1. Research Plan

Technical Area 2: Neuro-anatomical Data Collection

F. Overview of the neuroanatomical approach.

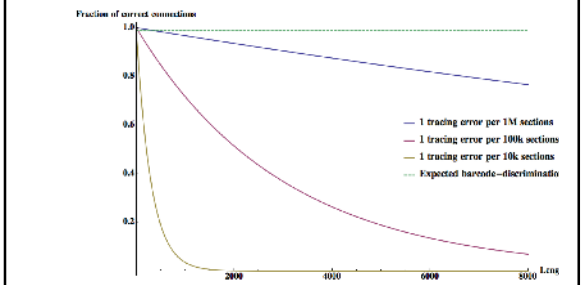

Figure 1 The probability of correctly ascribing a given neurite to its parent falls off exponentially with distance with serial reconstruction. By contrast, the error rate using barcode-based reconstruction is independent of distance.

Our neuro-anatomical approach is unique in that it relies on optical readout of cell-identifying nucleic acid barcodes, rather than on electron microscopic axon tracing. While highly novel in its principles and implications, this strategy is based on technologies that are now well established and rapidly advancing, including molecular barcoding (Lu, Neff, Quake, & Weissman, 2011; Peikon, Gizatullina, & Zador, 2014; Zador et al., 2012) and fluorescent in-situ sequencing (Lee et al., 2014, 2015; Mitra, Shendure, Olejnik, & Church, 2003) as well as expansion microscopy (Chen, Tillberg, & Boyden, 2015). ***Crucially, the approach treats long-range and short-range connections on an equal footing, and is applicable to arbitrarily large brain volumes without a drop in accuracy: error rates are constant as a function of the length of "traced" axon, rather than scaling linearly with it.*** The approach enables very high speeds: once the technology is in place, data collection of three mice for the Phase 1 milestone will take approximately one month on a single microscope. Moreover, the approach naturally integrates with measurement of cell morphology, and, for the first time, opens the way toward richly molecularly annotated connectomics. This transformative approach fundamentally changes the "governing laws" of connectomics technology, enabling massive scaling, redundant error-correction of multiple data-types, and integration of molecular annotation. Therefore, the approach pioneered here will establish the future trajectory of structural brain mapping methods.

From tracing to barcoding.

The brain is organized at the nanoscale, yet neural circuits span over centimeter-scale distances. Both long-range and short-range connections are integral to the brain's architecture: the algorithmic function of a piece of local cortical circuitry likely cannot be understood in isolation from its long-range cortico-cortical, cortico-striatal and cortico-thalamic inputs and outputs. In the theoretical framework of this proposal, long-range connections between areas are essential to a) connect a cortex-based slow learner with a basal ganglia based fast learner and b) implement Bayesian inference using top-down priors. *For our framework, and for most other interesting theories of brain architecture, the structures of the "local" networks do not make sense without understanding the detailed structures of the long-range connections.*

A key challenge for connectomics, therefore, is to enable the scalable, accurate, long-range tracing of axons in order to reconstruct *entire* circuits. Axons shrink in diameter down to tens of nanometers, yet often travel several millimeters along complex paths, with kilometers of axonal wiring present in a cubic millimeter of cortex. This makes accurate long-range axon tracing extremely challenging. The electron microscopy approach to connectomics views axon tracing as an image analysis problem: the axonal membrane is reconstructed through tens of thousands of thin cross-sections to identify its path. A crucial problem is the fragility of such analysis: each error affecting an axon can cause disproportionate damage to the reconstruction, by mis-labeling

each of the hundreds of downstream synapses in the connectivity matrix. If an error in an axonal trace occurs on average even once per the length of one axon, which is several millimeters in mouse brain, then 50% of all connections in the connectivity matrix will be incorrect; not even one ultrathin-section can be damaged or imperfectly stained, lest the entire reconstruction be lost. To be more rigorous, for an error model involving only two axons, the connection error rate would be 56.8% via sum of the even terms of the Poisson. If the model involves 3 or more axons then the connection error rate is roughly 38%. For an axon 5 mm long, and EM sections 50 nm thick, the accuracy per single axon section must be better than 99.999% in order to have a 36% chance of assigning a correct connection.

Our approach promises to define the future of connectomics for several reasons: (1) It is ultimately more scalable and cost-effective than the electron microscopic approach; (2) It enables the incorporation of multiple redundant forms of information that mutually error-correct one another, to allow accurate determination and cross-checking of connectomes; (3) It opens the path to *molecularly annotated* connectomes, which incorporate not only information about connectivity and morphology but also about the molecular identities of neurons and synapses, which are likely integral to a true understanding of the diverse learning rules and communication mechanisms used by the brain; (4) It can be readily validated on a small scale by the traditional methods of EM connectomics; (5) It is easy to register with functional imaging (e.g., 2P microscopy of calcium indicators). Specifically, we eschew electron microscopic image analysis and take a radically different and complementary approach, viewing long-range axon tracing not primarily as an image analysis problem but as a problem of genetic labeling and optical readout. By labeling each neuron with a unique nucleic acid barcode (which can be thought of as a "molecular ID card") which fills the axon and is transported to synapses, then any appearance of that barcode anywhere within the circuit identifies the corresponding neuron – the manual tracing of nanoscopic wires through tens of thousands of dense, greyscale image stacks is no longer required.

From bulk barcoding to in-situ barcoding.

A DNA barcode is a unique sequence of DNA used to "tag" an object of interest. The number of possible DNA barcodes grows exponentially with the DNA sequence length: there are 4^N possible sequences of length N. For example, a barcode of length 30 could has a potential diversity of $4^{30} = 10^{18}$, a number which vastly outstrips the number of neurons in the mouse cortex ($\sim 10^7$). Thus with sufficient diversity, each neuron will almost certainly express a unique tag even when barcodes are expressed at random.

Zador originally suggested (Zador et al., 2012) an approach to connectomics, called Barcoding of Individual Neuronal Connections (BOINC), which leverages large numbers of DNA barcodes. First, each neuron is given a unique DNA barcode. Copies of each neuron's barcode are exchanged with its immediate synaptic neighbors. A cell's own barcodes are then stitched together with barcodes received from its synaptic neighbors, forming a set of barcode pairs corresponding to synaptically connected neurons. The barcode-pair DNA strings are extracted, pooled, amplified (i.e., creating many identical copies of each barcode pair) and sequenced on a bulk DNA sequencing machine, such as an Illumina HiSeq. This results in digital data specifying a set of "on" matrix elements, corresponding to barcode pairs (synaptic neighbors) which are observed, and a set of "off" matrix elements, corresponding to barcode pairs which are not observed (e.g., due to the absence of a synapse between the corresponding two neurons). Because of dropping high-throughput sequencing costs, BOINC has the advantage of ultra-low cost (potentially on the order of $10k for the entire mouse cortex, see (Marblestone et al., 2013) for estimates, but in its original

form it also has several limitations. It does not include detailed spatial or morphological information, although, by dividing the brain into mesoscale cubes and appending additional barcodes to the DNA derived from each cube, it is possible to obtain coarse-grained spatial information. Below we describe how this limitation can be overcome by using in-situ sequencing (FISSEQ) on an optical microscope for barcode readout, rather than sequencing in a bulk sequencer.

Fluorescent in situ sequencing (FISSEQ).

As a solution to these problems, we have combined BOINC with in-situ multiplexed readout by fluorescent microscopy. The basic idea is to combine cellular barcoding with an in-situ optical readout, so that the barcode pairs can be read out optically without disrupting tissue structure (**Figure 4**).

Taking inspiration from BrainBow methods (Cai, Cohen, Luo, Lichtman, & Sanes, 2013; Livet et al., 2007), which give each neuron a unique color in the fluorescent microscope – an analog summation of three fluorescent protein expression levels – one can conceptualize our approach as a "4^N-color digital BrainBow". While fluorescence microscopy only has access to ~4-8 distinct color channels per image, our solution here is to encode cell identity into a *sequence* of colors which is revealed over time, in a sequence of biochemical reaction cycles. The linear sequence of an RNA molecule then determines the sequence of colors that appears in the microscope during successive cycles: thus the RNA barcode "blinks out its code" over time. With N cycles and 4 colors per cycle, we have 4^N distinguishable barcodes.

We will apply FISSEQ to read out neuronal connectivity. With FISSEQ, we sequence DNA or RNA molecules via fluorescent microscopy, in the context of intact, fixed tissue slices. The development of FISSEQ began in the Church laboratory over a decade ago (Mitra *et al.,* 2003). *The underlying principles are similar to those used by the Church laboratory and others to usher in the next-generation sequencing revolution (Church and Kieffer-Higgins 1988; Bentley et al., 2008; Peters et al., 2012).* In both bulk high-throughput sequencing and FISSEQ, short sequences are locally amplified on a substrate and then imaged one nucleotide at a time. However, the requirement to sequence RNA in intact tissue—rather than isolated and purified DNA, as in conventional bulk sequencing—posed additional challenges. These limitations have now been overcome (Lee *et al.,* 2014), and *FISSEQ is poised to transform many fields of biology by allowing the joint, high-throughput readout of sequence and spatial information*.

In FISSEQ, a series of biochemical processing steps, such as DNA ligations or single-base DNA polymerase extensions, are performed on a block of fixed tissue, interlaced with fluorescent imaging steps. The process is conceptually identical to the mechanism of fluorescent sequencing by synthesis in a commercial bulk DNA sequencing machine, except that it is performed in fixed tissue (we have typically used sequencing by ligation rather than sequencing by synthesis on our published work due to the wide availability of the reagents, but this is a technical issue of no importance here). Each DNA or RNA molecule in the sample is first "amplified" (i.e., copied) in-situ via rolling-circle amplification to create a localized "rolling circle colony" (rolony) consisting of identical copies of the parent molecule. A series of biochemical steps is then carried out. In the kth cycle, a fluorescent tag is introduced, the color of which corresponds to the identity of the kth base along the rolony's parent DNA strand. The system is then "paused" in this state for imaging. The entire sample can be imaged in each cycle. The fluorescent tags are then cleaved and washed away, and the next cycle is initiated. Each rolony – corresponding to a single "parent" DNA or

RNA molecule in the tissue – thus appears, across a series of fluorescent images, as a localized "spot" with a sequence of colors corresponding to the nucleotide sequence of the parent molecule. The nucleotide sequence of each DNA or RNA molecule is thus read out in-situ via fluorescent microscopy. FISSEQ is an example of an extremely dense form of in-situ nucleic acid readout: every letter along the RNA chain is read. Thus, barcodes for FISSEQ can be packed into a short string of DNA, as shorts as 15-20 nucleotides long for the mouse brain. As we will see below, this leads to a simple and powerful strategy for viral barcoding, in which random short strings (oligonucleotides) of DNA encode the barcodes.

G. Using in-situ multiplexing to map connectivity.

To read out synaptic connectivity in our approach, cell-identifying RNA barcodes are targeted to the pre-synaptic and postsynaptic membranes, and in-situ multiplexing is used to optically resolve and sequence the pre-synaptic and post-synaptic barcodes at a large fraction of synapses, thereby identifying connected pairs of cells in-situ.

Presynaptic, postsynaptic and proximity-ligation tags.

The key problems which must be solved are: 1) Detecting synapses, 2) Ruling out close appositions of cell membranes which are not synapses, 3) Attributing synapses to the correct pre-synaptic and post-synaptic neurons. In one strategy, barcodes localized to the fine neuronal processes adjacent to a synaptic connection are read separately, and are co-registered with a synaptic marker to both localize the synapse and identify the pre-synaptic and post-synaptic cells. Co-localization with a synaptic marker is important here, because a key challenge is to distinguish actual synapses from mere close appositions of neuronal processes. One trivial method for synapse tagging builds on the finding that synapsin is a fairly reliable marker of synapses, present at about 90% of them [2,3]. In the simplest method, then, we simply combine FISSEQ readout of the barcodes with conventional antibody staining of a synaptic marker. A more powerful method provides a nanoscopically precise marker of the location of the synaptic cleft by using the proximity ligation assay (PLA) (Söderberg et al., 2008) to detect the close co-localization of pre-synaptic and post-synaptic proteins (neurexin and neuroligin); this approach is shown in **Figure 5**. PLA can be used to detect protein-protein interactions, with single molecule resolution, without the requirement for modification of either protein target (termed **PLA-tagged**). PLA is as straightforward to perform as conventional immunostaining.

PLA offers two advantages over conventional dual-antibody stains. *First, it is more specific*. Since two binding events are required to form a signal, false positives occur only if there are two spurious binding events to two nearby off-target epitopes. Such "double errors" are quadratically rare; if the probability of a single off-target binding event is 5%, then to a first approximation the probability of a double event will be 0.25%. Furthermore, with PLA the signal is formed only when the two proteins are <40 nm from one another, a distance well below the limit of resolution of optical microscopy (200 nm) and only twice the size of the synaptic cleft (20 nm). With double antibody labeling, by contrast, any two overlapping spots are interpreted as representing co-localization, potentially increasing the rate of false positives. Second, positive signals—rolonies—formed in PLA are very bright because they consist of thousands of fluorophore-conjugated oligonucleotides. These rolonies are thus easily distinguished from background fluorescence, reducing the rate of false negatives (**Error! Reference source not found.B**).

This synapse-tagging strategy generalizes to allow tagging of many distinct structures – e.g., gap junctions, or particular *types* of chemical synapses – when combined with in-situ multiplexing:

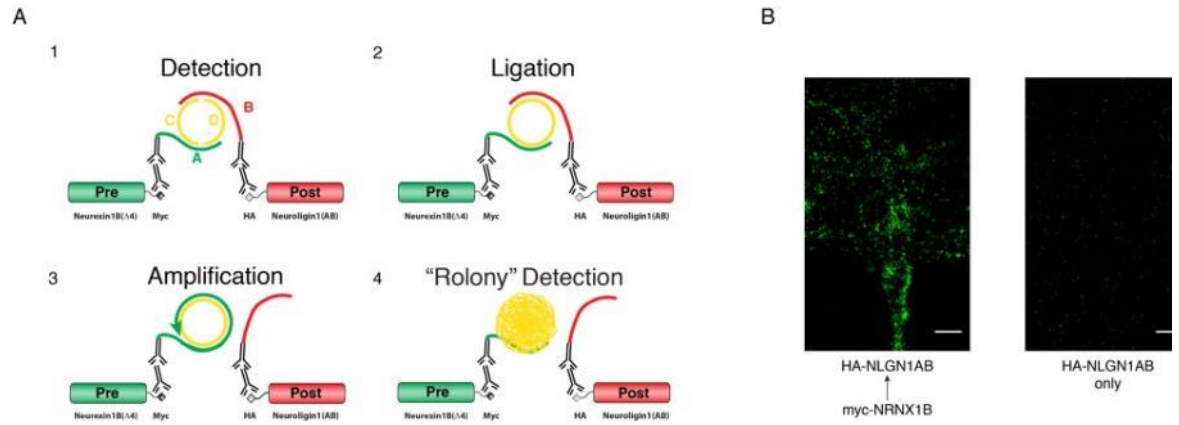

Figure 2 PLA for high-resolution detection of synapses in the context of FISSEQ (A1). Tissue is first fixed and permeabalized, then incubated with primary antibodies to (presynaptic) neurexin1B and (postsynaptic) neuroligin, after which secondary antibodies conjugated to oligos A and B are added. The primary antibodies must be derived from different hosts (e.g. goat and rabbit), to permit selective binding by the secondaries. Additional bridge oligos C&D are added, and form a circle if and only if the two proteins are within 40 nm. (A2) The bridge is ligated. (A3) Rolling circle amplification results in the formation of a "rolony" containing many copies of the template. (A4) The rolony is probed with a fluorescent oligo, resulting in a very bright signal. (B) Demonstration of synapse detection using PLA. Left, synapses detected by PLA between modified HA-NLGLN1AB expressed postsynaptically and MYC-NRXN1B expressed postsynaptically. Right, negative control demonstrating specificity of synapse detection.

for example, one could tag for synapsin (indicative of the presence of a synapse), acetylcholinesterase (indicative of the presence of a *cholinergic* synapse), connexin (indicative of the presence of a gap junction) or many other functionally decisive molecules, reading out all of them in parallel using in-situ multiplexing of short DNA oligonucleotides delivered via antibodies. Again, here we are not limited to only 4-8 colors of antibody staining, but rather have a multiplexing space of 4^N possible tags. Briefly, for FISSEQ readout, oligo-tagged antibodies to synapsin are introduced and fixed in place. Circularized oligonucleotides can be annealed to the antibody-conjugated oligo. This oligo will then serve as a primer for rolling circle amplification, creating a synapse-indicating rolony, which can be sequenced by FISSEQ. This technique can be extended to other synaptic proteins, for instance, to discriminate inhibitory versus excitatory synapses.

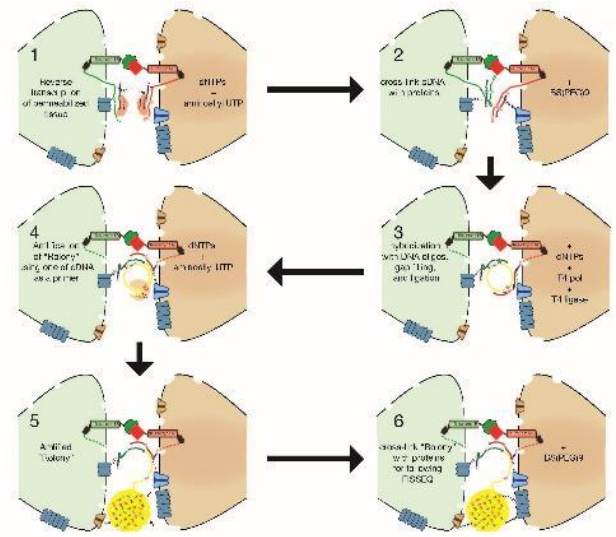

Figure 3 Barcode-entangeld PLA. Both the location of a synapse, and the identities of the pre- and post-synaptic cells, are encoded by a single FISSEQ rolony.

In a second strategy (**Figure 6**), pre-synaptic and post-synaptic barcodes will be joined across the synaptic cleft into a single barcode, which is then locally amplified **(termed PLA-linked)**. This is effectively a form of PLA in which the neuronal cell-identifying barcodes are themselves

used as the PLA partners, without the need for exogenous antibody staining. To achieve this, we attach the RNAs barcodes (initially present inside the pre-synaptic and post-synaptic compartments) via linkers which are sufficiently long to allow the barcodes to move into the synaptic cleft itself following tissue fixation and membrane permeabilization. At that point, the RNA barcodes are reverse transcribed to cDNA, and hybridized to universal probes. Primer extension leads to a gap-filling reaction that encodes **both** pre-synaptic and post-synaptic barcode sequences into a single rolony, which is then amplified. Appropriate sequence design allows direct determination of which barcode is pre-synaptic and which is post-synaptic, although we should not that this is also evident from the barcode sequence present in the nearby dendrite. This has the advantage of localizing and identifying synapses in an extremely precise manner, since all information needed to determine both the existence of the synapse and the identities of the pre-synaptic and post-synaptic cells are encoded in a single FISSEQ-compatible rolony.

At the beginning of Phase 1, an un-optimized PLA-tagged virus will be given to the neurophysiological team; in addition, we will start optimizing the FISSEQ PLA protocol investigating the efficacy of both PLA-tagged and PLA-Linked viruses. Criteria will include number of rolonies per cell and number of rolonies per volume detected in one sequencing cycle, and localization error of pre-synaptic and post-synaptic synapses. We will choose one of the viruses by Milestone 1.A Month 12 to deliver to the neurophysiological team.

Thus, the FISSEQ approach to connectomics has several key advantages. First, it allows cell positions and molecular profiles to be read out from the tissue in addition to the connectivity. Second, it can be used *in combination* with morphological reconstruction of some or all neurons (depending on the nature of the labeling), as well as the optical identification of synapses (e.g., with an anti-synapsin-1 antibody): these morphological readouts not only add information to the raw connectivity matrix – they can also be used as an independent validation of the connectivity pattern, i.e., a form of error correction. We have already demonstrated the integration of FISSEQ with other modalities of measurement such as protein tagging and membrane labeling.

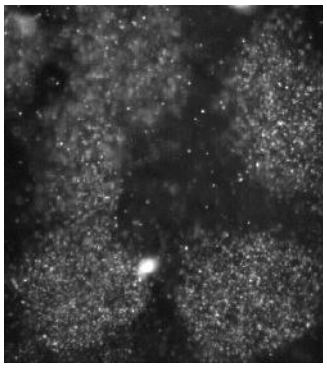

Figure 4 Integration of FISSEQ with ExM (EXSEQ), demonstrated on endogenous transcripts in cultured mammalian cells. RNA was captured into the hydrogel prior to 5x expansion, and then rolonies were generated following expansion. Each dot represents an EXSEQ rolony that has been fluorescently probed. The yield of rolonies appears to be strongly enhanced, compared to unexpanded FISSEQ, perhaps due to the ease of rolony formation in a low-density, water-like environment rather than inside heavily cross-linked fixed tissue.

Ultra-high-resolution in-situ multiplexing with expansion microscopy (ExM).

During FISSEQ-PLA protocol optimization, we will investigate the use of the most scalable form of optical super-resolution microscopy yet developed, Expansion Microscopy (ExM) as an alternative to straight tissue processing. Rather than using lenses and mirrors to create optical magnification in a microscope, we recently found (Chen et al., 2015) that physical magnification of the specimen itself is possible. Polymerizing electrolyte monomers directly within a sample to form an electrically charged polymer network, followed by solvent exchange, results in specimen expansion. By covalently anchoring specific molecules within the specimen to this polymer network and proteolytically digesting unwanted endogenous biological structure, we found that samples could be

expanded isotropically ~4.5-fold in linear dimension. We discovered that this isotropic expansion applies to nanoscale structures, and thus this method, which we call Expansion Microscopy (ExM), can effectively separate molecules located within a diffraction limited volume, to distances great enough to be resolved with conventional microscopes. Indeed, because such conventional microscopes can be inexpensively scaled up to very high speeds of imaging, without incurring the same kinds of hardware expenses associated with other super-resolution technologies, nor requiring custom chemicals, ExM not only represents a fundamentally new modality of magnification, but also enables scalable, multi-color super-resolution imaging of fixed cells and tissues. We have already demonstrated that this method applies to large volumes of intact brain tissue, revealing the nanoscale organization of synapses and dendritic spines across macroscopic volumes in the mouse brain (Chen et al, 2015). ExM can not only readily resolve synapses from their nearest neighbors, but also can resolve pre-synaptic proteins from post-synaptic proteins, as demonstrated by co-staining for the proteins Bassoon and Homer1.

To be clear, integration with ExM *is not necessary* to observe dense connectivity in our approach: the resolution needed to attribute tagged synapses to their pre-synaptic and post-synaptic barcode identities –for long-range as well as short-range connections alike – is much lower than that needed to resolve the fine structure of the *individual* synapse, as allowed by ExM. For comparison with other synaptic imaging methods used for deep characterization of individual synapses, note that, even in the absence of powerful super-resolution methods like ExM, it is possible to densely resolved every synaptic contact in the optical microscope. Typically, this is done at a spatial resolution of ~250 nm x ~250 nm x ~ 70 nm, using the high lateral resolution of a confocal microscope coupled with < 70 nm thin sections via Array Tomography (Micheva & Smith, 2007), which was specifically developed as a means to resolve individual synapses. For comparison, of the already-published version of ExM, demonstrated in the figures above, achieves <70 nm x <70 nm x <200 nm resolution, beating the resolution of Array Tomography along all three axes, yet without requiring physical thin sectioning of the tissue to do so. Thus, in combination with our in-situ multiplexing approach to reading synaptic connectivity, which does not itself rely on ExM, ExM can enable a truly comprehensive approach to molecular and structural characterization of dense neural circuitry, while automatically incorporating the long-range connections without the need for axon tracing. We have already integrated ExM with FISSEQ **(Figure 7).** Indeed, the integration of these methods is expected to broadly transform biology because it enables super-resolution localization and identification of thousands of distinct biomolecules, over large volumes of intact tissue. ExM increases the efficiency of FISSEQ rolony formation by removing steric hindrance of proteins and other molecules. The increase in efficiency reduces the number of low-quality non-sequencable samples, reducing wasted experimenter and device time. In addition, due to the optical clearing effect, background and non-tagged fluorescence is nearly removed, which decreases the burden on co-registration algorithms. In the FISSEQ-PLA protocol optimization we will investigate 2x super-resolution ExM achieving a resolution of ~125 nm x ~125 nm x ~250 nm, to provide higher detail and reduced error rates for the neural circuit reconstructions.

Barcode expression.

The core of our approach is to use viruses to barcode a population of neurons with random short sequences of RNA. The barcode, which is read out by FISSEQ, acts as a unique identifier for each neuron. To generate a library of barcoded viruses, we simply clone a commercially available N-mer random oligonucleotide into the Sindbis virus genome. We then infect a population of neurons in auditory cortex with this diverse viral barcode library via simple injection. A clear demonstration of this method is provided in Error! Reference source not found., wherein the axons of cultured neurons were separated from the cell bodies in a microfluidic device.

A barcode consisting of 30 random nucleotides has a theoretical sequence diversity of $4^{30}=10^{18}$, far more than the number of neurons infected in a typical experiment. With high probability the barcode in each neuron is unique, distinguishing that neuron from all other neurons. The barcodes fill the somatic and dendritic compartments, allowing us to determine laminar position, neuronal morphology, and local connectivity. The barcode is also engineered to bind to a modified presynaptic protein, which can be coexpressed to transport barcode to distal projections. In this way, in contrast to all existing approaches, *we can identify distant targets of many neurons in a single brain without tracing their entire axons*.

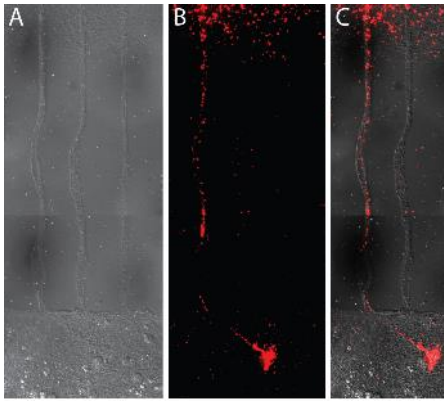

Figure 5 FISSEQ of viral barcodes can be used to trace the morphologies of neurons. (A) A neural culture in which axons are separated from cell bodies was sparsely infected with a virus encoding a barcode and presynaptic protein engineered to bind to the barcode. (B) Barcodes were amplified into rolonies and probed with a red fluorescent oligo. (C) Overlay of A and B. Expression of a postsynaptic protein engineered to bind the barcode allows reconstruction of dendrites. We expect that by optimizing the protein constructs and the imaging, we will be able to achieve even better reconstructions of neuronal processes.

To achieve high coverage of a circuit via virus infection, we infect at high viral titers. Under these conditions, some neurons are infected with multiple barcode-bearing viruses. Multiple infections per neuron are not a problem with FISSEQ, because we can read out each of the multiple barcodes from each neuron.

To enable the technique to ultimately scale to the whole brain level, we will move away from viral infection and develop inducible, transgenic barcodes that will then be present natively in every cell of the animal. We are already pursuing a strategy for barcoding every cell of **a transgenic mouse (named NHEJ).** The idea is very simple: to leverage the naturally error-prone nature of non-homologous end joining after a DNA break in the genome. To make the process non-toxic and targeted to an expression locus, we use the programmable DNA binding CRISPR protein Cas9, targeted via a guide RNA strand, to a specific genomic site. We will express the construct transiently in the adult to barcode each cell uniquely and safely. This strategy has the benefit of leveraging the huge investments now being made in using CRISPR for gene therapy, and indeed the design is almost identical to constructs used in gene therapy (Guilinger, Thompson, & Liu, 2014). We will begin development of the NHEJ transgenic mouse line in Phase 1 and deliver it to JAX by Waypoint 2.2 Month 27, so that they will be available for neurophysiology.

Data acquisition pipeline.

Existing automated FISSEQ instruments will enable large-scale, automated in-situ multiplexing for connectomic data collection. While the original sequencing and imaging was done manually, we already have advanced automation robotics generating FISSEQ data. We have accomplished this automation quickly due to our experience developing the Polonator platform (the first open source high-throughput sequencing automation system), and prior relationships with independent engineering and manufacturing firms (e.g. Dover Motion, a Danaher subsidiary manufacturing the sequencing robot for purchase). No *micro*-fluidics are required. The system is based on standard microscopy and macro-fluidic (not micro-fluidic) hardware and therefore the cost is comparable to commercial microscopes. Crucially, this system is manufactured by GT-Bioseq, LLC.

Imaging throughputs.

The throughputs of barcode-based connectomics are vastly higher than those of other approaches. Spinning disk confocal microscopy can process ~100k 4-color voxels per second. In each experiment, we will need to perform ~15 cycles of 4-color FISSEQ. This puts the overall imaging voxel throughput of FISSEQ in the range of ~5k voxels per second. Given this raw imaging throughput, and a ~250 nm x 250 nm x 60 nm resolution voxel size set by the optics of the confocal scan head, the 100 um x 500 um x 500 um Phase 1 neuroanatomical target in one mouse could be acquired in a matter of days on a single instrument, and the Phase 2 target of $(1000\ um)^3$ in ~3 months for one mouse. ExM throughputs are expected to be ~125x slower on a volumetric basis using current instrumentation, since the voxels are ~125x smaller.

Validation by EM.

It is straightforward to validate FISSEQ-based connectomics via EM. Rolonies formed by PLA cannot natively be visualized using EM. However, we have developed a method of visualizing the rolonies in EM by incorporating modified nucleotides (replacing 'T' with iodine-UTP) containing heavy-metal. Preliminary data (**Figure 9)** demonstrate that rolonies labeled in this way can incorporate heavy-metal labeled nucleotide analogs, that the subsequently formed rolonies can be visualized with EM, and that the rolonies are often found near synapses. To obtain quantitative estimates will require serial electron microscopy, but based on our preliminary data (several dozen PLA rolonies, no false positives detected), we predict the false positive rate to be very low. In this way we will validate both the localization of the synapse and the neural circuit reconstruction.

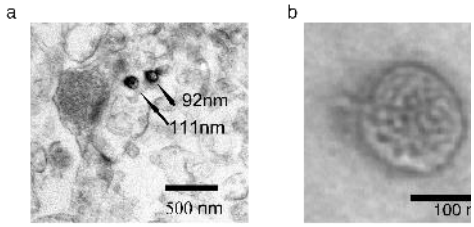

Figure 6 PLA signal can be directly visualized by electron microscopy. Cultured neurons expressed myc-NRXN1B on the presynaptic side and HA-NLGN1AB on the postsynaptic side. PLA was performed with a modified protocol so as to introduce 5-iodo-dUTP into the PLA 'rolony' during rolling circle amplification. The resulting heavy labeled balls of DNA are detectable proximal to synapses under TEM (a, detail of ball in b).

Technical Area 3: Neural circuit reconstruction

The algorithms needed to reconstruct a circuit from barcoded FISSEQ and PLA data are much less challenging than from conventional EM data for two reasons. First, synapses are brightly marked by PLA, and do not require sophisticated image processing to detect as they do in EM data. Second, because each neuronal process is filled with an identifying neuronal barcode, it is not necessary to trace each process with near perfect accuracy; indeed, even distant processes, millimeters away, can be correctly assigned to the soma of origin. Finally, we note that registration of the neuroanatomical data to the calcium-imaging data is relatively straightforward because both are optical measurements.

H. Co-registration of Functional and Structural Data

Extracting spatial (and temporal) information from calcium imaging data

Extracting information from large spatiotemporal calcium imaging movies has been challenging in the past, particularly in cases where many neuronal processes overlap spatially. Current approaches are typically based on off-the-shelf methods such as principal or independent components analysis (PCA or ICA); these approaches are based on an unstructured signal model and therefore fail to exploit the considerable prior knowledge we have about the underlying neural and calcium signals. Thus their performance is significantly suboptimal (these methods often fail to adequately demix spatially-overlapping data, for example), as we have shown quantitatively in our preliminary work [1].

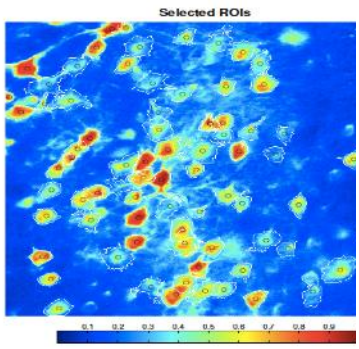

Figure 7 Automatic detection of neurons in two-photon calcium imaging movies. Regions of interest (ROIs) representing neurons are automatically extracted from data collected in the Yuste lab (layer 2/3 mouse visual cortex expressing GCaMP6s, imaged in vivo). Colorbar indicates the correlation of each pixel with its nearest neighbors. The white contours show the boundary of each inferred cell shape; black circles denote the corresponding centers of mass. Note that the method accurately and reliably extracts high-quality ROIs. The supplementary video provides a more complete depiction of the results, especially the dramatic denoising that the method provides: http://www.stat.columbia.edu/~eftychios/movies/Kira-vid2.mp4; see[1].

We have recently developed a new approach based on convex optimization and structured nonnegative matrix factorization. This approach relies on the observation that the spatiotemporal fluorescence activity can be expressed as a product of two matrices: a spatial matrix that encodes the location and shape of each neuron in the optical field and a temporal matrix that characterizes the calcium concentration of each neuron over time. This approach exploits the sparsity of both the underlying neural activity and the spatial shape of neurons. We have successfully applied this approach to

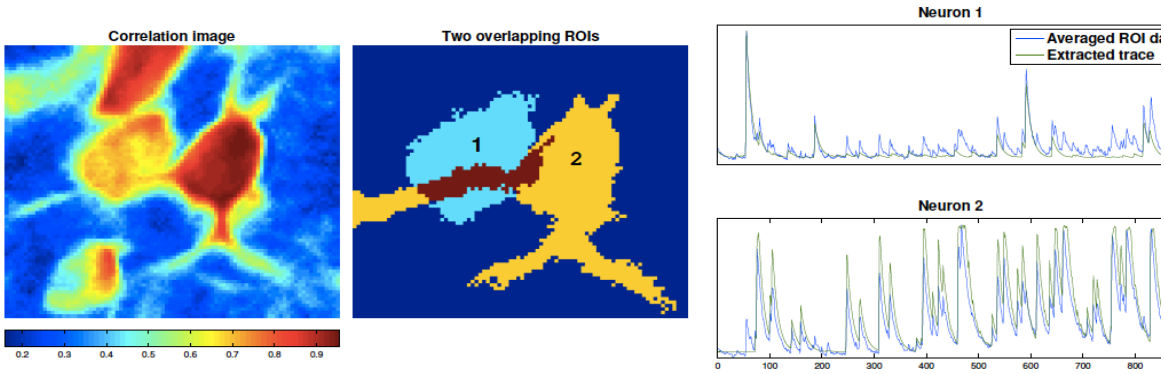

Figure 8 Importance of the demixing procedure. Left: Summary spatial correlation image zoomed into two overlapping neurons from the dataset analyzed in Fig X. The exact spatial masks cannot be readily resolved. Middle: Spatial masks for two identified overlapping ROIs after demixing (the brown region corresponds to the overlapping pixels). Right: Inferred calcium traces of the two neurons obtained by averaging over the spatial masks (blue traces) and by applying the proposed method (green traces). Simple averaging over the ROI of neuron 1 can be misleading if demixing is not performed.

a wide variety of calcium imaging data (both in vitro and in vivo in a number of different species and brain regions), achieving state of the art results in our preliminary work. See **Figure 10** and **Figure 11** for illustrative results.

Deconvolving rolony location and sequence estimates in FISSEQ imaging data

Our goal is to estimate the locations of the rolonies and the RNA sequences in each rolony. This can be cast as a structured deconvolution problem, similar in spirit to the calcium deconvolution problem but different in detail. Our method exploits the spatial sparsity of rolony locations (*i.e.*, that most pixels are not rolony centers), a simple linear convolutional model mapping the rolony signals into the observed blurred, noisy image data, and the non-negativity of the florescence signal (non-negative group-LASSO penalized optimization, [Beck2009]). Once the rolony locations and corresponding RNA sequences are obtained, we use standard sequence-comparison methods to determine the corresponding barcode for each rolony, and then "color" the rolonies by their barcodes to outline the shape of each neuron. We have tested the algorithm on the FISSEQ human dataset published in [4]; the image data in this case was of size 400 × 400 × 30 × 4 (width × height in number of pixels × sequence length × number of nucleotides). Our results reveal that the algorithm is easily parallelizable to scale up to larger volumes.

Co-registering FISSEQ and calcium imaging data volumes

The algorithms described briefly above will provide the shapes of the neurons observed in the neural volume, imaged via calcium fluorescence and then FISSEQ. To co-register these two data volumes, we will utilize a coarse-to-fine approach, using coarse anatomical landmarks, then cell bodies, then subcellular features to co-register the volumes. At each scale, we will initialize with linear registrations (initializing from the linearized registration inherited from the next-coarsest scale, where available) and then use penalized spline-based registration methods to handle any small deviations from linearity in the mappings between the two volumes.

At the coarsest scale we will register micro-anatomical landmarks, using a method inspired by Ko et al (2011 by marking recognizable landmarks (e.g., blood vessel bifurcations, fluorescent bead injections, and/or brightly labeled astrocytes) in both volumes; then use standard point-cloud registration methods to obtain a map from one volume to the other by solving a penalized least square problem for optimal registration. Ko et al successfully used linear landmark-based registration on spatial scales of 100-200 microns; we expect this approach to work well on similar

scales, though potentially nonlinear (spline-based) registration may be necessary at larger scales.

At the next finest scale we will register the locations of the somas located in each volume. If all the somas are visible in both volumes, this is again a straightforward penalized least squares problem. However, some cell bodies may be obscured in one or the other volume ([5] were unable to register 10 of 126 of their neurons). Thus we will use a more robust approach by optimize a similar penalized least-squares matching criterion as described above, but now simultaneously optimizing over the matchings between the collection of somas in each volume. A coordinate-descent approach will be used: holding the matching fixed while solving a penalized least-square problem to optimize over the linear or spline registration coefficients; then holding the registration mapping coefficients fixed and optimize (locally) over the matching.

Finally, at the finest scale we will register not just the cell body locations but also subcellular features such as soma shape and large dendritic branch points. This step will proceed as in the step described above, using the soma-based registration as an initialization. As an alternative, we will use automated image registration techniques based on maximization of pixel-based correlation between images without relying on computing landmarks – although these are not as computationally efficient as landmark-based methods.

I. Neural Circuit Reconstruction

To develop a set of computational algorithms for reconstructing connectivity.

We will design and apply a set of algorithms that can convert FISSEQ and PLA data into a neuronal connectivity matrix. The raw data in Phase I are in the form of a Z-stack of two-dimensional images, obtained from FISSEQ imaging, in which the barcode sequence associated with each rolony has already been determined using the algorithms outlined above. In addition, PLA spots, marking synaptic connections, will also be marked. Note that, because we will use targeted reverse transcription to form rolonies from the barcode, almost all the rolonies in a given neuron will have exactly the same sequence (modulo sequencing error; see below). This significantly simplifies the reconstruction problem. In subsequent phases the datasets will differ in two ways. First, in Phase 2 we will deploy the transgenic PLA mouse NHEJ, in which pre- and postsynaptic barcodes are entangled in a single rolony and read out. This further simplifies the task (described in more detail below) of ascribing a particular PLA dot to its pre- and postsynaptic partners. Thus the procedures we outline below provide the framework for most of the analyses required.

The set of algorithms that we will implement and apply can be broadly separated into three components: Preprocessing, connectivity reconstruction, and post-processing. Preprocessing encompasses various error correcting steps and other procedures that prepare the dataset for determination of the connectivity matrix. In the connectivity reconstruction step, we determine the connectivity matrix itself. Finally, the post-processing step includes building the database searchable by the rest of the team, and integrating data with the gene expressing profiles. Below, we describe these procedures in more detail.

Preprocessing. In the preprocessing step, we correct for errors that arise in determination of the sequence of each rolony. In FISSEQ, as in other forms of high-throughput DNA sequencing, several forms of error can arise. The most common forms include single nucleotide insertions, substitutions and deletions. There are a variety of standard algorithms in bioinformatics for correcting for such errors, usually by comparing a given sequence to a reference genome. This form of error correction can be seen as solving a generative model in which the observed sequence is derived from an underlying sequence by means of one or more generative processes that produce insertions, deletions and substitutions. In our specific problem conditions are particularly favorable

for such error-correcting algorithms because we can exploit a very prior that a given rolony has the same sequence as its neighbors. Only in cases where rolonies are from two nearby neurons, expressing different barcode sequences, will the sequences of nearby rolonies differ; such cases will be easily recognized because in those cases the typical Hamming difference between the rolonies will be large. Thus sequence correction will be relatively straightforward. The end result of this preprocessing step will be a data set with the same dimensions as the original data set in which the sequence of each rolony has been corrected.

Reconstruction. To reconstruct connectivity from this corrected data set, we must determine the pre- and postsynaptic barcodes associated with each synapse marked by PLA. In Phase I, this essentially amounts to determining the nearest-neighbor barcode pair, in three dimensions, for each PLA-marked synapse—a relatively straightforward computation. In Phase II, this requires applying the error-correcting methods described in the pre-processing step to correct errors in the transgenic mouse barcodes. The results of this step is a matrix describing the connectivity of neurons *1…N,* along with their position in 3D space and their registration to the calcium imaging data.

Post-processing. We will design a set of tools for displaying and analyzing connectivity. For example, on the basis of connectivity matrix, we will place network nodes into 3D space using an algorithm implementing force-directed graph drawing. We will compile a set of statistics that describe the connectivity matrix and make them available to the research team. We will link connection data to 3D position data, as described above, and to gene expression data which will allow identify cell types. Finally, we will make a searchable online database that allows other members of the team to have access to data for validation and hypothesis testing.

J. Neural Data Storage

Construct a Spatial Big Data (SBD) database for spatially resolved data

Spatially resolved data is massive in size and complexity. In this task we will extend and develop on the MICrONS GFE AWS API as necessary for storing and rapidly querying spatially resolved data including synaptic transcripts, c-registered functional data, experimental annotations, structural data, and neural circuit reconstructions. Given the size of each individual image, keeping all raw images for every FISSEQ base-pair for each sample would be prohibitive. We will therefore explore options for storing spatially-resolved transcriptome sequencing data using extensions of the FASTA and SAM/BAM formats for the sequencing data and a limited number of stained images for the morphological and biomarker data.

Spatial Big Data (SBD) Database Infrastructure for Spatially Resolved Data

Recent increases in the size of large-scale spatially resolved datasets (e.g. the massive datasets generated from Twitter, Facebook, and Google search activity) have led to significant recent developments in technology for storing, querying, and analyzing Spatial Big Data (SBD) [6]. We will utilize the MICrONS GFE AWS to store our data. If required, we will connect the GFE AWS through API to a database based on Hadoop-GIS also hosted in the GFS AWS, which has recently been shown to outperform Spatial Database Management System (SDMS) for compute-intensive queries[7] on large-scale spatial datasets. The Hadoop-GIS system parallelizes spatial queries and maps the queries onto MapReduce. The framework is integrated with HIVE [8] and supports both simple and complex spatial queries. Specifically, the cyber-infrastructure will consist of Hadoop and the Hadoop distributed file system (HDFS), will use GIS on Hadoop and Spatial Hadoop, and will enable storage and computation on spatially resolved transcriptome sequencing, morphological features, and cell and tissue morphology data. The Spatial Big Data database structure enables efficient computation of spatial relationships such as containment, neighbor

relations, and 3D distance, minimizing the data-footprint of pre-processed spatial relationships and providing additional flexibility to research applications. Data features include the development of data and metadata standards including the Digital Imaging and Communications in Medicine (DICOM) Working Group 26; the DICOM Structured Report standard[9], and the Annotation and Image Markup (AIM)[10].

Support of the use case specified in Neural Data Access Hadoop-GIS provides a robust and flexible API to allow database access Java, Python, and the Esri Geometry API (similar to SQL). This flexibility will allow us to open the database directly to users that a program savey, and facility the construction of the web interface described in Neural Data Access.

Risks and alternatives strategies If we have difficulty creating the Spatial Big Data database, we will use Bisque to manage our data. Bisque: Bio-Image Semantic Query User Environment [11] has an extensible data management system, and is suitable substrate for both development of the web-based interface as well as web-service tools, and we have already begun a productive collaboration with the director of Bisque (see Letter of Collaboration from Professor B.S. Manjunath). Bisque's web interface is utilized in our Neural Data Access plan.

K. Neural Data Access

Collaborator-available web-based application for processing and sharing data

We will utilize the MICrONS GFE AWS API with the database and service to allow rapid batch-mode querying of the image processing and visualization database via an extension to the Bio-Image Semantic Query User Environment (Bisque) architecture [11]. Bisque is a suitable substrate for both development of the web-based interface as well as web-service tools, and we have already begun a productive collaboration with the director of Bisque, Professor B.S. Manjunath. Bisque allows the visualization of images and overlays, and includes the ability to subselect regions, thus it contains all the necessary building blocks to meet the visualization criteria of BAA 1.B.3.c. We will extend the interface to support downloading of subsections and display neurophysiological data. Bisque is open-source and actively maintained and has an extensible and highly modular design using web-standard communication formats and contains an array of built-in analysis tools, such as CellProfiler[12].

Bisque contains Python and MATLAB interfaces to the widely used REST (Representational State Transfer) web service architecture. We will build on these interfaces to facilitate multiple forms of queries, such as positional requests (e.g., return all features that are adjacent to specific spatial co-ordinates); morphological requests (e.g. return all reads and spatial coordinates in all nuclei); gene requests (e.g. return all reads for a specific gene). We will deploy the entire software within the GFE AWS framework:

Risks and alternatives strategies If we have difficulty integrating Bisque with the GFE API we will use a simplified data storage interface, we have been working with the Bisque development team. Alternatively, Harvard has two Amazon Web Services technical representatives onsite that have helped in the past deploying applications and analyses in the cloud.

1 Pnevmatikakis, E. A. *et al.* Fast spatiotemporal smoothing of calcium measurements in dendritic trees. *PLoS Comput Biol* **8**, e1002569, doi:10.1371/journal.pcbi.1002569 (2012).

2 Micheva, K. D., Busse, B., Weiler, N. C., O'Rourke, N. & Smith, S. J. Single-synapse analysis of a diverse synapse population: proteomic imaging methods and markers. *Neuron* **68**, 639-653, doi:10.1016/j.neuron.2010.09.024 (2010).

3 Rah, J. C. *et al.* Thalamocortical input onto layer 5 pyramidal neurons measured using quantitative large-scale array tomography. *Front Neural Circuits* **7**, 177, doi:10.3389/fncir.2013.00177 (2013).

4 Lee, J. H. *et al.* Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc* **10**, 442-458, doi:10.1038/nprot.2014.191 (2015).

5 Ko, H. *et al.* Functional specificity of local synaptic connections in neocortical networks. *Nature* **473**, 87-91, doi:10.1038/nature09880 (2011).

6 Shekhar, S., Gunturi, V., Evans, M. R. & Yang, K. in *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access.* 1-6 (ACM).

7 Aji, A. *et al.* Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* **6** (2013).

8 Thusoo, A. *et al.* Hive: a warehousing solution over a map-reduce framework. *Proceedings of the VLDB Endowment* **2**, 1626-1629 (2009).

9 Clunie, D. A. DICOM structured reporting and cancer clinical trials results. *Cancer informatics* **4**, 33-56 (2007).

10 Channin, D. S., Mongkolwat, P., Kleper, V., Sepukar, K. & Rubin, D. L. The caBIG annotation and image Markup project. *Journal of digital imaging* **23**, 217-225, doi:10.1007/s10278-009-9193-9 (2010).

11 Kvilekval, K., Fedorov, D., Obara, B., Singh, A. & Manjunath, B. S. Bisque: a platform for bioimage analysis and management. *Bioinformatics* **26**, 544-552, doi:10.1093/bioinformatics/btp699 (2010).

12 Carpenter, A. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* **7**, R100 (2006).

13 Asari, H., Pearlmutter, B. A. & Zador, A. M. Sparse representations for the cocktail party problem. *J Neurosci* **26**, 7477-7490, doi:10.1523/JNEUROSCI.1563-06.2006 (2006).

14 Ma, W. J., Beck, J. M. & Pouget, A. Spiking networks for Bayesian inference and choice. *Current opinion in neurobiology* **18**, 217-222, doi:10.1016/j.conb.2008.07.004 (2008).

15 Uchida, N. & Mainen, Z. F. Speed and accuracy of olfactory discrimination in the rat. *Nature neuroscience* **6**, 1224-1229, doi:10.1038/nn1142 (2003).

16 Yang, Y., DeWeese, M. R., Otazu, G. H. & Zador, A. M. Millisecond-scale differences in neural activity in auditory cortex can drive decisions. *Nature neuroscience* **11**, 1262-1263, doi:10.1038/nn.2211 (2008).

17 Znamenskiy, P. & Zador, A. M. Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* **497**, 482-485, doi:10.1038/nature12077 (2013).

18 Xiong, Q., Znamenskiy, P. & Zador, A. M. Selective corticostriatal plasticity during acquisition of an auditory discrimination task. *Nature*, doi:10.1038/nature14225 (2015).

19 Jaramillo, S. & Zador, A. M. Mice and rats achieve similar levels of performance in an adaptive decision-making task. *Frontiers in systems neuroscience* **8**, 173, doi:10.3389/fnsys.2014.00173 (2014).

20 Sanders, J. I. & Kepecs, A. Choice ball: a response interface for two-choice psychometric discrimination in head-fixed mice. *Journal of neurophysiology* **108**, 3416-3423, doi:10.1152/jn.00669.2012 (2012).
21 Yuste, R., MacLean, J., Vogelstein, J. & Paninski, L. Imaging action potentials with calcium indicators. *Cold Spring Harbor protocols* **2011**, 985-989, doi:10.1101/pdb.prot5650 (2011).
22 Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature biotechnology* **29**, 928-933 (2011).
23 Peikon, I. D., Gizatullina, D. I. & Zador, A. M. In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic acids research* **42**, e127-e127 (2014).
24 Zador, A. M. *et al.* Sequencing the connectome. *PLoS biology* **10**, e1001411 (2012).
25 Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360-1363, doi:10.1126/science.1250212 (2014).
26 Mitra, R. D. *et al.* Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci U S A* **100**, 5926-5931 (2003).
27 Chen, F., Tillberg, P. W. & Boyden, E. S. Optical imaging. Expansion microscopy. *Science* **347**, 543-548, doi:10.1126/science.1260088 (2015).
28 Söderberg, O. *et al.* Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. *Methods* **45**, 227-232 (2008).