Reconstruction of complete connectivity matrix for connectomics by sampling neural connectivity with fluorescent synaptic markers

Yuriy Mishchenko
Columbia University, Department of Statistics, 1255 Amsterdam Ave
New York, New York, 10027

Abstract

Physical organization of the nervous system is a topic of perpetual interest in neuroscience. Despite significant achievements made here in the past, many details of nervous system organization and its role in animals' behavior remain obscure, while the problem of complete connectivity reconstructions has recently re-emerged as one of the major directions of modern neuroscience (i.e. connectomics). In this paper we describe a novel paradigm for neural connectivity reconstructions that can yield connectivity maps with high resolution, high speed of imaging and data analysis, and possesses significant robustness to errors. In essence, we propose that physical connectivity of a neural circuit can be sampled using anatomical fluorescent synaptic markers localized to different parts of the neural circuit with a technique for randomized genetic targeting, and that high-resolution connectivity maps can be extracted from such datasets. We describe in detail how this approach can be implemented experimentally and how neural connectivity matrix can be inferred subsequently statistically. We test utility of our approach on simulations of neural connectivity reconstruction experiments in C. elegans, where real neural wiring diagram is available from serial electron microscopy. Using such actual wiring diagram we find that complete connectivity matrix in C. elegans can be re-obtained using our approach in 1-7 days of imaging and data analysis. The best alternative connectomics approach would require at least 1-2 years for the same goal. We also describe applications with different connectivity probes and genetic targeting techniques that can allow connectome reconstructions also in larger organisms such as Drosophila.

1. Introduction

In the recent years the problem of complete reconstructions of neural connectivity has re-emerged as one of the major goals of modern neuroscience (Sporns et al., 2005; Briggman and Denk, 2006; Smith, 2007; Lichtman et al., 2008; Helmstaedter et al., 2009). Several "connectomics" projects had been thus proposed relying on classical electron microscopy (Briggman and Denk, 2006; Smith, 2007), injections of tracer viruses (Bohland et al., 2009), sparse expression of fluorescent cytoplasmic markers (Svoboda, Personal Communication), diffusion tensor imaging (Hagmann et al., 2007; Hagmann et al., 2008), and other. Despite significant progress made along these directions, this problem is far from being solved while majority of proposed approaches experience significant limitations in their resolution or possible reconstruction volume. E.g., diffusion tensor imaging (DTI) allows one to obtain connectivity maps for the entire human brain but only very large-scale projections, such as between entire cortical areas, can be mapped (Hagmann et al., 2007; Hagmann et al., 2008). Electron microscopy (EM), on the other hand, can produce reconstructions at the level of individual synapses but is very difficult to scale to large circuits (White et al., 1986; Briggman and Denk, 2006). Due to ability of EM to map neural circuits at the level of individual synapses, EM remains the favorite approach for the new connectomics reconstructions (White et al., 1986; Briggman and Denk, 2006; Smith, 2007). Yet, EM reconstructions are known to be extremely labor intensive (White et al., 1986; Briggman and Denk, 2006), while tracing of axons and dendrites over large distances through densely cluttered EM images has proven difficult to

automate (Jurrus et al., 2006; Jain et al., 2007; Macke et al., 2008; Mishchenko, 2009). Also, significant sensitivity of EM reconstructions to point errors during imaging or analysis had been recently revealed (Mishchenko, 2009).

In this paper, we describe a different paradigm for reconstruction of neural connectivity that can combine high level of detail of produced connectivity maps, high speed of imaging relying on fluorescent light microscopy, and significant robustness to point errors. In essence, we propose that neural connectivity can be sampled using anatomical fluorescent synaptic markers introduced to different parts of a neural circuit genetically or otherwise, and that high-resolution connectivity maps can be subsequently extracted from such datasets using statistical techniques. We describe a specific setup for such an experiment utilizing *fluorescent synaptic marker GRASP* (Feinberg et al., 2008) and *Cre/Lox system for stochastic gene expression* (Livet et al., 2007) to randomly express GRASP in a neural circuit and produce a sample of certain simple connectivity measurements (total count or total size of synapses labeled with GRASP in one animal). We show that synaptic connectivity matrix can be extracted from such data using techniques of *Compressive Sensing* (Donoho, 2006; Candes and Wakin, 2008). We test feasibility of this approach by simulating a hypothetical reconstruction experiment in *C. elegans*, a popular neuroscience model where real neural wiring diagram is available from EM (White et al., 1986). Using such actual wiring diagram we find that the connectivity matrix in *C. elegans* can be re-obtained with our approach in just 1-7 days of imaging and analysis, while the best alternative connectomics approach would require at least 1-2 years for a similar reconstruction. Finally, we describe applications of our approach with different connectivity probes and genetic targeting techniques that can allow connectome reconstructions in larger organisms such as *Drosophila*. Our results open new possibilities for data-rich, quantitative empirical studies of neural circuits' structure and function in the brain.

2. Materials and Methods

2.1. Probing synaptic connectivity with fluorescent synaptic marker GRASP

We suggest to use fluorescent synaptic marker GRASP (Feinberg et al., 2008) to conduct a large scale probe of physical connectivity in a neural circuit. GRASP relies on recombination of two split-GFP molecules to label synaptic contacts between selected neurons. Split-GFP molecules are a pair of molecules that by themselves are not fluorescent but can chemically recombine into a functional GFP when in proximity from each other (Sarkar and Magliery, 2008). In GRASP such split-GFP fragments are genetically tailored to the proteins normally present on the external sites of pre- and post-synaptic neural surfaces and are expressed in selected neurons using gene-fusion (Fig. 1A). When two such neurons form a synapse, split-GFP fragments recombine over the synaptic cleft and produce a functional GFP, thus, fluorescently labeling that synapse.

The connectivity probe considered here is essentially a replica of GRASP construct above (Feinberg et al., 2008) with one difference. Because GRASP originally was designed to only label synapses of selected neurons, corresponding cell bodies in GRASP remain unaffected. For our purposes, however, it will be necessary also to label cell bodies of the neurons expressing GRASP. For that we introduce a "helper" element to GRASP construct, consisting of a fluorescent nuclei marker, that can be used to identify GRASP expressing neurons post-factum, Fig. 2A-B.

2.2. Probing synaptic connectivity stochastically with GRASP and Cre/Lox system

We suggest to use Cre/Lox system in order to express GRASP stochastically in a neural circuit, and thus allow sampling different connections in a high-throughput manner. Cre/Lox system is a well known genetic technique for manipulating gene expression and, in particular, for expressing genes stochastically in a population of cells. A particularly ingenious tour-de-force demonstration of this latter aspect of Cre/Lox system had been recently produced with the Brainbow mouse (Livet et al., 2007; Lichtman et al., 2008; Lu et al., 2009a; Lu et al., 2009b). In the Brainbow mouse (or rather one of its implementations) a cassette of several genes encoding different emission spectra fluorochromes (e.g., GFP, YFP, etc.) is flanked in the genome with inversely oriented loxP-sites. Cre-recombinase allows one to flip orientation of such sequences stochastically, due to recombinase action. In this process loxP-enclosed sequences can assume either direct or reverse orientations stochastically in different cells (Fig. 1B). Since cells where orientation of the sequence is reversed cannot transcribe respective genes successfully, Brainbow thus succeeds to randomly deactivate different fluorochromes in different cells. Different cells then attain distinctive cytoplasmic colors depending on the combination of the fluorochromes active there, and can be distinguished merely by color. This makes analysis and tracing of projections of such cells significantly simpler and at the same time more reliable (Livet et al., 2007): a number of larger axons and synapses in several neural circuits had been reconstructed using this approach (Lu et al., 2009a; Lu et al., 2009b).

In our settings we use Cre/Lox to randomly deactivate GRASP in different cells of a neural circuit. In one possible implementation the sequences for post- and pre-synaptic GRASP split-GFP constructs (*sA* and *sB*, respectively) are placed into reverse-tandem flanked jointly with loxP-sites, Fig. 2A. Two outcomes of Cre-recombination can then occur with equal probability. In outcome (i) sequence *sA* is transcribed leading such neuron to express post-synaptic split-GFP fragment. Respective nuclei tag *nA* is also transcribed, thus, labeling the nucleus of such neuron with distinctive color. In outcome (ii) sequence s*B* is transcribed leading such neuron to express pre-synaptic split-GFP fragment and nuclei tag *nB*. Neurons always express either *sA* or *sB* and never *sA* and *sB* together. This may be advantageous if the expression of both split-GFP fragments simultaneously in the same cell is undesirable (Bargmann, Personal communication). GRASP puncta are formed by all synapses established by neurons that express complementary either pre- or post-synaptic split-GFP fragments, Fig. 2C.

In another possible implementation, Fig. 2B, sequences for post- and pre-synaptic split-GFP constructs are flanked with loxP-sites individually. Four outcomes of Cre-recombination are then possible with equal probability. In outcome (i) both s*A* and *sB* are transcribed and both pre- and post-synaptic sites of the neuron are tagged and visualized. In outcome (ii) only *sA* sequence is transcribed and, thus, only post-synaptic sites of such neuron are visualized. In outcome (iii) only *sB* sequence is transcribed and only pre-synaptic sites of such neuron are visualized. And in outcome (iv) neither *sA* or *sB* are transcribed and synapses of such neuron are not observed.

Other implementations of GRASP-Cre/Lox system are obviously possible, while Fig. 2 provides two possible “real life” examples of such implementations together with all Cre-recombination outcomes.

2.3. Sampling connectivity with stochastic GRASP

We suggest to use combination of GRASP and Cre/Lox system (stochastic GRASP), described in Sections 2.1-2.2, to sample synaptic connectivity in a neural circuit as illustrated schematically in Fig. 3. In Fig. 3A three subsets of neurons from a hypothetical neural circuit are shown expressing at random post- and pre-synaptic GRASP split-GFP fragments and respective nuclei tags, Sections 2.1-2.2. These three samples can correspond to three different specimens from a stochastic GRASP genetic line, i.e., three *stochastic GRASP phenotypes*. Whenever two neurons that express complementary pre- and post-synaptic split-GFP fragments form a synapse, a fluorescent punctum is also formed allowing one to observe such synapse with fluorescent microscopy. Obviously, different synapses are visualized in different specimens.

Because GRASP labels synapses with a single color fluorochrome (i.e., GFP), in general, it is impossible to distinguish which of the labeled synapses belong to given neuron. Instead, a cumulative measure of all labeled synapses should be obtained such as *the total number* of all puncta or *the combined fluorescent size* of all puncta. We denote such measurement with symbol *n*. One can also record identity of the neurons that expressed GRASP in each experiment via associated nuclei tags. We denote such expression patterns for post- and pre-synaptic GRASP split-GFP fragments with symbols *a* and *b*, respectively. In other words, *a* is just a list of all neurons that expressed given "post-synaptic" nuclei tag in particular specimen, and *b* is that for the neurons expressing "pre-synaptic" nuclei tag. The set of observations that is thus procured using stochastic GRASP is that of triples {(*n*, *a*, *b*)}. Such sample is a characteristic of the stochastic GRASP phenotype in given animal (Fig. 3B). As we will show below, it is possible to exactly reconstruct underlying neural connectivity matrix from such collection of measurements.

2.4. Reconstruction of neural connectivity matrix from stochastic GRASP measurements

The key observation of this work is to recognize that the measurements that can be collected with stochastic GRASP, i.e., the total count or the combined size of GRASP-labeled synapses in different specimens, can be interpreted as certain sums over neural connectivity matrix (Fig. 3C). Specifically, if C_{ij} is an element of such neural connectivity matrix, e.g., describing the number or the size of all synapses between given pair of post-synaptic and pre-synaptic neurons *i* and *j*, respectively, measurement *n* in specimen *k* can be mathematically described as the following sum,

$$n(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i^{(k)} b_j^{(k)} C_{ij} \ . \qquad (1)$$

Here, *N* is the total number of neurons in the circuit, $a_i^{(k)}$ and $b_j^{(k)}$ are the indicator functions characterizing post- and pre-synaptic GRASP expression patterns in animal *k*, respectively. More specifically,

$$a_i^{(k)} = \begin{cases} 1, \text{neuron } i \text{ expressed post - synaptic GRASP in specimen } k \\ 0, \text{neuron } i \text{ did not express post - synaptic GRASP in specimen } k \end{cases}$$
$$b_j^{(k)} = \begin{cases} 1, \text{neuron } j \text{ expressed pre - synaptic GRASP in specimen } k \\ 0, \text{neuron } j \text{ did not express pre - synaptic GRASP in specimen } k \end{cases} \qquad (2)$$

(See Fig. 3A-C for illustration). In stochastic GRASP $a_i^{(k)}$ and $b_j^{(k)}$ are essentially random vectors with approximately 50% of ones and 50% of zeros.

Using Eq. (1) we can now propose that it may be possible to estimate complete neural connectivity matrix C_{ij} statistically from a collection of highly general measurements $\{(n(k), \{a_i^{(k)}\}, \{b_j^{(k)}\}), k=1,\ldots,K\}$. Specifically, it is advantageous to focus on finding unknown matrix C_{ij} by searching for C_{ij} that can best describe available set of measurements (1) in the class of sparse matrices. We know that connectivity in neural systems is typically very sparse and sparse priors previously were shown to significantly accelerate inference of unknown signals.

While formulated problem may appear hopelessly ill-defined (e.g., all expression patterns are broad and nonspecific, connectivity matrix prior is also very nonspecific, etc.), contrary to this impression it is possible to recover the connectivity matrix under these conditions quite accurately. In fact, it can be mathematically shown that the connectivity matrix can be extracted *exactly* from relatively small number of above measurements using Bayesian approach or methods of Compressive Sensing (Candes and Romberg, 2005; Donoho, 2006). In subsequent sections we describe mathematical procedures due to Compressive Sensing that can be used to find the connectivity matrix from stochastic GRASP data.

2.5. Reconstruction of neural connectivity matrix using Compressive Sensing

Compressive Sensing (CS) (Candes and Romberg, 2005; Candes et al., 2006; Donoho, 2006; Candes and Wakin, 2008) is a recently emerged field of signal processing that deals with decoding of linearly encoded sparse signals. Linear encoding here means that a signal, $f(t)$, is encoded by a collection of linear measurements $y(k) = \phi_k * f = \sum \varphi_k(t) f(t)$ (sum is over $t=1,\ldots,T$). Sampling $f(t)$ at a set of predefined points $\{t_i\}$ or observing Fourier components of $f(t)$ would be examples of linear encoding. The goal of CS is to accurately reconstruct $f(t)$ from a set of such linear measurements while knowing that the original signal is sparse (i.e. that $f(t)=0$ for most t, but not knowing where $f(t)$ is zero). For example, we know that the connectivity matrix in a neural system should be sparse but don't know exactly which neurons are connected. CS is a perfect approach to address this reconstruction problem.

Certain remarkable mathematical properties of such reconstruction problem had been recently rigorously established and should be mentioned (Candes and Romberg, 2005; Candes et al., 2006; Candes and Wakin, 2008). First, it was shown that it is possible to reconstruct sparse signal *exactly* from a small number of such linear measurements $K \sim \log(T) T_{sp}$, where $T_{sp} << T$ is the number of nonzero elements in $f(t)$. Second, appropriate reconstruction procedure is tractable: with probability approaching to 1 exponentially in T the sought signal is the smallest l_1-norm solution to the set of linear equations $\{y(k) = \sum \varphi_k(t) f(t),\ k = 1,\ldots, K\}$. Third, these remarkable properties nearly any set of probing waveforms $\varphi_k(t)$.

Recall now that measurements of the count or combined sizes of labeled synapses in stochastic GRASP, n, can be represented with sums over neural connectivity matrix as follows,

$$n(k) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i^{(k)} b_j^{(k)} C_{ij} ,$$

where $a_i^{(k)}$ and $b_j^{(k)}$ are the indicator functions for post- and pre-synaptic GRASP expression patterns in animal *k*, Sections 2.3-2.4. We recognize in these settings the linear encoding problem described above. According to CS (Candes et al., 2006), then, the connectivity matrix can be accurately recovered from such sample of measurements by solving a linearly constrained l_1-optimization problem,

$$\min\|C\|_{l_1} = \min \sum_{i=1}^{N}\sum_{j=1}^{N}\left|C_{ij}\right|, \text{ subject to} \tag{3a}$$

$$n(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i^{(k)} b_j^{(k)} C_{ij}, k = 1,...,K\,. \tag{3b}$$

This is a standard linear-program (LP) and powerful methods exist for solving it efficiently, e.g., such as the interior point method (Wright, 1997; Vanderbei, 2001; Boyd and Vandenberghe, 2004). For large N^2, however, the issue of computational scalability should be dealt with. In particular, Matlab LP solver fails to solve problem (3) when *N* exceeds *N*~50 neurons due to computer memory limitations.

In (Candes and Romberg, 2005) alternative method for solving Eqs. (3) for very large *N* was proposed. Specifically, if l_1-norm of the connectivity matrix, $S = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|C_{ij}\right|$, is known a-priori, solution of problem (3) can be found as the intersection of two convex sets – l_1-cube $\|C\|_{l_1} = S$ and the hyper-plane (3b). In our case S can be estimated directly from the measurements (*n*,*a*,*b*), i.e. $S \approx \langle n(k)\rangle \big/ \left\langle \left|a^{(k)}\right\| b^{(k)}\right| \right\rangle$ (the average is over all samples *k* and $\left|a^{(k)}\right| = \frac{1}{N}\sum_{i=1}^{N} a_i^{(k)}$). According to the central theorem of CS, such intersection will be unique and correspond to the exact solution of problem (3) when *K* is sufficiently large (Candes and Romberg, 2005).

To actually find said intersection the method of alternate projections can be used (Bregman, 1965). In this algorithm one starts with a random guess for the connectivity matrix, $C^{(0)}$, and then repeats two steps of consequently projecting current guess, $C^{(l)}$, onto the hyper-plane (3b) and then onto the l_1-cube $\|C\|_{l_1} = S$,

$$C^{(l+1/2)} := C^{(l)} + P^T\left(PP^T\right)^{-1}\left(n - P\cdot C^{(l)}\right). \tag{4a}$$

$$C_{ij}^{(l+1)} := \max\left(0, C_{ij}^{(l+1/2)} - \gamma^{(l)}\right), \tag{4b}$$

In Eq. (4a) we used vector notation for clarity, i.e., *N*x*N* connectivity matrix C_{ij} is represented by a N^2x1 vector over joint indices (*ij*), *C*, and *P* denotes *K*xN^2 matrix $P(k;ij) = a_i^{(k)} b_j^{(k)}$. In such notation Eq. (3b), clearly, corresponds exactly to the dot-product of *P* and *C*, $n = P\cdot C = \sum_{(ij)} P(k;ij) C_{ij}$. Eq. (4a) describes the step of projecting $C^{(l)}$ onto the hyper-plane (3b): it can be trivially checked that $P\cdot C^{(l+1/2)} = n$. Eq. (4b) describes the step of projecting

$C^{(l+1/2)}$ onto the l_1-cube $\|C\|_{l_1} = S$ and $\gamma^{(l)}$ is chosen at each iteration to achieve $\|C^{(l+1)}\|_{l_1} = S$ and controls retraction of $C^{(l+1/2)}$ to the nearest face of that cube. Steps (4a) and (4b) are repeated until convergence that is typically achieved rapidly, because the process typically proceeds on two hyper-planes: that of constraints and one of the faces of the l_1-cube.

2.6. A hypothetical neural connectivity reconstruction experiment using stochastic GRASP in *C. elegans*

We evaluate performance of described approach using simulations of a hypothetical neural connectivity reconstruction experiment in *C. elegans*. *C. elegans* is a popular neuroscience model organism and the only organism where real wiring diagram is known from prior electron microscopy work (White et al., 1986). In particular, this motivated our attention specifically to that system although our method obviously is not restricted to applications in *C. elegans*.

We simulated connectivity reconstruction experiments involving from 500 to 10,000 stochastic GRASP measurements. For each observation we generated pre- and post-synaptic GRASP expression patterns according to Section 2.2. For construct in Fig. 2A each neuron was assumed to express pre- and post-synaptic split-GFP fragments mutually exclusively with probability 50%. For construct in Fig. 2B each neuron was assumed to express pre- and post-synaptic split-GFP fragments non-exclusively with probability 50%. Using thus generated $a_i^{(k)}$ and $b_j^{(k)}$ we constructed measurements $n(k)$ according to Eq.(1).

We also considered a number of noise factors that can affect actual experiment. One of the factors considered was biological variability, i.e., that actual connectivity matrix can vary from one stochastic GRASP animal to another. In that case for each animal k a different connectivity matrix $C_{ij}^{(k)}$ was prepared using formula $C_{ij}^{(k)} = C_{ij}^{e}(1-a_b) + v[a_b C_{ij}^{e}]$. Here C_{ij}^{e} was the test wiring data from (White et al., 1986), i.e. for each pair of neurons i and j C_{ij}^{e} described the number of synaptic contacts between these neurons per EM data (White et al., 1986). $v[x]$ was a Poisson-distributed random variable with mean x and modeled variation in C_{ij} from animal to animal. Parameter a_b specified the degree of such variation - a_b=0 corresponded to no variability and a_b=1 corresponded to the case where synapses were formed completely at random with certain number of synapses between given neurons on average. The objective of reconstruction in this case was to recover the *average* connectivity matrix C_{ij}^{e}.

Second, we considered noise that could be added into observations during the measurement process itself, e.g. due to imperfections of the experimental setup. For that we altered measurements $n(k)$ by adding white noise $n(k) \to n(k)(1+a_o v)$. Here, v was a Normally-distributed random variable with zero mean and unit variance. a_o specified the relative strength of such "measurement" noise.

Finally, we considered errors that could be made during measurements of the expression patterns $a_i^{(k)}$ and $b_j^{(k)}$. Since such errors typically would lead reconstruction algorithm to use wrong matrix P, such errors clearly could result in deviations of the reconstruction from the truth. To include this factor in our model we corrupted "true" expression patterns by randomly shuffling identities of a small number of neurons. In that sense, we assumed that the total number of

neurons was known a-priori (as is in fact in *C. elegans*) and the only error that could be committed was confusion of nearby neurons' by some cell-identification algorithm.

3. Results

3.1. Feasibility of wiring diagram reconstruction in *C. elegans* using stochastic GRASP

To test our approach we simulated a hypothetical neural connectivity reconstruction experiment in *C. elegans*. *C. elegans* is a popular neuroscience model and the only animal where complete wiring diagram is known from serial electron microscopy (White et al., 1986). This, in particular, motivated our attention specifically to this system although our approach obviously is not restricted to *C. elegans*.

The neural wiring diagram in *C. elegans* in (White et al., 1986) was produced as follows. Entire body of one *C. elegans* specimen was imaged at extremely high resolution using serial electron microscopy. (More accurately, several specimens were imaged partially in order to achieve a dataset equivalent to one full coverage.) Synapses were consequently manually found in the images and associated with corresponding axons and dendrites; and axons and dendrites were traced to respective cell bodies through multiple EM images, also manually. Total count of synapses for different pairs of neurons was then tabulated and the table describing these counts was recorded as the wiring diagram for the neural circuit in *C. elegans* (now available from *wormatlas.org*). We used this actual wiring diagram as the ground truth for simulated connectivity reconstruction experiment, see Section 2.6 for further details. This wiring diagram contains N~300 neurons connected by 2500 connection with grand total of about 6000 synapses.

We simulated stochastic GRASP experiment involving from K~500 to K~10,000 animals and inspected reconstructions obtained with stochastic GRASP both as a matrix and as a scatter plot of reconstructed vs. actual connection weights (Fig. 4). Quantitatively we characterized reconstructions using correlation coefficient between reconstructed and actual connection weights, r^2. r^2=0 would correspond, naturally, to no correlation between reconstructed and true connection weights and r^2=1 would correspond to reconstructions that were perfect (exact). Reconstructions with r^2=0.5, we observed in particular, were already good enough to provide a practically meaningful estimate of neural connectivity matrix in a neural circuit (Fig. 4).

In Fig. 5 we show r^2 for stochastic GRASP reconstructions vs. the number of measurements K. As can be seen from this figure, perfect reconstructions are obtained already from ~10,000 measurements. This is in excellent agreement with CS theory and far below ~90,000 measurements that can be naively expected to be necessary to at least once completely characterize the connectivity matrix of 300x300 neurons in *C. elegans*. Practically meaningful reconstructions are obtained already from ~5,000 measurements. This result is obtained either using stochastic GRASP construct in Fig. 2A or 2B, also in agreement with predictions of CS theory. Compressive Sensing is critical for this procedure as, e.g., a naïve solution using more conventional L_2 regularization (i.e., $\min\|C\|_{l_2} = \min\sum\sum|C_{ij}|^2$ s.t. $n(k) = \sum\sum a_i^{(k)} b_j^{(k)} C_{ij}$) leads to practically useless results with this number of measurements.

Our results indicate that complete reconstructions of wiring diagram in *C. elegans* with stochastic GRASP indeed can be feasible and can be performed, in fact, over a short amount of

time with already existing technologies (i.e., Cre/Lox system, GRASP, and low-end light microscopy).

3.2. Impact of biological variability on wiring diagram reconstruction using stochastic GRASP

It may be expected that connectivity matrix in different animals of the same specie may vary from animal to animal. In this case in each stochastic GRASP experiment a slightly different connectivity matrix can be probed. Although it is not yet known to what degree biological variability should be expected in real neural circuits (in fact directly opposite opinions exist on this topic among neuroscientists), we inspected potential impact of such variability on neural connectivity reconstructions with stochastic GRASP, see Section 2.6. for details. Results of this study are shown in Fig. 6. In every case we find that impact of biological variability on connectivity reconstructions will be insubstantial.

3.3. Impact of added noise on wiring diagram reconstruction using stochastic GRASP

It can be expected that perfect measurement of total puncta count/size may not be possible to obtain in realistic conditions. Such noise added during the process of collecting the data itself may obviously distort results of the connectivity reconstruction. We inspected potential impact of such noise on neural reconstructions with stochastic GRASP, see Section 2.6. for details. Results of this study are shown in Fig. 7. Reconstructions were found to be stable under small additions of noise, i.e. small amounts of noise caused only proportionally small deviations in the reconstruction result (Candes et al., 2006; Candes and Wakin, 2008). Yet, sensitivity to noise additions was substantial and only up to 3-4% of added noise could be tolerated for the minimal K~10,000. Robustness improved with the number of measurements K so that higher levels of noise could be potentially mitigated by collecting datasets with larger number of measurements. Although we do not extend this study to larger K, because such a study would be very computationally expensive, specifications for particular target reconstruction accuracy and given noise levels can be straightforwardly computed using the methods from Section 2.5 and 2.6.

3.4. Impact of misidentifications in detected GRASP expression patterns on wiring diagram reconstruction using stochastic GRASP

It can be expected that perfect measurement of GRASP expression patterns, a and b, may not be possible to obtain in realistic conditions, so that certain amount of errors in detected expression patterns should be anticipated. Clearly such errors can result in deviations of reconstructed connectivity matrix from the truth. We inspected potential impact of such errors on neural reconstructions with stochastic GRASP in greater detail, see section 2.6. Results of this study are shown in Fig. 8. Although reconstructions were found to be stable again, sensitivity to errors in the expression patterns was substantial and only up to 5-6% of misidentified neurons in detected expression patterns (i.e. 15-20 misidentified neurons per 300 in *C. elegans*) could be tolerated for the minimal K~10,000. As before, reconstructions with larger number of measurements K were more robust to errors, so that higher error levels could be tolerated with datasets of larger size. Although we do not extend this study to K greater than 10,000, again because that would be very expensive computationally, specifications for particular target reconstruction accuracy and given errors levels can be computed using the methods from Section 2.5 and 2.6.

3.5. Comparison with alternative approaches for reconstructions of neural connectivity

In this section we briefly survey alternative methods for connectomics reconstructions. Such comparison is complicated by the diversity of existing approaches, making it difficult to bring them to a common denominator, and the fact that none of the existing approaches (except for serial electron microscopy) is yet capable of detailed reconstructions of neural connectivity in a system as large as *C. elegans*. For comparison here we chose two approaches that in our opinion have the best prospects for producing complete neural circuit reconstructions in the foreseeable future – serial electron microscopy (EM) (White et al., 1986; Briggman and Denk, 2006; Mishchenko, 2009) and pair-wise cell recordings (PCR) (including optically stimulated and ChR2 assisted circuit mapping (Bureau et al., 2004; Baker et al., 2005; Petreanu et al., 2007)). We left out a number of promising approaches for estimating neural connectivity from functional correlations between neurons, e.g., observed using calcium imaging (Broome et al., 2006; Jones et al., 2007; Pillow et al., 2008). Although these potentially allow reconstructions at the level of individual neurons for circuits as large as hundreds and thousands of neurons (Stevenson et al., 2009; Mishchenko and Paninski, 2011), only effective connectivity matrix is produced whose relationship to physical circuit structure still could not be clearly elaborated.

EM reconstruction paradigm, as was briefly mentioned in the Introduction, comprises: a) directly imaging a block of neural tissue with electron microscopy; b) finding synapses in the images and associating them with corresponding axons and dendrites; c) tracing associated axons and dendrites through multiple images to corresponding cell bodies; d) establishing a connection between a pair of cells. The main advantage of EM reconstructions is their ability to assess physical structure of a neural circuit directly, as individual synapses, axons, and dendrites are all directly observed via EM images. Furthermore, details of neural morphology such as arbor shape, branching structure, synapses bunching, etc., can be also extracted from EM images. Such data can be used to directly model physical processes inside neurons as well as to study sub-cellular organization of neurons and micron scale architecture of neuropil.

The main disadvantage of EM reconstructions is low speed with which data can be acquired and extreme difficulty of subsequent image analysis. EM reconstructions are characterized by very high labor-intensity – manual reconstructions in one *C. elegans* specimen in (White et al., 1986) took over 10 years to complete. Modern automation can reduce this time but has to rely on complex and brittle image-understanding algorithms, making reliable scaling of such approaches difficult (Jurrus et al., 2006; Jain et al., 2007; Macke et al., 2008; Helmstaedter et al., 2009; Mishchenko, 2009). In (Mishchenko, 2009), in particular, a complete automation approach was presented and reconstruction speed of about 5 μm^3/man-hour in dense neuropil in rat hippocampus was demonstrated, estimated to correspond to a 10-fold improvement over purely manual analysis. Such automation in *C. elegans* potentially could allow complete connectome reconstructions in 1-2 years of work by a single operator, with subsequent speedup possible by parallelizing. Of course, this estimate should be only viewed as a rough, order of magnitude estimate since many different factors will play role in actual performance, including simpler, longitudinal organization of axons in *C. elegans* (which may help reconstructions) and their smaller size (which may make reconstructions more difficult), etc.

PCR reconstruction paradigm comprises: a) stimulating different neurons individually electrically or optically; b) observing sub-threshold electrical responses in selected neuron using electrode patch or (potentially) voltage sensitive fluorescent probe. The advantage of this

approach is in its ability to directly “listen” to electrical activity in individual neurons, which is how neurons exchange information. While this approach is responsible for the bulk of information available today about neurons and their circuits, such measurements are also notoriously difficult to perform. Normally, only as few as 10-100 neural pairs can be patched and tested a day. In *C. elegans*, in particular, this implies that a complete scan of the connectivity matrix would take anywhere from 3 to 30 years to complete, although we should also note that patching neurons in *C. elegans* is admittedly more difficult due to their extremely small size. With optical stimulation and optical readout techniques such as ChR2 and voltage sensitive dyes, however, this time can be dramatically reduced by a large factor, making complete scans possible in the span of several minutes to hours.

For a comprehensive comparison of these approaches we inspected their performance in terms of reconstruction accuracy and robustness to noise as a function of the degree of *reconstruction effort*. In EM, specifically, reconstruction effort would refer to the volume of neural tissue that had to be imaged and in PCR this would refer to the number of pairs of neurons tested. Results of this study are shown in Fig. 9.

For EM accuracy of reconstruction as a function of reconstruction effort is shown in Fig. 9A and robustness to errors is studied in Fig. 9B. Noise is typically introduced in EM reconstructions as point errors in traces of thin axons, either due to local mistakes in interpretation of images or small defects during imaging. Because each such error can affect a large number of downstream synapses on corresponding axon (e.g., causing all such synapses to be mis-assigned to a different neuron or lost from reconstruction), EM reconstructions can be very sensitive to point errors. In Fig. 9B, in particular, the quality of final reconstruction in *C. elegans* vs. the rate of such errors is shown. This is characterized by the probability to commit one error in reconstruction of one axon between two consecutive synapses. For instance, if EM reconstruction makes on average one error per every 100 µm of reconstructed axon that would correspond to roughly 10% chance to find an error on 10 µm axonal segment constituting typical distance between two consecutive pre-synaptic boutons in mammals. Thus, when tracing an axon from one synaptic connection to the next, there would be a 10% chance to make a point error affecting final reconstruction. From Fig. 9B we observe that the largest error rate that can be tolerated in *C. elegans* is just about 10% (for the largest dataset) or one error per ~50-100 µm trace. This corresponds to approximately one error per 1000-2000 typical serial EM sections. Note that this limit depends on the size of reconstructed neural circuit and will become progressively smaller for circuits of the size larger than that in *C. elegans*.

For PCR accuracy of reconstruction as a function of reconstruction effort is shown in Fig. 9C and robustness to errors is studied in Fig. 9D. In PCR errors typically correspond to failure to observe a connection when stimulating or recording from a neuron. This can occur, e.g., due to failure of target neuron to get stimulated or failure to observe weak post-synaptic sub-threshold response, etc. Such errors mathematically correspond to point errors in reconstructed connectivity matrix and can be characterized by their probability per one tested pair of neurons. There will be also certain variance in measured connection weights due to intrinsic fluctuations in sub-threshold membrane potential. In Fig. 9D we show quality of final reconstructions in *C. elegans* vs. the rate of such error. We see that PCR, indeed, is very robust to errors with noise levels up to 40-60% that can be easily tolerated.

Similarly to our analysis of stochastic GRASP above, either in case of EM or PCR we found that the impact of biological variability on reconstruction accuracy was insignificant (data not shown).

4. Discussion

We describe new approach for reconstructions of neural connectivity that utilizes fluorescent synaptic marker GRASP (Feinberg et al., 2008) with Cre/Lox system for random gene expression (Livet et al., 2007; Lichtman et al., 2008) to randomly sample synaptic connectivity in a neural circuit (stochastic GRASP). Complete connectivity matrix is reconstructed from thus produced fluorescent measurements using the methods of *Compressive Sensing* (Candes and Romberg, 2005; Candes et al., 2006; Donoho, 2006; Candes and Wakin, 2008).

We use real neural wiring data for *C. elegans*, available from EM (White et al., 1986), to show that described approach is feasible and can be executed using already existing technologies. In *C. elegans* we find that 5,000-10,000 stochastic GRASP measurements will suffice to recover complete wiring diagram. Given small size (100x100x1000 µm) and fast development of *C. elegans* (2-3 days), 10,000 animals can be incubated on a single Petri-dish in the span of several days. Modern *C. elegans* phenotype screens, in fact, routinely work with populations that large. Fluorescent measurements can be obtained by performing 3D-scans of specimen bodies while GRASP expression patterns can be obtained at the same time using associated nuclei-targeted fluorescence. Computer algorithms such as (Long et al., 2008) can be used to automate the process of determining GRASP expression patterns from produced imaging data. Assuming that imaging data at resolution ~0.5 µm/pixel can be acquired using a confocal microscope at speed of at least 10MHz (10MHz acquisition rate can be achieved already with a 1000x1000 pixel CCD camera mounted on a confocal microscope and taking images at frequency ~10 frames per second), it should be possible to perform one full scan of specimen's body using a specialized setup in 1 minute or less (Kerr, Personal communication). (Really, at resolution 0.5 µm/pixel entire body of one *C. elegans* (100x100x1000 µm) will contain approximately 80 million pixels. By using a CCD camera with 2000x2000 pixels and taking one image of longitudinal section of *C. elegans* of 2000x200 pixels in 1/5 of a second, 80 million pixels can be imaged in [$80*10^6$ pixels]/[$4*10^5$ pixels/frame]/[5 frame/second] = 40 seconds. We may also note that entire body of the specimen may not need to be imaged since neural connections will typically be only localized to small regions in its body.) The total time of imaging then can be estimated in 1-7 days. Subsequent reconstruction procedure is computationally straightforward and can be performed in several hours on single laptop computer.

Several remarkable features of such reconstructions should be explicitly mentioned. Complete connectivity can be recovered using what can be considered a grossly incomplete dataset, with the number of measurements only $K \sim N_{sp} \log N << N^2$, where N_{sp} is the number of connected neural pairs in the circuit and N is the total number of neurons (Candes and Romberg, 2005). Clearly, should we have known the identity of all connected neural pairs in the circuit, we could use $K \sim N_{sp}$ measurements to determine the connectivity matrix, e.g., using PCR. With stochastic GRASP and Compressive Sensing reconstruction algorithm we can achieve similar performance without any prior knowledge about identity of connected neural pairs. Furthermore, reconstructions with smaller datasets will produce the best possible approximations to true

connectivity matrix with ~K terms (Romberg, 2008). Stochastic GRASP reconstructions do not involve tracing of neurons, recordings from individual neurons, targeting individual neurons or their small neural groups, or other explicit association of synaptic connection with a remotely located neurons, which is the main difficulty of EM and many other techniques. Synapses are associated with relevant neurons a-posteriori, the only information that needs to be collected at the time of experiment is an overall measure of connectivity and the set of "activated" neurons. Detailed connectivity matrix can be obtained even when all "activation" patterns are broad and nonspecific (e.g., stochastic).

While we show that our approach can be successful in *C. elegans* already today, it is important to think about scaling it to larger systems, e.g., such as *Drosophila*. In *Drosophila*, in particular, $N \sim 10^5$ neurons and on average 100 synapses per neuron result in circuit complexity of $N_{sp} \sim 10^7$, and an experiment such as we described above may appear prohibitive.

We must note that this problem is not confined to our method per se. Even though reconstruction of a circuit with 300 neurons and 2500 connections in one *C. elegans* had been completed with EM (White et al., 1986), and connectivity studies involving ~1000 neural connections had been conducted in the past with other techniques (Ikegaya et al., 2005; Song et al., 2005; Broome et al., 2006; Pillow et al., 2008; Mishchenko et al., 2010), scaling any of these approaches to a circuit of $\sim 10^7$ connections is highly nontrivial to say the least.

Notwithstanding, three general directions can be proposed that can allow applying our strategy to a circuit as large as that in *Drosophila*. First, while we assumed that only a single measurement can be obtained from one stochastic GRASP animal, this does not need to be the case in principle. Whereas GRASP utilizes chemical reconstitution of a particular protein across synaptic cleft, it is bound to be essentially a single-color marker. However, recently the author had shown that mere spatial proximity of two different emission wavelength pre- and post-synaptic fluorescent markers can allow identifying synapses without the need for chemical reconstitution (Mishchenko, 2010). In this case, fluorochromes can be multiplexed onto synapses in large numbers leading to up to 2^{N_c} different synaptic labels for N_c distinct color fluorochromes [similar in essence to the idea employed in the Brainbow mouse (Lichtman et al., 2008)]. This case can be accommodated by our framework without any modifications: the number of measurements obtained from one stochastic "GRASP" animal will be 2^{N_c} instead of one. This will allow accelerating data acquisition dramatically and allow reconstructions of neural circuits as large as that in *Drosophila*, e.g., by imaging 10 synaptic fluorochromes in 1000 animals.

Second, one may consider formulating connectivity reconstructions in terms of certain neural populations rather than individual neurons (Luo et al., 2008). As the role of individual neurons vs. neural classes in animal's behavior is still being debated, such approach may prove to be not only less practically involving but also more empirically relevant. The connectivity matrix elements in that case would correspond to the couplings between different populations of neurons and can be "sampled" using the same general approach described here. In *Drosophila*, in particular, libraries of Gal4 lines provide particularly convenient tool for accessing as well as cataloguing neural populations. The connectivity matrix can be accessed by conducting experiments where GRASP, or its analog, is targeted using different Gal4 lines and their combinations, e.g., using methods from (Luo et al., 2008). The expression patterns *a* and *b* in

that case will be known a-priory from the identity of Gal4 lines used in different animals and need not to be measured. The only measurement that needs to be collected is that of the overall GRASP fluorescence strength. Such measurement can be obtained extremely quickly and for entire population of animals at once. The connectivity matrix in terms of *smallest resolved groups of neurons* for the set of used Gal4 lines, as described below, will be produced even if none of the populations were targeted specifically in any of the experiments. The best approximation to the true connectivity matrix given available data can be recovered and subsequently improved as more data is obtained.

Third, our strategy need not be constrained to a specific synaptic marker such as GRASP or synaptic Brainbow but can be used with other modalities that may ultimately lead to faster neural circuit reconstructions. Of these, pair-wise cell recordings using optical stimulation [e.g., ChR2 (Petreanu et al., 2007)] and optical observation of changes in membrane potential [e.g, using voltage sensitive dyes (Baker et al., 2005)] appears now as the most promising approach. In particular, if stimulation and recording of response can be performed optically, the data acquisition time as low as several hours for a circuit as large as that in *C. elegans* can be expected. The data produced by such experiment can be also represented as linear measurements over the connectivity matrix (at least in the first approximation), allowing recovery of neural connectivity using the same formalism as we developed above. Specifically, we can represent sub-threshold post-synaptic response in given neuron *i* observed after stimulation of a group of neurons $b_j(t)$ as follows,

$$V_i(t) = \sum_j C_{ij} b_j(t). \qquad (5)$$

Here, $V_i(t)$ can be, e.g., a measure of the integrated evoked post-synaptic response (iEPSP) in the imaged neuron after stimulation episode, *t* describes different stimulation episodes, and C_{ij} is the matrix of iEPSPs for different post- and pre-synaptic neurons. As is easy to see, Eq. (5) is identical to Eq. (1) and can be analyzed using the same formalism. Note that ability to observe sub-threshold responses is critical in this design and, thus, one needs to use voltage-sensitive dye and not calcium sensitive dye. This approach can allow mapping of very large neural circuits in a meaningful amount of time.

Another important question that can be asked is whether our strategy can be applied in organisms where neurons are not individually identifiable. Our method calls for identification of neurons affected by GRASP in each imaged animal. While it is known how to identify neurons in *C. elegans* and some other organisms, in general methods for, and even possibility of, identification of individual neurons in larger organisms are subject of significant debate. An important question therefore is how our approach can be applied in such organisms where neurons cannot be individually identified.

This question, again, is not specific to our approach per se. A degree of “identifiability” is required for results of any neural connectivity experiment to be comparable or generalizable across different animals. Really, if neurons in an organism cannot be identified, this also means that it is impossible to co-relate neurons in different animals. In turn, this means that neural connectivity reconstructions obtained from one animal cannot be compared with that obtained in another animal, and, even more generally, any similar experimental result from one animal

cannot be generalized or compared with that in another. “Identifiability” in some form should always be available and, indeed, that is always the case in neuroscience. In *C. elegans*, in particular, neurons can be identified individually; in *Drosophila* genetic lineages and Gal-4 classes are used to identify neurons and their populations; in higher organisms morphological classes and anatomical landmarks such as cortical layers, neural nuclei, or cortical areas are used to characterize connectivity. If any such system is available for the target organism, the framework that we developed here can be also applied, now in the context of such system. E.g., if one has a system of genetic classes in place to label neural structures in the brain, e.g. lineages or Gal-4 lines in *Drosophila*, the reconstruction program that we described can be carried out by delivering GRASP to different such genetic classes and performing the measurements and the calculations that we describe. The connectome can be recovered in terms of such neural populations corresponding to known lineages or Gal-4 lines.

While one may think that “class-wise” reconstructions that can be obtained with such strategy fall far short from ideal neuron-wide connectome that EM can potentially deliver, such class-wise reconstructions already contain significant amount of information and can be of substantial interest. In fact, class-wise connectomes have been already used widely in neuroscience research in systems where individual neurons cannot be identified, including EM studies of optic lobe in *Drosophila* (Meinertzhagen and Sorra, 2001), optical studies of cortical micro-circuits in mouse (Shepherd and Svoboda, 2005), etc. One can also argue that class-wise reconstructions may be more physiologically meaningful and relevant in the settings where neurons may have no explicit “individual identity” (Luo et al., 2008).

More specifically, consider a case where one has a system of different anatomical morphological classes for characterizing neural populations. One can then count neurons expressing stochastic GRASP in different such morphological classes in different animals. Such counts will then replace indicator functions $a_i^{(k)}$ and $b_j^{(k)}$ used above. Connectivity matrix C_{ij} then can be obtained using exactly the same procedures that we described in Section 2.5. In this case, clearly, i and j will refer to such different morphological classes and C_{ij} will describe the number of synaptic connections existing in the circuit between such classes. Similarly, if one has an atlas of anatomical landmarks available for target organism, such as a map of brain areas or neural nuclei, one can count neurons expressing stochastic GRASP in different such areas and use these counts in place of indicator functions $a_i^{(k)}$ and $b_j^{(k)}$, as above.

Libraries of genetic classes, such as Gal4 lines or lineages in *Drosophila*, have recently emerged as another popular way for describing neural structures in large neural systems (Phelps and Brand, 1998; Luo et al., 2008). Such libraries can provide a powerful tool for identifying and cataloguing neurons even when morphology and location cannot be used for that reliably. Having a library of genetic lines, consisting of certain lines *A*, *B*, *C*,…, one can associate with each neuron a pattern of 1’s and 0’s describing whether that neuron is present in expression patterns of lines *A*, *B*, *C*, etc. For example, pattern “110…” would correspond to a neuron that is present in the expression patterns of lines *A* and *B* but not line *C*, etc. Different such patterns will constitute an identification system for neural structures whereas different neurons will be identified by the combinations of genetic lines that contain them. One does not even need the ability to detect or identify these neurons anatomically: as long as such neurons can be repeatedy accessed genetically by manipulating Gal4 lines experiments involving such defined groups of

neurons can be conducted and their connectivity and properties studied. Note that neurons present in the same lines will be indistinguishable and equivalent in terms of such identification system, and will constitute the *smallest resolved group of neurons*, or equivalence class, for such identification system (Table 1). By delivering GRASP using combinations of such genetic lines and methods from (Luo et al., 2008), connectome can be obtained in terms of such smallest resolved populations and connectivity between them, as was described above in this section.

5. Conclusion

We describe new connectivity reconstruction approach that utilizes random or pseudo-random sampling of neural connectivity with anatomical fluorescent synaptic markers localized genetically or otherwise to different parts of a neural circuit. Synaptic connectivity matrix then can be recovered statistically from a collection of fluorescent measurements that can be obtained with such probe. While conventional reconstruction approaches focus on finding individual synapses in neuropil and explicitly relating them with remotely positioned neurons, we propose here to collect large sample of simple fluorescent measurements of connectivity and combine them using minimal model assumptions to determine configuration of the probed circuit.

The key contribution of this paper is to recognize that physical connectivity in a neural circuit can be recovered using a sample of simple observations produced with anatomical fluorescent synaptic markers such as GRASP. We observe that certain type of fluorescent measurements that can be obtained with such markers can be mathematically represented as particular linear sums over connectivity matrix. This allowed us to propose an algorithm to extract connectivity matrix from such data accurately and efficiently. *Compressive Sensing* (CS), in particular, is especially attractive framework for such analysis since neural connectivity is expected to be sparse: in *C. elegans*, e.g., total number of connections is ~2500 out of total ~90,000 and sparseness is 0.03; in *Drosophila* the number of connections per neuron is ~100 out of total ~100,000 and sparseness is $\sim 10^{-3}$, etc. CS algorithm is mathematically tractable and provably nearly optimal for reconstructions of sparse signals from ensembles of linear measurements.

The method that we proposed potentially can allow high speed, high detail connectome reconstructions in many model organisms. The advantages of our method are: the method relies on relatively easy to obtain measurements performed with light microscopy, furthermore, the measurements are simple - total count or total fluorescence strength of all synapses labeled in one animal and the expression pattern for affected neural nuclei; analysis of data is computationally straightforward and uses well understood methodologies; tracing of neural projections over macroscopic distances or other explicit association of individual synapses with remote neurons is not required; individual neurons or small groups of neurons need not be precisely targeted for highly detailed connectivity matrix. Practically meaningful strategies are available for larger organisms such as *Drosophila*, including use of multiplexed fluorescent synaptic markers (Mishchenko, 2010), controlled targeting protocol using libraries of genetic lines (Luo et al., 2008), or using probes such as voltage sensitive dyes (Baker et al., 2005). The method yields average connectivity matrix for a population of animals as well as variability in the connectivity that can be calculated from the same data, e.g., using bootstrapping – producing reconstructions from different parts of the same dataset.

Among the disadvantages of presented method is its relatively high sensitivity to noise. In particular, stringent conditions on data quality should be maintained if one wishes to recover

useable reconstructions with near-minimal set of measurements. Discussion that we presented here is theoretical and no actual implementations or experimental data are shown. Although we consulted to the fullest degree with the latest literature and experimentalists working with *C. elegans* to realistically assess feasibility and practical implementations of our approach (Livet et al., 2007; Micheva and Smith, 2007; Feinberg et al., 2008; Long et al., 2008; Bargmann, Personal communication; Kerr, Personal communication), impact of a number of different specific experimental factors remains unknown and cannot be assessed due to lack of experimental data. Specifically, it is not yet known quantitatively how specific GRASP labeling can be, how effective GRASP can be in inhibitory vs. excitatory synapses, how well GRASP can be expressed pan-neuronally, and how uniform such expression can be, etc. Unfortunately, answering such questions is beyond limited capabilities of the author, who does not have access to experimental facilities or resources needed to carry out extensive experimental program associated with answering these questions. While recognizing the many uncertainties that remain, the author believes that the data available in the literature today supports a very favorable view on the prospects for practical implementation and utilization of described approach (Livet et al., 2007; Micheva and Smith, 2007; Feinberg et al., 2008; Long et al., 2008; Bargmann, Personal communication; Kerr, Personal communication).

Our work opens many exciting possibilities for new data-rich, quantitative studies of neural connectivity in large neural circuits. In *C. elegans*, in particular, described approach allows reconstruction of complete wiring diagram in the span of 1-7 days using existing genetic and off-the-shelf fluorescent light microscopy tools, and straightforward data processing. Both the average connectivity matrix and variability in the connectivity matrix can be estimated at the same time, e.g., using bootstrapping. For comparison, alternative approach of serial electron microscopy would require at least 1-2 years of imaging and analysis to obtain reconstruction of just a single neural circuit of similar size, employing sophisticated and expensive high-end imaging equipment and complex and fragile image understanding algorithms. Described approach can be also employed in larger organisms such as *Drosophila* using different connectivity probes and/or genetic targeting techniques.

Acknowledgements

The author would like to acknowledge the essential contribution from Shiv V. Vitaladevuni, and many discussions with Max Nikitchenko, Liam Paninski, and Veit Elser.

References

Baker BJ, Kosmidis EK, Vucinic D, Falk CX, Cohen LB, Djurisic M, Zecevic D (2005) Imaging brain activity with voltage- and calcium-sensitive dyes. Cellular and Molecular Neurobiology 25:245-282.

Bargmann CI (Personal communication).

Bohland J et al. (2009) A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale. arXiv:0901.4598.

Boyd S, Vandenberghe L (2004) Convex Optimization: Oxfor University Press.

Bregman LM (1965) The method of successive projection for finding a common point of convex sets. Soviet Math Dokl 6:688-692.

Briggman KL, Denk W (2006) Towards neural circuit reconstruction with volume electron microscopy techniques. Curr Opin Neurobio 16:562.

Broome BM, Jayaraman V, Laurent G (2006) Encoding and decoding of overlapping odor sequences. Neuron 51:467-482.

Bureau I, Shepherd GM, Svoboda K (2004) Precise development of functional and anatomical columns in the neocortex. Neuron 42:789-801.

Candes EJ, Romberg J (2005) Practical signal recovery from random projections. In: SPIE Conference on Wavelet Applications in Signal and Image Processing XI.

Candes EJ, Wakin M (2008) An introduction to compressive sampling. IEEE Signal Processing Magazine 25:21-30.

Candes EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory 52:489-509.

Donoho D (2006) Compressed sensing. IEEE Trans on Information Theory 52:1289-1306.

Feinberg EH, Vanhoven MK, Bendesky A, Wang G, Fetter RD, Shen K, Bargmann CI (2008) GFP Reconstitution Across Synaptic Partners (GRASP) defines cell contacts and synapses in living nervous systems. Neuron 57:353-363.

Hagmann P, Kurant M, Gigandet X, Thiran P, Wedeen V, Meuli R, Thiran J-P (2007) Mapping Human Whole-Brain Structural Networks with Diffusion MRI. PLoS ONE 2:e597.

Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey C, Wedeen V, Sporns O (2008) Mapping the Structural Core of Human Cerebral Cortex. PLoS Biology 6:e159.

Helmstaedter M, Briggman KL, Denk W (2009) 3D structural imaging of the brain with photons and electrons. Curr Opin Neurobio:Epub.

Ikegaya Y, Le Bon-Jego M, Yuste R (2005) Large-scale imaging of cortical network activity with calcium indicators. Neuroscience Research 52:132-138.

Jain V, Murray J, Roth F, Turaga S, Zhigulin V, Briggman K, Helmstaedter M, Denk W, Seung H (2007) Supervised learning of image restoration with convolutional networks. . In: International Conference on Computer Vision, pp 1-8.

Jones LM, Fontanini A, Sadacca BF, Miller P, Katz DB (2007) Natural stimuli evoke dynamic sequences of states in cortical ensembles. Proceedings of the National Academy of SciencesJournal of Neuroscience 104:18772-18777.

Jurrus E, Tasdizen T, Koshevoy P, Fletcher P, Hardy M, Chien C, Denk W, Whitaker R (2006) Axon tracking in serial block-free scanning electron microscopy. In: MICCAI 2006 workshop. Copenhagen.

Kerr R (Personal communication).

Lichtman JW, Livet J, Sanes JR (2008) A technicolour approach to the connectome. Nat Rev Neurosci Epub ahead of print.

Livet J, Weissman TA, Kang H, Draft RW, Lu J, Bennis RA, Sanes JR, Lichtman JW (2007) Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature 450:56-62.

Long F, Peng H, Liu X, Kim S, Myers E (2008) Automatic recognition of cells (ARC) for 3D images of C. Elegans. In: Lecture Notes in Computer Science: Research in Computational Molecular Biology, pp 128-139: Springer Berlin

Lu J, Fiala JC, Lichtman JW (2009a) Semi-automated reconstruction of neural processes from large numbers of fluorescence images. PLoS ONE 4:e5655.

Lu J, Tapia J, White O, Lichtman JW (2009b) The interscutularis muscle connectome. PLoS Biology 7:e1000108.

Luo L, Callaway EM, Svoboda K (2008) Genetic dissection of neural circuits. Neuron 57:634-660.

Macke J, Maack N, Gupta R, Denk W, Scholkopf B, Borst A (2008) Contour-propagation algorithms for semi-automated reconstruction of neural processes. J Neurosci Methods 167:349-357.

Meinertzhagen IA, Sorra KE (2001) Synaptic organisation in the fly's optic lamina: few cells, many synapses and divergent microcircuits. Progress Brain Research 131:53-69.

Micheva KD, Smith SJ (2007) Array tomography: a new tool for imaging the molecular architecture and ultrastructure of neural circuits. Neuron 55:25-36.

Mishchenko Y (2009) Automation of 3D reconstruction of neural tissue from large volume of conventional Serial Section Transmission Electron Micrographs. Journal of Neuroscience Methods 176:276-289.

Mishchenko Y (2010) On optical detection of densely labeled synapses in neuropil and mapping connectivity with combinatorially multiplexed fluorescent synaptic markers. PLoS ONE 5:e8853.

Mishchenko Y, Paninski L (2011) Bayesian approach to reconstruction of neural connectivity using arrays of direct fluorescent connectivity probes. Preprint.

Mishchenko Y, Vogelstein J, Paninski L (2010) A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. Annals of Applied Statistics.

Petreanu L, Huber D, Sobczyk A, Svoboda K (2007) Channelrhodopsin-2-assisted circuit mapping of long-range callosal projections. Nature neuroscience 10:663-668.

Phelps CB, Brand AH (1998) Ectopic gene expression in Drosophila using GAL4 system. Methods 4:367-379.

Pillow J, Shlens J, Paninski L, Sher A, Litke A, Chichilnisky EJ, Simoncelli E (2008) Spatiotemporal correlations and visual signaling in a complete neuronal population. Nature 454:995-999.

Romberg J (2008) Imaging via compressive sampling. IEEE Signal Processing Magazine 25:14-20.

Sarkar M, Magliery TJ (2008) Re-engineering a split-GFP reassembly screen to examine RING-domain interactions between BARD1 and BRCA1 mutants observed in cancer patients. Mol Biosyst 4:599-605.

Shepherd GM, Svoboda K (2005) Laminar and columnar organization of ascending excitatory projections to layer 2/3 pyramidal neurons in rat barrel cortex. Journal of Neuroscience 25:6037-6046.

Smith SJ (2007) Circuit reconstruction tools today. Curr Opin Neurobiol 17:601-608.

Song S, Sjostrom PJ, Reiql M, Nelson S, Chklovskii DB (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biology 3:e68.

Sporns O, Tononi G, Kotter R (2005) The Human Connectome: A Structural Description of the Human Brain. PLoS Computational Biology 1:e42.

Stevenson IH, Rebesco JM, Hastopoulos NG, Haga Z, Miller LE, Kording KP (2009) Bayesian inference of functional connectivity and network structure from spikes. IEEE Trans Neural Systems and Rehab 17:203-213.

Svoboda K (Personal Communication) Cajal 2.0 Cajal 2.0 optical mapping of mouse brain. In.

Vanderbei RJ (2001) Linear Programming: Foundations and Extensions. New York: Springer-Verlag.

White J, Southgate E, Thomson JN, Brenner S (1986) The structure of the nervous system of the nematode caenorhabditis elegans. . Philosophical Transactions of Royal Society London Series B, Biological Sciences 314:1-340.

Wright S (1997) Prima-dual interior-point methods: SIAM.

Figures Legends and Tables Legends

Figure 1: Schematic description of GRASP system for visualization of synapses in vivo (Feinberg et al., 2008) and Cre/Lox system for stochastic expression of genes (Livet et al., 2007). A) Schematic description of GRASP. Two fragments of a split-GFP (sA and sB) are expressed separately at post-synaptic (left) and pre-synaptic (center) sites of different neurons. Separately,

split-GFP fragments do not produce fluorescence. At the location of synapses split-GFP can recombine into a functional GFP molecule and produce fluorescence, thus, fluorescently tagging synapses (right). B) Schematic description of Cre/Lox system for stochastic expression of genes. A cassette with two genetic sequences for two different wavelength fluorochromes (red and green, left) in mutually inverted orientations is introduced into the genome and flanked with inversely-oriented loxP-sites. When Cre-recombinase is introduced into cells, recombinase reacts with loxP-sites in such a way that orientation of the flanked sequence is flipped at certain rate. When Cre is removed, the sequence may be found in either original or reversed orientation (right). When fluorochrome sequences in the cassette are transcribed, only those in the direct orientation can be produced successfully. Thus, different cells start producing either green or red fluorochromes at random.

Figure 2: Schematic description of stochastic GRASP. Stochastic GRASP combines GRASP from Figure 1A with Cre/Lox system from Figure 1B to allow neurons express GRASP stochastically. A) In first implementation a cassette with two inversely oriented sequences, encoding pre- and post-synaptic GRASP fragments (sA and sB) and two associated nuclei-bound fluorescent proteins (nA and nB), is introduced into the genome and flanked with inversely-oriented loxP-sites. Effect of Cre is to randomize the orientation of such cassette in different neurons leading to two recombination outcomes: (i) post-synaptic GRASP and associated nuclei XFP are expressed tagging post-synaptic sites and nuclei of such neurons and (ii) pre-synaptic GRASP and associated nuclei XFP are expressed. A promoter (p) can be used to additionally restrict cassette expression to a defined population of neurons. B) In second implementation a cassette with two loxP-flanked sequences for pre- and post-synaptic GRASP fragments is introduced into the genome. Effect of Cre is to randomize orientation of both sequences leading to four recombination outcomes: (i) pre- and post-synaptic GRASP are expressed together in a neuron, (ii) post-synaptic GRASP is expressed only, (iii) pre-synaptic GRASP is expressed only, and (iv) neither of the GRASP fragments is expressed. C) When two neurons form a synapse while expressing GRASP complementary, a fluorescent punctum is produced by the recombined split-GFP. Nuclei of the cells expressing GRASP are labeled with associated nuclei XFP and also can be observed.

Figure 3: Schematic description of neural connectivity reconstruction using stochastic GRASP. A) With stochastic GRASP random neurons in a circuit express GRASP in different animals. Three random expression patterns (phenotypes) are shown for illustration. In each pattern all synapses formed by "activated" neurons produce same-color fluorescent puncta that can be observed with light microscope. An overall measure such as total count or combined size of all labeled puncta is produced. B) For reconstruction of the connectivity matrix it is sufficient to collect the above overall measurements, n, along with pre- and post-synaptic GRASP expression patterns, a and b, in a sample of stochastic GRASP phenotypes (A). C) Sample (n,a,b) can be mathematically represented as a sample of linear measurements over the connectivity matrix. Specifically, n can be related to certain sums of connectivity matrix elements that correspond to the intersection of the rows and columns corresponding to the neurons expressing pre- and post-synaptic GRASP. We recognize in these settings an instance of Compressive Sensing problem (Donoho, 2006; Candes and Wakin, 2008) that allows tractable recovery of synaptic connectivity matrix from a set of such data.

Figure 4: Result of a hypothetical neural connectivity reconstruction experiment in *C. elegans* simulated using real wiring diagram for that animal available from electron microscopy (White et al., 1986). Shown are the scatter plots of reconstructed connectivity weights versus that for true connectivity weights. Reconstructions with 4000-6000 measurements can be already practically meaningful and the reconstruction with 10,000 measurements is exact.

Figure 5: Result of a hypothetical neural connectivity reconstruction experiment in *C. elegans* simulated using real wiring diagram available from electron microscopy (White et al., 1986). Shown is the correlation coefficient square, r^2, for true and reconstructed connectivity weights as the function of the number of measurements K. Reconstructions from 4000-6000 measurements are already practically meaningful and reconstruction from 10,000 measurements is exact. No significant difference in performance is observed when using exclusive (Fig. 2A) or non-exclusive (Fig. 2B) stochastic GRASP construct. These results allow us to estimate that complete wiring diagram in *C.Elegans* can be re-obtained using stochastic GRASP in the span of 1-7 days. Result of a similar reconstruction using simpler L_2 regularized algorithm is shown to emphasize dramatic improvement conferred by Compressive Sensing algorithm.

Figure 6: Impact of biological variability on the reconstruction of neural connectivity using stochastic GRASP. Biological variability describes possible variation in the connection matrix from one animal to another. Different degrees of variability, from 0 (no variability) to 100% (connection weights in different animals are purely random with a Poisson statistics) are shown. In all cases the impact of biological variability is found to be insignificant.

Figure 7: Impact of added noise in the measurements on the reconstruction of neural connectivity using stochastic GRASP. Although reconstructions are stable (i.e., they do not break down in the presence of small amounts of noise), the sensitivity is substantial and only up to 4-5% of added noise can be tolerated at the minimal $K \approx 10{,}000$. Larger sample size can be used to mitigate the impact of higher levels of added noise.

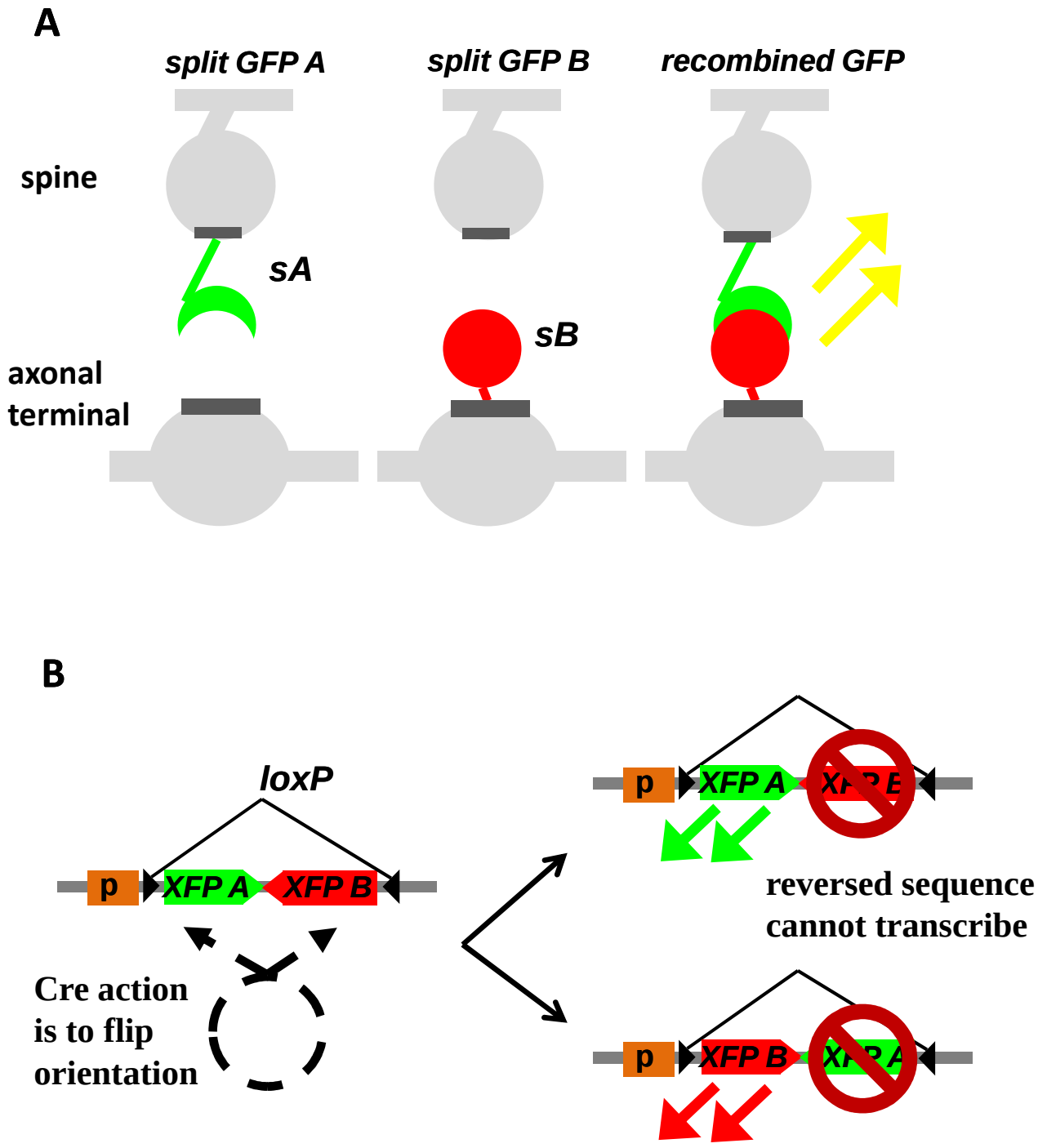
Figure 8: Impact of errors in GRASP expression patterns on the reconstruction of neural connectivity using stochastic GRASP. Although reconstructions are stable, the sensitivity is again high with only up to 5-6% of errors that can be tolerated at the minimal $K \approx 10{,}000$. Larger sample size can be used to mitigate the impact of such errors at higher rates.

Figure 9: Comparison of two other prospective approaches for connectome reconstructions, namely, serial electron microscopy (EM, panels A and B) and en-mass pair-wise cell recordings (PCR, panels C and D). Reconstructions are characterized by the amount of reconstruction effort, E, where $E=1$ corresponds to the effort necessary to obtain one complete reconstruction (i.e. image and analyze one full circuit). In *C. elegans* we estimate that $E=1$ for EM corresponds currently to ~1-2 man-years of work and for PCR ~3-30 man-years of work. A) Quality of EM reconstruction as the function of reconstruction effort in an ideal case (i.e., no noise, compare Figure 5). B) Quality of EM reconstruction as the function of reconstruction effort in the presence of noise (e.g., due to imaging or reconstruction mistakes, compare Figure 7 and 8). "Noise" strength is characterized by the probability of an error in an axonal trace between two subsequent synapses. Due to long reach of such point errors EM reconstructions are vulnerable to small amounts of such errors. C) Quality of PCR reconstruction as the function of the reconstruction effort in an ideal case. D) Quality of PCR reconstruction as the function of

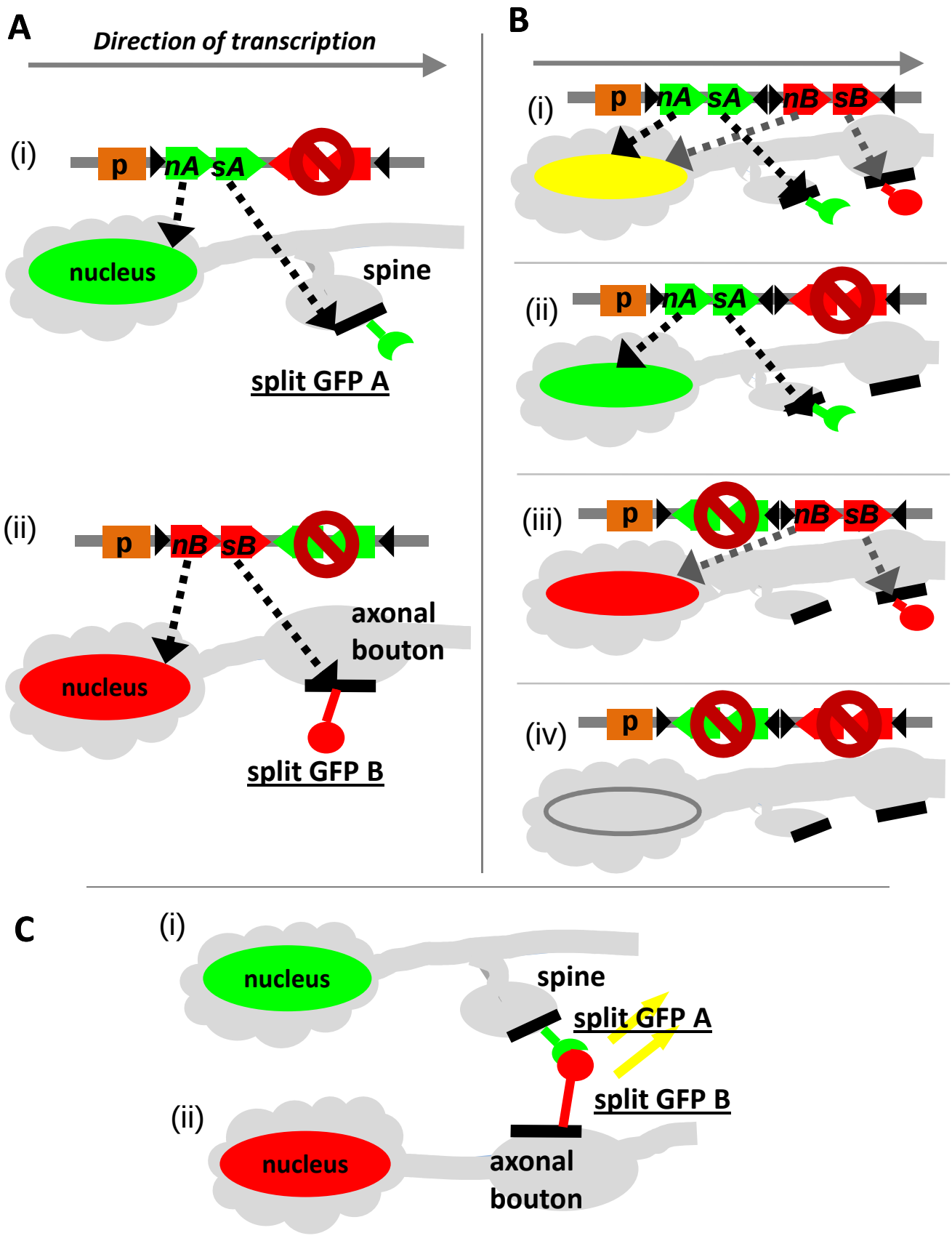
reconstruction effort in the presence of noise. “Noise” strength is characterized by the probability of making an error in detecting a connection between two cells (e.g., failure to stimulate a cell or failure to observe response from a cell). PCR reconstructions are found to be very robust to such point errors and can successfully tolerate large amounts of noise.

Table 1: Libraries of genetic lines such as Gal4 lines in *Drosophila* can provide a natural classification and identification system for neural circuits. Each neuron can be identified by the set of genetic lines in which it appears, i.e. for each neuron one can associate a pattern of 1’s and 0’s corresponding to the sets of genetic lines containing it. E.g., a pattern 110… identifies a neuron that is present in lines *A* and *B* but not line *C*, etc. Neurons that are present always together in the same genetic lines comprise groups of indistinguishable, equivalent neurons within such identification system. Such groups can be accessed and manipulated via associated combinations of genetic lines and the methods from (Luo et al., 2008). This allows conducting experiments involving such neurons and studying their properties even if it is not possible to locate and identify such neurons anatomically.

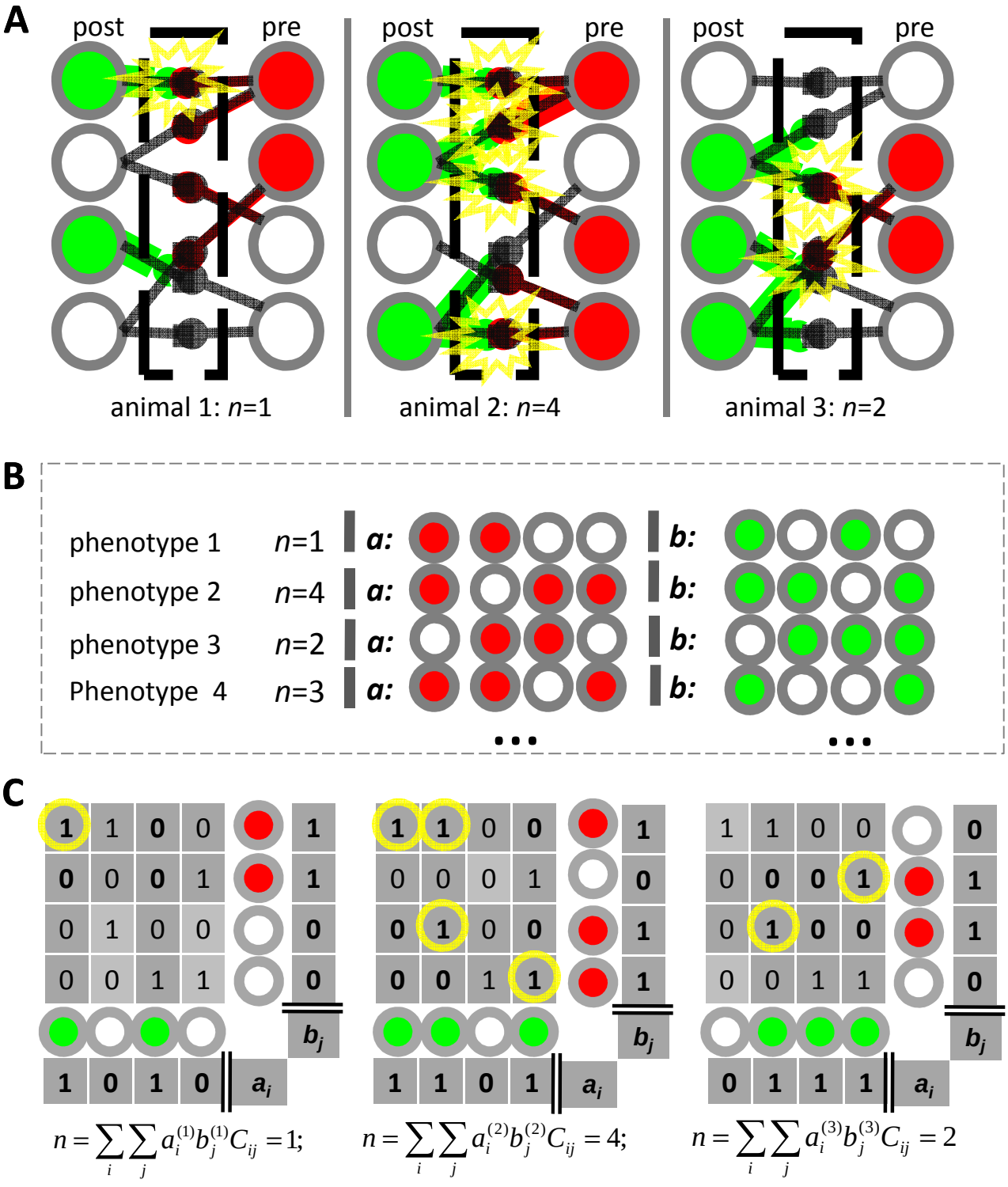
Mishchenko Figure 1



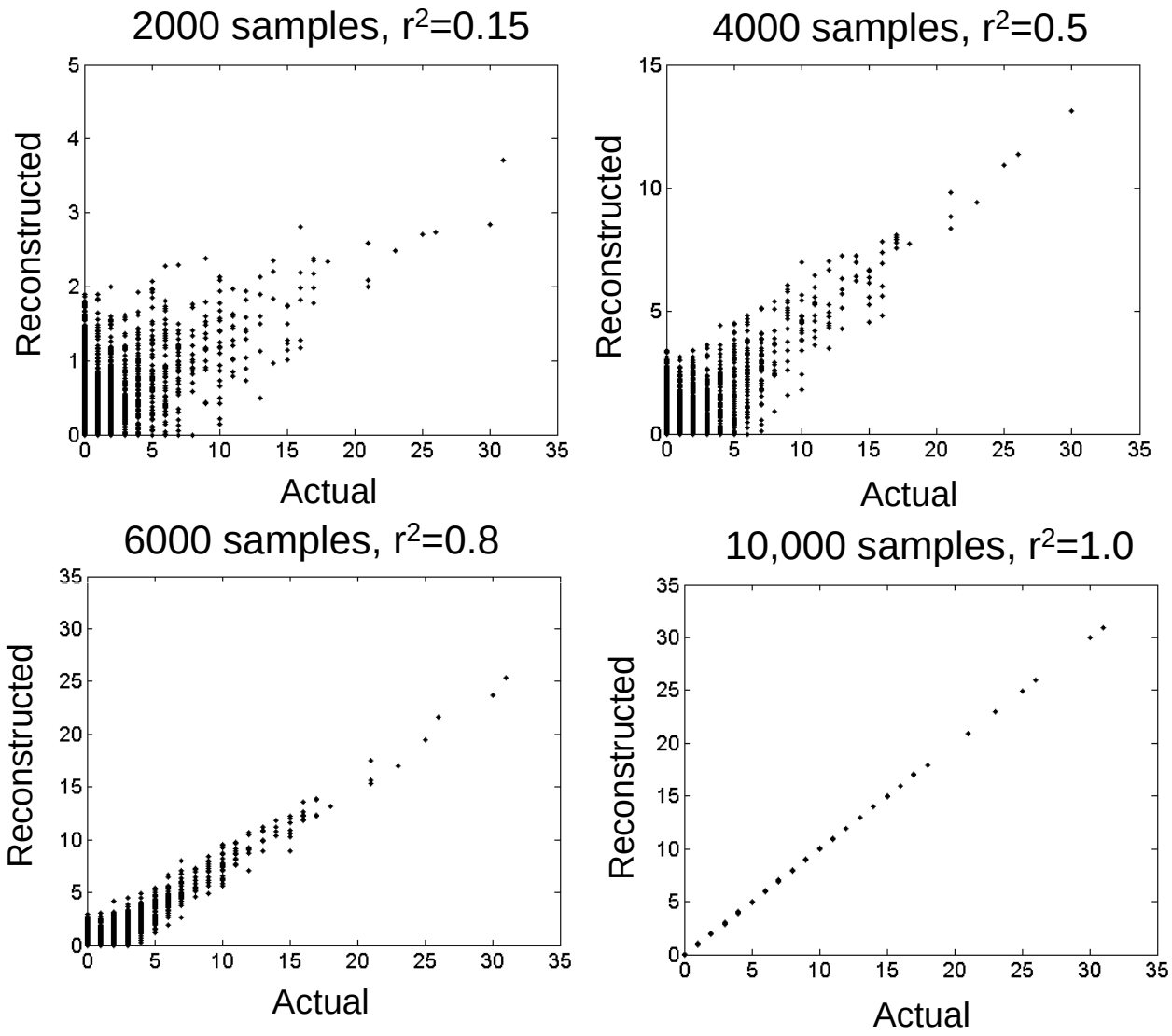
Mishchenko Figure 2



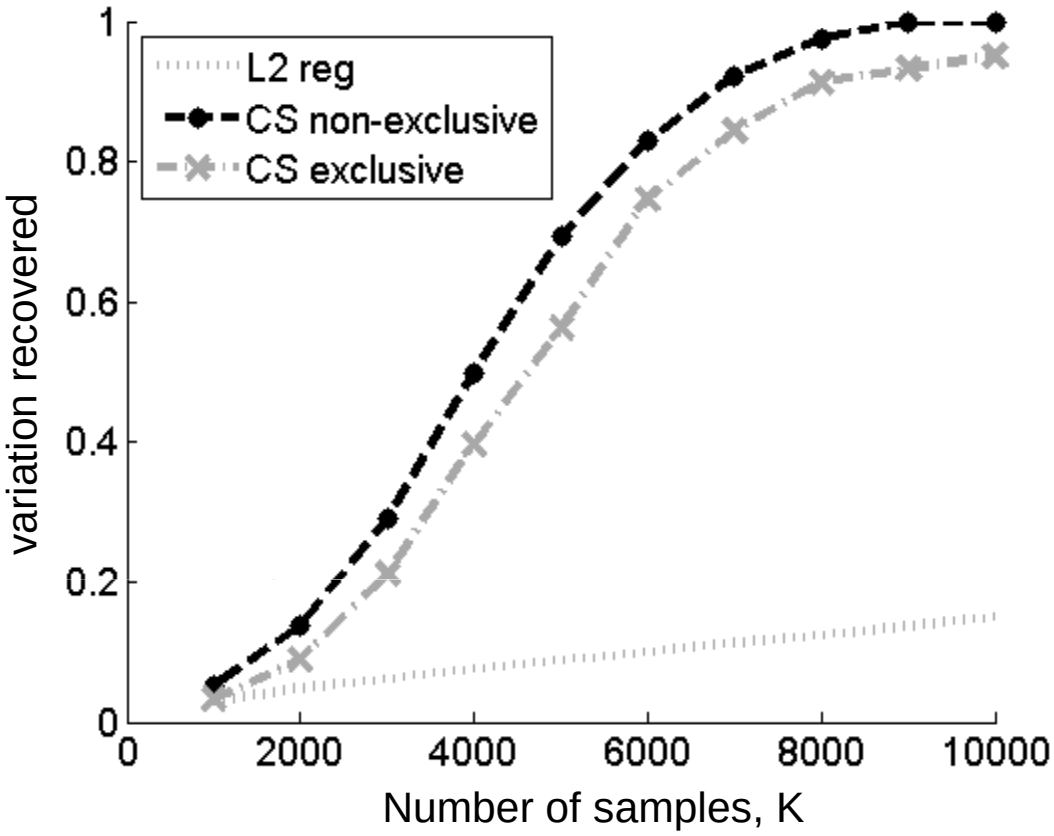
Mishchenko Figure 3



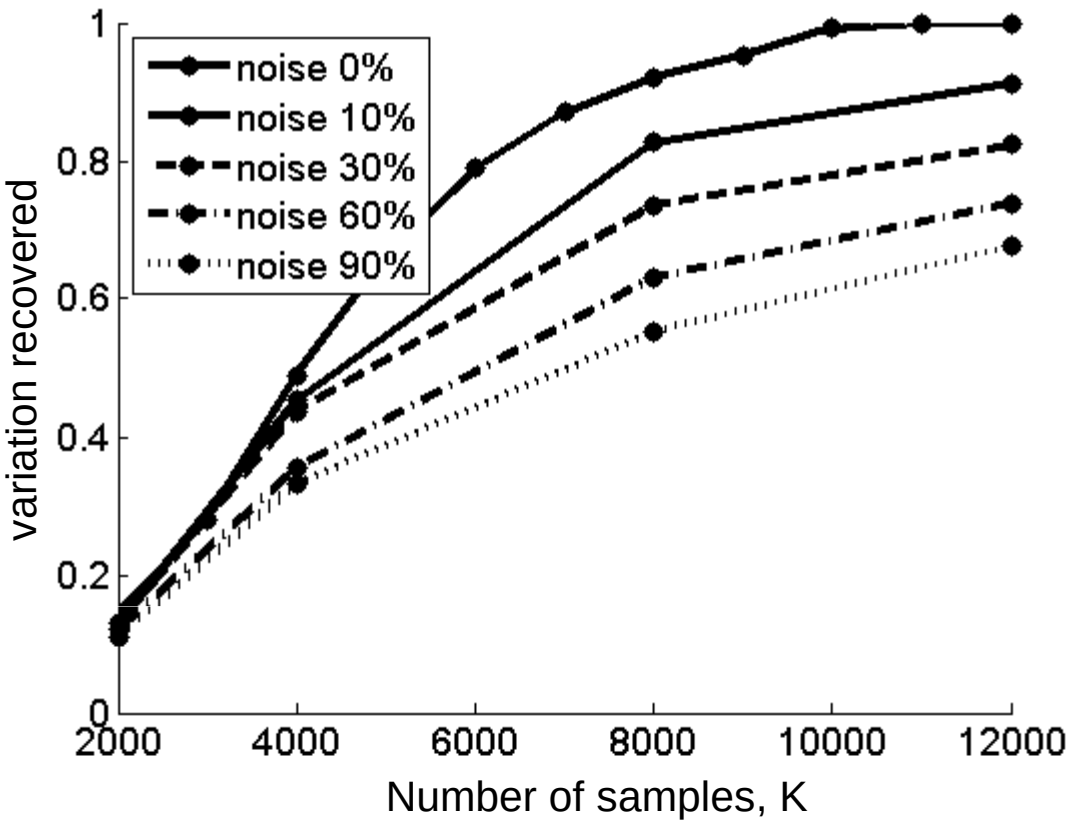
Mishchenko Figure 4



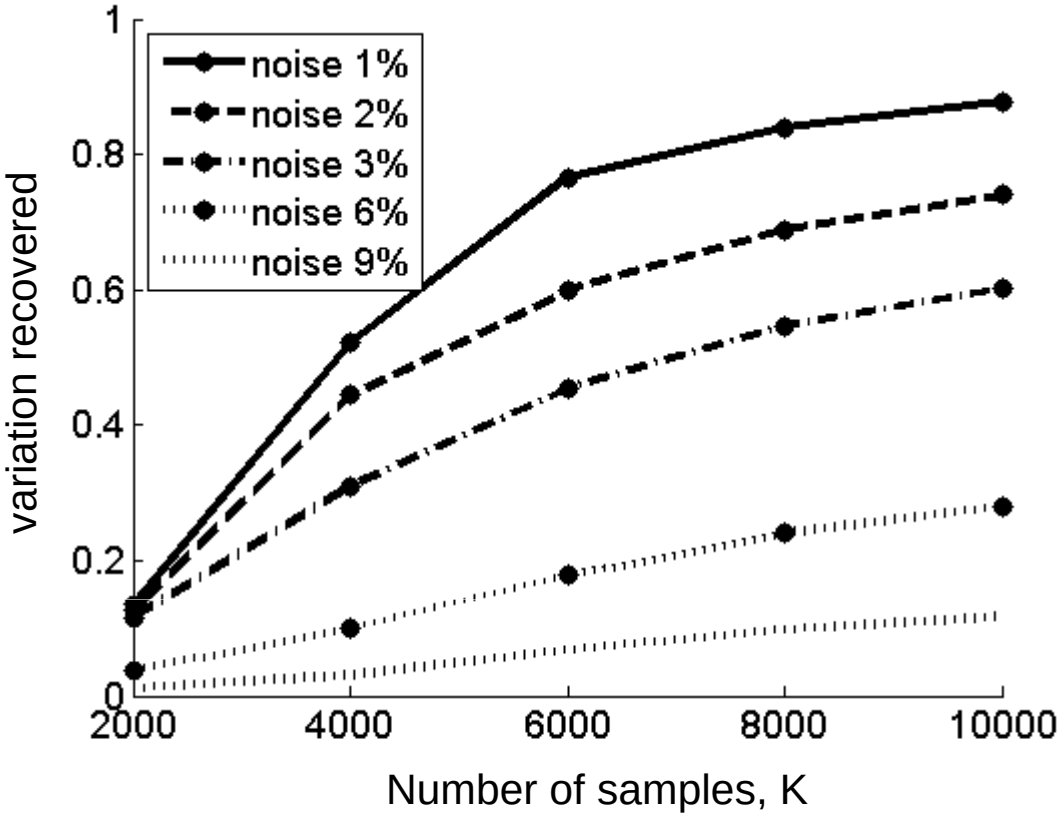
Mishchenko Figure 5



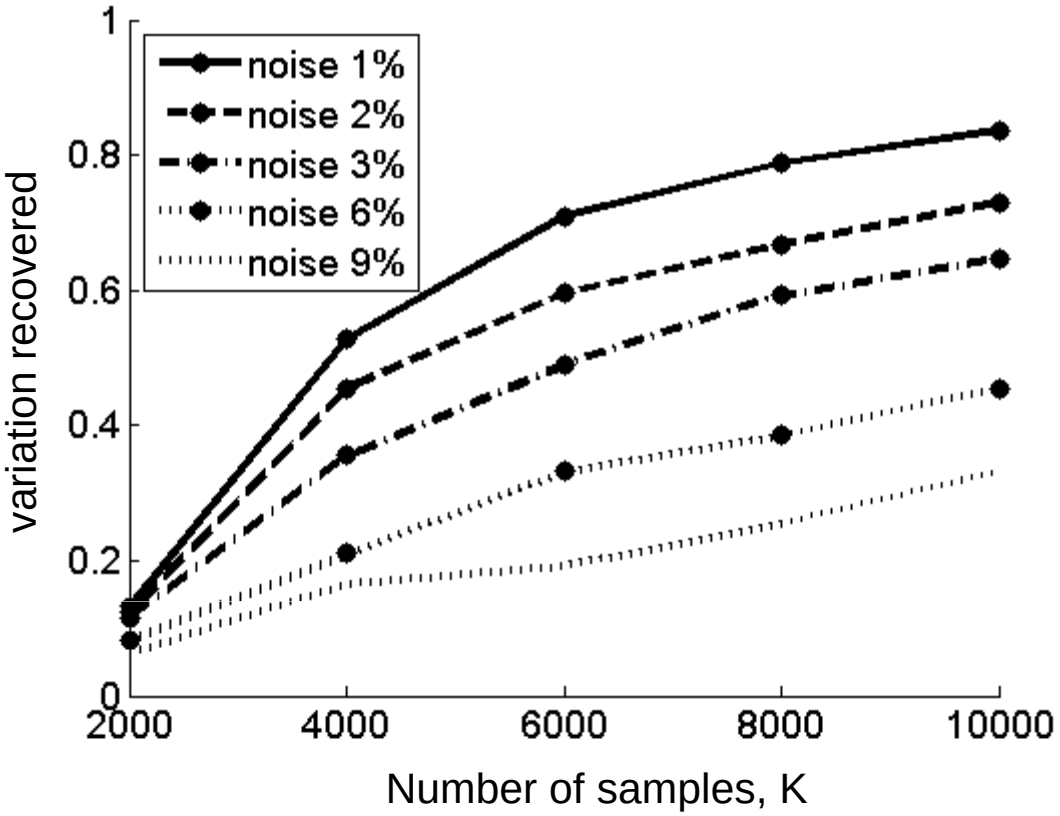
Mishchenko Figure 6



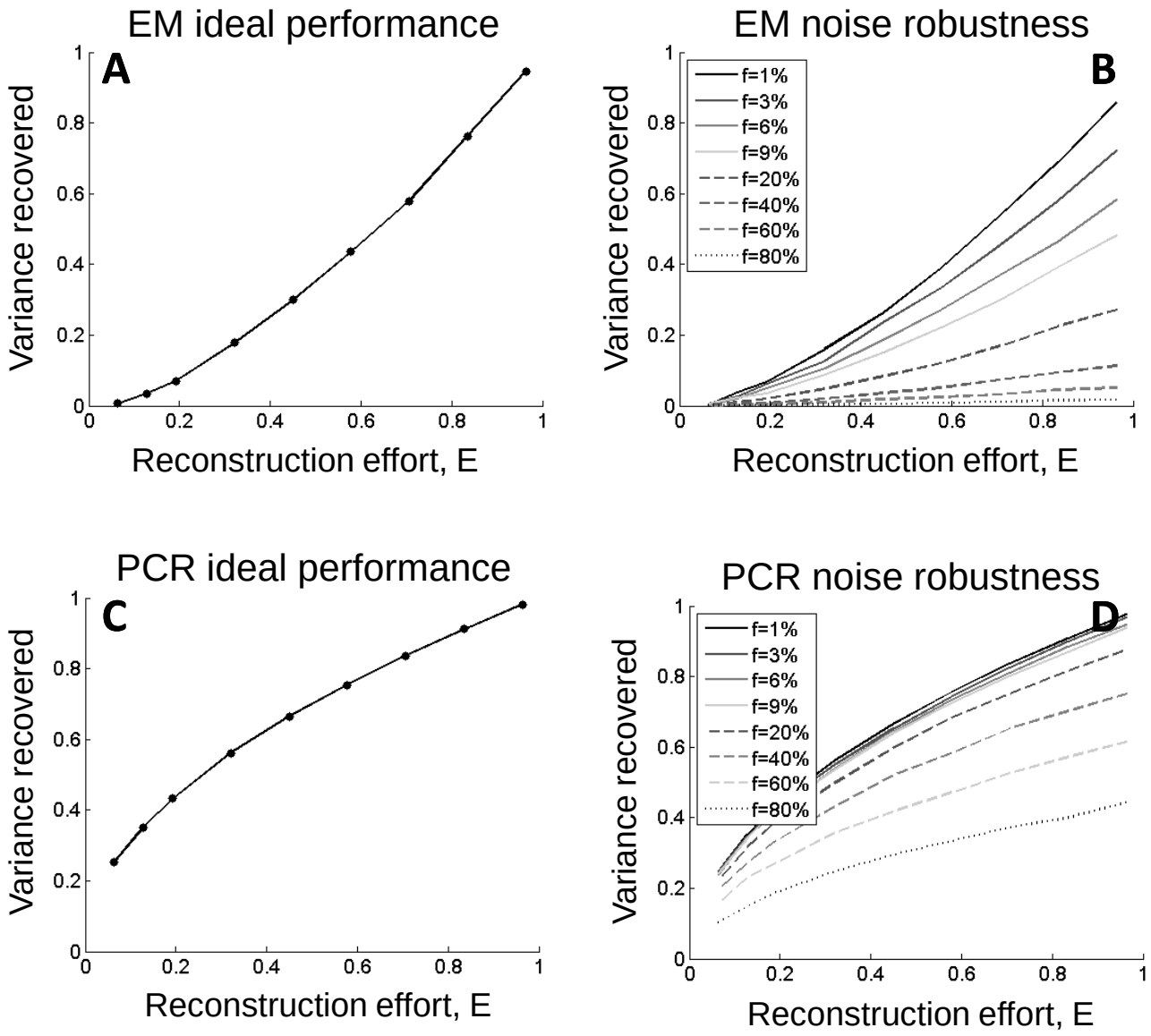
Mishchenko Figure 7



Mishchenko Figure 8



Mishchenko Figure 9



Mishchenko Table 1

Groups of neurons	Present in expression pattern of…			Code
	line A	*line B*	*line C*	
ABC…	yes	yes	yes	111…
BC…	no	yes	yes	011…
AC…	yes	no	yes	101…
AB…	yes	yes	no	110…
A…	yes	no	no	100…
B…	no	yes	no	010…
C…	no	no	yes	001…
…	no	no	no	000…