TECHNICAL REPORTS

nature neuroscience

High-accuracy neurite reconstruction for high-throughput neuroanatomy

Moritz Helmstaedter, Kevin L Briggman & Winfried Denk

Neuroanatomic analysis depends on the reconstruction of complete cell shapes. High-throughput reconstruction of neural circuits, or connectomics, using volume electron microscopy requires dense staining of all cells, which leads even experts to make annotation errors. Currently, reconstruction speed rather than acquisition speed limits the determination of neural wiring diagrams. We developed a method for fast and reliable reconstruction of densely labeled data sets. Our approach, based on manually skeletonizing each neurite redundantly (multiple times) with a visualization-annotation software tool called KNOSSOS, is ~50-fold faster than volume labeling. Errors are detected and eliminated by a redundant-skeleton consensus procedure (RESCOP), which uses a statistical model of how true neurite connectivity is transformed into annotation decisions. RESCOP also estimates the reliability of consensus skeletons. Focused reannotation of difficult locations promises a rather steep increase of reliability as a function of the average skeleton redundancy and thus the nearly error-free analysis of large neuroanatomical datasets.

The reconstruction of neuronal circuits has been a central approach toward understanding the function of the nervous system since some of the earlier studies[1,2]. Whereas many neurons extend over tens of centimeters, the caliber of thin neurites can be as small as 40 nm at spine necks[3]. This range of length scales challenges any method aimed at the extraction of neuron morphology from the data. For sparsely stained tissue with only a small fraction of all neurons labeled, such as in the Golgi method[2] or selective dye injection[4,5], imaging techniques operating at a resolution of ~1 µm are sufficient to follow all processes. This holds true even if the neurite caliber is much less than the imaging resolution, because in very sparsely stained data the identity of each neurite is easily established. Manual reconstructions of individual neurons from such data are therefore assumed to be highly reliable, even though little validation of this reliability has been reported. Almost all available neuroanatomical data at single-cell resolution stem from such experiments, but as fluorescence imaging data from samples with a much higher staining density (hundreds of neurons per 1 mm^3, labeled using various genetic or virus-based techniques[6,7]) are becoming available, high reconstruction reliability can no longer be presumed.

For the reconstruction of complete cellular wiring diagrams, also known as connectomes[8,9], assuring reconstruction reliability is even more difficult because the morphologies of all neurons must be extracted. This may eventually be possible at light-microscopic resolution by staining all neurons with a sufficient number of distinguishable colors[7,9], but it otherwise requires imaging at a resolution high enough to follow all neurites in densely packed neuropil[10]. Such a reconstruction has been done for the 302-neuron nervous system of the nematode *Caenorhabditis elegans*[11] using serial-section electron microscopy.

Recently developed techniques for automated volume electron microscopy[12–15] enable the imaging of volumes large enough to contain more complex neural circuits[16]. However, extracting information about neuron morphology and circuit structure from such data poses two major challenges. First, the total neurite path length in many neural circuits is typically in the range of meters (at least 0.3 m for small circuits such as a 100 × 100 × 100 $µm^3$ region of retina, and as much as 400 m for a mouse cortical column[10]). Using currently available software tools for neurite contouring (for example, Reconstruct[17]) the complete analysis of such circuits is very slow and thus prohibitively expensive. Contouring every neurite for a path length of 0.3 m would require an estimated 60,000 h (30 person-years) of annotation time. Reconstruction accuracy is the second major concern. For sparsely stained data the selectivity of the stain makes following the neurites easy, but connectomic reconstruction requires many decisions (as many as one every ~4 µm in the retina) about whether to continue, branch or terminate a neurite. Some of these decisions are difficult and, because they must be made constantly while annotating, their reliability depends on the uninterrupted attentiveness of the human annotator. Synapses must also be identified with sufficient accuracy.

Here, we describe a set of tools that substantially improve both the speed and accuracy of neurite reconstruction. We chose to annotate the data by following a single core line along the inside of each neurite, creating a 'skeleton' representation of each neuron's morphology. Using the KNOSSOS software tool, which we developed for the convenient browsing and annotation of large data sets, we observed a 50-fold (range 20–130-fold) increase in the amount of neurite path length reconstructed per unit time. We quantified discrepancies between multiple skeletons of the same neurite and, on the basis of their distribution, optimized the correction of errors and the creation of a consensus skeleton, which is a bundle of closely spaced skeleton pieces. Our method, RESCOP (**Fig. 1**), uses multiple redundant annotations to increase reliability. We show that the accuracy of the

Max Planck Institute for Medical Research, Heidelberg, Germany. Correspondence should be addressed to M.H. (moritz.helmstaedter@mpimf-heidelberg.mpg.de).

Received 28 February; accepted 23 May; published online 10 July 2011; doi:10.1038/nn.2868

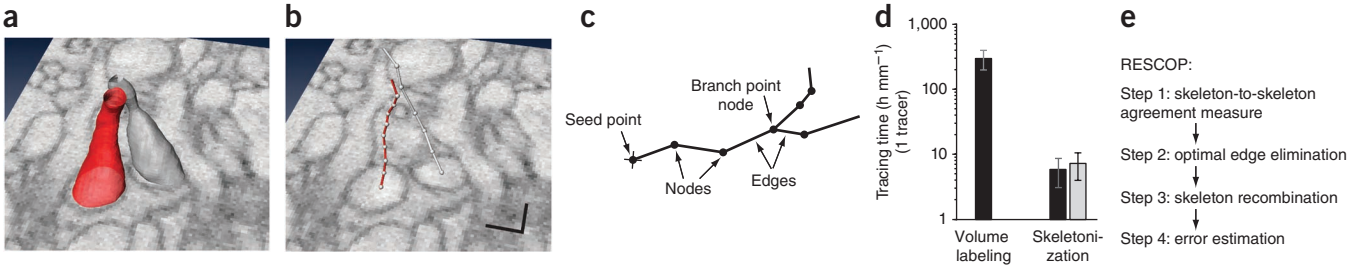

Figure 1 Comparison of volume and skeleton annotation. (**a**,**b**) Examples of volume labeling (**a**) and skeletonization (**b**) for the same two neurite fragments; cell-surface labeled data (data set E1088; Online Methods). Scale bars represent 250 nm. (**c**) Sketch of a neurite skeleton. (**d**) Rate of time consumption for volume labeling[10] and for skeleton annotation (data from this study; annotated using KNOSSOS, see **Supplementary Movie 1**), for both cell surface–labeled data (black) and conventionally stained data set (K0563, gray; see **Fig. 5d**). Error bars represent range for volume labeling and s.d. for skeletonization. (**e**) Outline of RESCOP.

consensus skeleton increases with the number of redundant skeletons, even when only slightly trained annotators are employed. We have used our set of reconstruction tools to skeletonize all rod bipolar cells in a block of mouse retina.

RESULTS

Browsing large-scale electron microscopy data

We first developed a software tool (KNOSSOS; **Supplementary Movie 1** and **Supplementary Fig. 1**) for browsing and annotating large-scale volume data. Such data are generated, for example, by serial block-face electron microscopy (SBEM[12]). At nanometer resolution, imaging volumes large enough to contain entire circuits yield data sets of at least several hundreds of gigabytes. We designed KNOSSOS to make three-dimensional navigating and viewing of such data sets convenient. KNOSSOS allows quick navigation along all axes by selectively loading only the data surrounding the currently viewed location. Neurites can be oriented along any direction in dense neuropil and can often be followed more conveniently using views other than the imaging plane (the block face in SBEM), in particular if the data, as is the case for SBEM, are nearly isotropic in resolution. KNOSSOS therefore displays three orthogonal views of the data (see also V3D[18]), which are essential for navigating along neurites oriented obliquely to the slice plane. KNOSSOS runs smoothly on laptops, even when the data are located on an external hard drive. This allowed us to distribute the workload to many nonexpert annotators (>80 undergraduate students).

Fast neurite reconstruction by skeletonization

To densely reconstruct even a local neuronal circuit, at least several hundred millimeters of neurite need to be correctly followed. This can be done by contouring, or volume labeling, of neurites (**Fig. 1a**). However, contouring is slow (200–400 h per mm neurite length[10]).

KNOSSOS therefore provides a skeletonization mode (**Fig. 1b** and **Supplementary Movie 1**). The user starts at a location within a neuron (called a "seed"), for example the cell body, and places a marker (called a "node," **Fig. 1c**). Then, the user advances through the data along a neurite, and places nodes at intervals of about 7–10 image planes, approximately at the center of the neurite. Notably, the user can move in any of the cardinal directions, and can place nodes in any of the three orthogonal view ports. Sequentially placed nodes are connected by line segments (called "edges," **Fig. 1c**). When a location at which the neurite branches is encountered, the user designates the current node as a branch point, and is later directed back to this branch point after completing one of the branches. Skeletonization allows the user to focus annotation on the core line of a neurite. We found that skeletonization reduced annotation time to 5.9 ± 2.8 h per millimeter path length, which is ~50-fold (range 20–130-fold) faster than fully manual volume labeling (**Fig. 1d** and **Supplementary Fig. 2**).

Discrepancies between skeletons

We next investigated how frequently annotators disagreed when skeletonizing the same neurite, starting from the same initial location. In an overlay of two skeletons generated by two experienced neuroscientists, both starting at the soma of an amacrine cell in a SBEM data set of rabbit retina (data set E1088; see Online Methods), the skeletons disagreed at 12 locations along the dendritic tree, which has a total path length of 0.8 mm (**Fig. 2**). Most of the disagreements (10 of 12) were caused by missed branch points (locations 1, 2, 4 and 6–12; **Fig. 2** and see **Supplementary Image Stacks 1–12**). Upon reinspecting these ten locations, both annotators quickly reached agreement, suggesting that missed branch points had been overlooked. This indicates

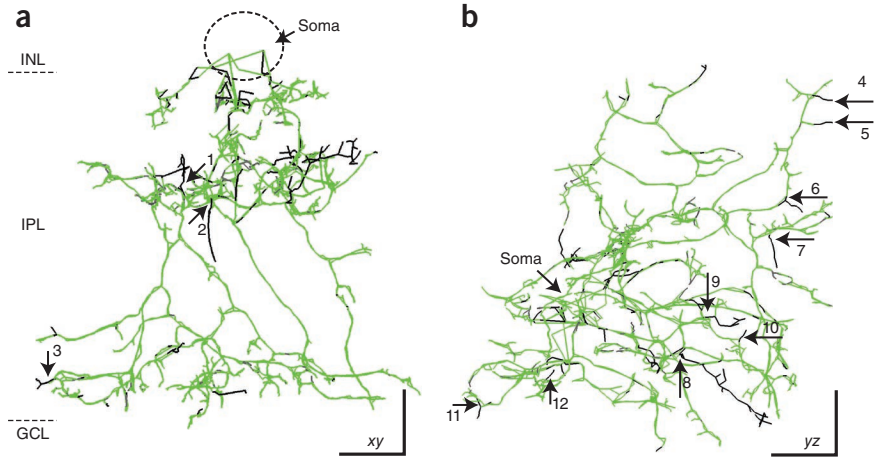

Figure 2 Skeletonization by expert annotators. (**a**) Two complete skeletons of the same amacrine cell annotated independently by M.H. and K.L.B., starting at the soma. (**b**) Same skeletons shown looking onto the plane of the retina. Green, agreement among annotators; black, disagreement; numbers, disagreement locations. Stacks of original data surrounding disagreement locations are shown in **Supplementary Image Stacks 1–12**. INL, inner nuclear layer; IPL, inner plexiform layer; GCL, ganglion cell layer. Scale bars, 5 µm.

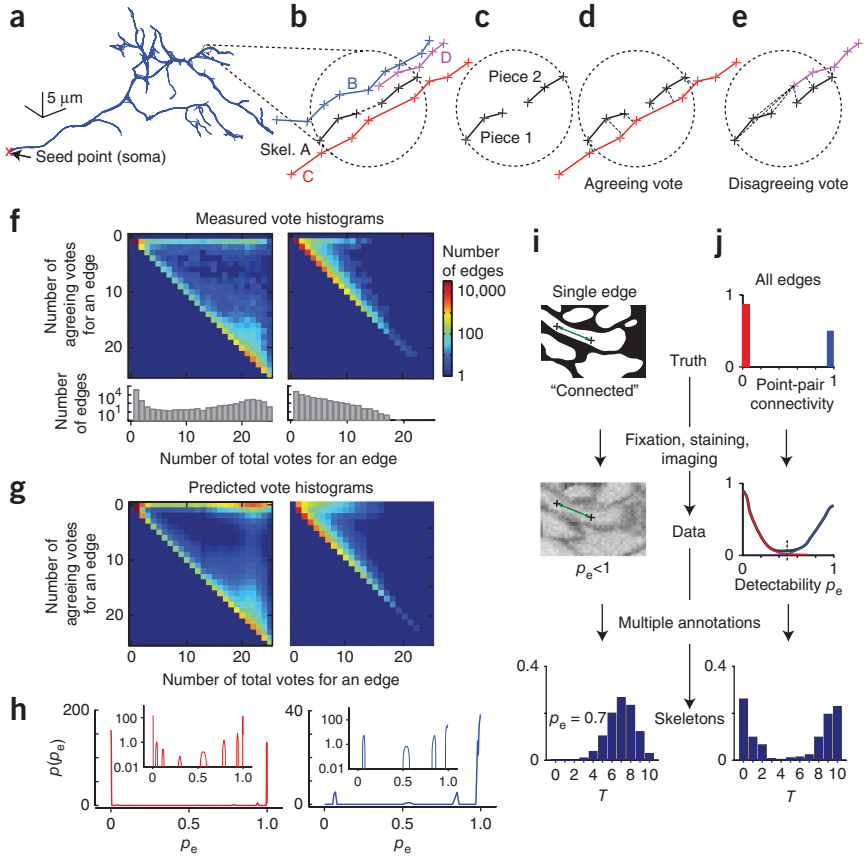

Figure 3 RESCOP step 1, skeleton-to-skeleton agreement measurement. (**a**) Overlay of seven independent skeletons of the same neurite (bipolar cell axon) annotated by slightly trained nonexperts, all starting at the soma (red cross). (**b**–**e**) Schematic of procedure for measuring agreement among multiple annotators for one skeleton edge (dashed line) in skeleton A. (**f**) Histograms of edge votes for 50-fold annotation of one cell (left) and dense skeletonization of 98 neurites (right). Bottom, vote count versus total votes (log scale). Histograms were corrected for multiple counting of the same location; see Online Methods. (**g**,**h**) Predicted vote histograms for single cell (left) and for dense skeletons (right) (**g**), using the distribution of edge detectabilities $p_{fit}(p_e)$ (**h**) that best predicted the respective histograms in **f**. (**i**,**j**) Schematic of how the truth (top) is converted to detection probability (middle). Bottom, probabilities for different T (number of agreeing votes) for one edge (**i**, binomial distribution for $p_e = 0.7$ and $N = 10$ annotators) and for all edges combined (**j**, schematic).

that annotators must pay attention continuously to avoid missing any of the branches along the neurite. Two of the disagreements (locations 3 and 5) were not missed branches but locations at which one annotator continued the neurite skeleton and the other annotator did not. Although one of these two locations (location 3) was easily resolved, agreement between the annotators was reached for location 5 only upon close inspection, suggesting that this location was difficult to annotate. In this case, the difficulty was caused by the local neurite geometry (a tip-to-tip contact). We similarly found both attention- and difficulty-related errors when annotators skeletonized axons in fluorescent data (imaged by confocal microscopy; data not shown). This variation in difficulty is captured by the statistical model of neurite detectability we introduce below.

These initial results indicated that even experts make annotation errors and that skeletons must be cross-checked. We therefore proceeded to further quantify skeleton accuracy across several annotators, and then developed an algorithm to find the consensus skeleton and to estimate its accuracy.

Error quantification

To detect errors in the skeletons, we asked multiple annotators to skeletonize the same neurite (**Fig. 3a**). For each decision by one of the annotators to create an edge, we measured how many of the other annotators agreed with the decision (**Fig. 3b**). Our agreement measure is based on the following reasoning: when one annotator skeletonized an edge, they decided that the neurite continued at the location of this edge. A second annotator agreed with this decision if his/her skeleton also reached the edge location and continued beyond it. Conversely, a second annotator disagreed with this decision if his/her skeleton reached this location but did not continue. To detect and distinguish these two cases we used the following procedure to evaluate the proximity between skeletons.

To evaluate an edge created by one of the annotators, we first considered only the edge in question plus a few edges on each side (skeleton A; **Fig. 3b**), yielding an evaluation spotlight moving along the skeleton (**Fig. 3b**). The size of the spotlight depended on how closely the annotator had placed the neighboring skeleton nodes; it was a sphere with a mean radius of 700 nm (**Fig. 3b**; see below and Online Methods). We next temporarily removed the edge in question, splitting the skeleton into two pieces (**Fig. 3c**), and then measured the distances between each of these two skeleton pieces and all the other annotators' skeletons (skeletons B, C, D and so on). If another annotator's skeleton (skeleton C in **Fig. 3b**,**d**) was close enough to both skeleton pieces, this annotator was considered to have voted for the edge in question (agreeing vote; **Fig. 3d**). Conversely, if another annotator's skeleton was close to only one of the skeleton pieces (skeleton D in **Fig. 3b**,**e**), this annotator was considered to have voted against the edge in question (disagreeing vote; **Fig. 3e**), because this corresponds to a skeleton reaching the location of the edge but not continuing. If the other skeleton was too distant from both skeleton pieces, it probably belonged to a different neurite and was therefore disregarded. Skeletons were considered close enough when the r.m.s. distance between the nodes of the skeleton piece and the edges of the other annotator's skeleton was <625 nm. The value of this maximal distance and the value of the spotlight radius used above were determined by searching for parameters that minimized the disagreements between 50-fold and 15-fold consensus skeletons (see below and Online Methods). This procedure for measuring the agreement between skeletons requires a sufficient node density but does not require the node density to be the same or the node locations to be in register for different skeletons.

After applying this distance measurement to all edges in all annotators' skeletons, we obtained for each edge the number of agreeing votes (T) and total votes (N, sum of agreeing and disagreeing annotators). We then counted the number of edges with a certain combination of T and N (for example, $N = 6$ and $T = 10$), and recorded these for

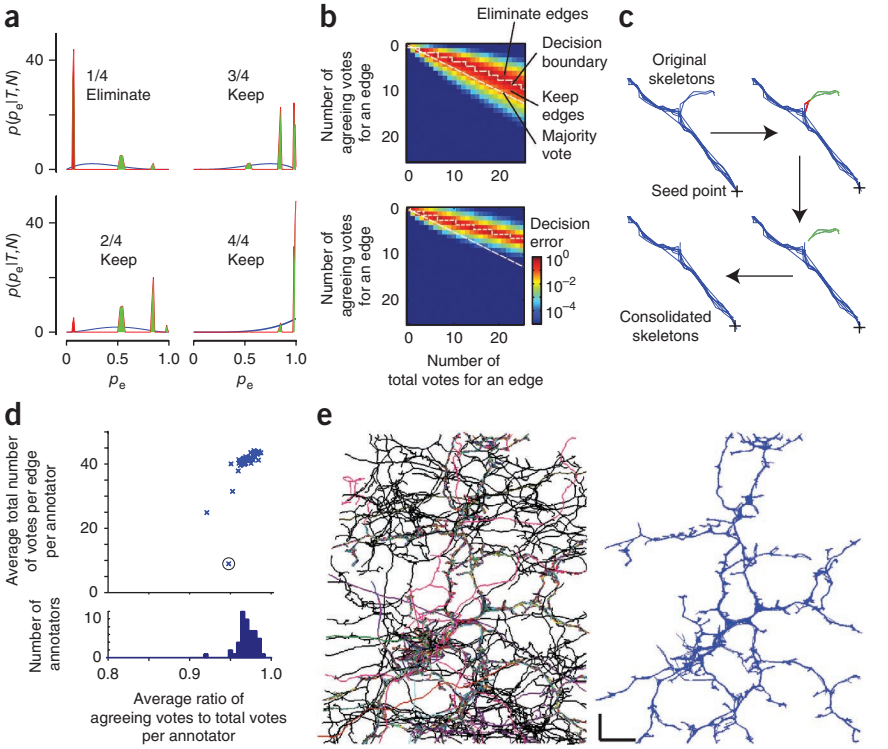

Figure 4 RESCOP steps 2 and 3, edge elimination and skeleton recombination. (**a**) Probability that edge detectability p_e has a certain value, given different edge votes, without prior knowledge (blue) and for the fitted distribution of edge detectabilities $p_{fit}(p_e)$ (red). Whether an edge is kept or eliminated depends on whether the integral of $p(p_e|T,N)$ for $p_e > 0.5$ (green shading) is larger or smaller than that for $p_e < 0.5$ (red shading). In this example, edges with $T = 1$ and $N = 4$ would be eliminated and those with $T = 2$ to 4 would be kept. (**b**) Decision error, $p_{err}(T,N)$, with optimal (stepped line) and majority vote (dashed straight line) decision boundaries for the single-cell (top) and dense skeletonization data (bottom). (**c**) Elimination procedure illustrated at a branch point. Red, eliminated edges. Green, discarded skeleton pieces. (**d**) Variation of annotator performance reflected in average total number of votes per edge and average ratio of agreeing to total votes for each annotator. Circle, worst-performing annotator who skeletonized black skeleton in **e**. (**e**) Fifty skeletons of one amacrine cell before (left) and after (right) edge validation and consensus computation. Scale bar, 5 µm.

all encountered combinations of T and N in a two-dimensional vote histogram (**Fig. 3f**).

The distribution of interannotator agreement

One amacrine cell (~600 µm total neurite path length) was skeletonized by 50 different annotators. Before voting, we divided the set of 50 skeletons three times into two subsets, to which skeletons were randomly assigned. This created six subsets of 25 skeletons each. We calculated their vote histograms separately to later assess the variability of our procedure, and first used the sum of these vote histograms (**Fig. 3f**). Most parts of the amacrine cell were found and annotated by all ($N = 25$) or almost all ($N = 20$–24) annotators (**Fig. 3f**). Because some branches were followed by only a few annotators, the vote histograms also contain entries for few total votes (**Fig. 3f**). In this histogram, we found complete agreement among annotators ($T = N$, evaluated for edges with at least three votes) for 68% of all locations. For 8% of locations only one annotator disagreed, and 10% of the locations were annotated by only one annotator. The locations at which one annotator disagreed can be interpreted, at least for many total votes, as having been missed because of inattention. We interpreted the locations found by only one annotator as erroneous continuations or branches. Most of the remaining 14% of locations, at which more than one annotator disagreed, are presumably more difficult locations in the data, because it is improbable that two or more attention-related mistakes occurred at the same location.

To measure annotation agreement for different kinds of neurites from different types of cells, we also calculated the vote histogram (**Fig. 3f**) for 98 skeletonized neurite fragments densely packed in another region of the same data set (166,472 annotated edges with a total path length of 43.2 mm; **Supplementary Figs. 3** and **4**). In this case N was lower on average (3.2 ± 2.9; **Fig. 3f**) and varied much more. In both cases most annotators agreed for most edges; that is, the votes were concentrated near the diagonal of the vote histogram. The vote histograms can be used to compare the difficulty of data sets, provided that the annotators were similarly trained and similarly attentive.

Skeleton consensus rules

Our next goal was to find the consensus skeleton using multiple annotations of the same neurite by eliminating edges that were probably not correct, on the basis of the number of agreeing and disagreeing votes. The intuitive choice for whether to accept or eliminate an edge is the majority vote, but it was not clear whether this leads to the optimal decision. We therefore analyzed the annotation process (**Fig. 3g**–**j**) to determine a rule to find the best consensus skeleton and to estimate the residual error rate of the consensus skeleton.

To describe the annotation process we used the following decision model, which reflects the fact that annotation difficulty varies with location (**Fig. 3i**,**j**). Although two intracellular voxels are either connected (that is, belong to the same neurite) or not connected (that is, belong to different neurites), this ground truth is to some degree obscured by fixation, staining and imaging of the sample at limited resolution and signal-to-noise ratio. This makes annotation an inherently noisy process, with a probability, p_e, for each pair of points that annotators will create an edge, that is, label the points as connected (we also refer to p_e as edge detectability; **Fig. 3i**,**j**). The edge detectability depends on whether the points are actually connected (see below), but it also varies as a consequence of the local neurite geometry (wide, straight or bundled neurites are, for example, easier to follow) and local staining quality.

In this model, the decision to create an edge between a pair of points corresponds to a biased coin toss, with the bias equal to the edge detectability p_e. Therefore, the decisions of the annotators will follow binomial statistics with a bias of p_e (**Fig. 3i**; Online Methods, equation (4)). Obvious neurite continuities (where $p_e \approx 1$) and neurite discontinuities (where $p_e \approx 0$) will both lead to a high agreement among annotators. Difficult locations have $p_e \approx 0.5$.

We cannot determine p_e at a given location directly except by annotating it many times. However, for any assumed distribution of edge detectabilities, $p(p_e)$, in the data, we can compute the expected distribution of agreeing and disagreeing votes (predicted vote histograms;

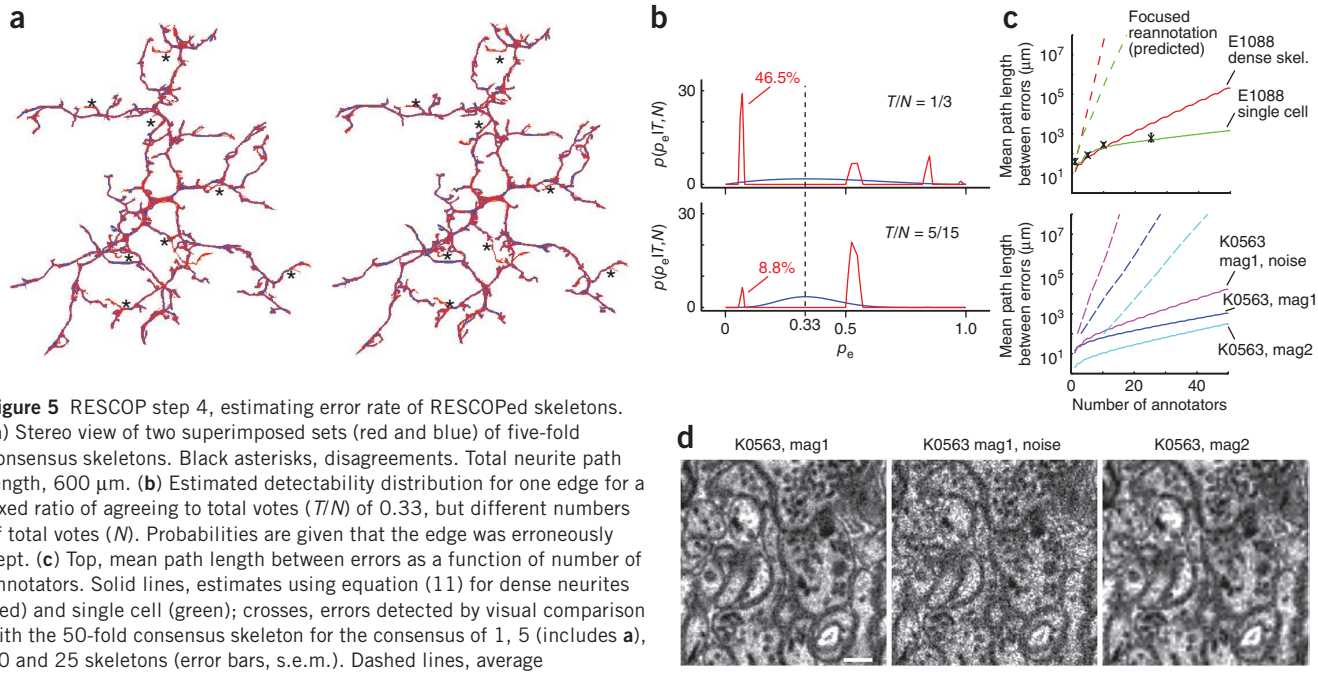

Figure 5 RESCOP step 4, estimating error rate of RESCOPed skeletons. (**a**) Stereo view of two superimposed sets (red and blue) of five-fold consensus skeletons. Black asterisks, disagreements. Total neurite path length, 600 µm. (**b**) Estimated detectability distribution for one edge for a fixed ratio of agreeing to total votes (*T*/*N*) of 0.33, but different numbers of total votes (*N*). Probabilities are given that the edge was erroneously kept. (**c**) Top, mean path length between errors as a function of number of annotators. Solid lines, estimates using equation (11) for dense neurites (red) and single cell (green); crosses, errors detected by visual comparison with the 50-fold consensus skeleton for the consensus of 1, 5 (includes **a**), 10 and 25 skeletons (error bars, s.e.m.). Dashed lines, average redundancy as a function of the target error rate for focused reannotation (Monte Carlo simulations). Bottom, same analysis for a conventionally stained data set annotated using the original data (blue, K0563, mag1, s.d.), data with added noise (magenta, K0563, mag1, noise) and data at half the resolution (cyan, K0563, mag2). (**d**) Examples from the original and degraded data sets. Scale bar, 250 nm.

equations (5) and (7), Online Methods). We compared measured and predicted vote histograms (**Fig. 3f**,**g**) to search for the optimal $p(p_e)$, that is, the distribution that best explained the measurements. We found that the optimal $p(p_e)$ consists of several peaks with a large peak near 1 (**Fig. 3h**), which reflects the high frequency of obvious neurite continuities. Because we cannot measure zero agreeing votes ($T = 0$), the fit is not well constrained near $p_e = 0$. In fact, a delta function at $p_e = 0$ can be added to $p(p_e)$ without changing the goodness of the fit and without affecting the following results.

To explore how variable $p(p_e)$ is for different annotations of the same cell, we separately fitted vote histograms for the six sets of 25 of 50 skeletons and found similar $p(p_e)$ distributions (**Supplementary Fig. 5c**,**d**); what varies is the exact location of the peaks in the middle part of the p_e range. We also determined the optimal $p(p_e)$ for the vote histogram of the dense annotation (**Supplementary Fig. 4**). Again, we found the same general structure, with a strong peak near 1 and several peaks throughout the rest of the range (**Fig. 3h**).

Computing the consensus skeletons

We next used the annotation-decision model to find the consensus skeletons (**Fig. 4**). Using equation (2), we estimated the distribution of p_e for each edge, given the agreeing and disagreeing votes. We made the assumption that true connectivity leads to above-chance edge detectability ($p_e > 0.5$). This implies that the annotation decisions will converge toward the ground truth as the number of redundant annotations increases. When training the annotators, we encouraged this by providing training examples rich in difficult locations.

This assumption about the relationship between the detectability, p_e, of an edge and its actual connectedness is probably not entirely correct. The number of locations for which this assumption is incorrect is, however, probably small (the crossover region between the sketched curves in **Fig. 3j**).

We therefore based our consensus rule for an edge on whether the estimated distribution of edge detectability given the agreeing and disagreeing votes cast for that edge, $p(p_e|(T,N))$, indicated that the edge at that location was more probable to be detected than not. (**Fig. 4a** and Online Methods, equation (3)). By evaluating this rule for all possible combinations of agreeing and disagreeing votes, we obtained the optimal decision boundary in the vote histogram between 'eliminate edge' and 'keep edge' (this optimal decision boundary was substantially below the majority rule, that is, edges with less than majority agreement are typically accepted; **Fig. 4b**). Because the consensus rule depends on $p(p_e)$, the optimal boundary is generally different for different neurite data sets (**Fig. 4b**).

Because edge elimination splits some skeletons (**Fig. 4c**), it is necessary to determine which skeleton pieces still belong together. Whenever annotators started from a soma, we checked whether there was still a connection between the skeleton pieces and a seed region in the proximal dendrite (**Fig. 4c**). For the 50-fold annotated cell, the consensus skeleton now lacks many presumably erroneous neurites (**Fig. 4e**). In other cases, multiple annotators were instructed to start at different seed points along the same neurite (**Supplementary Figs. 4** and **5** and Online Methods). In these cases, finding the consensus skeletons is substantially more complicated, but our model still yields reasonable consensus skeletons. Each consensus skeleton is a bundle of closely spaced skeleton pieces (**Fig. 4e**).

Annotator quality

So far we have assumed that the error rates of different annotators are similar. To determine how much error rates vary among annotators, we assessed for each annotator how close his/her skeletons were to those of others by calculating (i) the average number of total votes for or against that annotator. This value, when low, indicates that an annotator followed many neurites in little agreement with the other annotators; and (ii) his/her average ratio of agreeing to total votes (**Fig. 4d**).

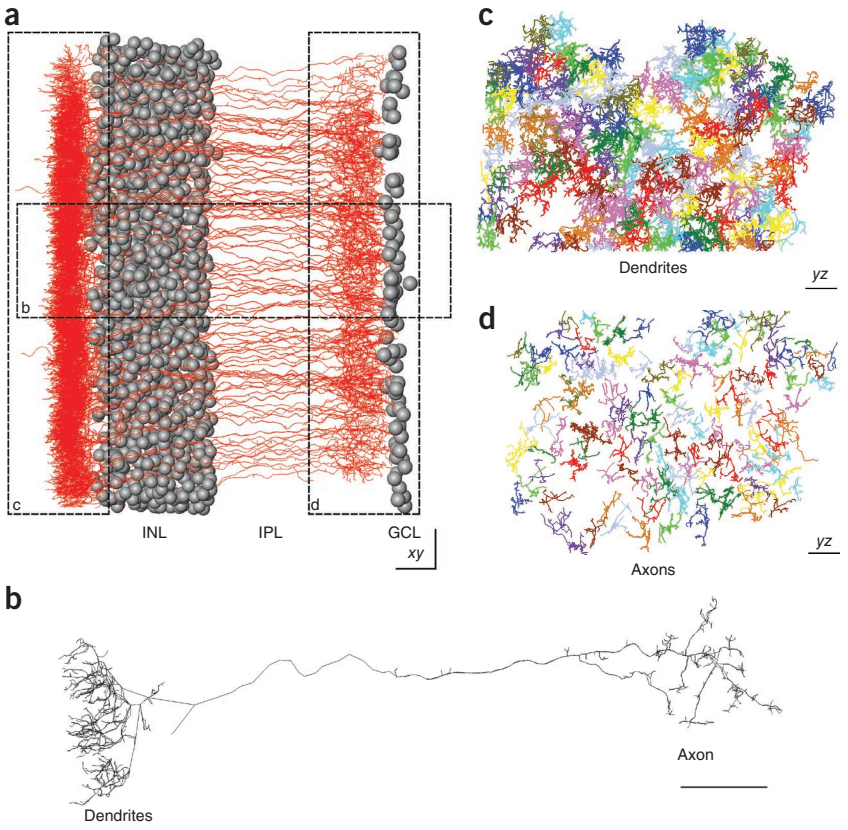

Figure 6 Doubly annotated skeletons of 114 putative rod bipolar cells in a block of mouse retina. (**a**) View onto the block face. INL, inner nuclear layer; IPL, inner plexiform layer; GCL, ganglion cell layer. Dashed lines indicate bounding boxes for **b**–**d**. (**b**) Two skeletons of a single rod bipolar cell. (**c**,**d**) View onto the plane of the retina confined to the dendrites (**c**) and axons (**d**) of bipolar cells, respectively. Cells are colored randomly in **c**,**d**. Scale bars, 10 µm.

For the majority of annotators, the average ratio of agreeing to total votes was 95–98% (**Fig. 4d**). The worst-performing annotator (**Fig. 4d,e**) generated a skeleton with >4× the total path length, even entering additional cells. The best annotators, in contrast, had as few as two disagreements with the 50-fold consensus skeleton.

The residual error rates of RESCOP'd skeletons

To estimate how many errors to expect in the consensus skeletons, we computed the error probabilities for each of the decisions to eliminate or accept an edge (**Fig. 5**). As we described above, an edge was eliminated whenever the vote count for this edge indicated that it was more probable than not that the edge was incorrectly annotated. However, there remains an error probability that the edge was, in fact, correctly annotated and should have been accepted. To calculate the error probabilities for eliminated edges and accepted edges, we integrated $p(p_e|(T,N))$ for $p_e > 0.5$ and $p_e < 0.5$, respectively (**Fig. 4a**). Because $p(p_e|(T,N))$ becomes more sharply peaked as the total number of votes increases (**Fig. 5b**), the error rate for a given ratio of agreeing to total votes decreases.

As the number of annotators rises, the accuracy of the consensus skeleton increases (**Fig. 5c**) initially steeply but then more slowly, because as the detectability of an edge approaches 0.5, the number of votes needed to achieve a given error rate diverges (edges with $p_e = 0.5$ are fundamentally undecidable). Therefore, near an edge detectability of $p_e = 0.5$ the error for a large number of votes N is very sensitive to the shape of $p(p_e)$, and the error predictions for a large N can scatter substantially for different neurites, or even different groups of annotators (**Supplementary Fig. 5d**).

We next compared this error-rate prediction with the accuracy of the consensus skeletons. We randomly selected from the 50 skeletons sets of 25, 10, 5 and 1 skeletons ($n = 6$, 15, 20 and 10, respectively) and computed the consensus skeleton for each set independently (**Fig. 5a**). Then we visually assessed the differences among all those consensus skeletons and the 50-fold consensus skeleton, which we used as a reference. The average number of disagreements was 1.0 ± 0.4, 2.1 ± 0.3, 7.2 ± 0.9 and 15.5 ± 3.5 (mean ± s.e.m.) for the 25-fold, 10-fold, 5-fold and single skeletons, respectively, corresponding to respective mean distances between errors of 600.2 µm, 281.3 µm, 83.4 µm and 38.7 µm (**Fig. 5c**).

So far we have considered the case in which the entire length of neurites is multiply annotated. Because for most locations connectedness is easy to determine, increasing the overall redundancy is inefficient. We therefore explored focused reannotation, that is, repeatedly examining each edge until a given accuracy is reached rather than annotating each edge a fixed number of times. This should concentrate the annotators' effort onto difficult locations. To determine the redundancy-accuracy tradeoff for focused reannotation, we carried out Monte Carlo simulations and found that for focused reannotation, the accuracy rises almost exponentially with the average redundancy (**Fig. 5c**).

Variation of error rate with data quality

To test how the error rate depends on the staining method and on the data quality, we annotated a conventionally stained data set (K0563, see **Supplementary Image Stacks 13–15**) at its original resolution (12 × 12 × 25 nm^3 voxels), with added noise (Gaussian, s.d. = 20), and at half the resolution (24 × 24 × 50 nm^3 voxels) (**Fig. 5d**). We found that error rates were slightly lower for the added-noise case, possibly owing to increased attention, but that for the reduced-resolution data, annotation reliability was substantially lower (**Fig. 5c**).

Dense reconstruction

To demonstrate dense neuron reconstruction from SBEM data using the tools presented here, we selected all rod bipolar cells (RBCs; **Fig. 6**) from a SBEM data set in the process of being skeletonized (data set E2006, currently at two-fold redundancy, M.H., K.L.B. and W.D., unpublished data). Compared with data set E1088, the E2006 data set covers a different block of tissue (80 µm × 117 µm × 135 µm, see Online Methods), comes from a mouse rather than a rabbit retina, was imaged at a higher resolution and was stained more intensely. RBCs were initially identified on the basis of geometrical parameters using automatic clustering (M.H., K.L.B. and W.D., unpublished data). We refined this selection by manually removing 23 of 137 cells because they were cone bipolar cells (14 cells) or had an aberrant morphology, indicating a substantial annotation error (9 cells) that had not yet been eliminated because of the only two-fold redundancy.

The remaining 114 cells had the tiling patterns of axons and dendrites expected for rod bipolar cells (**Fig. 6c**,**d**).

The annotation speed for these skeletons was 5.3 h per mm path length (RBC average neurite length, 368 ± 103 µm, mean ± s.d.). Using the model described above, we expect about ten errors per cell for double annotation. To reduce the error rate to 1 per cell, a redundancy of 18 or 19 (and a redundancy of 4 on average for focused reannotation) should be sufficient. These numbers indicate that it is feasible to reconstruct all bipolar cells and all the dendrites or dendrite fragments of all ganglion cells with their somata in such a block of tissue using ~7,500 work hours (Online Methods).

DISCUSSION

Dense versus sparse reconstruction

Our data show not only that neurites can be densely reconstructed using SBEM volume electron microscopy data, but also that manual annotations, even when carried out by experts, contain errors. Many of these errors are caused by insufficient attention, particularly where neurites branch (**Figs. 2** and **3**). This problem does not occur when labeling is sufficiently sparse, but is prevalent for densely stained tissue, which is needed for any kind of connectomic analysis of neurite networks. Branching-type errors occur even for light-microscopic data as soon as the stained-neurite density is so large that the frequency of close encounters between neurites becomes substantial, as it does with even a moderate fraction of neurons stained. Annotation errors are probably widespread, but they are rarely acknowledged, let alone quantified. Annotation error rates are related to the information content and quality of the staining (**Figs. 2** and **5**). For the study of local synaptic geometry, in which many serial electron microscopy studies exist, a modest error rate would only rarely affect the conclusions. Error rates need to be much lower for connectomic neuroanatomy, however, in which a single missed branch point typically leads to thousands of lost or wrongly attributed synapses. Other errors are less costly; a missed spine neck would lead to loss of a few synapses at most. We have so far quantified only errors caused by incorrect neurite reconstruction. Although the identification of synapses can be error-prone as well, one such error affects only one particular synapse, and much less severely affects the connectomic reconstruction error than the typical neurite continuity error.

The few published reconstructions of entire neurites from electron microscopy data have been carried out by highly trained and dedicated experts and extensively proofread by the same or other experts[3,11,19,20]. For the *C. elegans* connectome, several corrections have been made[21] using the original image series. Some form of proofreading is necessary during the connectomic reconstruction of neuronal networks[22,23]. However, proofreading existing skeletons not only is very tedious but may be less efficient than redundantly annotating the same neurites and detecting inconsistencies. In contrast to conventional proofreading, redundant annotation can be used to quantify annotation difficulty (**Fig. 5c**).

Mass annotation, distribution of skill and training levels

Finding the consensus of multiple annotations using RESCOP may reduce the error rate to a level sufficient for almost any application of connectomic circuit reconstruction. RESCOP can also be used to estimate the number of reconstruction errors remaining in the consensus skeleton, and to find probable locations for those errors, a prerequisite for focused reannotation. Our analysis also shows that the optimal vote threshold (the decision boundary) can be substantially different from majority voting (**Fig. 4b**).

RESCOP can be used to create connectomic reconstructions with a known accuracy by annotators that have no prior neurobiological knowledge and are only slightly trained. Even if the error rate is high for individual annotators, focused reannotation could be used to substantially reduce it in the consensus skeleton, with only a moderate increase in the average redundancy. Most of the effort could then be concentrated on difficult locations ($p_e \approx 0.5$), which require a higher redundancy to reach a given reliability. In our data, difficult locations seem to be rare, as the prevalence of vote ratios ~1 shows (**Fig. 3**). The low density of difficult locations also indicates that ambiguous vote ratios ($T/N \approx 0.5$) are rare and the fits for $p(p_e)$ are not very well constrained in the region around $p_e = 0.5$, making estimates of error rates for large N somewhat uncertain (**Supplementary Fig. 5**). Ultimately, the error rate will be affected by the assumption we made that an infinite number of annotators will converge to the correct decision. This limits the validity of the accuracy predictions (**Fig. 5c**) for very large N. We expect that the availability of improved staining and imaging methods will further reduce the frequency of locations at which the data biases even experts toward the wrong conclusion.

One advantage of using weakly trained annotators is that the increase in reliability can be achieved at a lower cost than with expert proofreaders, who might still make attention-related errors at an undesirable rate (**Fig. 2**). Also, requiring graduate students or postdoctoral fellows to annotate for several thousand hours is hardly a good use of their talents. Finally, untrained personnel can in many academic settings be recruited quickly and on a temporary basis. RESCOP can automatically direct annotator effort and assess annotator quality, making it well suited for Web-based crowdsourcing. This makes it practical to scale up annotation capacity to the limit of the available budget. RESCOP thus removes a major obstacle to high-throughput circuit reconstruction. This is demonstrated by our reconstruction of bipolar cells (**Fig. 6**), which took 5.3 h per mm of skeleton length, ~60× faster than volume labeling.

Skeletons and automated reconstruction algorithms

Computer algorithms, especially those using machine learning[24–26], can help reconstruct neural circuits. In the long run, such tools could replace or greatly reduce the need for manual annotation. However, automatic methods need to be evaluated through comparison to a reliable ground truth. The consensus among manual annotations can serve as such a ground truth, in particular when the error rate is known, as is the case for RESCOP. Such an estimation of annotator errors is not available for other major benchmark data sets in machine learning (for example, Berkeley Segmentation Data set[27]). In medical imaging (magnetic resonance imaging and computed tomography), expert annotation by trained radiologists is the gold standard, but expert annotations frequently differ substantially[28]. Therefore, algorithms to estimate optimal annotations have recently received more attention (for example, STAPLES[29]). Often, majority voting is close to optimal[30]. Our study presents a 'decision theoretic' approach, which involves finding the optimal decision criterion given a model of the belief formation or decision process[31].

A major shortcoming of skeleton annotations is that they do not produce a complete volume representation; this is especially important for detecting contacts between neurites, a prerequisite for synapses. This problem can be solved (M.H., K.L.B., V. Jain, S. Turaga, S. Seung and W.D., unpublished data) by combining high-accuracy long-range manual annotation, as we report here, with locally accurate but globally error-prone automated volume reconstructions[24–26]. Such hybrid techniques could reduce the manual effort to create full volume representations by as much as two orders of magnitude, and will enable researchers to carry out connectomic reconstruction on much larger volumes than before.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/natureneuroscience/.

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank B. Andres, F. Hamprecht, U. Köthe, V. Jain, S. Seung and S. Turaga for many fruitful discussions and comments on the manuscript, J. Bollmann and A. Schaefer for helpful comments on the manuscript, C. Roome for information technology support, J. Kornfeld and F. Svara for programming KNOSSOS, and J. Hanne, H. Jakobi and H. Wissler for help with annotator training. We thank N. Abazova, E. Abs, A. Antunes, P. Bastians, J. Bauer, M. Beez, M. Beining, S. Bender, S. Best, L. Brosi, M. Bucher, E. Buckler, J. Buhmann, C. Burkhardt, F. Drawitsch, L. Ehm, S. Ehm, C. Fianke, R. Foltin, S. Freiss, M. Funk, A. Gebhardt, M. Gruen, K. Haase, J. Hammerich, J. Hanne, B. Hauber, M. Hensen, L. Hofmann, P. Hofmann, M. Hülser, F. Isensee, H. Jakobi, M. Jonczyk, A. Joschko, A. Juenger, S. Kaspar, K. Kessler, A. Khan, M. Kiapes, A. Klein, C. Klein, S. Laiouar, T. Lang, L. Lebelt, H. Lesch, C. Lieven, D. Luft, E. Moeller, A. Muellner, M. Mueller, D. Ollech, A. Oppold, T. Otolski, S. Oumohand, S. Pfarr, M. Pohrath, A. Poos, S. Putzke, J. Reinhardt, A. Rommerskirchen, M. Roth, J. Sambel, K. Schramm, C. Sellmann, J. Sieber, I. Sonntag, M. Stahlberg, T. Stratmann, J. Trendel, F. Trogisch, M. Uhrig, A. Vogel, J. Volz, C. Weber, P. Weber, K. Weiss, L. Weisshaar, E. Wiegand, T. Wiegand, M. Wiese, R. Wiggers, C. Willburger and A. Zegarra for neurite skeletonizations. This work was funded by the Max Planck Society.

AUTHOR CONTRIBUTIONS

M.H. and W.D. designed the study and devised the analysis algorithms; K.L.B. carried out the SBEM experiments; M.H., K.L.B. and W.D. specified the KNOSSOS software; M.H. analyzed the data; M.H., K.L.B. and W.D. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/natureneuroscience/.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ramón y Cajal, S. *Textura del Sistema Nervioso del Hombre y de los Vertebrados* (Moya, Madrid, 1899).
2. Golgi, C. Sulla struttura della sostanza grigia del cervelo. *Gazzetta Medica Italiana Lombardia* **33**, 244–246 (1873).
3. Harris, K.M. & Stevens, J.K. Dendritic spines of CA 1 pyramidal cells in the rat hippocampus: serial electron microscopy with reference to their biophysical characteristics. *J. Neurosci.* **9**, 2982–2997 (1989).
4. Horikawa, K. & Armstrong, W.E. A versatile means of intracellular labeling: injection of biocytin and its detection with avidin conjugates. *J. Neurosci. Methods* **25**, 1–11 (1988).
5. Stretton, A.O. & Kravitz, E.A. Neuronal geometry: determination with a technique of intracellular dye injection. *Science* **162**, 132–134 (1968).
6. Wickersham, I.R. *et al.* Monosynaptic restriction of transsynaptic tracing from single, genetically targeted neurons. *Neuron* **53**, 639–647 (2007).
7. Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56–62 (2007).
8. Sporns, O., Tononi, G. & Kotter, R. The human connectome: A structural description of the human brain. *PLoS Comput. Biol.* **1**, e42 (2005).
9. Lichtman, J.W., Livet, J. & Sanes, J.R. A technicolour approach to the connectome. *Nat. Rev. Neurosci.* **9**, 417–422 (2008).
10. Helmstaedter, M., Briggman, K.L. & Denk, W. 3D structural imaging of the brain with photons and electrons. *Curr. Opin. Neurobiol.* **18**, 633–641 (2008).
11. White, J.G., Southgate, E., Thomson, J.N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
12. Denk, W. & Horstmann, H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biol.* **2**, e329 (2004).
13. Hayworth, K.J., Kasthuri, N., Schalek, R. & Lichtman, J.W. Automating the collection of ultrathin serial sections for large volume TEM reconstructions. *Microsc. Microanal.* **12**, 86–87 (2006).
14. Knott, G., Marchman, H., Wall, D. & Lich, B. Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *J. Neurosci.* **28**, 2959–2964 (2008).
15. Briggman, K.L. & Denk, W. Towards neural circuit reconstruction with volume electron microscopy techniques. *Curr. Opin. Neurobiol.* **16**, 562–570 (2006).
16. Briggman, K.L., Helmstaedter, M. & Denk, W. Wiring specificity in the direction-selectivity circuit of the retina. *Nature* (in the press) (2011).
17. Fiala, J.C. Reconstruct: a free editor for serial section microscopy. *J. Microsc.* **218**, 52–61 (2005).
18. Jeong, W.K. *et al.* Ssecrett and NeuroTrace: interactive visualization and analysis tools for large-scale neuroscience data sets. *IEEE Comput. Graph. Appl.* **30**, 58–70 (2010).
19. Trachtenberg, J.T. *et al.* Long-term *in vivo* imaging of experience-dependent synaptic plasticity in adult cortex. *Nature* **420**, 788–794 (2002).
20. Stevens, J.K., McGuire, B.A. & Sterling, P. Toward a functional architecture of the retina: serial reconstruction of adjacent ganglion cells. *Science* **207**, 317–319 (1980).
21. Chen, B.L., Hall, D.H. & Chklovskii, D.B. Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci. USA* **103**, 4723–4728 (2006).
22. Chklovskii, D.B., Vitaladevuni, S. & Scheffer, L.K. Semi-automated reconstruction of neural circuits using electron microscopy. *Curr. Opin. Neurobiol.* **20**, 667–675 (2010).
23. Mishchenko, Y. *et al.* Ultrastructural analysis of hippocampal neuropil from the connectomics perspective. *Neuron* **67**, 1009–1020 (2010).
24. Jain, V. *et al.* Supervised learning of image restoration with convolutional networks. *IEEE 11th Int. Conf. Comput. Vis.* 1–8 (2007).
25. Andres, B., Köthe, U., Helmstaedter, M., Denk, W. & Hamprecht, F. Segmentation of SBFSEM volume data of neural tissue by hierarchical classification. in *Pattern Recognition: Lecture Notes in Computer Science* (ed. Rigoll, G.) 142–152 (Springer-Verlag, 2008).
26. Turaga, S.C. *et al.* Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Comput.* **22**, 511–538 (2010).
27. Martin, D., Fowlkes, C., Tal, D. & Malik, J. A Database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *IEEE 8th Int. Conf. Comput. Vis.* **2**, 416–423 (2001).
28. Warfield, S.K., Zou, K.H. & Wells, W.M. Validation of image segmentation by estimating rater bias and variance. *Philos. Transact. A Math. Phys. Eng. Sci.* **366**, 2361–2375 (2008).
29. Warfield, S.K., Zou, K.H. & Wells, W.M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* **23**, 903–921 (2004).
30. Wang, W. *et al.* A classifier ensemble based on performance level estimation. *IEEE Int. Symposium on Biomed. Imaging: from Nano to Micro*, 342–345 (2009).
31. Körding, K. Decision theory: what "should" the nervous system do? *Science* **318**, 606–610 (2007).

ONLINE METHODS

Serial block-face electron microscopy. Retinas from a 6-week-old rabbit (for E1088, **Figs. 1–5** and **Supplementary Figs. 3** and **4**), from a P30 C57BL/6 mouse (for E2006, **Fig. 6** and **Supplementary Fig. 2**) and from a P30 C57BL/6 mouse (for K0563, **Fig. 5c**,**d**) were prepared for E1088 and E2006 to selectively enhance cell outlines by using HRP-mediated precipitation of DAB as described[16] and stained with osmium alone (E1088) or in conjunction with lead citrate (E2006); a more conventional stain was used for K0563 (same data set as in ref. 16). All procedures were approved by the local animal care committee and were in accordance with the law of animal experimentation issued by the German Federal Government.

The embedded tissue was trimmed to a block face of about 200 µm × 300 µm, and imaged in a scanning electron microscope with a field-emission cathode (QuantaFEG 200, FEI Company) and a custom-designed backscattered electron detector based on a silicon diode (AXUV, International Radiation Detectors) combined with a custom-built current amplifier. The incident electron beam energy was 3.6 keV for E1088, 3.0 keV for E2006 and 2.0 keV for K0563, respectively; its current was ~100 pA for all three data sets. At a pixel dwell time of 8 µs and a pixel size of 22 nm × 22 nm (E1088), 6 µs and 16.5 nm × 16.5 nm (E2006) and 5 µs and 12 nm × 12 nm (K0563), this corresponds to doses of about 10 (E1088), 14 (E2006) and 22 (K0563) electrons nm^{-2}, not accounting for skirting due to low vacuum operation. The chamber was kept at a pressure of 75 Pa of water vapor (E1088) or 130 Pa of hydrogen (E2006) to prevent charging. K0563 was conducting enough to be imaged in high vacuum. The electron microscope was equipped with a custom-made microtome[12], which allows the repeated removal of the block surface at a cutting thickness of ~30 nm (E1088) or ~25 nm (E2006 and K0563). Consecutive slices (1,999, E1088; 3,200, E2006; 5,765, K0563) were imaged, leading to data volumes of 2,048 × 1,768 × 1,999 voxels (E1088), 8,192 × 7,072 × 3,200 voxels (4 × 4 mosaic of 2,048 × 1,768 images, E2006) and 4,096 × 5,304 × 5,760 voxels (2 × 3 mosaic of 2,048 × 1,768 images, K0563), corresponding to volumes of 45 × 39 × 60 μm^3, 135 × 117 × 80 μm^3, and 50 × 65 × 145 μm^3, respectively. For E1088, the imaged region spanned the inner plexiform layer of the retina and included parts of the inner nuclear and of the ganglion cell layers. E2006 spanned the retina from the ganglion cell layer to the cell bodies of photoreceptors. K0563 spanned the inner plexiform layer of the retina and included the ganglion cell layer and part of the inner nuclear layer. The noise-degraded dataset (**Fig. 5d** and **Supplementary Image Stack 14**) was generated by addition of Gaussian noise (s.d. = 20) to the original dataset, which had a gray value range of 101 to 196, 3rd and 97th percentile, respectively. Consecutive slices were aligned offline to subpixel precision by Fourier shift–based interpolation, using cross-correlation-derived shift vectors.

Reconstruction software. Neurite skeletons were annotated using KNOSSOS (written in C by J. Kornfeld and F. Svara (Max Planck Institute for Medical Research) according to specifications by the authors). KNOSSOS (**Supplementary Movie 1**) is available at http://www.knossostool.org/.

Skeletonization. Data were annotated using KNOSSOS and skeletons were saved in an .xml format called .nml, which is similar to the NeuroML format[32]. Each file contains a list of the skeleton nodes. For each node, parameters including index, coordinates, radius, viewport used for node placement and time stamp are given, as well as a list of the edges between nodes, and a list of nodes tagged as branch points (example file is in **Supplementary Data 1**). Annotators were instructed as follows: (i) start at a given seed point, typically inside the soma of a neuron (for randomly dense seeding strategies, see below and **Supplementary Figs. 3** and **4**); (ii) follow the neurite from that location; the neurite generally continues in two (if the seed point is in an axon or dendrite) or more (if the seed point is in a soma with more than two primary neurites) directions; (iii) while annotating, focus on the viewport that is most orthogonal to the current neurite axis (for versions after v3.0435 of KNOSSOS, the appropriate viewport, is automatically highlighted based on the vector between the two most recently placed nodes, and highlight it); (iv) accuracy is more important than speed; (v) place a node about every seven to ten planes (corresponding to ~200–300 nm edge length for SBEM data); (vi) generously place branch point flags, so as not to miss branches. Annotators were trained on at least three neurons during 10–40 h of training. Their training results were compared to annotations of the same neurons by experts, and disagreements were inspected and discussed. Annotators were allowed to continue with novel tasks only when their training performance was sufficient as judged by the trainer.

Speed measurement. To measure the speed of skeletonization we initially asked annotators to report the time spent annotating. This yielded annotation time of 5–10 h per mm path length. Then we included a time stamp feature in KNOSSOS that recorded the time when each skeleton node was placed (**Supplementary Fig. 2a**). To determine the effective annotation rate, we summed the internode time intervals and excluded intervals >7 min to account for breaks taken by the annotators. This assumes that no single location takes that long to contemplate (**Supplementary Fig. 2b**,**c**).

Edge validation algorithm. To test each edge in a given set of skeletons $\{S_\alpha, S_\beta, S_\gamma, \ldots\}$, created by multiple annotators (α, β, γ, and so on) starting at the same seed point (for different reseeding strategies, see below), we used the following procedure. To test, for example, edge $E_{\alpha ij}$, (which connects nodes N_i and N_j in skeleton S_α), S_α was first pruned beyond a sphere of radius r_p around the center of the edge $E_{\alpha ij}$, yielding two skeleton pieces, $S_{\alpha i1}$ and $S_{\alpha i2}$, starting at the ends of $E_{\alpha ij}$ ($N_{i\alpha}$ and $N_{j\alpha}$, respectively; pieces 1 and 2 in **Fig. 3b**). The cutoff radius r_p was set to ensure that at least one further edge was included at each end of the tested edge:

$$r_p = \max\left(\frac{|E_{ki\alpha}|}{2} + \max\left(\min_k\left(|E_{ki\alpha}|\right), \min_k\left(\left|E_{kj\alpha}\right|\right)\right), 625\,\text{nm}\right) \quad (1)$$

where $\min_k(|E_{kj\alpha}|)$ is the length of the shortest of the edges connected to node $N_{j\alpha}$ (for the 50-fold single cell voting, r_p was on average 28 voxels, or ~700 nm). Next, one of the other skeletons, S_β, was taken and the r.m.s. node-to-edge distances were calculated between each of the skeleton pieces $S_{\alpha i1}$, $S_{\alpha i2}$ and S_β using all nodes of $S_{\alpha i1}$ and $S_{\alpha i2}$. When both r.m.s. distances were less than the set threshold θ = 625 nm, this was a vote in favor of the tested edge (the agreeing vote count $T_{\alpha ij}$ and the total vote count $N_{\alpha ij}$ for edge $E_{\alpha ij}$ were both increased by 1); if only one but not the other distance was less than the threshold, this counted against $E_{\alpha ij}$ (only the total vote count $N_{\alpha ij}$ for edge $E_{\alpha ij}$ was increased by 1). For both distances above the threshold, no vote was counted because this indicated that S_β was not near the tested edge. θ was on the order of the typical neurite radius, which, however, varies widely; both θ and r_p were selected to minimize the difference between the 50-fold consensus skeleton and sets of 10-fold consensus skeletons. If the edge was within three nodes of a neurite ending, we used $\theta_{end} = 2r_p$ as the threshold for agreement to account for the variability in the placement of terminal nodes. This procedure was repeated for all remaining skeletons S_γ, S_δ and so on, and T and N were finally both increased by 1 to account for the tested edge itself (seen as agreeing with itself). Although the reliability of consensus skeletons is probably lower near endings, errors near endings are also less consequential because the number of misallocated nodes is small.

Finding the consensus skeleton. After validating all edges, the consensus skeleton was computed. Finding the consensus skeletons means eliminating edges that are more probable to be wrong than correct. To decide whether to eliminate or keep an edge, given a vote (T,N), we calculated the conditional probability distribution of the hidden parameter p_e, toward which T/N would converge for an infinite number of annotators:

$$p(p_e \mid T, N) = \frac{p(T \mid N, p_e)p(p_e)}{p(T \mid N)} \quad (2)$$

Because we assumed that the annotators have no additional bias, an edge should be eliminated if and only if

$$\int_0^{0.5} p(p_e \mid T, N)\mathrm{d}p_e > \int_{0.5}^{1} p(p_e \mid T, N)\mathrm{d}p_e. \quad (3)$$

For independent annotators the model for the likelihood is the binomial distribution

$$p(T \mid N, p_e) = \binom{N}{T} p_e^T (1 - p_e)^{N-T} \quad (4)$$

To determine the most probable $p(p_e)$, we computed the predicted vote histograms, $hist_{pred}$, while varying $p(p_e)$, and compared $hist_{pred}$ to the measured vote

histogram, $hist_{meas}$, in the following way. First we corrected for the fact that if, at one given location, T of N skeletons agreed, there would be a vote entry at (T,N) in the histogram for each of the T skeletons. We therefore divided the vote counts by T ($hist^{*}_{meas}(T,N) = hist_{meas}(T,N)/T$). Because we cannot measure edges with $T = 0$, the predicted vote distribution was normalized for $T = 1 \ldots N$:

$$hist_{pred}(T,N) = p(T \mid N, p_{fit}) \frac{\sum_{T=1}^{N} hist^{*}_{meas}(T,N)}{\sum_{T=1}^{N} p(T \mid N, p_{fit})} \quad (5)$$

whereby $p(T \mid N, p_{fit}) = \int_0^1 p(T \mid N, p_e)\, p_{fit}(p_e)\, dp_e$ is the probability that an edge sampled N times has T agreeing votes.

We then assumed $p_{fit}(p_e)$ to be a function that is piecewise linear between the points $\rho_i = f(i / 80)$, with i varying from 0 to 80, and $f(x) = 2x^2$ for $x < 0.5$ and $f(x) = 1 - 2(1 - x)^2$ for $x \geq 0.5$. This ensures that $p_{fit}(p_e)$ is more finely sampled near 0 and also near 1, where the bulk of the probability mass is expected. We can write $p_{fit}(p_e)$ as a sum over triangle-shaped basis functions g_i with peaks at the points ρ_i and weights w_i

$$p_{fit}(p_e \mid w_0 \ldots w_{80}) = \sum_{i=0\ldots80} w_i g_i(p_e) \quad (6)$$

leading to a vote prediction of

$$hist_{pred}(T,N) = \sum_{i=0}^{80} \left(w_i c_{i,T,N} \right) \sum_{T=1}^{N} hist^{*}_{meas}(T,N) \Bigg/ \left(1 - \sum_{i=0}^{80} \left(w_i c_{i,T=0,N} \right) \right) \quad (7)$$

whereby $c_{i,T,N} = \int_0^1 p(T \mid N, p_e) g_i(p_e) dp_e$. We varied all w_i to maximize the probability

$$\prod \frac{e^{-\lambda} \lambda^{k}}{k!} = \prod_{T>0,N} \frac{e^{-hist_{pred}(T,N \mid w_0 \ldots w_{80})} \left(hist_{pred}(T,N \mid w_0 \ldots w_{80}) \right)^{hist^{*}_{meas}(T,N)}}{hist^{*}_{meas}(T,N)!} \quad (8)$$

that a given prediction leads to the observed vote distribution, assuming a Poisson distribution for the individual votes, where λ is the expected number of events and k is the actual number of events. This correctly weights even small histogram numbers, including zero. Fitting was implemented in both Matlab (Mathworks) and Mathematica (Wolfram Research), yielding identical results.

After edge elimination, we collected all skeleton nodes for all redundantly annotated skeletons that were still connected to a source seed area near the soma by a continuous path of edges, using connected components. This constituted the RESCOP consensus skeleton. The remaining skeleton pieces were discarded. For methods to reuse the discarded skeleton pieces, especially for locally dense skeletonization, see below and **Supplementary Figures 3** and **4**.

Accuracy of RESCOPed skeletons. The calculation made to decide whether to eliminate an edge can be extended to calculate the probability that the decision was wrong and that the RESCOPed consensus skeleton therefore contains an error at that point.

For a given (T,N) the probability that $p_e > 0.5$ is

$$p_{keep}(T,N) = \int_{0.5}^{1} p(p_e \mid T,N)\, dp_e \quad (9)$$

If the edge is kept, the probability of having done so erroneously is $1 - p_{keep}(T,N)$. Conversely, if the edge is eliminated, the error probability is $p_{keep}(T,N)$. The decision rule (equation (3)) to keep an edge if and only if $p_{keep}(T,N) > 0.5$ minimizes the error probability

$$p_{err}(T,N) = \min\left(p_{keep}(T,N), \left(1 - p_{keep}(T,N)\right) \right) \quad (10)$$

and is thus optimal. To calculate the error rate for a given N, we sum $p_{err}(T,N)$ over T weighted by the probability ($\int_0^1 p(T \mid N, p_e)\, p(p_e)\, dp_e$) of T occurring.

$$p_{err}(N) = \sum_{T=0}^{N} \left(p_{err}(T,N) \int_0^1 p(T \mid N, p_e)\, p(p_e)\, dp_e \right) \quad (11)$$

This is the probability that there is still an error after finding the consensus of N skeletons at a given location. The mean path length between errors is then $r_p / p_{err}(N)$. r_p was used rather than the edge length because our voting procedure creates a correlation between errors of neighboring edges (**Fig. 3b–e**).

Focused reannotation. To estimate the average annotation redundancy for the case in which each edge is reannotated until a given accuracy goal is reached, we ran a Monte Carlo simulation as follows. We picked a p_e using $p(p_e)$ as the probability density, repeatedly tossed a coin biased with p_e, and incremented T and N accordingly with each toss, until $p_{err}(T,N)$ was less than the set accuracy goal or N_{max} was reached. The set accuracy goal was then corrected for the residual errors for those runs that reached N_{max} (N_{max} = 6,000, E1088 single-cell data and K0563 data, **Fig. 5c** and **Supplementary Fig. 5d**), with the exception of the dense skeletonization data in which the number of runs that reached N_{max} was small (N_{max} = 200, **Fig. 5c** and **Supplementary Fig. 5b**).

Random skeleton reseeding. For the nearly dense reconstruction of neurites in a limited region (**Fig. 3f**,**g**) we did not seed at the somata, because they were not contained in the region, but used a strategy of random seeding and iterated reseeding (**Supplementary Figs. 3** and **4**). Briefly, annotation was restricted to a sphere around a seed point, but seed-point placement was iterated several times, each time using as new seeds the end points of the skeletons from the previous iteration. We modified RESCOP so that an enforced ending near a tested edge did not count against that edge, whereas a natural end point did, and placed the skeletons remaining after edge elimination into clusters on the basis of the proximity of the skeleton pieces. We also accounted for the possibility that some of the randomly placed initial seed points were in the same neurite. For details see **Supplementary Figures 3** and **4**.

Reconstruction cost estimation. To calculate reconstruction costs, we estimated that a block of mouse retina sized 120 × 80 × 130 μm^3 contains ~460 bipolar cells with ~0.3–0.8 mm path length each and ~40 ganglion cell somata with 1–2 mm dendritic path length each, which in most cases is only part of the dendrite. Annotating these at 6 h mm^{-1} with four-fold redundancy would take 7,500 work hours. In our institution, each undergraduate student works ~27 h per month. The reconstruction of all bipolar and ganglion cells at four-fold redundancy would thus take 3 months with a team of 120 annotators.

32. Crook, S., Gleeson, P., Howell, F., Svitak, J. & Silver, R.A. MorphML: level 1 of the NeuroML standards for neuronal morphology data and model specification. *Neuroinformatics* **5**, 96–104 (2007).

doi:10.1038/nn.2868