remaining either didn't have a corresponding projection density volume, alternatively target and source cubelets were not connected in the projection density volume. Each projection path was then smoothed as a spline using a Generalized Additive Model (GAM) (*7*). Each path was rendered in 3D with a unique color given by the position of the path's source cubelet. The color-coding of target cubelet locations was based on a red-green-blue (RGB) spatial color cube code where red represents medio-lateral, green represents anterior-posterior, and blue represents dorso-ventral axis.

Supplementary Notes

Supplementary Note 1: Potential MAPseq artifacts are minimal and MAPseq efficiency is high

Below we discuss several classes of potential MAPseq artifacts.

Degenerate and double barcode labeling. In MAPseq, barcodes drawn from a high-diversity viral pool are used to uniquely label neurons. Ideally, every infected neuron would have a single, unique barcode. As discussed in detail in ref. (*1*), there are two potential deviations from this ideal scenario: (i) multiple neurons per barcode, and (ii) multiple barcodes per neuron. The former, multiple neurons per barcode, i.e. re-used barcodes, is problematic as it leads to incorrect results. Because the probability that two neurons are infected by the same barcode is determined by the diversity of the barcode library and the total number of infected cells, we generated a high diversity viral library with over 8×10^6 barcodes for muMAPseq (about 5×10^4 neurons). We also computationally inferred and calculated false positive projections caused by re-used barcodes (see details in Supplementary Note 2). The consistency between muMAPseq and Allen Connectivity Atlas (Fig. 2D,E) confirmed that the effects were minimum. The second scenario, multiple barcodes per neuron, has much less severe consequences. If a neuron expresses more than one barcode, muMAPseq will overestimate the number of traced neurons. However, the relative abundance of each projection type and the bulk connection strength remain unchanged. In practice, we also aimed to infect neurons at close to 1 barcode per neuron by using appropriate volume and titer of Sindbis viruses.

Non-uniform barcode transport. In MAPseq, we interpret the number of barcodes from source cubelet X in target cubelet Y as a measure of the strength of projection from X to Y, analogous to GFP intensity. For this assumption to be valid, we must implicitly assume that barcode transport is uniform; in particular, we must assume that nearby and distal targets are equally filled. This assumption was rigorously validated in previous work (see Figure 2 in Ref (*1*)).

Fibers of passage. Although the MAPseq carrier protein is derived from the synaptic protein neurexin (*1*), it does not exclusively target to presynaptic terminals. Thus MAPseq does not distinguish between synaptic connections (axon terminals) and fibers of passage. In this respect it is analogous to using GFP intensity in a conventional connectivity Atlas to measure the strength of the connection from the injection site to a target. To minimize potential confounds due to fibers of passage, we avoided white matter when dissecting cortical cubelets.

MAPseq efficiency. The efficiency of MAPseq has also been quantified in (*1*). By injecting red retrobeads into the olfactory bulb, and infecting the locus coeruleus with GFP-barcode Sindbis viruses, 91.4±6% of all barcodes from cells that projected to the olfactory bulb as determined by bead labeling also appeared to project to the bulb by sequencing (Fig. S6 in (*1*)).

End-to-end assessment of artifacts. The agreement between muMAPseq and the Allen Atlas reported in Figure 2D,E also confirmed that potential MAPseq artifacts above were minimal, consistent with previous validations (*1*, *8*, *9*).

Supplementary Note 2: Sources of errors and calculation of cubelet-to-cubelet connections in muMAPseq

List of variables in Supplementary Notes 2

l_1	Number of molecules in cubelet 1
l_2	Number of molecules in cubelet 2
c	Template switching rate constant
h_{12}	Number of cubelet 1-cubelet 2 hybrid molecules
$N(i)$ or $N_1(i)$	Number of projection neurons (type 1 neurons, Supplementary Note 2.2) residing in cubelet i
$N_3(i)$	Number of type 3 neurons (Supplementary Note 2.2) residing in cubelet i
N_t	Total number of barcodes in the muMAPseq result (type 1-4, Supplementary note 2.2)
N_{re}	Total number of re-used barcodes
$n(i,j)$	Total number of molecules of jth neuron in ith cubelet (soma molecules + all axon molecules)
$n_{soma}(i,j)$	The number of soma molecules of jth neuron in ith cubelet
$p(i,j,k)$	The probability that molecules of jth neuron in ith cubelet were detected in kth cubelet due to template switching
m_k	Number of error molecules from neurons in experimental cubelets that were detected in kth control cubelet due to template switching
b	Number of error molecules in each cubelet due to baseline contamination
$P_\theta(i,j,k)$	The probability that > θ error molecules of jth neuron in ith cubelet were detected in kth cubelet due to template switching
$r_{ts}(i,k)$	The average probability that a false projection from a neuron in ith cubelet to kth cubelet was detected due to template

	switching
$r_{re}(i,k)$	The average probability that a false projection from a neuron in ith cubelet to kth cubelet was detected due to re-used barcodes
$r_{ba}(i,k)$	The average probability that a false projection from a neuron in ith cubelet to kth cubelet was detected due to baseline contamination
v_{ik}	p value (false positive probability) of cubelet i-to-cubelet k projection
$N_{pro}(i,k)$	Observed number of neurons in cubelet i that projected to cubelet k
C	Cubelet-to-cubelet connection matrix

There are two major error sources that affected muMAPseq data: template switching and re-used barcodes. We have tried to reduce them both experimentally and computationally. The significance level of each cubelet-to-cubelet connection was also evaluated based on the false-positive error rate.

The following terms are defined before further discussion. 1) Barcode: a barcode is a unique 32nt sequence delivered by the Sindbis virus. One barcode theoretically corresponds to a neuron. 2) Molecule: here a molecule is defined as a unique BC-CSI-UMI (32nt + 8nt + 12nt) sequence. A molecule should correspond to a single RT product. Due to barcode amplification in a neuron, one barcode has multiple molecules. 3) Molecule copy: a molecule copy is defined as a final product after PCR. A large number of molecule copies are generated from one molecule during PCR. 4) Read: reads are the sequencing product. Not considering sequencing errors, all the reads constitute a subset of all the molecule copies.

2.1 Template switching

Template switching may occur when DNA templates share a common sequence during PCR (Fig. S3A). In muMAPseq, cDNA from all the cubelets was pooled together for PCR, and they all shared a common RT primer annealing sequence. Template switching is usually considered to be rare, and might be corrected by setting a read threshold for molecules (*10*). However, low sequencing depth disabled the use of read threshold to remove error molecules. Moreover, as molecules of a barcode in a soma usually outnumbered molecules in axons by ~100 fold, template switching molecules might constitute a large proportion in axon barcodes, albeit rare compared to total molecules. Thus, template switching had a significant influence in measuring projection strengths in muMAPseq.

As DNA concentration is a major factor determining the template switching rate, we proposed we could reduce template switch molecules by increasing the PCR volume. To systematically evaluate template switching and test our hypothesis, we designed an experiment to perform muMAPseq from two brains. We injected similar amount of barcoded viruses into two animals, collected cubelets, and performed RT from individual cubelets. Then single-strand DNA molecules were pooled (48 cubelets from each animal, 96 in total) for second-strand synthesis, PCR and sequencing. Thus 'inter-brain' projection molecules reflected template switching. To measure the effect of DNA concentration on template switching, the same sample was separated to perform PCR either in 25 μL or in 2 mL. In the 25 μL PCR experiment, a large number of molecules that were detected in both brains ('inter-brain' molecules) as well as stripe-like patterns indicated a high template switching level (Fig. S3B, left). By increasing PCR volume to 2 mL, 'inter-brain' molecules were dramatically decreased (Fig. S3B, right). The rate of template switching could be further reduced by raising the UMI threshold that was used to determine a real projection (Fig. S3C). In addition to the high reaction volume, we also set the PCR extension time in each cycle to 2min to reduce incompletely elongated products, another possible source of template switching.

To reduce template switching, we chose to perform PCR in 12 mL systems for muMAPseq experiments LJ7, LJ9 and BTBR. While Sindbis viruses harboring barcode library 2 were used to label experimental animals, we also injected Sindbis viruses harboring barcode library 1 into a few brain areas in a separate animal. After RT and second-strand synthesis, DNA molecules in experimental animals (261 cubelets in LJ7, 262 cubelets in LJ9 and 258 cubelets in BTBR) were mixed with DNA molecules in control animals (21 cubelets in LJ7 control, 12 cubelets in LJ9 control, and 12 cubelets in LJ10 BTBR) for PCR and sequencing, as an internal measurement for template switching. In LJ7, when we set UMI threshold to 1 (i.e. a projection was positive when its UMI count was greater than 1), 4794 out of 63107 barcodes were detected in the control brain (Fig. S3D). Similar numbers were also found in LJ9 and BTBR (data not shown).

With PCR volume = 12mL and UMI threshold = 1, the probability that a barcode was detected in a non-projecting cubelet due to template switching on average was reasonably low ($\frac{4794}{63107\times21}$< 1%). To further determine whether a bulk projection was significant, we determined the noise floor for each cubelet-to-cubelet projection by calculating the number of false projection neurons in a given source cubelet to a given target cubelet predicted by template switching. The calculation was as follows:

Let l_1 denote the number of molecules in cubelet 1, and l_2 denote the number of molecules in cubelet 2. If we pool these molecules to perform PCR, we assume the number of hybrid molecules after PCR h_{12} can be written as:

$$h_{12} = 2cl_1l_2 \quad (1)$$

, where c is called template switching rate constant, and should be dependent on PCR cycle numbers and PCR volume. As we pooled all the samples together for PCR, c was a constant in one muMAPseq experiment.

Similarly, in muMAPseq, let $N(i)$ denote the number of neurons in cubelet i, $n(i,j)$ denote the number of molecules (including both soma molecules and axon molecules) for the jth neuron in cubelet i, and $n_{soma}(i,j)$ denote the number of soma molecules for the

jth neuron in cubelet i. The probability that the jth neuron in cubelet i had a false positive molecule in cubelet k, $p(i,j,k)$ was:

$$p(i,j,k) = cn(i,j)\sum_{l=1}^{N(k)} n_{soma}(k,l) \quad (2)$$

.

As we performed PCR and sequencing by pooling molecules from two brains, we were able to estimate c by counting ‘inter-brain’ molecules. If we considered template switching across two brains, then the number molecules that were from neurons residing in the experimental brain and found in the control brain cubelet k, m_k was:

$$m_k = c\sum_{i}\sum_{\substack{j\ in\\ exp.}} n(i,j)\sum_{l=1}^{N(k)} n_{soma}(k,l) \quad (3)$$

, where j visited all the cubelets in the experimental brain and i visited all the neurons in each experimental brain cubelet.

In the real experiment, there was an extra baseline contamination term (this term can also be inferred from molecules in additional control cubelets from a brain without viral injection), so Eq (3) was modified as:

$$m_k = c\sum_{i}\sum_{\substack{j\ in\\ exp.}} n(i,j)\sum_{l=1}^{N(k)} n_{soma}(k,l)\ + b \quad (4)$$

, where b was the baseline contamination constant.

To estimate the constants c and b, a linear regression model was used to fit Eq. (4) to the observed data set. As an example, in LJ7, we got:

$$c = 1.12 \times 10^{-11}$$
$$b = 3.90 \times 10^{3}$$

.

With estimated c and b, we could predict intra-brain template switching probability, $p(i,j,k)$ with Eq. (2) when i and k were both from the experimental brain. However, as we further filtered the data by setting a UMI threshold θ, a false-positive projection was detected only when at least $(\theta + 1)$ template switching molecules from a given neuron to a given cubelet were seen. Let $P_\theta(i,j,k)$ denote the probability that the jth neuron in cubelet i falsely projected to cubelet k with UMI threshold = θ, then according to Poisson distribution, we had

$$P_\theta(i,j,k)\ =\ \sum_{l=\theta+1}^{\infty} e^{-p(i,j,k)}\frac{p(i,j,k)^l}{l!} \quad (5)$$

.

As an approximation, we only calculated the first three terms when $\theta = 1$, as $p(i,j,k)$ was small enough and false projections with high UMI counts were extremely rare. We got:

$$P_1(i,j,k) \approx e^{-p(i,j,k)}\frac{p(i,j,k)^2}{2!} + e^{-p(i,j,k)}\frac{p(i,j,k)^3}{3!} + e^{-p(i,j,k)}\frac{p(i,j,k)^4}{4!} \approx \frac{p(i,j,k)^2}{2!} + \frac{p(i,j,k)^3}{3!} + \frac{p(i,j,k)^4}{4!} \quad (6)$$

. With Eq. (6), we were able to calculate the probability that a given neuron in cubelet i falsely 'projected' to cubelet k. However, as cubelet i consisted of $N(i)$ neurons, and each neuron had a different template switching probability, the total number of i-to-k false-positive neurons caused by template switching obeyed a Poisson binomial distribution. Note it was neither a Poisson distribution nor a binomial distribution, but a distribution of the sum of Bernoulli trials with different probabilities.

To calculate the rates of false positive connections, we sought to calculate the Poisson binomial cumulative probability distribution. In muMAPseq, there were over 30000 possible cubelet-to-cubelet projections, and for each of these projections, there were 500~1000 cells in the source cubelet (corresponding to 500~1000 Bernoulli trials). To our knowledge, there did not exist a fast and precise way to calculate the cumulative probability of the Poisson binomial distribution for each cubelet-to-cubelet projection. This was even difficult for a p value as small as $0.05/36018 \approx 1.66 \times 10^{-6}$ when multiple comparison correction was considered. Thus, we chose to use binomial distributions to approximate Poisson binomial distributions, assuming the probability of any given neuron in cubelet i falsely projected to cubelet k, $r_{ts}(i,k)$, was the mean probability over all the neurons in cubelet i:

$$r_{ts}(i,k) = \frac{\sum_{j=1}^{N(i)} P_1(i,j,k)}{N(i)} \quad (7)$$

. $r_{ts}(i,k)$ was next used to calculate the net false positive probability for individual cubelet-to-cubelet connections (Supplementary Note 2.5).

Note when the required p value was not too small (for example, p = 0.05, without multiple comparison), we used Monte-Carlo method (10000 trials each) to estimate the cumulative probability of the Poisson binomial distribution for each cubelet-to-cubelet projection.

To summarize, template switching could be a detrimental error source when DNA concentration during PCR is high and sequencing depth is low. By using a large volume of the reaction system for PCR, setting a UMI threshold, and rejecting false positive projections, we have greatly reduced template switching errors to a very low level.

2.2 Re-used barcodes

To scale up MAPseq, it is crucial to use a barcode library whose diversity is high enough. Otherwise, the same barcode used in two cells would cause misinterpretation of the data (Fig. S3E). The rate of re-used barcodes was determined by barcode diversity and the total number of infected neurons. In muMAPseq, the diversity of the barcode library was no less than 8.26×10^6, according to the viral library sequencing result. However, the total number of neurons expressing barcodes was much higher than the number of recovered neurons (~50000) due to a large number of 'non-projection' neurons. For example, in LJ7, over 600000 'non-projection neurons' were recovered. Some of these 'non-projection' neurons might belong to local inhibitory or excitatory neurons, but a large number of them expressed RNA barcodes at very low levels. It was likely that due to variations of RNA expression levels, some projection neurons expressed very small amount of RNA barcodes, which couldn't be efficiently trafficked to axon terminals. These low expressed barcodes were almost all in the right cortical cubelets (injection site), and usually fewer than 20 molecules were detected in somas, and no molecules above the UMI threshold (=1) were detected in axons. Although these 'non-projection' neurons were not included for data analysis, they might harbor re-used barcodes shared with other projection neurons, resulting in false projections (Fig. S3E).

To solve this problem, an additional set of thresholds was used to reduce re-used barcode errors. For each barcode, we defined its firstmax and secondmax as the highest and second highest abundance among all the cubelets. If a barcode corresponded to one neuron, then its firstmax was the count of molecules in its soma, and its secondmax was the count of molecules in its strongest axon. If a barcode was used in two neurons, then firstmax and secondmax were the highest two of UMI counts in two somata and two strongest axons. As the molecules in somata statistically outnumbered molecules in axons, secondmax of a re-used barcode was likely to be the amount of molecules in one of the two somata. According to this, we reasoned that re-used barcodes might have distinct distribution in the (firstmax, secondmax) space from barcodes used only once. To quantify this, we simulated the barcode sampling process (details in Supplementary Note 1.3, we modeled viral infection as a process where neurons randomly sampled barcodes from the barcode library), and calculated the number of re-used barcodes in the (firstmax, secondmax) space, given the observed joint distributions of (firstmax, secondmax) and the known barcode library. The ratio of simulated re-used barcodes to the total barcodes in the (firstmax, secondmax) space (Fig. S3F). Not surprisingly, a higher ratio of re-used barcode was present close the diagonal line in the (firstmax, secondmax) space.

We next set a soma threshold (=250) and an axon threshold (=20) (Fig. S3F), and defined 4 types of barcodes according to the thresholds:
Type 1 barcode: firstmax > soma threshold AND secondmax > UMI threshold AND secondmax < axon threshold.
Type 2 barcode: firstmax > soma threshold AND secondmax ≤ UMI threshold.
Type 3 barcode: firstmax < axon threshold AND firstmax > UMI threshold.
Type 4 barcode: secondmax > axon threshold.

To reduce the effect of re-used barcodes, we only included type 1 barcode for projection pattern analysis. Based on simulation results, in LJ7, 8.77% of type 1 barcodes were re-used barcodes (8.27% in LJ9 and 8.62% in BTBR). As there were 115 cubelets in the injection site of LJ7, if a source cubelet and a target cubelet were both in the injection site (right hemisphere), then the probability of a type 1 neuron in the source cubelet that falsely projected to the cubelet was on average $\frac{8.77\%}{115} \approx 0.0763\%$, which was reasonably low.

To quantify the significance level for each cubelet-to-cubelet connection, we calculated $r_{re}(i,k)$, the probability that a type 1 neuron in cubelet i that falsely projected to cubelet k due to re-used barcodes. In LJ7, for example, because a re-used type 1 barcode could only occur when a type 1 or type 2 neuron in the source cubelet and a type 3 neuron in a target cubelet shared the same barcode, we could estimate $r_{re}(i,k)$ with:

$$r_{re}(i,k) = \frac{8.77\% * N_3(k)}{\sum_l N_3(l)} \quad (8)$$

, where $N_3(k)$ represents the number of type 3 barcodes in cubelet $k.$ $r_{re}(i,k)$ were next used to calculate the net false positive probability for each cubelet-to-cubelet connection (Supplementary Note 2.5).

To conclude, with the current viral barcode diversity, the probability of re-used barcodes cannot be ignored. By setting thresholds for soma and axon identification, and rejecting false positive projections, we have removed most of errors due to re-used barcodes. However, as an ideal way to solve all these problems, a barcode library of much higher diversity should be made for high throughput muMAPseq in future.

2.3 Simulating re-used barcodes

The distribution of barcode abundance in the barcode library was not uniform, so barcodes with higher abundance in the library were more likely to be re-used multiple neurons. Moreover, as we did not sequence the full viral barcode library, we also found barcodes present in the muMAPseq result but absent in the viral library sequencing result. We set a viral barcode threshold (=4), and classified barcodes according to their abundance: high-abundance barcodes (present and over 4 counts in the library sequencing result), low-abundance barcodes (present but no-greater-than 4 counts in the library sequencing result), and non-sequenced barcodes (absent in the library sequencing result, but present in the muMAPseq result). To reduce the chance of re-used barcodes, we included low-abundance barcodes and non-sequenced barcodes for neuronal projection analysis. But for re-used barcodes simulation as follows, we only included low-abundance barcodes as the re-used barcode chance in the non-sequenced barcodes should be lower than the chance in the low-abundance barcodes.

To simulate re-used barcodes, we assumed 1) most of the observed barcodes were single-cell barcodes (re-used barcodes were rare) and 2) the barcode expression level and projection patterns of an infected cell were independent of the barcode sequence itself. The simulation steps were:

1. Estimate total number of re-used barcodes. We calculated the total amount of low-abundance barcodes in the MAPseq results (including type 1-4 barcodes) N_t, and sampled N_t barcodes from low-abundance barcodes in the barcode library sequencing result using the observed abundance distribution. We estimated the total number of re-used barcodes, N_{re} from the sampling simulation.
2. Estimate the distribution of re-used barcodes in the (firstmax, secondmax) space. As most re-used barcodes were used twice, we randomly sampled firstmax and secondmax from the measured distributions for two neurons related to each re-used barcode, and calculated the new firstmax and secondmax for this barcode. By doing this N_{re} times, we generated a distribution of re-used barcodes in the (firstmax, secondmax) space.

3. Given the simulated distribution of re-used barcodes, we calculated the ratio of re-used barcodes to the total number of barcodes in the muMAPseq result in the (firstmax, secondmax) space. (Fig. S3F).

2.4 Calculating cubelet-to-cubelet connection strength

Projection strength from a source cubelet to a target cubelet was defined as the total count of UMIs in the target cubelet from all the neurons residing in the source cubelet divided by total number of projection neurons in the source cubelet. Considering the projection from cubelet i to cubelet j, let $N(i)$ denote number of projection neurons in cubelet i and $UMI(i,j,k)$ denote the UMI count in cubelet k from jth neuron in cubelet i, then the UMI count in cubelet k from an average neuron in cubelet i, $UMI(i,*,k)$ could be written as:

$$UMI(i,*,k) = \frac{\sum_{j=1}^{N(i)} UMI(i,j,k)}{N(i)} \quad (9)$$

. However, noise caused by template switching, re-used barcodes, and baseline contaminations could also contribute to $UMI(i,*,k)$. The noise level of the i-to-k projection, $Noise(i,k)$ was calculated as:

$$Noise(i,k) = UMI_{ts}(i,*,k) + r_{re}(i,k) * UMI_{type3}(k) + r_{ba}(i,k) * UMI_{ba} \quad (10),$$

where $UMI_{ts}(i,*,k)$ is the expected UMI count in cubelet k from an average neuron in cubelet i due to template switching, $UMI_{type3}(k)$ is the average UMI count of type 3 neurons in cubelet k (after UMI thresholding), UMI_{ba} is the average UMI count of a barcode in cubelets from the uninjected control brain (baseline contamination, after UMI thresholding), and the $r_{ba}(i,k)$ is the probability that a neuron in cubelet i falsely projected to cubelet k due to baseline contaminations, (estimated from non-injected control cubelets). These three terms corresponded to the template switching noise, re-used barcode noise, and baseline contamination noise. Particularly, $UMI_{ts}(i,*,k)$ was calculated with:

$$\begin{aligned} UMI_{ts}(i,*,k) &= \sum_{l=\theta+1}^{\infty} l e^{-p(i,j,k)} \frac{p(i,j,k)^l}{l!} \\ &\approx 2e^{-p(i,j,k)} \frac{p(i,j,k)^2}{2!} + 3e^{-p(i,j,k)} \frac{p(i,j,k)^3}{3!} + 4e^{-p(i,j,k)} \frac{p(i,j,k)^4}{4!} \\ &\approx \frac{p(i,j,k)^2}{1!} + \frac{p(i,j,k)^3}{2!} + \frac{p(i,j,k)^4}{3!} \quad (11) \end{aligned}$$

. The projection strength from cubelet i to cubelet j, $C(i,j)$ was then calculated with:

$$C(i,k) = \max\{UMI(i,*,k) - Noise(i,k), 0\} \quad (12)$$

. In addition to calculate the projection strength, we also calculated p value for each cubelet-to-cubelet connection, as noted in Supplementary Note 2.5.

2.5 Calculating p values

In addition to removing the noise estimate from the projection strength, we also calculated the p value for each cubelet-to-cubelet projection. For a source cubelet i and a target cubelet k, we calculated the probability that a neuron in cubelet i falsely projected to cubelet k due to template switching, $r_{ts}(i,k)$ (Supplementary Note 2.1), the probability that a neuron in cubelet i falsely projected to cubelet k due to re-used barcodes, $r_{re}(i,k)$ (Supplementary Note 2.2), and the probability that a neuron in cubelet i falsely projected to cubelet k due to baseline contaminations, $r_{ba}(i,k)$. Note that $r_{ts}(i,k)$, $r_{re}(i,k)$, and $r_{ba}(i,k)$ were all very small, so the overall false-positive probability could be calculated additively. If there were $N(i)$ neurons in cubelet i, and $N_{pro}(i,k)$ neurons in cubelet i were found to project to cubelet k, then the p value of i-to-k connection, v_{ik} was calculated with:

$$v_{ik} = 1 - f\left(N_{pro}(i,k), N_i, r_{ts}(i,k) + r_{re}(i,k) + r_{ba}(i,k)\right) \quad (13)$$

, where f was the binomial cumulative distribution function:

$$f(n,N,p) = \sum_{l=0}^{n} \binom{N}{l} p^l (1-p)^{N-l} \quad (14)$$

.

With p-values, we were able to determine whether a given cubelet-to-cubelet connection was significant. Volcano plots of ipsilateral connections and contralateral connections in LJ7 are shown in Fig. S3G. All the data in Fig. 2B; Fig. 4B; Fig. S4A show significant projections (Bonferroni correction for multiple comparison, p-value < $\frac{0.05}{N}$, N is total number of possible projections).

Summary of error sources

Error sources	Effects	Solutions
Barcode base substitution	Generate barcodes with 1 or very few counts in 1 or very few cubelets	Collapse barcodes with up to 3 mismatches. Set UMI threshold. Set soma threshold.
Barcode base insertion/deletion	Generate barcodes with 1 or very few counts in 1 or very few cubelets	Set UMI threshold. Set soma threshold.
CSI sequencing errors	Generate barcodes in ‘non-existing’ cubelets	CSIs that did not match any of the 288 used CSIs were excluded for further analysis
UMI sequencing errors	Cause overestimated barcode counts	Not corrected (But errors should be rare and uniformly randomly distributed)
Template switching	False projections	PCR with a large volume. Set UMI threshold. Calculate false-positive rates.
Re-used barcodes	False projections	Use a high diversity barcode library. Exclude over-represented barcodes in the barcode library. Set axon/soma threshold. Calculate false-positive rates

Non-collected soma	Strongest projections were detected as somas	Set soma threshold.

Supplementary Note 3: Comparing muMAPseq projectome with Allen Projectome, and comparing between muMAPseq mapped brains

List of variables in Supplementary Notes 3

A_k	Brain area-to-brain area connection matrix, type k (k=1,2,3,4)
C_k	Cubelet-to-cubelet connection matrix, type k (k=1,2,3,4)
P_k	Cubelet-to-brain area connection matrix, type k (k=1,2,3,4)
M	Cubelet-to-brain area mapping matrix
M_a	Cubelet-to-brain area mapping matrix, normalized to total size of each brain area
M_c	Cubelet-to-brain area mapping matrix, normalized to total size of each cubelet
S_a	Brain area size matrix, diagonal
S_c	Cubelet size matrix, diagonal

With muMAPseq, we were able to map cubelet-to-cubelet connections from one individual brain. In order to compare between muMAPseq data and Allen data, we utilized brain registration results to infer cubelet-to-brain area connections and brain area-to-brain area connections from cubelet-to-cubelet connections. Here we describe and discuss various models underlying connection inference.

The following terms and variables are defined before further discussion:
Considering the connection from cubelet i to cubelet j, $\{C\}_{ij}$, we could quantify its strength by calculating the average counts of UMIs (molecules) in cubelet j per neuron in cubelet i (See Supplementary Note 2.4). This described the projection strength (axon volume) from an average neuron in cubelet i to the whole cubelet j, and thus was called 'unit-to-total' connection here. By considering the physical sizes of cubelet i and cubelet j, we could also define and calculate 'unit-to-unit' connection (connection from a neuron in cubelet i to a unit area size in cubelet j), 'total-to-unit' connection (connection from the whole cubelet i to a unit area size in cubelet j), and 'total-to-total' connection (connection from the whole cubelet i to the whole cubelet j), as summarized in the table below (similar to Supplementary Fig.2 in (*5*)).

Connection type	Connection source	Connection target	Definition	Formula
Type 1, C_1	Cubelet	Cubelet	Unit-to-unit	C_1

Type 2, C_2	Cubelet	Cubelet	Unit-to-total	$C_2 = C_1 S_c$
Type 3, C_3	Cubelet	Cubelet	Total-to-unit	$C_3 = S_c C_1$
Type 4, C_4	Cubelet	Cubelet	Total-to-total	$C_4 = S_c C_1 S_c$

Here S_c is a diagonal matrix, and its element $\{S_c\}_{ii}$ represents the physical size of cubelet i.

In conventional fluorescence tracing, projection strength is usually quantified as the normalized fluorescence intensity in the target area to the fluorescence intensity in the injection area (*5*). This was analogous to the type 2 connection, as defined above. Connections mentioned in this manuscript all referred to type 2 connections, unless otherwise stated.

Similar to cubelet-to-cubelet connections, C_k (k=1,2,3,4), we also defined 4 types of brain area-to-brain area connections, A_k (k=1,2,3,4), and cubelet-to-brain area connections, P_k (k=1,2,3,4), as summarized below.

Connection type	Connection source	Connection target	Definition	Formula
Type 1, A_1	Brain area	Brain area	Unit-to-unit	A_1
Type 2, A_2	Brain area	Brain area	Unit-to-total	$A_2 = A_1 S_a$
Type 3, A_3	Brain area	Brain area	Total-to-unit	$A_3 = S_a A_1$
Type 4, A_4	Brain area	Brain area	Total-to-total	$A_4 = S_c A_1 S_a$

Connection type	Connection source	Connection target	Definition	Formula
Type 1, P_1	Cubelet	Brain area	Unit-to-unit	P_1
Type 2, P_2	Cubelet	Brain area	Unit-to-total	$P_2 = P_1 S_a$
Type 3, P_3	Cubelet	Brain area	Total-to-unit	$P_3 = S_c P_1$
Type 4, P_4	Cubelet	Brain area	Total-to-total	$P_4 = S_c P_1 S_a$

Here S_a is a diagonal matrix, and its element $\{S_a\}_{ii}$ represents the physical size of brain area i.

We also calculated a cubelet-to-brain area mapping matrix, M, based on cubelet registration results. $\{M\}_{ij}$ represents the physical size of the intersection of cubelet i and brain area j. The mapping matrix M was also normalized to either the total size of each brain area or to the total size of each cubelet:

$$M_a = M S_a^{-1} \quad (15)$$

$$M_c = S_c^{-1} M \quad (16)$$

. In M_a, the sum of each column is 1; in M_c, the sum of each row is 1.

3.1 Inferring cubelet-to-brain area connections/brain area-to-brain area connections by weighted averaging (Fig. 2C-E; Fig. S4B-D; Fig. S8H).

While we have dissected the cortex into ~ 230 cubelets, there are ~ 70 brain cortical areas according to Allen atlas (2011 version). The size of a cortical area was much larger than a cubelet, and an area on average consisted of 10 cubelets. Thus we considered the cubelets as building blocks of brain connectivity and assumed connections between brain areas were weighted averages of cubelets contained (Fig. S4B,D). With such an assumption, we had:

$$P_2 = C_1M \quad (17)$$
$$A_3 = M^TP_1 \quad (18)$$

, where M^T denotes the transpose of M.

With Eq. (16) and (17), we got

$$P_2 = C_1M = C_2S_c^{-1}M = C_2M_c \quad (19)$$

. With Eq. (15) and (18), we got

$$A_2 = S_a^{-1}A_3\,S_a = S_a^{-1}M^TP_1S_a = (MS_a^{-1})^T\,P_1S_a = M_a^TP_2 \quad (20)$$

. With Eq. (19) and (20), we got

$$A_2 = M_aP_2 = M_a^TC_2M_c \quad (21)$$

. We inferred cubelet-to-brain area connections with (19) in Fig. 2D,E; and inferred brain area-to-brain area connections with Eq. (21) in Fig. 2C, Fig. S8H.

To reduce the variations brought by dissection and registration errors, we downsampled the cubelet-to-cubelet connection matrix for analyses here. If α_0 and β_0 were two cubelets, $\alpha_1, \alpha_2 \ldots \alpha_m$ were neighbors of α_0, and $\beta_1, \beta_2 \ldots \beta_n$ were neighbors of β_0, then the projection strength from α_0 to β_0, $C_{\alpha_0-\beta_0}$ was downsampled as:

$$C_{\alpha_0-\beta_0} = \begin{pmatrix} 0.9 & \dfrac{0.1}{m} & \cdots & \dfrac{0.1}{m} \end{pmatrix} \begin{pmatrix} C_{\alpha_0-\beta_0} & C_{\alpha_0-\beta_1} & \cdots & C_{\alpha_0-\beta_n} \\ C_{\alpha_1-\beta_0} & C_{\alpha_1-\beta_1} & \cdots & C_{\alpha_1-\beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{\alpha_m-\beta_0} & C_{\alpha_m-\beta_1} & \cdots & C_{\alpha_m-\beta_n} \end{pmatrix} \begin{pmatrix} 0.9 \\ \dfrac{0.1}{n} \\ \vdots \\ \dfrac{0.1}{n} \end{pmatrix} \quad (22).$$

For the analysis in 3.1, all the non-significant cubelet-to-cubelet connections were set to 0. As multiple comparison had a high false negative rate particularly for weak projections, p value = 0.05 (no multiple comparison) was used for the criterion of significance here. For comparison between cubelets and injections in the same source brain area (Fig. 2D,E), we require the cubelets reside primarily (>70%) in the brain area.

3.2 Inferring brain area-to-brain area connections by constrained optimization (Fig. S4E-G).

In contrast to assuming cubelets, which were smaller in size, were building blocks of brain connections, connections of brain areas could also be inferred assuming input and output of cells within each brain area were homogeneous (Fig. S4E) (*5*). With this assumption, we had:

$$P_3 = MA_1 \quad (23)$$
$$C_2 = P_1M^T \quad (24)$$

. The Eq. (23) and (24) corresponded to output homogeneity and input homogeneity, respectively.

With Eq. (16) and (23), we got

$$P_2 = S_c^{-1}P_3S_a = S_c^{-1}MA_1S_a = M_cA_2 \quad (25)$$

. With Eq. (15) and (25), we got

$$C_2 = \ P_1M^T = P_2S_a^{-1}M^T = P_2M_a^T \quad (26)$$

. With Eq. (25) and (26), we got

$$C_2 = M_cA_2M_a^T \quad (27)$$

.

According to Eq. (27), we could estimate A_2 (least-squares solution) with:

$$\widetilde{A_2} = M_c^+C_2(M_a^+)^T \quad (28)$$

, where $\widetilde{A_2}$ is estimated A_2, and M_c^+ (M_a^+) is the pseudo-inverse matrix of M_c (M_a). However, this might result in negative connection values. Thus, we determined to estimate A_2 with constrained optimization:

$$\widetilde{A_2} = argmin_{A_2}(\|C_2 - M_cA_2M_a^T\|) \ (29)$$

, with the constraint

$$A_2 \geq 0 \quad (30)$$

. With Eq. (29) and formula (30), we inferred brain area-to-brain area connections in Fig. S4F,G.

To reduce the variations brought by registration errors, downsampling was also performed here for the cubelet-to-cubelet connection matrix with Eq. (22).

For the analysis in 3.2, all the non-significant cubelet-to-cubelet connections were set to 0. As multiple comparison had a high false negative rate particularly for weak projections, p value = 0.05 (no multiple comparison) was used for the criterion of significance here.

3.3 Discussions

It remains a challenge to infer the underlying connections between brain areas from neural tracing experiments (*11, 12*). In the real scenario, neither cubelets nor brain areas were necessarily homogeneous, and thus we did not aim to develop a method to precisely quantify brain area-to-brain area connection patterns here. However, by inferring brain area connections with abovementioned assumptions, we argue that it provided a fair approach to validate muMAPseq and screen for long-range connection disruptions in neuropsychiatric disorders.

In the 'weighted averaging' approach, we assumed that individual cubelets were homogeneous, and broke down brain areas into cubelets contained. Ideally this would be correct if the size of cubelets was small enough so that each cubelet was a homogeneous unit. However, with the current experiment protocol, inferring connections with this method was still an estimation. In the 'constrained optimization' approach, the assumption that input and output of all the regions within a brain area were homogeneous was also imprecise. For example, the connections between primary visual cortex and higher visual areas are organized according to retinotopic maps (*13*). Moreover, our results also indicated the rank of C was much higher than the total number of brain areas (data not shown), arguing against the brain area homogeneity assumption. Future work should be done to better define, quantify and calculate brain connections.

Supplementary Note 4. Supplementary results and discussions

4.1 Single cell projection data

In principle, by determining soma locations of individual barcodes, we were able to reconstruct projection patterns at single cell resolution. As an example, Fig. S2A shows projection patterns of 712 cells in one source cubelet (cubelet #53 in PTLp); the cells are clustered with non-negative least squares-based sparse non-negative matrix factorization (cluster number is 10). Three of them are highlighted in dorsal views (Fig. S2B). In the current data, individual cells displayed a wide range of node degrees (i.e. numbers of projecting target; Fig. S2C). However, due to re-used barcodes (Supplementary Note 2.2), template switching (Supplementary Note 2.1) and inadequate sequencing depth (Fig. S2D), further analyses regarding single cell projections were not performed with the current data.

4.2 Comparing MAPseq data with functional imaging data

To compare MAPseq data with functional imaging data, we trained animals with a perceptual decision making task (Fig. S5A; see details in methods) (*4*), and calculated the reciprocal connection strength, input correlation (Pearson correlation of input vectors), activity correlation, noise correlation and spontaneous correlation for each pair of cubelets (Supplementary Note 5.10). Activity correlations (Fig. 3C; Fig. S5B), noise correlations (Fig. S5F) and spontaneous correlations (Fig. S5D) were all strongly correlated with reciprocal connection strengths or input correlations. As the distance between cubelets had a large effect on the connection strength (Fig. 3B), input correlation (Fig. S6B) and activity correlation (data not shown), we further removed distance-dependent components from connection strengths, input correlations, and activity correlations (Supplementary Note 5.10). The residual distance-independent components showed weaker, but still significant correlations (Fig. S5C). Moreover, similar results were found between connection strength/input correlation and noise/spontaneous correlation (Fig. S5E,G). The consistency between connection data and functional data not only validated MMAPseq as a functionally relevant measure of cortical connectivity, but also suggested that intracortical connections are highly related to cortical activity.

4.3 Input/output correlation

The input correlation (Pearson correlation of a pair of input vectors) and output correlation (Pearson correlation of a pair of output vectors) were calculated for each pair of cubelet. We found that both the input correlation and output correlation showed asymmetric distributions and no strongly negatively correlated input/output patterns were observed. This is consistent with sparseness of connections in the cortex. Furthermore, not surprisingly, both input correlation and output correlation decayed as the distance between two cubelets increased (Fig. S8B,C,F,G), suggesting that proximal cubelets share more similar connection patterns than distal cubelets.

4.4 Motif analysis

The statistical properties of a connected network can be decomposed into a series of mathematical terms that quantify increasingly complex motifs in the network, analogous to a Taylor expansion. Specifically, considering a binary directed network $\boldsymbol{G}$ (represented by its adjacent matrix), we define its following properties with increasing statistical orders:

1st order property: connection probability. Let $f_{1,1}(\boldsymbol{G})$ denote the connection probability in network $\boldsymbol{G}$.

2nd order properties: probability of 2-node motifs. There are 3 possible 2-node motifs: non-connected 2-node motif, uni-connected 2-node motif, and bi-connected 2-node motif, so (3-1=2) degrees of freedom are needed to describe 2nd order properties. Let $f_{2,1}(\boldsymbol{G})$, $f_{2,2}(\boldsymbol{G})$ and $1 - f_{2,1}(\boldsymbol{G}) - f_{2,2}(\boldsymbol{G})$ denote the probabilities of 2-node motifs in $\boldsymbol{G}$.

3rd order properties: probability of 3-node motifs. There are 16 possible 3-node motif, so (16-1=15) degrees of freedom are needed to describe 3rd order properties. Let $f_{3,1}(\boldsymbol{G})$ … $f_{3,15}(\boldsymbol{G})$ and $1 - \sum_{i=1}^{15} f_{3,i}(\boldsymbol{G})$ denote the probabilities of 3-node motifs in $\boldsymbol{G}$.

…

nth order properties: probability of n-node motifs. Let M_n denote the number of n-node motifs, then we need ($df_n = M_n - 1$) degrees of freedom to describe nth order properties. Let $f_{n,1}(\boldsymbol{G})$ … $f_{n,df_n}(\boldsymbol{G})$ and $1 - \sum_{i=1}^{df_n} f_{n,i}(\boldsymbol{G})$ denote the probabilities of n-node motifs in $\boldsymbol{G}$.

Given nth order properties of $\boldsymbol{G}$, we are able to infer the probability distribution of $\boldsymbol{G}$, $P_n(\boldsymbol{G})$, using the maximum entropy principle (*14*):

$$P_n(\boldsymbol{G}) = \frac{1}{\ln Z_n} \exp\left(- \sum_{i=1}^{df_n} \mu_{n,i} f_{n,i}(\boldsymbol{G}) \right)$$

, where

$$Z_n = \sum_{\boldsymbol{G}} \exp(- \sum_{i=1}^{df_n} \mu_{n,i} f_{n,i}(\boldsymbol{G}))$$

$$\frac{\partial Z_n}{\partial \mu_{n,i}} =< f_{n,i}(\boldsymbol{G}) >$$

. Here $< f_{n,i}(\boldsymbol{G}) >$ represents observed probability of a given n-node motif.

Define

$$I_n = \begin{cases} \sum_{i=1}^{df_n} \mu_{n,i} f_{n,i}(\boldsymbol{G}) - \sum_{i=1}^{df_{n-1}} \mu_{n-1,i} f_{n-1,i}(\boldsymbol{G}) \quad (n > 1) \\ \mu_{1,1} f_{1,1}(\boldsymbol{G}) \quad (n = 1) \end{cases}$$

, then we get an expansion form of $P_n(\boldsymbol{G})$:

$$P_n(\boldsymbol{G}) = \frac{1}{\ln Z_n} \exp\big(-(I_1 + I_2 + \cdots + I_{n-1} + I_n)\big)$$

. In the above formula, if we only take the first k terms and recalculate the partition function Z, we get $P_k(\boldsymbol{G})$, i.e., the probability distribution of $\boldsymbol{G}$ given only k^{th} order statistical properties. Studying $P_k(\boldsymbol{G})$ with increasing k reveals the effect of higher order statistical properties on $\boldsymbol{G}$. Particularly, the second and third terms represent the interactions between pairs and triplets of elements, respectively.

We first studied 2-cubelet motifs in the C57BL/6J cortical network. Random networks, $\mathrm{RN_g}$ were generated based on observed global first-order properties ($< f_{1,1}(\boldsymbol{G}) >$). In our C57BL/6J cortical network, the probability that a pair of cubelets x and y was reciprocally connected was greater than predicted by the null hypothesis (Fig. 4H; Fig. S9B-D). Three-cubelet motifs were also highly non-random, with a tendency for densely connected motifs to be particularly overrepresented (Fig. 4I; Fig. S9B-D), compared to $\mathrm{RN_g}$, random networks generated based on observed global second-order properties (i.e. probabilities of 2-cubelet motifs, $< f_{2,i}(\boldsymbol{G}) >, i = 1,2$). The most under-represented motif was a unidirectional cycle, similar to what has been reported at the cellular level (*15*). Consistent with 3-cubelet motif statistics, the observed clustering coefficient was high compared to random networks (Fig. S9G). Interestingly, the distribution of 3-cubelet motifs was strikingly similar to statistics of connections among single neurons in the rat visual cortex (*16*), suggesting that a common rule might govern the organization of neural circuits at both microscale (intra-neuron) and mesoscale (intra-area) levels. These analyses reveal that the network architecture was highly structured, deviating sharply from simple random connectivity.

As the distances between cubelets may affect the probability of connections or motifs, we also considered random networks generated with the observed distance-dependent low-order properties, $\mathrm{RN_{dd}}$. Comparing $\mathrm{RN_{dd}}$ to observed networks (Supplementary Note 5.7), similar overrepresented/underrepresented motifs were found (Fig. S9A).

4.5 Module analysis

Previous analyses of the connectivity between cytoarchitecturally defined brain areas (*11, 17*) revealed “modules”— regions of the brain within which connections are dense, and which may reflect functional units. Because the basic unit in muMAPseq is a cubelet, defined by dissection without regard to functionally defined regions, we wondered whether similar modules would emerge, or whether we could reveal structure within classical brain areas that were previously obscured by their labeling as one homogeneous area.

To analyze modules of the ipsilateral cubelet-to-cubelet connection matrix, we utilized a community structure-finding algorithm (*18*). In the algorithm, a resolution parameter, γ can be tuned to get smaller/more or larger/fewer modules. To choose a proper γ, we undersampled from all the projection neurons, and used the algorithm to find modules. The optimal γ was chosen so that the Rand index (inconsistency) was low and the average number of modules was stable (Fig. S10A; Supplementary Note 5.6).

In the C57BL/6J mouse #LJ7, with the optimal γ=0.87, we recovered four major modules (Fig. 4J,K), of which module 1 belonged to visual-auditory areas, modules 2 and 3 belonged to somatosensory/motor areas, and module 4 belonged to anterior cingulate/retrosplenial areas. Interestingly, the two modules belonging to somatosensory/motor areas were not clustered according to brain areas defined in the Allen atlas (i.e. SSp, SSs, MOp, MOm), but were clustered according to the represented body parts. Roughly, module 2 corresponded to somatosensory and somatomotor areas associated with sensation and movement of limbs, trunk and whiskers, whereas module 3 corresponded to areas associated with mouth and nose. Interestingly, the modules obtained by this analysis of connectivity closely match those obtained by brain-wide calcium imaging and clustering, but not the partitioning based on cytoarchitecture(*19*).

In addition to the connection matrix, we could perform a similar analysis on the input (or output) correlation (Pearson correlation between input to a pair of cubelets or between output from a pair of cubelets) matrix. The modular organization of these matrices was similar to that of the connections themselves (Fig. S10B,C), suggesting that inter-connected modules tend to receive similar inputs and send similar outputs.

As the distance between two cubelets strongly affected their connection strength (Fig. S8A), we asked whether the observed connection modules in the connection matrix were a result of simple distance dependence, or due to specific connection patterns between brain areas. To address this question, we generated a distance-dependent connection matrix (M_{dd}) with observed average connection strengths at various distances, and performed clustering analysis (Fig. S10D, middle; Fig. S10E, left). A distance-independent connection matrix (M_{di}) was also generated by subtracting the distance-dependent connection matrix from the original connection matrix, and analyzed in a similar way (Fig. S10D, right; Fig. S10E, right). The original connection modules were more similar to the distance-independent connection modules, suggesting the cortical modules were not simply organized with a distance-dependent rule, but reflected specific connection patterns between certain brain areas.

We also performed module analysis with various γ. With lower γ, we found fewer modules, which consisted of one or more modules that were determined with γ=0.83. However, we failed to detect fine modules from the connection network with higher γ, probably due to limited spatial resolution of dissected cubelets (300μm×1mm×1mm).

Similar analysis was done in mouse LJ9 (Fig. S10G-I). With the optimal resolution parameter γ=0.8, five major modules were recovered: two somatosensory-somatomotor area modules (similar with the two modules in LJ7), a visual-auditory area module, an anterior cingulate area module, and a retrosplenial area module. The global modular organizations between LJ7 and LJ9 were similar, and the subtle differences may result from limited spatial resolution and cubelet dissection variations.

4.6 Analysis of contralateral projections

Commissural connections play important roles in regulating a variety of behaviors, and disruption of these projections might play a role in autism and other neurological dysfunction (*20, 21*). However, the network properties of commissural projections are much less studied than their ipsilateral counterparts. We therefore examined the structure of commissural connections in our dataset. In the C57BL/6J mouse #LJ7, homotopic projections, i.e. projections from one area of cortex to the corresponding area in the other hemisphere (Fig. 4D, left), are the most likely to be positive (Fig. 4D, middle; Fig. S7A, blue; 47.2±0.2% of all possible homotopic projections are positive), and made up a substantial fraction of all commissural projections (36.1±2.1% of all non-zero commissural projections are homotopic), consistent with previous reports (18-21). The remaining projections were heterotopic projections from one area of cortex to a non-corresponding area in the contralateral hemisphere, but not to its corresponding contralateral area. To understand the structure of these projections, we further divided heterotopic projections into those with projections to both the ipsi and contralateral versions of a target area (heterotopic ipsi+) and those that projected only to the contralateral version of one area, but not the ipsilateral one (heterotopic ipsi-; Fig. 4D, left). 73.1±7.2%of positive heterotopic projections were of the ipsi+ kind. Heterotopic ipsi- projections accordingly made up only 17.1%±3.9% of the all commissural projections. Our findings therefore support a largely symmetric model of the mouse cortex, where a given area in one hemisphere often directly projects to its corresponding area in the contralateral hemisphere, and moreover preferably projects to both ipsi- and contralateral versions of other target areas.

Heterotopic ipsi+ commissural projections constituted a substantial fraction of total commissural projections and they represented bifurcated projections to two hemispheres. To further study these projections, we defined the source area S, the ipsilateral target area T, and the contralateral target area T’. Note T and T’ are homotopic, according to the definition. For all the positive heterotopic ipsi+ commissural projections, the correlation between S-T projection strength and S-T’ projection strength was weak (Fig. S7B).

All the previous analysis of commissural projections was based on projections that have passed the significance test with multiple comparison. As commissural projections were usually weaker than association projections (Fig. 4E), the false negative rate for commissural projections might be higher. To examine how the false negative rate may affect results, we also did parallel analysis with projection data that have passed the significance test but without multiple comparison. Homotopic commissural projection was still the major commissural projection type, while the number of heterotopic ipsi+ commissural projection increased (Fig. S7C,D). For heterotopic ipsi+ projections, weak but significant correlation was observed between the association projection strength and the commissural projection strength (Fig. S7E).

4.7 The BTBR brain

In the BTBR brain, the connection strength and the input/output correlation between a pair of cubelets decreased as the distance increased (Fig. S8G), similar to what was observed in LJ7 and LJ9.

By performing module analysis, we found modules in the connection matrix (M_c; Fig. S10L). Four major modules were found in the connection matrix at the optimal resolution parameter γ=1.1 (Fig. S10J, left): three modules in the somatosensory-somatomotor area (roughly corresponding to orofaciophryngeal, upper limb, lower limb – whisker areas respectively), and one module in the anterior cingulate-retrosplenial-visual area. The differences in the somatosensory-somatomotor areas between the BTBR brain and C57BL/6J brains might be due to limited spatial resolution and cubelet dissection variations. The failure to get the visual-auditory area module might be explained by injection artifacts. Note that in the BTBR brain, probably due to lack of corpus callosum, the two cortical hemispheres are physically separated much more rostrally than a C57BL/6J mouse. Thus some cortical brain areas including the auditory cortex and part of the visual cortex are more lateralized and difficult to be targeted by viral injection from the dorsal surface. Actually, very few, if any, somata were found in these areas from the sequencing results. As cubelets with too few infected cells (less than 50) were excluded for analysis, not surprisingly the visual-auditory area was not recovered as a module, as seen in LJ7. We also performed module analysis with the input correlation matrix (M_{ic}; Fig. S10J, middle) and the output correlation matrix (M_{oc}; Fig. S10J, right), and the results were similar to modules in the connection matrix (Fig. S10K).

Topological properties of the ipsilateral cubelet-to-cubelet connectivity network was also examined in the BTBR brain. All the results were very similar to C57BL/6J brains (Fig. S9E-G). Briefly, among all the 2-node motifs, bidirectional connections were overrepresented; the clustering coefficient was significantly higher than random networks generated with the same second-order properties; the distribution of 3-node motifs was highly non-random, and densely connected motifs were overrepresented. The results suggested that these topological properties of the ipsilateral connection network were not disrupted in the BTBR brain.

Supplementary Note 5: Bioinformatics, statistics and computational methods

5.1 Processing of raw sequencing data.

Raw Illumina sequencing results consisted of two .fastq files: 32-nt BC sequences were in paired end 1, and 12-nt UMI and 8-nt CSI sequences were in paired end 2. The full BC-UMI-CSI sequences were merged and then de-multiplexed based on CSIs (cubelets). All the sequences with ambiguous bases (shown as N in the sequencing results) were removed. We then collapsed all the identical reads. As the current sequencing depth was too low and most of the sequences only had 1 read each, we didn't set any threshold for read counts to remove errors (but see Supplementary Note 1). Unique sequences were next sorted into barcode library 1 (BC ended with 2 purines), barcode library 2 (BC ended with 2 pyrimidines), and spike-in (BC ended with ATCAGTCA). We then counted the number of unique UMIs for each BC-CSI, which represented the molecule count of a given barcode in a given cubelet.

5.2 Substitution error correction.

Base substitution is one of the major error sources. As the theoretical diversity of a random barcode of $N_{30}YY$ or $N_{30}RR$ is $4^{30}\times2^{2} \approx 10^{18}$, an error barcode due to substitution should be very similar to one of the real barcodes, while any two real barcodes should be very different. To correct substitution errors, we first found all the barcode pairs with up to 3 mismatches using the short read aligner *bowtie* (http://bowtie-bio.sourceforge.net/index.shtml). We next collapsed all the barcodes into a large number of clusters, such that for any barcode (BC1) in a given cluster, there existed another barcode (BC2) in the same cluster with less than 3 mismatches. As a simple algorithm, mathematically it could cause very different barcodes to be collapsed into the same cluster; however, this did not happen in the real scenario due to the high theoretical diversity. The barcode with the highest UMI counts in each cluster was used to represent the cluster, and the summed UMI count of all the barcodes in the cluster was calculated as the corrected UMI count of the barcode. After substitution correction, we generated a barcode-cubelet matrix, where each element represented the molecule count of a given barcode in a given cubelet after collapsing.

5.3 Reconstruction of single cell projections.

5.3.1 Viral abundance thresholding. To reduce re-used barcode errors, barcodes whose counts were greater than 4 in the viral library sequencing result were excluded for analysis in the barcode-cubelet matrix. See full details in Supplementary Note 1.3.

5.3.2 UMI thresholding. To remove noises, we set all the no-greater-than-1 (UMI threshold) elements in the matrix to 0.

5.3.3 Soma/axon thresholding. After barcode abundance thresholding and UMI thresholding, we determined the soma location of each barcode using the ‘soma-max’ strategy. To exclude local dendritic innervations, for each barcode, the UMI counts of all the cubelets neighboring to the soma cubelet were set to 0. Firstmax and secondmax were then calculated as the highest and second highest UMI counts for each barcode. We chose soma threshold to be 250 and axon threshold to be 20, and only analyzed barcodes whose firstmax was greater than soma threshold and secondmax was between UMI threshold and axon threshold. See full details in Supplementary Note 1.2.

5.3.4 Filter right cortical neurons. We remove the barcodes whose somas did not reside in the right cortical hemisphere. Cells not in the right cortex were extremely rare, and they were likely due to virus spread.

With these steps, we were able to determine each cell’s location and its projection pattern.

5.4 Calculating bulk projection patterns.

To calculate bulk projection patterns, we pooled all the projection cells that resided in the same cubelets together, and calculated their average projection patterns. Projection strengths due to noise were evaluated and subtracted from the uncorrected projection strengths. P values were also calculated for each projection. See details in Supplementary Note 1.

In the manuscript, ‘(non-)significant connections (no multiple comparison)’ refer to connections with p value (≥) < 0.05; ‘(non-)significant connections (multiple comparison)’ refer to connections with p value (≥) < 0.05/N, where N is total number of possible connection (the number of right cortical cubelets times the number of all the cortical and subcortical cubelets).

Some of the RT primers were found to be cross-contaminated at low levels post hoc. Thus, we didn't analyze the projections between these contaminated cubelets. These projections include: LJ7, cubelet 97-to-cubelet 68, cubelet 21-to-cubelet 268; LJ9, cubelet 75-to-cubelet 13, cubelet 13-to-cubelet 75; LJ10, cubelet 60-to-cubelet 81, cubelet 81-to-cubelet 60.

5.5 Distribution of connection strengths and distance-dependent connection properties

To calculate distributions of connection strengths and distance-dependent connection properties, only significant non-zero (with Bonferroni multiple comparison correction) cubelet-to-cubelet connections were included. The distance between 2 cubelets was defined as the distance of their centroids.

5.6 Analysis of modules.

We utilized the Brain Connectivity Toolbox (https://sites.google.com/site/bctnet/) for module analysis in Matlab. *modularity_dir.m* was used to find modules in the connectivity matrix (directed graph), and *modularity_und.m* was used to find modules in the input/output correlation matrix (undirected graph). In input/output correlation matrix, negative values were set to 0 before clustering. A resolution parameter γ can be tuned to get smaller/more or larger/fewer modules. To determine the optimal γ, we undersampled half of the total projection neurons for 100 times, and performed clustering with various γ. For each γ, we calculated the average number of modules over 100 undersampling trials, and quantified the inconsistency of clustering that was defined as the mean of Rand indices between pairwise trials' clustering results. The optimal γ was chosen so that the inconsistency was low and the average number of modules was stable (Fig. S8A). All the analyses were done with the optimal γ unless otherwise stated.
To generate the distance-dependent connection matrix, we first calculated connection strengths and physical distances for all cubelet pairs. We next grouped cubelet pairs into bins according to the distances (50 μm each bin), and calculated the mean connection strength in each bin. Then in the distance-dependent connection matrix, each element was set to the mean connection strength of the bin it belonged to. To calculate the distance-independent connection matrix, the distance-dependent connection matrix was subtracted from the original connection matrix. Negative values in the distance-independent connection matrix were set to 0 before clustering. The distance between 2 cubelets was defined as the distance of their centroids.

Clustering results were compared with Rand indices (*22*).

For module analysis, non-significant (with Bonferroni multiple comparison correction) cubelet-to-cubelet projections were set to 0.

5.7 Analysis of motifs.

clustering_coef_bd.m in the Brain Connectivity Toolbox was used to calculate the clustering coefficient. The connection matrix was binarized for this analysis. For comparison, we generated random connection networks based on distance-dependent connection probability rule: in the real network, we calculated the probability that cubelet i projected to cubelet j if their distance was d (in 50 μm bins); then the measured probabilities were used to generate 10000 random networks assuming each connection was independent.

3 types of 2-node motifs and 16 types of 3-node motifs were counted in real cortical networks. Random networks were also simulated to calculate the relative abundance of each motif in real networks. The relative abundance was calculated with:

$$\frac{Count_{real}(motif\ i) - Count_{random}(motif\ i)}{Count_{random}(motif\ i)}.$$

Different models were used to generate random networks, and 10000 random networks were generated each:
In 2-node motif comparison, $\mathrm{RN_g}$ was generated based on a global connection probability rule: in the real network, we calculated the probability that cubelet i projected to cubelet j; then the measured probability was used to generate $\mathrm{RN_g}$ assuming each connection was independent.
In 2-node motif comparison, $\mathrm{RN_{dd}}$ was generated based on a distance-dependent connection probability rule: in the real network, we calculated the probability that cubelet i projected to cubelet j if their distance was d (in 50 μm bins); then the measure probabilities were used to generate $\mathrm{RN_{dd}}$ assuming each connection was independent.
In 3-node motif comparison, $\mathrm{RN_g}$ was generated based on a global 2-node motif probability rule: in the real network, we calculated the probability of each 2-node motif between cubelet i and cubelet j, then the measured probability was used to generate $\mathrm{RN_g}$ assuming each 2-node motif was independent.
In 3-node motif comparison, $\mathrm{RN_{dd}}$ was generated based on a distance-dependent 2-node motif probability rule: in the real network, we calculated the probability of each 2-node motif between cubelet i and cubelet j if their distance was d (in 50 μm bins), then the measured probabilities was used to generate $\mathrm{RN_{dd}}$ assuming each 2-node motif was independent.

For all the analysis in 5.7, the distance between 2 cubelets was defined as the distance of their centroids.

For motif analysis, non-significant (with Bonferroni multiple comparison correction) cubelet-to-cubelet projections were set to 0.

5.8 Analysis of contralateral projections.

As we dissected most cubelets symmetrically, we were able to find the contralateral homotopic cubelet for a given cubelet. Considering sectioning/dissection variations, we also generalized contralateral homotopic cubelets to include all the neighbor cubelets of the exact contralateral homotopic cubelet. Generalized contralateral homotopic cubelets were used for all the analysis regarding contralateral projections.

For contralateral projection analysis, non-significant cubelet-to-cubelet projections were set to 0. When calculating log values, zeroes were set to 1/10 of the smallest nonzero values.

5.9 Node degree analysis

In Fig. 4G, the normalized node degree is defined as the number of non-zero projection target cubelets divided by the total number of possible target cubelets in ipsilateral cortex.

5.10 Analysis of function-connection relationship

With the functional imaging data, we first performed singular-value decomposition with the activity matrix (pixel-by-time). The first 500 components were included for following analysis. To determine the activity of each cubelet, we calculated the mean activity over all pixels belong to the same cubelet. The activity correlation was calculated using activity data in all the time frames of all the trials. The spontaneous correlation was calculated using activity data from 0-1s of all the trials (note the initialization of each trial was at 2±0.2 s). To calculate the noise correlation, we grouped them into auditory-left-correct (modality-choice-result), auditory-right-correct, visual-left-correct, visual-right-correct, auditory-left-incorrect, auditory-right-incorrect, visual-left-incorrect, visual-right-incorrect trial groups. The mean activity at a given time point over all the trials in the same group was subtracted from the original activity data belonging to the corresponding trial group to calculate noises. All the correlations were calculated as Pearson correlations.

The down-sampled connection matrices (Supplementary Note 2.1) were used for the connection analysis. The reciprocal connection strength was calculated as the mean of logarithm of connection strengths in two directions. To compare function data with connection data, we only included cubelet pairs that satisfied 1) number of infected cells in both cubelets were greater than 50 in MAPseq, 2) both cubelets were well imaged (excluding non-surface areas like orbitofrontal cortex/anterior cingulate cortex/retrosplenial cortex, and lateral areas like insular cortex), 3) the two cubelets in a pair were not neighbors (neighbor connections were not analyzed in MAPseq).

To remove distance-dependent components from activity correlations, spontaneous correlations, noise correlations, connection strengths, and input correlations, we grouped cubelets pairs into bins according to the distances (50 μm each bin), and calculated the mean value of each variable in each bin. The mean value of each variable was then subtracted from the original data in the corresponding bins to calculate distance-independent components. The averaging and subtraction of connection strengths were performed in the logarithm form. The distance between 2 cubelets was defined as the distance of their centroids.

For the analysis in 5.9, all the non-significant cubelet-to-cubelet connections were set to 0. As multiple comparison had a high false negative rate particularly for weak projections, p value = 0.05 (no multiple comparison) was used for the criterion of significance here.

Supplementary Table 1. Sindbis viral injections
Supplementary Table 2. Cubelet-to-cubelet projection strengths
Supplementary Table 3. Cubelet registration result
Supplementary Table 4. Inferred brain area-to-brain area connection strengths

Supplementary Movie 1. Summary of muMAPseq results

1. J. M. Kebschull *et al.*, High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA. *Neuron* **91**, 975-987 (2016).
2. J. M. Kebschull, P. Garcia da Silva, A. M. Zador, A New Defective Helper RNA to Produce Recombinant Sindbis Virus that Infects Neurons but does not Propagate. *Front Neuroanat* **10**, 56 (2016).
3. J. Morris, J. M. Singh, J. H. Eberwine, Transcriptome analysis of single cells. *J Vis Exp*, (2011).
4. S. Musall, M. T. Kaufman, S. Gluf, A. K. Churchland, Movement-related activity dominates cortex during sensory-guided decision making. *bioRxiv*, (2018).
5. S. W. Oh *et al.*, A mesoscale connectome of the mouse brain. *Nature* **508**, 207-214 (2014).
6. P. E. Hart, N. J. Nilsson, B. Raphael, A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* **4**, 100-107 (1968).
7. T. J. Hastie, R. J. Tibshirani, *Generalized additive models*. (Chapman & Hall, Boca Raton, 1999).
8. X. Chen, J. M. Kebschull, H. Zhan, Y.-C. Sun, A. M. Zador, Spatial organization of projection neurons in the mouse auditory cortex identified by in situ barcode sequencing. *bioRxiv*, (2018).
9. Y. Han *et al.*, The logic of single-cell projections from visual cortex. *Nature* **556**, 51-56 (2018).
10. J. M. Kebschull, A. M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res* **43**, e143 (2015).
11. J. A. Harris *et al.*, The organization of intracortical connections by layer and cell class in the mouse brain. *bioRxiv*, (2018).
12. J. E. Knox *et al.*, High resolution data-driven model of the mouse connectome. *bioRxiv*, (2018).
13. Q. Wang, A. Burkhalter, Area map of mouse visual cortex. *J Comp Neurol* **502**, 339-357 (2007).
14. E. T. Jaynes, Information Theory and Statistical Mechanics. *Phys Rev* **106**, 620-630 (1957).
15. J. B. Dechery, J. N. MacLean, Functional triplet motifs underlie accurate predictions of single-trial responses in populations of tuned and untuned V1 neurons. *PLoS Comput Biol* **14**, e1006153 (2018).
16. S. Song, P. J. Sjostrom, M. Reigl, S. Nelson, D. B. Chklovskii, Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol* **3**, e68 (2005).

17. B. Zingg *et al.*, Neural networks of the mouse neocortex. *Cell* **156**, 1096-1111 (2014).
18. M. Rubinov, O. Sporns, Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**, 1059-1069 (2010).
19. M. P. Vanni, A. W. Chan, M. Balbi, G. Silasi, T. H. Murphy, Mesoscale Mapping of Mouse Cortex Reveals Frequency-Dependent Cycling between Distinct Macroscale Functional Modules. *J Neurosci* **37**, 7513-7533 (2017).
20. R. W. Sperry, Citation Classic - Interhemispheric Relationships - the Neocortical Commissures - Syndromes of Hemisphere Disconnection. *Cc/Life Sci*, 21-21 (1985).
21. L. K. Paul, Developmental malformation of the corpus callosum: a review of typical callosal development and examples of developmental disorders with callosal involvement. *J Neurodev Disord* **3**, 3-27 (2011).
22. W. M. Rand, Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* **66**, 846-850 (1971).