Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data

Fernando Amat[1], William Lemon[1], Daniel P Mossing[1], Katie McDole[1], Yinan Wan[1], Kristin Branson[1], Eugene W Myers[2] & Philipp J Keller[1]

The comprehensive reconstruction of cell lineages in complex multicellular organisms is a central goal of developmental biology. We present an open-source computational framework for the segmentation and tracking of cell nuclei with high accuracy and speed. We demonstrate its (i) generality by reconstructing cell lineages in four-dimensional, terabyte-sized image data sets of fruit fly, zebrafish and mouse embryos acquired with three types of fluorescence microscopes, (ii) scalability by analyzing advanced stages of development with up to 20,000 cells per time point at 26,000 cells min^{-1} on a single computer workstation and (iii) ease of use by adjusting only two parameters across all data sets and providing visualization and editing tools for efficient data curation. Our approach achieves on average 97.0% linkage accuracy across all species and imaging modalities. Using our system, we performed the first cell lineage reconstruction of early *Drosophila melanogaster* nervous system development, revealing neuroblast dynamics throughout an entire embryo.

Following the dynamic behavior of cells in complex organisms, such as fruit fly, zebrafish and mouse embryos, is a fundamental goal of developmental biology[1,2]. The ability to accurately reconstruct the positions, movements and divisions of cells is crucial for understanding the morphogenesis of tissues and organs[3–6], linking developmental history to cell function[7], dissecting the role of differential gene expression in directing cell-fate decisions[8,9], quantitatively analyzing mutant defects[10], determining stereotypy in cell behavior[11] and experimentally testing models of development[12,13].

State-of-the-art live imaging technologies, such as light-sheet and confocal fluorescence microscopes, are capable of recording development at the cellular level for several days[14]. Frequently, nucleus-specific fluorescent labels are employed to reveal cellular dynamics across cell populations or even entire embryos[3,4,15–17]. Such imaging experiments easily generate terabytes of image data, comprising up to tens of thousands of cells imaged for thousands of time points. However, existing manual or semiautomated approaches to reconstructing cell lineages do not scale to data sets of such complexity and size[18].

Automated computational approaches have been developed to analyze such image data for small model organisms such as *Caenorhabditis elegans* embryos[15] and for early developmental stages of more complex organisms such as the early zebrafish blastula[19,20] and the *Drosophila* blastoderm[16,20]. However, a method for accurate, automated cell lineaging in later stages of development does not currently exist.

Four major challenges complicate automated cell segmentation and tracking in advanced developmental stages. First, image data are complex, i.e., the specimen often comprises a large number of densely packed cells with different shapes and complex behaviors. Second, image quality varies markedly across the specimen because of the limited physical penetration depth of the microscope. Third, data sets are large, and thus scalability is indispensable for any general computational approach. Fourth, high accuracy and robustness are required, because a few errors can fundamentally alter lineage results.

In general, computational approaches to cell segmentation and tracking can be divided into three main categories: contour evolution (for example, level sets)[21], state-space models (for example, particle filters)[22] and data association methods (for example, graph matching)[20,23]. All of these approaches have individual strengths and weaknesses, and none solves all of the challenges outlined above. For example, contour evolution methods can be slow to compute for thousands of three-dimensional (3D) objects, but they provide detailed cell outlines, even during cell division. State-space models and data association methods do not scale well with increasing object density. However, they can model complex spatiotemporal knowledge about the system.

Here we present a new hybrid approach to segmenting and tracking cell nuclei, which facilitates fast and accurate reconstruction of cell lineages. Our method operates on three principles. First, we reduce data size and complexity by using low-level segmentation to generate supervoxels. Second, we use parametric contour evolution, which provides good segmentation and tracking results in easy cases. Third, we flag areas where contour evolution might have failed and use spatiotemporal local data association rules to improve results. This division of labor provides higher accuracy and orders-of-magnitude faster computation

[1]Howard Hughes Medical Institute, Janelia Farm Research Campus, Ashburn, Virginia, USA. [2]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. Correspondence should be addressed to F.A. (amatf@janelia.hhmi.org) or P.J.K. (kellerp@janelia.hhmi.org).

RECEIVED 8 JANUARY; ACCEPTED 16 JUNE; PUBLISHED ONLINE 20 JULY 2014; CORRECTED ONLINE 20 AUGUST 2014 (DETAILS ONLINE); DOI:10.1038/NMETH.3036

time compared to existing methods. We also augment our automated computational pipeline with visualization and editing methods to efficiently curate the results and obtain error-free reconstructions.

We show that our framework rapidly and accurately reconstructs cell lineages from large-scale image data sets obtained with different microscopy methods and for different model organisms. As a case study, we systematically analyzed the lineages and cell behavior of neural progenitors in the entire early *Drosophila* ventral nerve cord.

RESULTS

Automated segmentation and tracking of cell nuclei

We developed an automated computational pipeline to efficiently and accurately segment and track cell nuclei in 3D time-lapse fluorescence microscopy recordings with terabytes of image data (**Fig. 1** and **Supplementary Software 1**). In addition to high accuracy, our design goals included scalability, speed and generality.

First, we partitioned the 3D image volume recorded at each time point into supervoxels (**Fig. 1a**). A supervoxel is a connected set of voxels in space that all belong to a single nucleus, and each nucleus can be represented by multiple supervoxels. The use of supervoxels as the smallest image unit (instead of voxels) reduces complexity by approximately three orders of magnitude while retaining morphological information[24]. In addition, we applied a conservative global intensity threshold that further reduced the data size by at least threefold. By using watershed techniques and persistence-based clustering (PBC) agglomeration[25,26] (**Supplementary Fig. 1** and Online Methods), we created a hierarchical representation of all possible partitions of the image into supervoxels (**Fig. 1a**). Thus, only two parameters needed to be provided: the global background intensity threshold and the level, τ, at which the hierarchical representation is cut to generate the initial set of supervoxels (**Supplementary Software 1**). As each time point can be processed independently, this processing step is trivially parallelized.

Second, we connected supervoxels in space (segmentation) and time (tracking) to recover full cell lineages. We developed a sequential Bayesian approach with Gaussian mixture models (GMMs) to perform both tasks simultaneously using parametric contour evolution (**Fig. 1b**, **Supplementary Fig. 2**, Online Methods and **Supplementary Note 1**). The intensity profile of each nucleus can be modeled as a Gaussian distribution in 3D (**Supplementary Fig. 3**), and thus each image volume is described as a GMM. We applied local background estimation to each supervoxel using Otsu's method[27] to tighten the nuclear segmentation mask and improve fitting (Online Methods). The parametric model reduces the segmentation and tracking problem to finding ten parameters per nucleus: its 3D center, 3D covariance matrix (shape) and parent identity. Fitting a GMM to image data is a well-established estimation problem with efficient computational solutions[28] (Online Methods and **Supplementary Note 1**). The supervoxel image partition is also advantageous in this second step: it improves global convergence, processing speed and memory efficiency[29], as each Gaussian only has to cluster a few supervoxels instead of thousands of voxels. Moreover, instead of fitting a GMM at each time point independently, we propagated the solution obtained for time point t to the next time point, $t+1$. Because of the temporal coherence between consecutive time points, we imposed the *a priori* knowledge (Bayesian approach) that the location, shape and intensity of nuclei cannot change abruptly between time points.

Robust detection of cell divisions is a key requirement for the reconstruction of cell lineages. However, achieving high accuracy for this task is complicated by the sparse occurrence of cell divisions in the image data. To maximize detection accuracy, we again took advantage of the supervoxel partition: if two supervoxels are clustered into the same Gaussian but are not spatially connected, the cluster is flagged as a cell division candidate (**Fig. 1b**

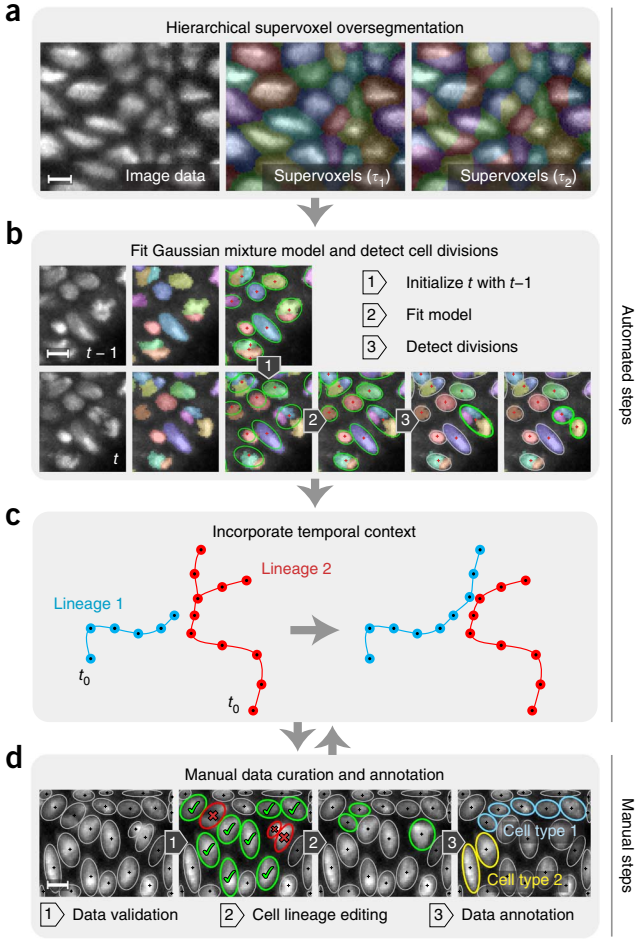

Figure 1 | Computational framework for nuclei segmentation and tracking. Schematic representation of the cell lineage reconstruction pipeline on example image data. For simplicity, only 2D slices of the 3D data are shown. (**a**) A hierarchy of possible segmentations (from coarse to fine) is calculated from the image data using watershed and PBC. The PBC distance threshold, τ, controls the merging of neighboring image regions and thereby defines which segmentation in the hierarchical model is selected to generate supervoxels. The higher the value of τ ($\tau_1 > \tau_2$), the coarser the segmentation, as more image regions are merged. (**b**) A GMM (green ellipsoids) is fit to the gray-scale image data to group supervoxels that belong to the same nucleus. For cell tracking, the GMM is sequentially propagated forward in time (steps 1 and 2). A module for cell division detection (step 3, green compared to gray ellipsoids) is used to control the number of objects in the GMM. (**c**) The algorithm determines local spatiotemporal windows in which the sequential GMM model might be erroneous; heuristic rules are applied to improve accuracy. (**d**) The results of the automated tracking and segmentation are output into a framework for data visualization, editing and annotation. Scale bars, 5 µm.

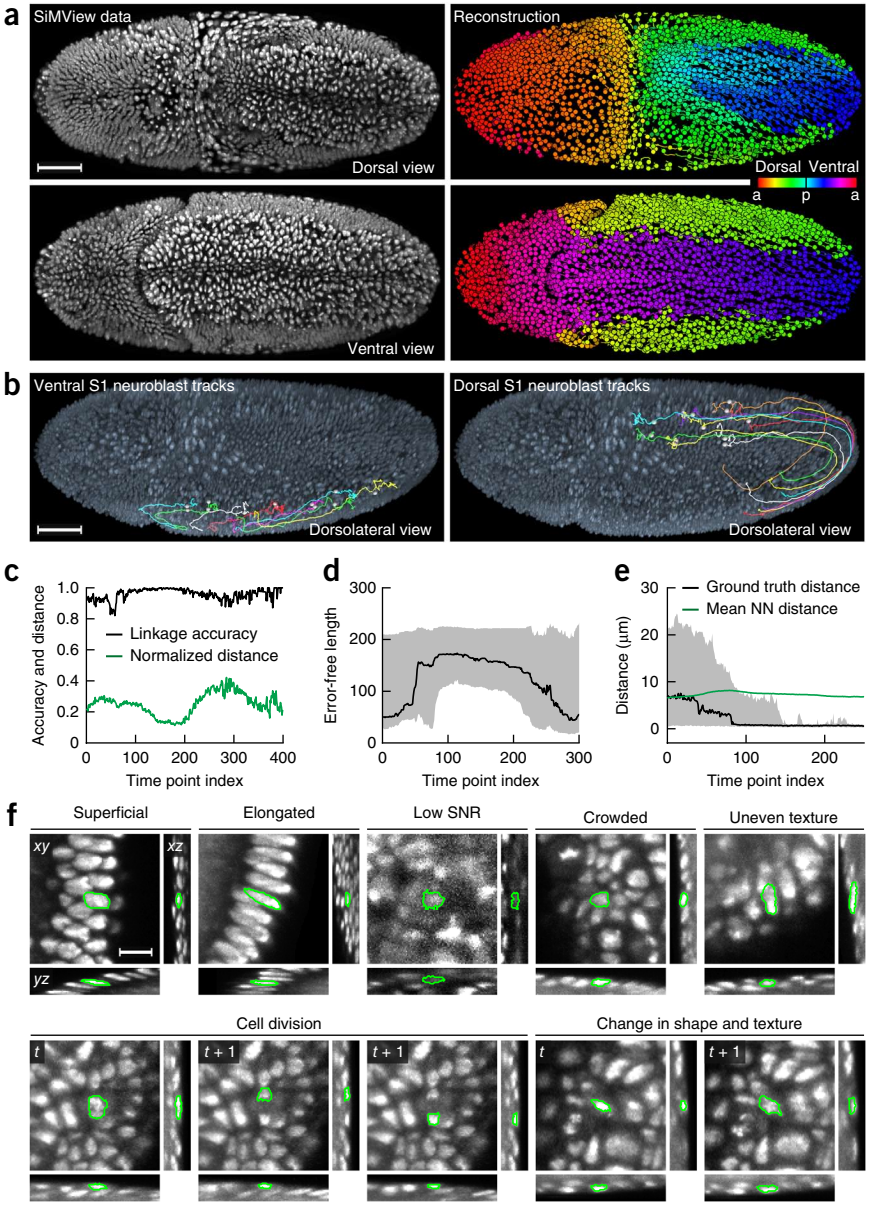

Figure 2 | Automated cell lineaging in *Drosophila* embryos. (**a**) Left, maximum-intensity projections of a SiMView recording of a nuclei-labeled (His2Av-mRFP1) stage 9 *Drosophila* embryo. Right, automated reconstruction of cell positions (spheres) and movements (lines, last 5 min) using a color code that indicates blastoderm lineage origin (**Supplementary Video 2**). The reconstruction comprises 501 time points (3–7 h AEL, stages 6–11). a, anterior; p, posterior. (**b**) Tracks of eight ventral (left) and dorsal (right) S1 neuroblasts superimposed with images at 4.9 h AEL (stage 10). The tracks are 270–450 time points long and show cell movements during germ band extension. (**c**) Average linkage accuracy (black) and average normalized Euclidean distance between automatically and manually detected nuclei centroids (green) for 235 neuroblasts. (**d**) Average error-free length of neuroblast cell lineages (black, median; gray, 25th and 75th percentiles; $n = 262$ on average). Rapid germ band elongation starts at time point 40. (**e**) Average distance between curated and corresponding uncurated cell tracks (black, mean; gray, s.d.) measured by backtracking cells from time point 250 (stage 10, $n = 241$ cells). Gastrulation starts at time point 0 (stage 6), and internalization of neural precursors occurs on average at time point 120 (stage 9, $n = 295$). The average nuclei neighbor (NN) distance is shown as a reference (green). (**f**) Orthogonal image slices, each centered on a nucleus representing a different type of image analysis challenge (green, automated segmentation results). SNR, signal-to-noise ratio. Scale bars, 50 µm (**a**,**b**); 10 µm (**f**).

and **Supplementary Table 1**). We then analyzed a local spatiotemporal window around each flagged object to determine whether a true cell division had occurred (**Fig. 1c**). These spatiotemporal windows are essential to incorporate information to which the sequential GMM model does not have access. Within each local window, we used data association approaches to cell tracking to consider multiple segmentation and linkage hypotheses. The hierarchical segmentation structure (Online Methods and **Supplementary Fig. 1**) is crucial to efficiently enumerate all possible solutions. This strategy allowed us to accurately reconstruct cellular dynamics in challenging regions of the image volume while avoiding an increase in computational complexity. In particular, we incorporated data association heuristic rules that identified background detections and corrected false track termination and wrong linkages in the sequential GMM (**Supplementary Fig. 4**, Online Methods and **Supplementary Note 2**).

Performance of the automated cell lineaging pipeline

We assessed the performance of the automated segmentation and tracking algorithm across three different model systems (fruit fly, zebrafish and mouse embryos) and three different types of fluorescence microscopes (a custom SiMView light-sheet microscope[16], a Carl Zeiss Lightsheet Z.1 microscope and a Carl Zeiss LSM 710 confocal microscope) (Online Methods). Seven users annotated a total of 42,947 ground truth data points to generate accuracy metrics for all model systems and microscopes. Moreover, we curated 116,820 data points across different developmental stages in *Drosophila*. We adjusted only two parameters of the pipeline, the background intensity threshold and the PBC threshold τ, across data sets to obtain the results presented in this section (**Supplementary Software 1** and **Supplementary Fig. 5**).

First, we analyzed performance on the SiMView data set of *Drosophila* embryonic development in detail (**Fig. 2a–e** and **Supplementary Videos 1–6**). Because of the complexity of the image data and the associated computational challenges (**Fig. 2f** and **Supplementary Videos 5** and **6**)[30], we used several complementary metrics to evaluate different types of errors (**Supplementary Note 3**). First, we reported the Euclidean distance between manually and automatically marked nuclei centroids as a measure of segmentation accuracy. Second, we normalized this distance by the nearest neighbor distance to

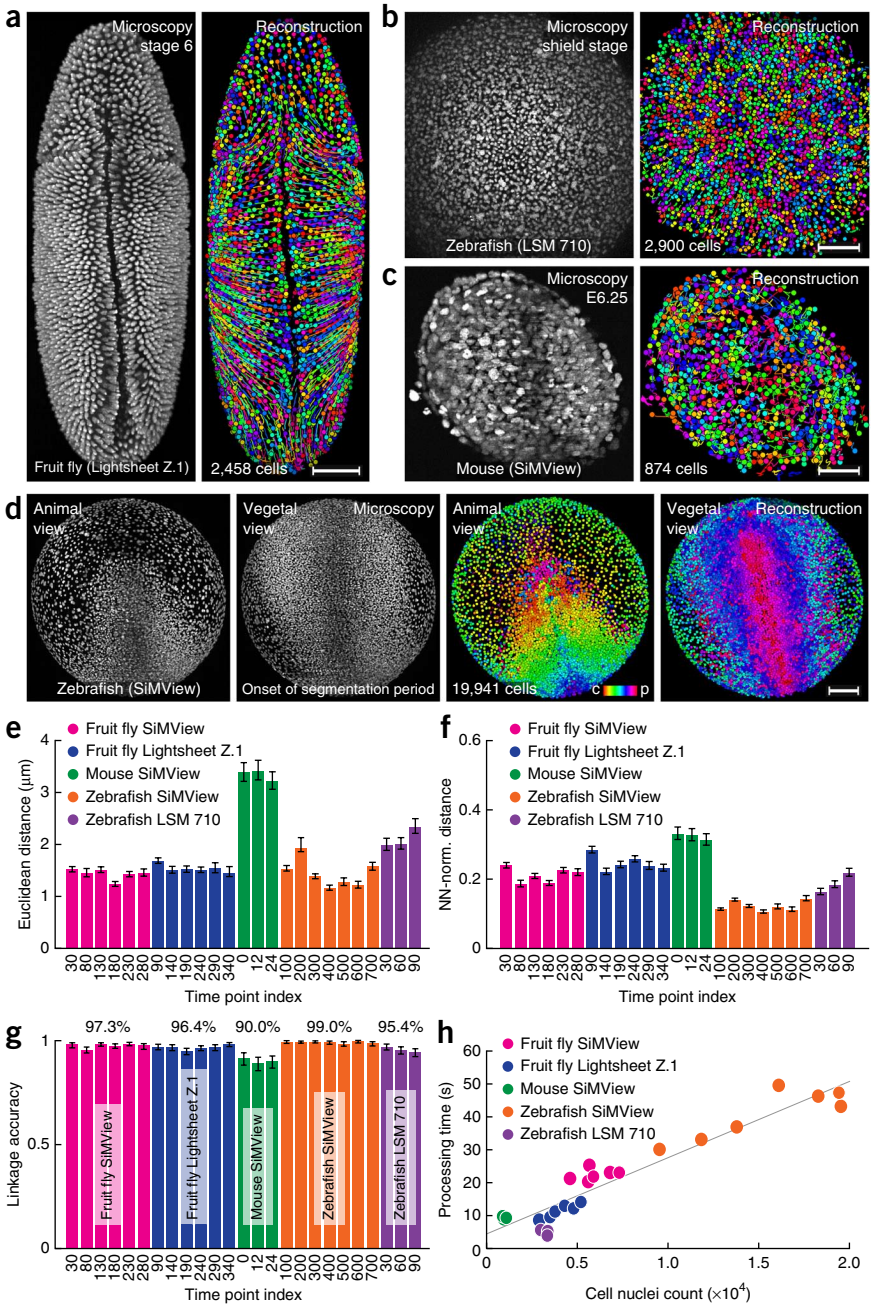

Figure 3 | Automated cell lineaging in zebrafish, fruit fly and mouse embryos using various microscopes. (**a**) Left, maximum-intensity projection of a 4-h Lightsheet Z.1 microscopy recording of the ventral half of a nuclei-labeled (His2Av-mRFP1) stage 6 *Drosophila* embryo. Right, reconstruction of cell positions (circles) and movements (lines, last ten time points) using random colors. 501 time points (stages 5–11) were reconstructed. (**b**) As in **a** but for a 3.4-h confocal microscopy recording of a nuclei-labeled (H2B-mCherry) zebrafish embryo during early gastrulation (101 time points from 50% epiboly). (**c**) As in **a** but for a 2-h SiMView microscopy recording of a nuclei-labeled (H2B–enhanced GFP (eGFP)) mouse embryo (26 time points from embryonic day (E) 6.25). (**d**) As in **a** but for a 13.4-h SiMView microscopy recording of a nuclei-labeled (H2B-eGFP) zebrafish embryo during late gastrulation (801 time points from the sphere stage). The radial gradient color code indicates the blastoderm lineage origin. c/p, blastoderm center/periphery. (**e**) Average Euclidean distance between automatically and manually detected nuclei centroids (from left to right: n = 1,211; 1,718; 2,408; 2,008; 2,041; 1,436; 1,330; 1,365; 1,744; 1,998; 1,327; 1,594; 715; 740; 841; 1,375; 2,655; 2,416; 1,469; 1,210; 1,369; 1,246; 1,206; 1,219; 1,202; error bars, 5th and 95th percentile confidence intervals). (**f**) As in **e** but normalizing individual distances to the local nearest-neighbor nuclei (NN-norm.) distance (error bars as in **e**). (**g**) Average linkage accuracy (error bars as in **e**). (**h**) Scatter plot of cell count versus computation time. Computation time scales linearly with cell counts (R^2 = 0.90, linear regression). The average processing speed is 26,000 cells min^{-1}. Scale bars, 50 µm (**a**,**c**); 100 µm (**b**,**d**).

account for bias toward oversegmentation. Third, we evaluated tracking accuracy by determining the fraction of correct linkage assignments in consecutive time points. The average Euclidean distance was below 50% of the nucleus radius, and the average pairwise linkage accuracy was above 95% for all time points (**Supplementary Tables 1**–**3**). Image quality, cell density and the magnitude of cell movements generally varied across the embryo (**Fig. 2e** and **Supplementary Fig. 6**) and constituted the most important factors influencing tracking accuracy (**Supplementary Figs. 7** and **8**). Particularly challenging scenarios included crowded populations of internalized cells and fast cell movements **Supplementary Fig. 6** and 7).

We also measured the average error-free length of cell lineages (**Fig. 2d**), i.e., the number of consecutive time points over which we encountered no segmentation or tracking errors. These segments, with an average length of 122 time points, were often separated by a segmentation mistake that propagated on average to the next two time points (**Supplementary Fig. 9**). Such mistakes do not propagate indefinitely, as the pipeline is capable of self-correction (**Supplementary Fig. 6f** and **Supplementary Note 3**). Moreover, the majority of cell tracks affected by reconstruction errors stayed within the median nearest neighbor distance of the true cell identity, even when cells moved hundreds of micrometers over hundreds of time points (for example, during *Drosophila* germ band extension; **Fig. 2b**,**e**).

A comparative analysis of the data sets for all biological model systems and microscopes (**Figs. 2a** and **3a**–**d** and **Supplementary Videos 1**–**23**) showed that the average Euclidian distance between manually and automatically determined centroids was below the nuclear radius (**Fig. 3e**,**f**), and the average linkage accuracy was between 90% and 99% (**Fig. 3g**). The mouse data set (**Fig. 3c** and **Supplementary Videos 16**–**18**) was an exceptionally challenging case because of the crowded cells, low temporal sampling, low image contrast and specimen rotation over time.

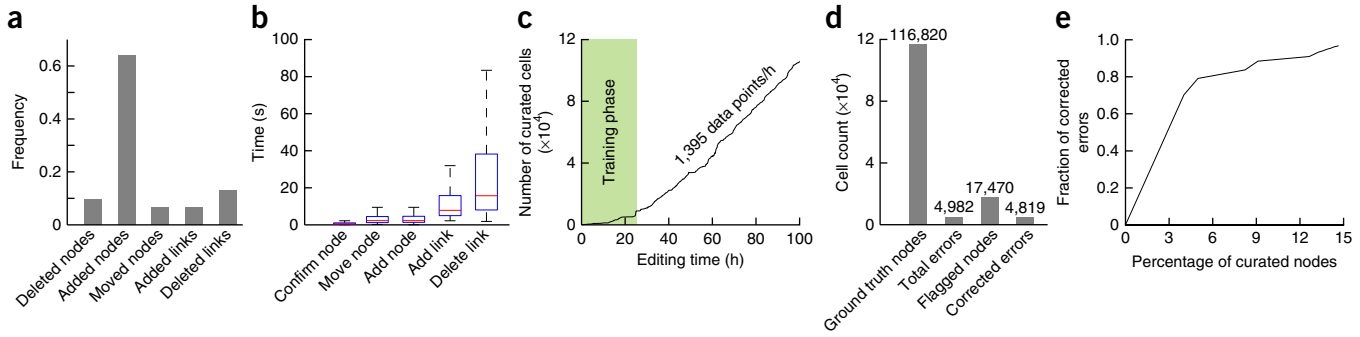

Figure 4 | Manual data curation and annotation. (**a**) Frequency of editing actions when correcting cell lineages with the manual curation framework (n = 40,310 operations). Confirmation actions are not shown. (**b**) Average time per editing action (red, median; blue, 25th and 75th percentiles; black, full data range; $n_{\text{Confirm Node}}$ = 16,330, $n_{\text{Move Node}}$ = 2,647, $n_{\text{Add Node}}$ = 25,498, $n_{\text{Add Link}}$ = 2,044, $n_{\text{Delete Link}}$ = 4,806). (**c**) Curation speed of a novice user using the manual curation framework. The user curated a data set with 116,820 data points within 100 h, including the initial training phase. (**d**) Bar plot showing (from left to right) the number of data points in the curated data set, data points edited by the user to fully curate the data set, data points flagged by the error guidance system as nodes potentially affected by reconstruction errors and errors in the uncurated data set that are corrected when manually validating nodes flagged by the error guidance system. Although the error guidance system flags only 15% of all data points for manual inspection, it makes the user aware of 97% of all errors in the data set. (**e**) Total fraction of corrected errors versus the percentage of manually curated data points for the cell lineage data set analyzed in **d** when using the error guidance system to sort nodes on the basis of curation priority.

We then evaluated the processing speed of our pipeline. The hybrid approach and parallelization on multicore CPU and graphics processing unit (GPU) platforms resulted in a linear scaling of computation time with the number of cells tracked (**Fig. 3h**). We measured an average processing speed of 26,000 cells min^{-1} on a single computer workstation (Online Methods). In all presented scenarios, the computation time was thus substantially shorter than the image acquisition time (**Fig. 3h** and **Supplementary Fig. 10**).

Visualization and manual curation of lineaging results

To meet the requirements of applications that demand error-free lineage reconstructions, we integrated a module for manual data inspection and editing in our reconstruction pipeline (**Fig. 1d**). For this purpose, we extended the CATMAID (Collaborative Annotation Toolkit for Massive Amounts of Image Data) platform for neuron tracing in large electron microscopy data sets (A. Cardona, Janelia Farm, personal communication)[31,32] by adding data handling of 3D time-lapse images with multiple color channels, orthogonal image slicing and temporal logic for enforcing biological constraints during data annotation (**Supplementary Software 2**). We used this module to store, curate and annotate the cell lineage data from all reconstructions performed in this study, thereby demonstrating its ability to handle multi-terabyte data sets with more than 10 million data points (**Supplementary Fig. 11**).

A novice user was able to curate the entire cell lineage reconstruction of early *Drosophila* nervous system development presented in the next section (**Supplementary Videos 24–28**). All editing actions were recorded in the database and used to estimate key data curation performance metrics (**Fig. 4a**,**b**). Within a 24-h period, the user reached maximum efficiency and achieved a sustained curation rate of 1,395 data points per hour (**Fig. 4c**). The user performed the full validation and curation of the 116,820 data points in approximately 100 h, including the initial training phase.

We also developed an approach for reducing the overall number of data points that need to be validated. Our computational framework automatically outputs a confidence score for each data point (Online Methods) and thereby guides the user to data points that are the most likely to be affected by errors (**Fig. 4d**). By focusing on data points flagged by this error guidance system, inspection and curation of only 15% of all data points was required to correct 97% of all errors in the automated *Drosophila* nervous system reconstruction (**Fig. 4d**,**e**). We measured a 28% improvement in data curation speed when using the error guidance system (n = 2,332 data points, two annotators) (Online Methods).

Cell lineage reconstruction of the early nervous system

To demonstrate the capabilities of our cell lineaging framework, we performed a cell lineage reconstruction of early *Drosophila* nervous system development (**Supplementary Videos 24–28**)[33]. Neuroblasts, the neural precursors, differentiate from neighboring blastoderm cells by lateral inhibition and delaminate from the surrounding cell sheet[34,35]. These cells divide asymmetrically to self-regenerate and produce a ganglion mother cell (GMC), which in turn divides symmetrically to generate two neurons or a neuron and a glial cell.

Comprehensive tracking of neuroblasts has so far been technically impossible. Using our framework, we reconstructed, curated and analyzed the cell tracks and divisions of 295 neuroblasts (**Supplementary Videos 25–27**), including 92.4% of all neuroblasts in the first delamination wave (S1 neuroblasts). We used our data visualization module to identify neuroblasts on the basis of their arrangement in a stereotypic array at 4.4 h hours after egg laying (AEL; **Fig. 5a**). We further annotated ten neuroblast subtypes on the basis of their relative positions within the neuroblast array (**Supplementary Fig. 12**) using definitions from earlier studies[36]. We followed these neuroblasts backward in time to their original location in the blastoderm and forward in time through two divisions for S1 neuroblasts and one division for S2 neuroblasts (**Fig. 5a**,**b**, **Supplementary Fig. 13** and Online Methods).

To our knowledge, our cell lineage reconstruction, which comprises 116,820 fully validated data points (**Supplementary Video 28** and **Supplementary Data 1**), provides the first single-cell–resolution blastoderm fate map for almost all S1 neuroblasts

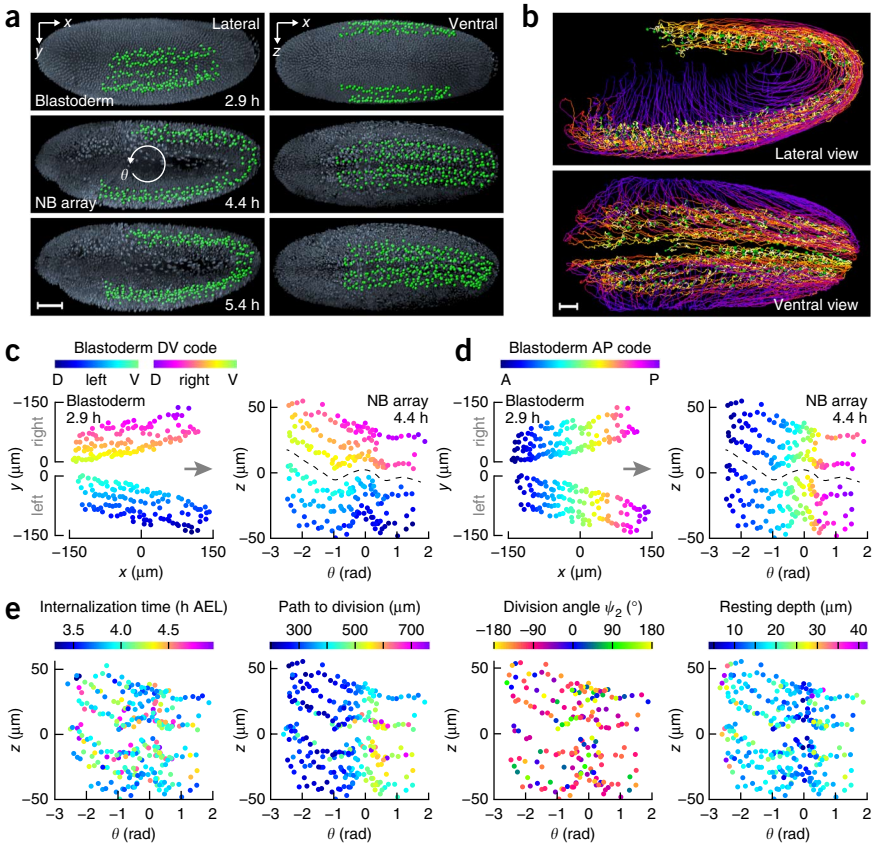

Figure 5 | Cell lineage reconstruction of the early *Drosophila* embryonic nervous system. (**a**) Location of neural precursors in the *Drosophila* embryo at the indicated time points (h AEL). NB, neuroblast. (**b**) Lateral and ventral views of all neural precursor cell tracks included in the cell lineage reconstruction (n = 235). Tracks are color coded for time (purple to yellow, 2.9–5.4 h AEL). (**c**) Single-cell–resolution fate maps of neuroblast precursors using a color code for dorsoventral (DV) position in the blastoderm. The color code is propagated forward in time from the blastoderm stage (left) to the neuroblast array at 4.4 h AEL (right; dashed line, midline). The x origin marks the center of the ventral neurectoderm along the a-p axis, and the y origin marks its ventral end. (**d**) As in **c** but using a color code for anteroposterior (AP) position in the blastoderm. (**e**) Features of cell behavior measured for all neural precursors using a color code in the neuroblast array at 4.4 h AEL. The z origin marks the average midline offset. The θ origin marks the embryo posterior end. $\hat{\theta}$ points to the end of the germ band. Scale bars, 50 μm (**a**); 20 μm (**b**).

in a *Drosophila* embryo. This map reveals an almost perfect neighborhood preservation between the ventral neurogenic regions in the blastoderm and the neuroblast array (**Fig. 5c**,**d**).

We also analyzed neuroblast movements and divisions (**Figs. 5e** and **6a**,**b**, **Supplementary Figs. 14** and **15** and **Supplementary Note 4**). Control over the cell division angle has been implicated in cell layer formation[37], and the first neuroblast divisions are an early stage in the creation of 3D structure from the two-dimensional (2D) ectoderm[38]. We therefore systematically analyzed cell division orientation[39,40] relative to the local embryo surface and the embryo midline. Our results show a preference of the newly formed GMCs to come to rest not only deeper but also more lateral relative to their sister cells, the self-renewed neuroblasts (polar angle $\eta = 28 \pm 15°$, azimuthal angle $\psi = 271 \pm 65°$, mean ± s.d., n = 415; **Fig. 6c**,**d** and **Supplementary Fig. 14**).

We then investigated whether different cell types in the early nervous system exhibit distinct signatures in their dynamic cell

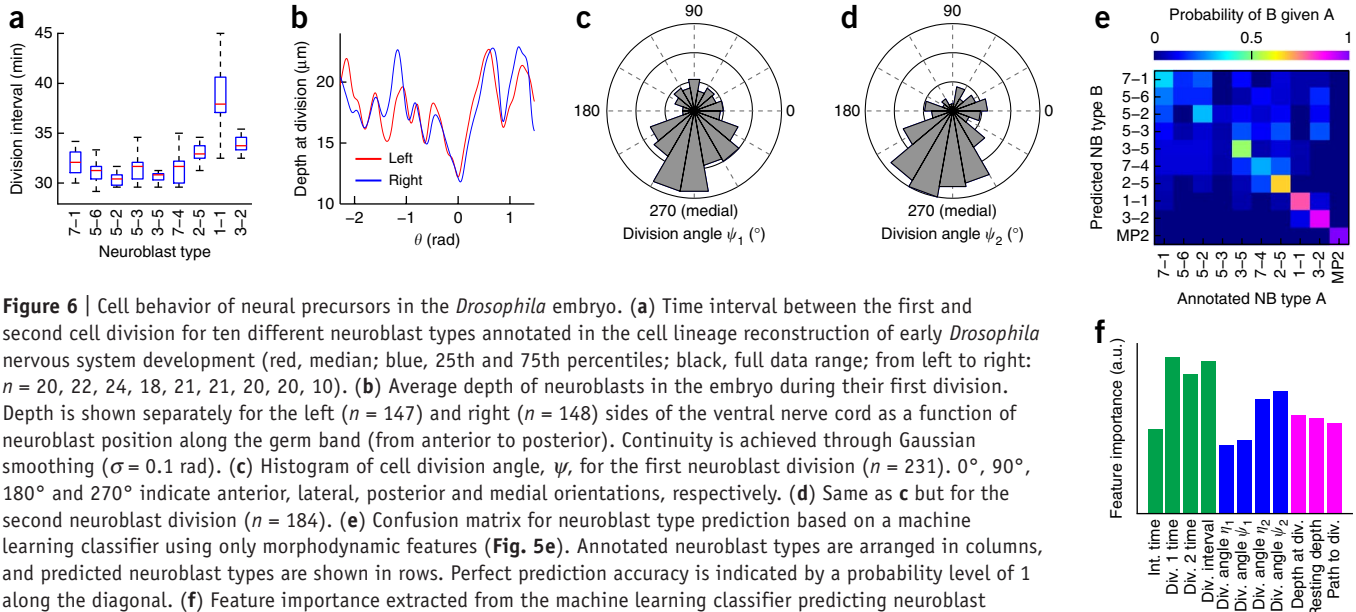

Figure 6 | Cell behavior of neural precursors in the *Drosophila* embryo. (**a**) Time interval between the first and second cell division for ten different neuroblast types annotated in the cell lineage reconstruction of early *Drosophila* nervous system development (red, median; blue, 25th and 75th percentiles; black, full data range; from left to right: n = 20, 22, 24, 18, 21, 21, 20, 20, 10). (**b**) Average depth of neuroblasts in the embryo during their first division. Depth is shown separately for the left (n = 147) and right (n = 148) sides of the ventral nerve cord as a function of neuroblast position along the germ band (from anterior to posterior). Continuity is achieved through Gaussian smoothing (σ = 0.1 rad). (**c**) Histogram of cell division angle, ψ, for the first neuroblast division (n = 231). 0°, 90°, 180° and 270° indicate anterior, lateral, posterior and medial orientations, respectively. (**d**) Same as **c** but for the second neuroblast division (n = 184). (**e**) Confusion matrix for neuroblast type prediction based on a machine learning classifier using only morphodynamic features (**Fig. 5e**). Annotated neuroblast types are arranged in columns, and predicted neuroblast types are shown in rows. Perfect prediction accuracy is indicated by a probability level of 1 along the diagonal. (**f**) Feature importance extracted from the machine learning classifier predicting neuroblast cell types (green, cell division timing; blue, cell division orientation; magenta, cell movement). A high level of feature importance indicates a strong effect on the prediction accuracy of the machine learning classifier. Int., internalization; Div., division.

behavior. Using machine learning models (Online Methods), we found that four of the ten annotated neuroblast cell types could be predicted with high accuracy from their behavior: using information primarily about the timing and orientation of cell divisions (**Fig. 6f**), we obtained prediction accuracies of 100% for MP2, 90% for 3–2, 79% for 1–1 and 67% for 2–5 neuroblasts (**Fig. 6e**), which are sixfold to tenfold higher than the probability of assigning the correct cell identity at random (10%). Notably, the models made these predictions without access to the spatial position information used for manual neuroblast annotation.

DISCUSSION

The automated cell lineaging framework presented in this study has three main strengths. First, its segmentation and tracking performance is robust across different biological model systems and fluorescence microscopes. Second, only a single computer workstation is needed, and computation time scales linearly with cell counts. Third, the framework requires adjustment of only two parameters and is thus easy to use by non-experts. In addition, we provide tools for efficient data visualization and editing. All software is open source and freely available for download at http://www.janelia.org/lab/keller-lab/.

Our performance results suggest that we did not sacrifice accuracy for speed, as our pipeline is more accurate and faster than state-of-the-art methods for cell lineaging in fluorescence microscopy (**Supplementary Tables 4–6**). In general, previous algorithms have been designed for early developmental stages, with up to a few thousand cells and low cell density. We found that these methods fail to correctly segment densely populated regions in more complex developmental stages and/or require a computation time that scales nonlinearly with the number of cells (**Supplementary Note 5**).

The fact that the processing speed of the presented pipeline exceeds the speed of image acquisition means that our framework is in principle suitable for image segmentation and cell tracking in real time. This opens up the prospect of building 'smart microscopes', in which real-time cell lineage reconstructions are used to design interactive experiments. For example, real-time cell tracking in the developing embryo could be used for optical manipulation at the single-cell level, as well as for real-time predictions of cell type identity on the basis of automatically reconstructed morphodynamic cell behavior.

The ability to efficiently perform system-level cell lineaging in complex multicellular organisms brings us closer to the fundamental goal of reconstructing the developmental building plans of vertebrates and higher invertebrates. We envision that the computational framework presented here will enhance the speed and accuracy of investigations requiring cell tracking information, enable quantitative comparisons of cell lineage reconstructions within and across species and provide crucial data for the development of computer models of embryonic development.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Cardona and the participants of the Janelia CATMAID hackathon for help with modifying the open source code of CATMAID; R. Chhetri and A. Pavlopoulos for invaluable contributions to ground truth annotations of the microscopy data sets; the Ilastik development team for help using Ilastik; M. Schroeder, H. Lacin, J. Truman and T. Lee for helpful discussion about the *Drosophila* nervous system; S. Srinivas and T. Watanabe (University of Oxford) for their generous help in exploring imaging assays for mouse embryonic development, helpful discussions about mouse embryo culturing and providing the CAG-TAG1 transgenic mouse strain; A. Denisin for her outstanding help developing SiMView live imaging assays; C. Akitake (Carl Zeiss) for her generous help executing the Lightsheet Z.1 experiments; S. Olenych and O. Selchow (Carl Zeiss) for supporting the Lightsheet Z.1 experiments; and C.-P. Heisenberg (Institute of Science and Technology Austria) for kindly providing the Tg(β-actin:H2B-mCherry) and Tg(β-actin:H2B-eGFP) zebrafish lines.
This work was supported by the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

P.J.K. and F.A. conceived of the research with input from E.W.M. F.A. developed the cell lineaging framework and performed the cell lineage reconstructions with input from P.J.K. and K.B. W.L. performed the *Drosophila* imaging experiments. D.P.M. curated and analyzed the reconstruction of early *Drosophila* nervous system development. K.M. performed the mouse imaging experiments. Y.W. and W.L. performed the zebrafish imaging experiments. F.A. and P.J.K. analyzed the data with input from K.B. F.A. and P.J.K. wrote the manuscript with input from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Megason, S.G. & Fraser, S.E. Imaging in systems biology. *Cell* **130**, 784–795 (2007).
2. Khairy, K. & Keller, P.J. Reconstructing embryonic development. *Genesis* **49**, 488–513 (2011).
3. Keller, P.J., Schmidt, A.D., Wittbrodt, J. & Stelzer, E.H.K. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science* **322**, 1065–1069 (2008).
4. McMahon, A., Supatto, W., Fraser, S.E. & Stathopoulos, A. Dynamic analyses of *Drosophila* gastrulation provide insights into collective cell migration. *Science* **322**, 1546–1550 (2008).
5. Fernandez, R. *et al.* Imaging plant growth in 4D: robust tissue reconstruction and lineaging at cell resolution. *Nat. Methods* **7**, 547–553 (2010).
6. Bosveld, F. *et al.* Mechanical control of morphogenesis by Fat/Dachsous/Four-jointed planar cell polarity pathway. *Science* **336**, 724–727 (2012).
7. Lemon, W.C. & Keller, P.J. Live imaging of nervous system development and function using light-sheet microscopy. *Mol. Reprod. Dev.* doi:10.1002/mrd.22258 (2014).
8. Murray, J.I. *et al.* Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* **5**, 703–709 (2008).
9. Liu, X. *et al.* Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. *Cell* **139**, 623–633 (2009).
10. Held, M. *et al.* CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat. Methods* **7**, 747–754 (2010).
11. Amat, F. & Keller, P.J. Towards comprehensive cell lineage reconstructions in complex organisms using light-sheet microscopy. *Dev. Growth Differ.* **55**, 563–578 (2013).
12. Trichas, G. *et al.* Multi-cellular rosettes in the mouse visceral endoderm facilitate the ordered migration of anterior visceral endoderm cells. *PLoS Biol.* **10**, e1001256 (2012).
13. Xiong, F. *et al.* Specified neural progenitors sort to form sharp domains after noisy Shh signaling. *Cell* **153**, 550–561 (2013).
14. Keller, P.J. Imaging morphogenesis: technological advances and biological insights. *Science* **340**, 1234168 (2013).
15. Bao, Z. *et al.* Automated cell lineage tracing in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **103**, 2707–2712 (2006).
16. Tomer, R., Khairy, K., Amat, F. & Keller, P.J. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nat. Methods* **9**, 755–763 (2012).

17. Swoger, J., Muzzopappa, M., Lopez-Schier, H. & Sharpe, J. 4D retrospective lineage tracing using SPIM for zebrafish organogenesis studies. *J. Biophotonics* **4**, 122–134 (2011).
18. Giurumescu, C.A. *et al.* Quantitative semi-automated analysis of morphogenesis with single-cell resolution in complex embryos. *Development* **139**, 4271–4279 (2012).
19. Olivier, N. *et al.* Cell lineage reconstruction of early zebrafish embryos using label-free nonlinear microscopy. *Science* **329**, 967–971 (2010).
20. Kausler, B.X. *et al.* A discrete chain graph model for 3D+t cell tracking with high misdetection robustness. *ECCV* 144–157 (2012).
21. Li, K. *et al.* Cell population tracking and lineage construction with spatiotemporal context. *Med. Image Anal.* **12**, 546–566 (2008).
22. Smal, I. *et al.* Multiple object tracking in molecular bioimaging by Rao-Blackwellized marginal particle filtering. *Med. Image Anal.* **12**, 764–777 (2008).
23. Jaqaman, K. *et al.* Robust single-particle tracking in live-cell time-lapse sequences. *Nat. Methods* **5**, 695–702 (2008).
24. Amat, F., Myers, E.W. & Keller, P.J. Fast and robust optical flow for time-lapse microscopy using super-voxels. *Bioinformatics* **29**, 373–380 (2013).
25. Skraba, P., Ovsjanikov, M., Chazal, F. & Guibas, L. Persistence-based segmentation of deformable shapes. *Proc. CVPR* 45–52 (2010).
26. Vincent, L. & Soille, P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 583–598 (1991).
27. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
28. Bishop, C.M. *Pattern Recognition and Machine Learning* (Springer, 2007).
29. Dasgupta, S. & Schulman, L.J. A two-round variant of EM for Gaussian mixtures. *Proc. UCAI* 152–159 (2000).
30. Milan, A., Schindler, K. & Roth, S. Challenges of ground truth evaluation of multi-target tracking. *Proc. CVPR* 735–742 (2013).
31. Saalfeld, S., Cardona, A., Hartenstein, V. & Tomančák, P. CATMAID: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics* **25**, 1984–1986 (2009).
32. Cardona, A. Collaborative annotation toolkit for massive amounts of image data. *CATMAID GitHub Repository* https://github.com/acardona/CATMAID (2014).
33. Hartenstein, V. & Camposortega, J.A. Early neurogenesis in wild-type *Drosophila melanogaster*. *Rouxs Arch. Dev. Biol.* **193**, 308–325 (1984).
34. Doe, C.Q. Molecular markers for identified neuroblasts and ganglion mother cells in the *Drosophila* central nervous system. *Development* **116**, 855–863 (1992).
35. Hartenstein, V., Younossi-Hartenstein, A. & Lekven, A. Delamination and division in the *Drosophila* neurectoderm: spatiotemporal pattern, cytoskeletal dynamics, and common control by neurogenic and segment polarity genes. *Dev. Biol.* **165**, 480–499 (1994).
36. Broadus, J., *et al.* New neuroblast markers and the origin of the aCC/pCC neurons in the *Drosophila* central nervous system. *Mech. Dev.* **53**, 393–402 (1995).
37. Wang, F., Dumstrei, K., Haag, T. & Hartenstein, V. The role of DE-cadherin during cellularization, germ layer formation and early neurogenesis in the *Drosophila* embryo. *Dev. Biol.* **270**, 350–363 (2004).
38. Technau, G.M., Berger, C. & Urbach, R. Generation of cell diversity and segmental pattern in the embryonic central nervous system of *Drosophila*. *Dev. Dyn.* **235**, 861–869 (2006).
39. Siegrist, S.E. & Doe, C.Q. Extrinsic cues orient the cell division axis in *Drosophila* embryonic neuroblasts. *Development* **133**, 529–536 (2006).
40. Bowman, S.K., Neumuller, R.A., Novatchkova, M., Du, Q. & Knoblich, J.A. The *Drosophila* NuMA homolog Mud regulates spindle orientation in asymmetric cell division. *Dev. Cell* **10**, 731–742 (2006).

ONLINE METHODS

Sample preparation and imaging of *Drosophila* embryos using SiMView microscopy. *Drosophila* live-imaging experiments were performed with embryos homozygous for the nuclear label histone 2A-mRFP (w-; P{w[+mC] = His2Av-mRFP1}; +, stock number 23560 from the Bloomington *Drosophila* Stock Center). Fluorescently labeled *Drosophila* embryos were dechorionated with 50% sodium hypochlorite solution (Sigma-Aldrich, 425044) and embedded in 1% low–melting temperature agarose (Lonza, SeaPlaque) in a 1.5-mm inner diameter (ID) × 20-mm glass capillary (Hilgenberg GmbH). After polymerization, the agarose cylinder was extruded just enough to expose the embryo outside of the glass capillary. The capillary holding the embryo was mounted vertically within the water-filled recording chamber of the SiMView light-sheet microscope[16] (temperature, 21.5 °C) so that the agarose was supported from below with the dorsal and ventral sides of the embryo facing the cameras. RFP was excited with scanned light sheets[3] using a 594-nm laser. Emitted light was imaged with Nikon 16×/0.8 numerical aperture (NA) water immersion objectives, 594-nm long-pass detection filters (Semrock) and Hamamatsu Orca Flash 4.0 sCMOS cameras (lateral pixel size in the acquired images, 406 nm). Using bidirectional illumination and bidirectional detection[16], four complementary views of the embryo were recorded. Image stacks of 154 planes encompassing the entire volume of the embryo with an axial step size of 2.03 µm were acquired at 30-s intervals for all four views.

The experiment shown in **Supplementary Video 1** captured *Drosophila* development in 2,881 time points for a 24-h period from 2.9 h AEL to larval hatching. Each multifused image stack is 183 Mb, resulting in a total data set size of 515 Gb.

Sample preparation and imaging of *Drosophila* embryos using a Lightsheet Z.1 microscope. *Drosophila* embryos imaged with the Lightsheet Z.1 (Carl Zeiss) commercial light-sheet microscope had the same genotype as the *Drosophila* embryos imaged with the SiMView microscope. Embryos were embedded in 1% low–melting temperature agarose in a 1.2-mm ID glass capillary. The polymerized gel was extruded to expose the embryo, and the capillary was mounted in the microscope so that the agarose cylinder was supported from above. In this microscope, the histone 2A-mRFP nuclear label was excited with a 561-nm light sheet and imaged with a Plan Apochromat 20×/1.0 NA water immersion objective (Carl Zeiss), a 585-nm long-pass detection filter and a PCO.edge sCMOS camera (lateral pixel size in the acquired images, 333 nm). The embryo was oriented so that the ventral side faced the camera. In order to image the ventral and dorsal hemispheres, the embryo was rotated by 180° between successive volume acquisitions. Each image stack contained 124 planes with an axial step size of 2.03 µm. Both views, ventral and dorsal, were acquired at 30-s intervals. Note that in order to avoid fusion artifacts arising from sequential multiview imaging, the dorsal and ventral image stacks were not fused in the subsequent image processing, and cell dynamics were only reconstructed for the ventral half of the embryo (**Fig. 3a**). This is in contrast to SiMView microscopy (as described above), where two cameras capture images of the dorsal and ventral sides simultaneously, and dorsal and ventral image stacks are readily fused without fusion artifacts[16].

The experiment shown in **Supplementary Video 7** captured *Drosophila* development in 1,304 time points for an 11-h period from 2 to 13 h AEL. Each single-view image stack is 228 Mb, resulting in a total data set size of 595 Gb.

Sample preparation and imaging of zebrafish embryos using SiMView microscopy. Zebrafish-line maintenance and SiMView experiments were performed according to the Institutional Animal Care and Use Committee of Janelia Farm (Howard Hughes Medical Institute). Fluorescently labeled zebrafish embryos imaged on the SiMView microscope were heterozygous for the nuclear label H2B-GFP expressed under the control of the β-actin promoter. Embryos were embedded in a 2.0-mm ID glass capillary filled with 0.5% low–melting temperature agarose prepared in E3 buffer (for 60× stock solution, 34.8 g NaCl, 1.6 g KCl, 5.8 g $CaCl_2 \times 2H_2O$ and 9.78 g $MgSO_4 \times 6H_2O$ are dissolved in 2 l H_2O, then the pH is adjusted to 7.2 with NaOH, and finally the solution is autoclaved). The polymerized agarose cylinder was extruded from the capillary to expose the embryo outside of the glass. The recording chamber of the microscope was filled with E3 buffer and equilibrated to a temperature of 21.5 °C. The capillary was then mounted vertically in the recording chamber so that the agarose was supported from below with the animal and vegetal poles of the embryo facing the cameras. GFP was excited with scanned light sheets using a 488-nm laser, and the images were acquired with Nikon 16×/0.8 NA water immersion objectives, 525/50-nm band-pass detection filters (Semrock) and Hamamatsu Orca Flash 4.0 sCMOS cameras (lateral pixel size in the acquired images, 406 nm). Using bidirectional illumination and bidirectional detection[16], four complementary views of the embryo were recorded. Image stacks of 262 planes encompassing the entire volume of the embryo with an axial step size of 3.25 µm were acquired at 60-s intervals for all four views.

The experiment shown in **Supplementary Video 19** captured zebrafish development in 1,170 time points for an 18-h period starting at the sphere stage at 6 hours post-fertilization (hpf). Each multifused image stack is 1.5 Gb, resulting in a total data set size of 1.7 Tb.

Sample preparation and imaging of zebrafish embryos using confocal microscopy. Zebrafish-line maintenance and confocal microscopy experiments were performed according to the Institutional Animal Care and Use Committee of Janelia Farm (Howard Hughes Medical Institute). Zebrafish embryos imaged with the Carl Zeiss LSM 710 laser-scanning confocal microscope were heterozygous for the fluorescent nuclear label H2B-mCherry expressed under the control of the β-actin promoter. Embryos were mounted in a deep-well microscope slide, embedded in 0.4% low–melting temperature agarose prepared in E3 buffer and covered with a #1 coverslip. The embryos were oriented so that the animal pole was nearest the cover slip. mCherry was excited with a 561-nm laser, and images were acquired with a Plan Apochromat 20×/0.8 NA air objective (Carl Zeiss). The lateral pixel size in the acquired images (1,024 × 1,024 pixels each) was 590 nm. Each image stack contained 35 planes with an axial step size of 3.25 µm, which covered about 15% of the volume of the embryo. Each plane was imaged in approximately 2 s, and image stacks were acquired at 2-min intervals.

The experiment shown in **Supplementary Video 12** captured zebrafish development in 101 time points for a 3.4-h period (temperature, 21.5 °C) starting at the 50%-epiboly stage. Each image stack is 96 Mb, resulting in a total data set size of 9.4 Gb.

Sample preparation and imaging of mouse embryos using SiMView microscopy. Mouse line maintenance and SiMView experiments were performed according to the Institutional Animal Care and Use Committee of Janelia Farm (Howard Hughes Medical Institute). Mouse embryos were obtained from natural matings by crossing female CD-1 mice with male CAG-TAG1 mice[41] maintained on a C57BL/6J background. Embryos were dissected at E5.5 in DMEM and F-12 (Invitrogen, 21041-025) and 10% FBS (Invitrogen, 10082-147) and imaged in DMEM and F-12 and 50% rat serum (WEC, Harlan) and maintained at 37 °C, 5% CO_2 and 5% O_2 on the SiMView light-sheet microscope as described above. Using bidirectional illumination and bidirectional detection[16], four complementary views of the embryo were recorded. Nuclei expressing histone 2B-eGFP from the CAG-TAG transgene were imaged using a 488-nm laser for fluorescence excitation, providing image stacks containing 356 planes with a step size of 2.03 µm every 5 min.

The experiment shown in **Supplementary Video 16** captured mouse development at E6.25 in 26 time points for a 2-h period. Each multifused image stack is 332 Mb, resulting in a total data set size of 8.4 Gb.

Multiview image fusion and data handling of SiMView microscopy data sets. SiMView four-view image data sets were processed and fused with our SiMView image processing pipeline, as described previously[16]. For efficient long-term data storage, background regions in the image stacks were removed using an image mask generated by adaptive thresholding of the Gauss-convolved image volume using a conservative threshold setting and retaining all foreground regions in the image volume. The foreground regions were subsequently compressed using lossless 3D wavelet compression following the JPEG2000 standard. The combined effect of these two post-processing steps is data reduction by a factor of 10 to 100 without affecting image quality or data sampling in any foreground region of the raw volumetric image data set. Together with the fourfold data size reduction in the multiview fusion process, total lossless data compression ratios are thus on the order of 40:1 to 400:1 compared to the raw, unfused SiMView image data, enabling efficient routine acquisition and storage of multi-terabyte SiMView image data sets.

Components of the processing workstation used for all cell lineage reconstructions. All computational reconstructions were performed on a single computer workstation with the following hardware components: two Intel Xeon E5-2687W CPUs, 192 GB DDR3 memory, an Nvidia Tesla Kepler K20 GPU, six Seagate Savvio 10K.5 ST9900805SS hard disks combined in a RAID-6 disk array and an Intel RMS25CB080 RAID module. For optimal processing speed in the cell lineaging framework, a good GPU and sufficient memory are of primary importance. The Tesla GPU can be replaced by a lower-cost GeForce GTX Titan GPU with little impact on performance. For a particularly cost-efficient build, slower CPUs and hard disks will generally suffice, as these components will only have a minor impact on processing speed.

For data visualization, editing and annotation, a CATMAID server with the following hardware components was used: two Intel Xeon E5-2690 CPUs, 128 GB DDR3 memory, six Intel 520 Series 480 GB SSDs combined in a RAID-6 disk array, an Intel RMS25CB080 RAID module and an Intel X520-SR1 10Gb fiber network adapter. In this case as well, slower CPUs and storage hardware will generally only have a minor impact on performance.

Supervoxel formation with watershed persistence-based agglomeration. We used a modified watershed algorithm[26] to group voxels into coherent regions belonging to the same nucleus. We incorporated three modifications to the classical watershed algorithm in order to avoid excessive oversegmentation due to Poisson noise in the light microscopy images. First, we preprocessed each 2D image in the 3D image stack with a 5×5 median filter. Second, we considered an anisotropic neighborhood of $2 \times 2 \times 1$ (74 elements in total) to avoid an overabundance of local minima. Third, and most importantly, we use an agglomeration technique based on PBC[25] to merge regions extracted from the watershed approach. The number of merged regions is defined by a parameter, τ. If the difference between the minimum intensity in a region and the intensity at the contact point with another region is below τ, then these two regions are merged. The merging has to be performed sequentially, as each merge affects other possible merges of neighboring regions (**Supplementary Fig. 1**). At negligible computational cost[25], we constructed a binary tree with all possible segmentations for all possible values of τ. This hierarchical segmentation was used to split and merge supervoxels during run time. Initially, τ should be set conservatively by the user (**Supplementary Software 1**) to avoid undersegmentation.

Sequential Gaussian mixture model. We modeled the intensity profile of each nucleus as a 3D Gaussian bell-shaped curve (**Supplementary Fig. 3**). The complete volumetric image data at each time point t was then modeled as a GMM:

$$I^t[n] \propto \sum_{k=1}^{K^t} \pi_k^t \mathcal{N}(x_n; \mu_k^t, \Sigma_k^t) \quad (1)$$

where K^t is the number of nuclei at time t, x_n are the 3D coordinates for the n^{th} voxel, and π_k^t, μ_k^t and Σ_k^t define the k^{th} Gaussian mixture at time t. Because information between two consecutive time points is correlated, we used a full Bayesian approach (**Supplementary Note 1**) in order to incorporate priors for each parameter in equation (1). In particular, three hyperparameters (α, β and ν) define the distributions that model the expected changes in intensity, mean location and shape (covariance), respectively, between consecutive time points. The larger the values of these hyperparameters, the more weight is given to each prior at time point t in order to estimate parameters at time point $t + 1$. For all data sets presented in this work, these parameters were set to $\alpha = 0.8 \times 10^{-5}$, $\beta = 0.1$ and $\nu = 1.0$.

The linkage of objects between time points (tracking information) is automatically obtained by following the evolution of a Gaussian in time. In order to handle cell divisions, we first solved equation (1) with a fixed K^t (although Gaussians can be removed if they occupy an image volume with low sum intensity). Then, we determined which Gaussians might contain two nuclei and split them into two Gaussians. Finally, we solved equation (1) again with the new value of K^t obtained by estimating cell divisions.

Initialization for tracking and segmentation. Because we did not have prior information at the first time point, we initialized the GMM with one Gaussian per supervoxel. The variational inference

doi:10.1038/nmeth.3036

solving equation (1) and the temporal logic rules successfully disregard excess Gaussians resulting from oversegmentation.

Detection of cell divisions. To detect cell divisions, we determined whether all supervoxels assigned to the same Gaussian formed a single connected region after applying Otsu's threshold[27] to each supervoxel individually. If multiple unconnected regions existed, we split the group of supervoxels into subgroups corresponding to connected components and assigned one new Gaussian to each subgroup. This split mechanism is important to detect true biological cell divisions as well as to correct grouping mistakes that occurred at an earlier time point. Ensuring that Gaussians never contain multiple unconnected supervoxels is generally important to prevent individual Gaussians from covering large regions of the image volume.

In order to avoid degenerate cases in the Otsu's thresholding process (for example, all voxels belonging to foreground or vice versa), we defined three advanced parameters. minNucleiSize defines the minimum number of voxels that should be considered foreground. If fewer voxels than minNucleiSize are foreground, the entire supervoxel is considered background. maxNucleiSize defines the maximum number of voxels (on an absolute scale) that are allowed to be considered as foreground, and maxPercentileTrimSV defines this limit on a relative scale with respect to the total number of voxels in the current supervoxel. For all experiments presented in this work, these three parameters were kept constant at minNucleiSize = 50, maxNucleiSize = 3000 and maxPercentileTrimSV = 0.4. However, these parameters can be modified in the advanced parameters section of the software configuration file (**Supplementary Software 1**).

Spatiotemporal context rules. The sequential GMM only incorporates first-order Markovian information. However, when performing cell tracking, we have further *a priori* temporal information on cell behavior that we can exploit. For example, after a cell division, the daughter cells should not divide again for a certain number of time points. Thus, we defined three temporal logic rules that are applied using a sliding time window over the last T time points processed in the sequential approach. For all experiments presented in this work, we used a temporal window of size ten.

Applying these combinatorial rules to all cells, however, is unnecessary, because usually only a small fraction of cells exhibits challenging dynamic behaviors at any given time point. Broad application of these rules would also consume large amounts of computational resources. Thus, we only applied combinatorial rules in critical cell lineaging events, mainly cell deaths and cell divisions. These events typically occurred for less than 3% of all cells per time point and were locally isolated in space.

We refer to the first rule as 'short-lived daughter'. Many cell divisions are artifacts due to spurious changes in image data appearance between consecutive time points that affect the coherence of supervoxels across time. For example, changes in fluorescence intensity, scattering effects or autofluorescent elements can produce a scenario that resembles a cell division event. However, these effects are temporary, and one of the tracks will die shortly after such a misdetection. We therefore eliminated any track that ended within L time points after the last cell division. For all reconstructions in this work, L was set to 5.

The second heuristic is 'track death extension with hierarchical segmentation'. Because we set a global parameter, τ, for the persistence-based clustering of watershed regions, undersegmentation can occur in low-contrast or crowded image regions (at a rate well below 1%). Undersegmentation at a given time point results in a cell death in the reconstruction, as the corresponding Gaussian cannot be propagated in time. Thus, for each ending track, we determined whether neighboring supervoxels in the next time point could be split into smaller regions. We used our hierarchical segmentation to formulate a local graph-matching problem between different possible levels of segmentation across time. The weights between edges corresponded to Jaccard distances between supervoxels, and the graph-matching problem could thus be solved efficiently using the Hungarian algorithm[42].

The third rule is 'distance of mother cell to division plane'. In order to determine incorrectly detected cell divisions (false positives), we considered that the dividing mother cell is usually located in between the two daughter cells. We calculated the plane equidistant to the centroids of the two daughters and computed the distance of the mother cell to this plane. For true cell divisions, this metric yields relatively low values, whereas for false positives, such as a recovery from undersegmentation, this metric typically yields relatively high values. Thus, to discriminate between true cell divisions and false positives, we defined a threshold (**Supplementary Software 1**) and broke the linkage between mother and the more distant daughter for any cell division with a distance of mother to division plane above this threshold. After analyzing a small training set, the threshold was set to 3.2 for all reconstructions in this study.

The fourth rule is a background detector, which will be explained in the next subsection.

Background detector. Modeling the gray-scale intensity information of each image volume as a GMM implies that all image parts with nonzero intensity levels need to be accounted for. However, some specimens contain many autofluorescent parts that do not correspond to nuclei. Therefore, we built a machine learning classifier for detecting background objects that uses several features within the same temporal windows used to evaluate the logic rules (**Supplementary Note 2**). In general, the properties of background objects are not coherent over short periods of time, in contrast to those of actual cell nuclei. The background detector, which takes advantage of these differences, is provided as an optional post-processing module in the cell lineaging pipeline. Note that all quality metrics presented in this study were evaluated before applying the background detector. We used a two-threshold hysteresis approach (thrHigh and thrLow) on the outcome of the classifier probability to remove tracks corresponding to autofluorescent structures instead of true cell nuclei. When the classifier detects an object with a background probability above thrHigh, it keeps deleting connected data points forward in time until the probability decreases below thrLow or until it encounters a cell division event. For all post-processed reconstructions presented in this work (**Supplementary Videos 2**, **8**, **13** and **20**), thrHigh was set to 0.7 and thrLow was set to 0.2 unless otherwise indicated. The value of thrHigh represents a tradeoff between the number of false positives and number of false negatives and can be changed in the advanced parameters section of the software configuration file.

Confidence score for guiding users to possible lineaging mistakes. Some of the prior information on temporal cell behavior can be used to detect possible mistakes in the cell lineage reconstruction, even if it is not straightforward to automatically correct such mistakes without affecting other correct lineages. Thus, in those cases where we estimated a high error probability but could find a spatiotemporal context rule to fix the issue, we labeled the output lineage data accordingly. Using this information and the CATMAID data curation interface, the user can be efficiently directed to local spatiotemporal windows that are likely to contain mistakes (**Fig. 4d**,**e**).

In particular, we used two simple rules to flag data points that should be manually revisited. First, we flagged all cell divisions and cell deaths, as these are rare events with high importance for an accurate cell lineage reconstruction. Second, we estimated the mean and s.d. of cell displacements at each time point. We used these statistics to flag trajectories with unusually large displacements (above 4 s.d.), indicating abnormal cell behavior.

Data curation time with the error guidance system. In order to estimate the improvement in curation time obtained using the error guidance system, we randomly selected neuroblast lineages from our reconstruction of early *Drosophila* nervous system development (**Fig. 5b**) at time point 200. The annotators then proceeded to curate the selected lineages backwards in time until reaching time point 0 using both curation methods presented in the main text: (i) a computer-guided curation using the error guided system, followed by (ii) a full curation with CATMAID. We implemented the error guidance system in CATMAID by introducing a hot key ('c') that directs the user to the next predicted low-confidence region for manual data curation. Each time an annotator started curating a new lineage, we measured the time that elapsed until the completion of data curation for the respective lineage. Using this methodology, two different annotators reconstructed a total of 12 lineages with both curation methods. The total time required in full curation mode was 1,625 s compared to 1,169 s using the error guidance system. This number shows a 28% improvement in data curation speed with the error guidance system. The annotators corrected a total of 34 mistakes using the full curation mode and 32 mistakes using the error guidance system.

Nervous system reconstruction. Neuroblasts were morphologically identified in the image data at a time point after delamination for each wave (**Supplementary Video 24**). The automatically tracked paths were proofread forward and backward in time (**Supplementary Videos 25–27**) so that the first asymmetric divisions were found wherever the cells could be followed unambiguously. Relative positions of neuroblasts based on literature were used to assign putative cell identities (**Supplementary Fig. 12**). Features of the paths taken by these cells, including internalization time point, resting depth of the nucleus, maximal migration velocity, path length to the resting point, time point of cell divisions and division angles, were calculated from annotated nucleus locations.

In order to describe the location of the embryo surface, which is required for definitions of several dynamic features, we constructed the 3D convex hulls of all nuclei at each time point. Body axes in the anterior-posterior, medial-lateral and ventral-dorsal directions were manually aligned for the initial time point and adjusted using the iterative closest point algorithm to correct for gradual drift of the specimen.

For the visualization of neuroblast patterns along the length of the germ band, we defined cylindrical coordinates (z, ϕ) with z, the transverse axis coordinate, and ϕ, the polar angle running from the ventral side, around the posterior tip at $\phi = 0$ up to the dorsal side.

In order to efficiently calculate the distance between each point and the convex hull, we first used a 3D k-d tree to store the points on the hull. On this structure, we found the k nearest neighbors. We found the triangles containing these points and calculated the distance from the point to the plane containing each triangle as

$$d_s(\vec{p}_0) = \frac{(\vec{p}_1 - \vec{p}_2)\times(\vec{p}_1 - \vec{p}_3)}{||(\vec{p}_1 - \vec{p}_2)\times(\vec{p}_1 - \vec{p}_3)||} \cdot (\vec{p}_1 - \vec{p}_0) \qquad (2)$$

for a nucleus at $\vec{p}_0$ and triangle vertices at $\vec{p}_1$, $\vec{p}_2$ and $\vec{p}_3$.

The distance from the neuroblast to the convex hull over time through the course of delamination follows a characteristic shape, reflecting an internalization phase between the embryo surface phase and the delaminated phase. A logistic fit of the form

$$d_{fit}(t) = d_0 + \frac{d_\infty}{1 + e^{-\mu(t - t_{1/2})}} \qquad (3)$$

was used to incorporate the global behavior of the trajectory to robustly identify the geometric center of the descent to identify a time point of delamination. Notably, this fit shows considerably better performance than a simple threshold criterion and defines both delamination depth (d_∞) and internalization time ($t_{1/2}$). Note that the nuclear label does not allow us to discuss canonically defined 'delamination' but only internalization of the nucleus, as the nucleus moves basally before the segregation of the neuroblast from the outer cell sheet.

In order to fully parameterize the asymmetric neuroblast division orientation, we calculated two spherical angles. Because neuroblasts divide approximately orthogonally to the outer cell sheet, one axis of the cell division coordinate system was defined along the local surface normal. The other axis was chosen as the transverse body axis projected onto the plane locally tangent to the surface. The origin for each division angle coordinate system was placed in the GMC daughter cell, identified as the one further from the embryo surface. The polar angle was then calculated as

$$\eta = \arccos(\hat{n} \cdot \hat{r}) \qquad (4)$$

for the unit surface normal $\hat{n}$ and the unit vector between daughter cells $\hat{r}$. The second, azimuthal angle was calculated as

$$\psi = sgn\left((\hat{r}_{\perp\hat{n}} \cdot \hat{z})\arccos\left(\frac{\hat{r}_{\perp\hat{n}}}{||\hat{r}_{\perp\hat{n}}||} \cdot \frac{\hat{\phi}_{\perp\hat{n}}}{||\hat{\phi}_{\perp\hat{n}}||}\right)\right) \qquad (5)$$

where $\hat{\phi}$ and $\hat{z}$ are the angular and radial unit vectors of the cylindrical coordinate system (**Fig. 5a**). The angle ψ was then reflected across the midline and the posterior tip of the egg to account for bilateral symmetry as well as the change in orientation of the germ band. Thus, the coordinate system is consistently defined relative to lateral and medial directions as well as relative to the directionality of the germ band. Because the neuroblasts are located

doi:10.1038/nmeth.3036

on the germ band, away from the lateral surfaces of the embryo, the transverse axis will contain a large component within the tangent plane, and thus the projection will remain robust. Final angles were calculated by averaging the vector between daughter cells between the first two time points. The calculation remained robust when averaging one to three time points (data not shown). We note that the limited spatial resolution in the dorsoventral direction (2 μm) created an up to 10% uncertainty in the angle calculations.

Predicting neuroblast type and transcription factor annotations using machine learning. We used the Matlab implementation of boosting with classification trees to learn models that can predict neuroblast type or transcription factor annotations on the basis of the morphodynamic features obtained from our cell lineage reconstruction. Because our data set contained only 295 samples (total number of neuroblasts in the cell lineage reconstruction), we used fivefold crossvalidation and 500 weak classifiers to estimate the prediction accuracy results. Each weak classifier in the boosting model was grown and then pruned until each leaf had at least five samples, and we used the surrogate techniques implemented in Matlab to accommodate missing values for cells that were not tracked up to the second cell division. The feature importance for each model was returned by Matlab in the same boosting function. For each tree, the importance was calculated by summing changes in the risk due to splits on every feature and dividing the sum by the number of branch nodes.

41. Trichas, G., Begbie, J. & Srinivas, S. Use of the viral 2A peptide for bicistronic expression in transgenic mice. *BMC Biol.* **6**, 40 (2008).
42. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Log. Quart.* **2**, 83–97 (1955).

Erratum: Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data

Fernando Amat, William Lemon, Daniel P Mossing, Katie McDole, Yinan Wan, Kristin Branson, Eugene W Myers & Philipp J Keller
Nat. Methods; **doi:10.1038/nmeth.3036; published online 20 July 2014; corrected online 14 August 2014**

In the version of this article initially published online, equation (2) in the Online Methods, which describes the expression used to calculate the distance between the location of a nucleus ($\vec{p}_0$) and the plane defined by the vertices of a triangle on the convex hull ($\vec{p}_1$, $\vec{p}_2$ and $\vec{p}_3$), was incorrect. The operations between the positions of the nucleus and triangle vertices ($\vec{p}_0$, $\vec{p}_1$, $\vec{p}_2$ and $\vec{p}_3$) were incorrectly shown as a scalar product. The correct operation is a subtraction. The operations between the pairwise differences of the positions of the triangle vertices ($\vec{p}_1$, $\vec{p}_2$ and $\vec{p}_3$) were also incorrectly shown as a scalar product. The correct operation is a vector product. These errors have been corrected for the print, PDF and HTML versions of this article.