A novel ground truth dataset enables robust 3D nuclear instance segmentation in early mouse embryos

Hayden Nunley[1,☯], Binglun Shao[1,2,☯], Prateek Grover[1], Jaspreet Singh[1], Bradley Joyce[3], Rebecca Kim-Yip[3], Abraham Kohrman[3], Aaron Watters[1], Zsombor Gal[3], Alison Kickuth[3], Madeleine Chalifoux[2,3], Stanislav Shvartsman[1,2,3,4], Eszter Posfai[3,*], Lisa M. Brown[1,*]

1 Center for Computational Biology, Flatiron Institute - Simons Foundation, New York, United States of America
2 Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey, United States of America
3 Department of Molecular Biology, Princeton University, Princeton, New Jersey, United States of America
4 The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America

☯These authors contributed equally to this work.
* eposfai@princeton.edu, lbrown@flatironinstitute.org

Summary

For investigations into fate specification and cell rearrangements in live images of preimplantation embryos, automated and accurate 3D instance segmentation of nuclei is invaluable; however, the performance of segmentation methods is limited by the images' low signal-to-noise ratio and high voxel anisotropy and the nuclei's dense packing and variable shapes. Supervised machine learning approaches have the potential to radically improve segmentation accuracy but are hampered by a lack of fully annotated 3D data. In this work, we first establish a novel mouse line expressing near-infrared nuclear reporter H2B-miRFP720. H2B-miRFP720 is the longest wavelength nuclear reporter in mice and can be imaged simultaneously with other reporters with minimal overlap. We then generate a dataset, which we call BlastoSPIM, of 3D microscopy images of H2B-miRFP720-expressing embryos with ground truth for nuclear instance segmentation. Using BlastoSPIM, we benchmark the performance of five convolutional neural networks and identify Stardist-3D as the most accurate instance segmentation method across preimplantation development. Stardist-3D, trained on BlastoSPIM, performs robustly up to the end of preimplantation development (> 100 nuclei) and enables studies of fate patterning in the late blastocyst. We, then, demonstrate BlastoSPIM's usefulness as pre-train data for related problems. BlastoSPIM and its corresponding Stardist-3D models are available at: blastospim.flatironinstitute.org.

Introduction

During preimplantation development of the mouse embryo, two consecutive cell fate decisions, coupled to cellular rearrangements, set aside precursors of extraembryonic tissues from cells which will form the body of the embryo. Live images of embryos expressing fluorescently tagged proteins are particularly useful for learning the rules by which cells in the embryo dynamically interact with each other to specify these fates during development; however, deriving mechanistic insights from these images depends on extraction of quantitative information about cellular features, such as the position of each cell or the expression levels of specific proteins within each cell. Accurate segmentation of nuclei is a first step towards such a goal, as a cell's nucleus is a good proxy for cell position relative to its neighbors and can contain information about cell-fate-specifying protein expression. To quantify these features, the segmentation must not only classify each voxel as foreground (i.e., belonging to nuclei) or background, but also assign each "instance" (i.e., nucleus) with a distinct label (S1 Intro Fig).

Studying the dynamics of development requires instance segmentation not for a single frame, but for a (3+t)-D series of images of a developing embryo. To observe both fate decisions in preimplantation embryos,

these movies start at the early morula stage (8-cell embryo) and end at the late blastocyst stage (>100-cell embryo), encompassing approximately 48 hours of development. Acquisition of a time lapse at sufficient spatial and temporal resolution to follow individual cells through 48 hours yields nearly 200 3D images (each composed of ≈60 2D slices), containing a total of ≈ 10,000 individual instances of nuclei; thus, manual segmentation of every instance in every frame is not feasible. Although classical image analysis methods have had success in automated nuclear segmentation [1–3], these methods often require high signal-to-noise ratio (SNR) images and tuning of parameters by hand. Shallow-learning methods, such as ilastik, offer an alternative solution for instance segmentation [4]; however, since these methods have relatively few trainable parameters, their performance saturates as the training set's size grows [4]. Supervised deep-learning methods have many trainable parameters; thus, the performance of these networks benefit greatly from large ground-truth sets, which allow the networks to learn salient features. Relative to classical and shallow-learning methods, deep-learning methods often generalize better across biological conditions and microscopy types [5].

Currently several supervised deep learning methods are available for nuclear segmentation of 3D images (S2 Intro Fig). Many widely used methods have architectures that form a U-shape and predict, at the network output, one or more semantic maps (S2 Intro Fig). The U-shape comes from progressive downsampling of the resolution, then upsampling to the original resolution. For example, QCANet, specifically designed to segment nuclei in early mouse embryos, uses two separate 3D U-Nets for instance segmentation. The two U-Nets are trained to predict the foreground-background segmentation and the center of each nucleus, respectively, and with these two outputs, marker-controlled watershed produces instances. Another method, U3D-BCD, uses a single modified U-Net to predict three different outputs: a foreground-background map, a map of instance contours, and a signed-distance map. The first two outputs are used to locate seeds for watershed performed on the signed-distance map. These examples illustrate that U-Nets predict semantic maps by which post-processing methods like marker-controlled watershed decode the instances.

Not all 3D instance segmentation methods require the U-shape though. Though the user can choose a U-Net backbone instead, Stardist-3D's default is to avoid downsampling by using a linear arrangement of convolutional blocks, followed by residual blocks that allow for deeper nets without performance loss. Stardist-3D predicts two outputs: whether a given voxel belongs to a nucleus and a set of distances to the nucleus boundary, assumed to be star-convex. Instead of a U-Net, another method, called RDCNet, applies the same block iteratively to refine an output (S2 Intro Fig). The output includes a foreground-background mask and a vector pointing from each voxel to its instance's center. Thus, these two methods avoid the U-shape and predict vectors, either to the boundary or the center. For a more detailed discussion of relevant methods, see the Methods section and S1 Table.

Since various deep-learning methods have different number of trainable parameters, different architectures and different assumptions about nuclear shape (*e.g.*, star convexity), it is difficult to know a priori which method will segment nuclei most accurately for any biological system of interest. To answer this question, ground-truth data is needed to a) train each network on relevant image annotations and b) to comprehensively test network performance by quantifying overlap between each instance in the ground-truth test set and each instance in the model output.

A study by Tokuoka *et al.* documented one of the first attempts to compare the performance of different deep-learning methods (including their method QCANet, another 3D U-Net [6], and 3D Mask R-CNN [7]) on a large ground-truth dataset of nuclear instance segmentation in the mouse embryo [8]. Their ground-truth dataset spans from the 2-cell stage to at most the 53-cell stage (S2 Table) and enabled state-of-the-art performance for QCANet on the early stages of development, up to approximately the 16-to-32-cell stage. The deterioration in performance of their model for later stages of development is likely due to the scarcity of training data past the 32-cell stage. The packing of nuclei in space becomes denser as development progresses, which leads to improper merging of two or more nuclei into a common instance.

Tokuoka *et al.*'s study demonstrates a clear need for improved nuclear instance segmentation that would perform accurately up to the end of preimplantation development (> 100-cell blastocyst stage) in live images. For example, the first fate decision in mammalian preimplantation, though initiated at the 8-to-16-cell transition, does not generate two fate-committed populations until the ≈ 64-cell stage [9]. Thus, to quantify this fate decision – which differentiates those cells on the surface of the embryo (the trophectoderm, or TE) from those on the inside (the inner cell mass, or ICM) – requires extending accurate instance segmentation from the 8-to-64 cell stages. Studying the next fate decision, in which ICM cells differentiate into epiblast and primitive endoderm cell populations that spatially segregate, requires accurate segmentation for later

stages, up to peri-implantation (>100 nuclei).

To this end, here we first generate a mouse line that expresses a near-infrared nuclear reporter H2B-miRFP720. H2B-miRFP720 is well suited for live imaging because of its reduced phototoxicity and its lack of spectral overlap with reporters in the visible range. Then, we generate a large dataset, called BlastoSPIM (1.0), of light-sheet images of H2B-miRFP720-expressing preimplantation embryos with corresponding ground-truth for nuclear instance segmentation. We use this dataset – that extends from the 8-cell stage to the >100-cell stage – to train and test five different deep-learning methods, including Cellpose, Stardist-3D, RDCNet, U3D-BCD, and UNETR-BCD. We find that the Stardist-3D model, trained on BlastoSPIM 1.0, achieves state-of-the-art performance, detecting nuclei with high accuracy in early to mid-stage preimplantation embryos, even with low SNR images. Next, to further improve segmentation accuracy at later embryonic stages, we generate a new ground truth dataset, termed BlastoSPIM 2.0, on blastocyst embryos and show that Stardist-3D trained on this dataset achieves similarly high nuclear segmentation accuracy even for embryos with >100 cells. Using the two (early and late) Stardist-3D models, we quantify how ICM-TE differences in nuclear aspect ratio become larger between the 32- and 64-cell stages, then decrease slightly by the $>$ 100-cell stage; by contrast, ICM-TE nuclear volume differences – non-existent at the 32-cell stage – develop by the 64-cell stage and persist to the $>$ 100-cell stage. We close by demonstrating that our dataset and corresponding models can aid in nuclear segmentation in other system, as in *Platynereis dumerilli* embryos.

Materials and methods

Transgenic mouse line generation

The H2B-miRFP720 transgenic mouse line was generated by targeting the TIGRE locus using the 2C-HR-CRISPR method [10]. Two targeting plasmids were constructed with InFusion cloning (Takara Bio), one consisting of 5' and 3' homology arms (each 1kb in length), surrounding H2B-miRFP720 driven by a CAG promoter and flanked by rabbit beta globin polyA sequence; the second construct contained an additional ORF-2A preceding H2B-miRFP720 flanked by a bGH polyA sequence. A single guide RNA (sgRNA) designed using CRISPOR [11] was used to target the TIGRE locus: CAUCCCAAAGUUAGGUGUUA (Synthego). CD1-IGS mice (Charles River strain 022) were used as embryo donors. Briefly, female CD1-IGS were superovulated at 5-7 weeks of age using 7.5IU PMSG (Biovendor) administered by IP injection followed by 7.5IU HCG (Sigma) by IP injection 47 hours post PMSG. Super ovulated females were mated to CD1-IGS stud males and checked for copulatory plug the following morning.

Cytoplasmic microinjection of 2-cell embryos was performed as previously described [10, 12]. Briefly, embryos were harvested at the 2-cell stage on E1.5 by flushing the oviducts with M2 Media (Cytospring) and each cell was microinjected with 100ng/ul Cas9 mRNA (made by IVT (mMESSGAE mMACHINE SP6 transcription kit, Thermo Fisher) using Addgene plasmid 122948), 30ng/ul donor plasmid and 50ng/ul sgRNA, using a Leica Dmi8 inverted epifluorescent microscope, an Eppendorf Femtojet and a Micro-ePore (WPI). Embryos were immediately transferred into the oviducts of pseudopregnant female CD1 mice. N0 pups were identified using over-the-arm PCR primers (Fwd:tcagcctacctcaccaactg, and Rev:ccccatcgctgcacaaaata) and outcrossed to CD1-IGS mice. N1 animals were genotyped using the same primers and the transgene was Sanger sequenced. The N1 generation was further outcrossed twice before incrossing the line to obtain homozygous mice. Homozygous and heterozygous offspring were distinguished using a wild-type PCR of the TIGRE locus (TIGRE WT Fwd:CTTTCCAGTGCTTCCCCAAC and TIGRE WT Rev: CCCTTTCCCAAGTCATCCCT).

The first mouse line showed varying levels of H2B-miRFP720 fluorescence in cells of preimplantation embryos, while the second ORF-2A-H2B-miRFP720 mouse line showed ubiquitous expression. Therefore, the ORF sequence was deleted in 2-cell embryos isolated from this mouse line using the following sgRNAs: GGUGACGCGGCGCUGCUCCA and CAUGCCCAUUACGUCGGUAA, resulting in a truncated ORF with a functional 2A peptide. Founders and subsequent generations were established from this line, herein referred to as the H2B-miRFP720 mouse line, and ubiquitous H2B-miRFP720 fluorescence was confirmed once again in embryos.

Other transgenic mouse strains used in this study include Cdx2-eGFP [13] and mT/mG [14]. The unpublished Halo-Yap mouse line was generated in the lab of Janet Rossant by targeting a Halo tag to the 5' end of the endogenous Yap allele. The Halo tag is visualized by adding JF646 Halo tag ligand to the

culture media. Note, in this work the Halo-Yap mouse line is only used to demonstrate that a combination of spectrally distinct reporters can be used together with H2B-miRFP720.

Dataset Acquisition

Embryos were obtained from naturally mated or superovulated H2B-miRFP720 females mated to either wild-type (CD1) or H2B-miRFP720 males. Embryos were isolated at E1.5 (2-cell), or E2.5 (8-cell) in M2 media and were cultured in Embryomax KSOM (Sigma-Aldrich) under paraffin oil (Life Global Paraffin Oil - LGPO from Cooper Surgical) in a V-shaped imaging chamber at 37°C, with 5% O_2 and 5% CO_2. Images were acquired on an InVi SPIM (Luxendo/Bruker). To limit light exposure to the embryo, we acquired a full 3D image of each embryo at 15- minute intervals, with 2.0 µm z-axis resolution and 0.208 µm x- and y-axis resolution. Typically, the embryos were imaged from the 8-cell stage until the 64-cell stage or to the >100-cell stage, resulting in a duration of 48 hours or more. Raw time-lapse images were compressed to keller-lab-block (klb) format, on the fly.

Dataset Annotation

Raw 3D images of developing embryos were manually annotated using AnnotatorJ, an ImageJ plugin that supports both semantic and instance annotation. Images were loaded into the tool as Z stacks in .tiff format. For all images, brightness and contrast were adjusted by using the 'auto' and 'reset' functions in ImageJ. 'Instance' was selected as the annotation type. For each nucleus, the top or the bottom slice was found by comparing consecutive Z slices, and a contour was drawn for every slice that contained the nucleus. The coordinates of the regions of interest (ROIs) enclosed by the contours were then saved in an individual file. After each instance was annotated, the contours were overlaid on the image to distinguish the instance from unannotated ones. Five individuals annotated, and an expert checked for annotation errors, via a custom MATLAB code, before incorporation into the dataset.

Dataset Characteristics

The BlastoSPIM 1.0 dataset includes 573 fully annotated 3D images of nuclei in mouse embryos, each manually curated for annotation. Across all images, there are 11708 individual nuclear instances annotated and 116 annotated polar bodies. Not all of these 3D images come from different time series. For example, for one developing mouse embryo, we annotated 89 consecutive time-points, and for another embryo, we annotated 100 consecutive time-points. Both of these time-lapse annotations extend from the 16-cell stage to the approximately 64-cell stage. The total number of distinct embryos imaged and annotated is 31.

Aside from diversity in developmental stage, the embryos in this dataset express different H2B-miRFP720 alleles (see details in mouse line generation) and were also imaged with different laser intensities. This diversity in SNR allows us to test whether model performance degrades significantly as SNR decreases. We quantify SNR in our case by calculating mean foreground intensities and mean background intensities. We report the distribution of SNRs, one point for each fully annotated 3D image, as the difference between mean foreground and mean background in (S1 Fig). For comparison, the background intensity – in gray values – typically has a mean of 118 and a variance of 10-14.

The BlastoSPIM 2.0 dataset consists of 80 annotated images of late-stage embryos (S2 Fig). This set includes 6628 nuclear instances. Because the lowest SNRs in the last blastocyst set were higher than the low SNR cutoff used for the original set (S1 Fig), we simply selected a few of the lowest SNR images from the late blastocyst set to incorporate into the existing low SNR set. When added to our original ground-truth set, the final number of annotated images is 653, and the number of annotated nuclear instances is 18336.

Dataset Splits and Evaluation Metric

When splitting our dataset into a training set, a test set, and a validation set, our main objective was to quantify how model performance varies as a function of both developmental stage and SNR. For BlastoSPIM 1.0 we created two separate test sets, one for low SNR and one for moderate SNR, each of which contained a diversity of developmental stages. We define "low SNR" and "moderate SNR" by comparing the mean

foreground intensity to the mean background intensity. The "low SNR" images all have a mean foreground intensity which is at most 134 gray values, approximately 15 gray values above the typical mean background intensity. For reference, the background intensity – in gray values – typically has a variance of 10-14 (S1 Fig). Within both the moderate SNR and the low SNR sets, we group annotated embryos based on their developmental stage, estimated by the number of nuclei (*i.e.*, ≈8-cell, ≈16-cell, ≈32-cell, ≈64-cell, >100-cell). Each set deliberately contains more images from earlier stages than later stages so that the total of number of nuclei per developmental stage is at least partially equalized across stages. Table 1 and S6 Table specify the composition of the moderate SNR and low SNR sets, respectively. From the BlastoSPIM 2.0 dataset, 8 embryos from various stages were used as a test set, as early as the 48-cell stage and as late as the 107-cell stage (S2 Fig). The remainder of the data, 72 annotated embryos, were either for validation or training. The exact breakdown is specified at blastospim.flatironinstitute.org .

To evaluate how well the models performed on the test sets, we computed the intersection-over-union (IoU) between the models' segmentation and the ground truth. We considered an instance in the models' segmentation to match an instance in the ground truth if the IoU between the two was at least 0.1. We acknowledge that this IoU cutoff is small; nonetheless, because one of our main future goals is the tracking of instances over time, an instance that weakly overlaps with the true 3D nuclear instance is preferred to having no instance at all. We also provide how performance varies as a function of this IoU threshold. For the sake of reproducibility, the train-test-validation split for all the models in Table 1 is specified on the BlastoSPIM website. We additionally specify the model hyperparameters used for each evaluation table.

Statistical Comparison of TE and ICM nuclei

The ICM-TE comparisons in this study are based on aspect ratio and volume. For the quantification of aspect ratio, we use the same procedure – based on calculating the moment of inertia tensor – described in [15]. The quantification of volume relies on directly the number of voxels in the instance and multiply the sum by the voxel volume.

For the nuclear properties we measured, such as the aspect ratio, we compared on an embryo-by-embryo basis, via a Mann-Whitney U-test, the median of the TE nuclei and the median of the ICM nuclei. The null hypothesis is that for the nuclear property being measured, like volume, the distribution of that property for ICM nuclei has the same median as the distribution for TE nuclei. Since this test returns a p-value for each embryo, we combine the results of the different embryos via a Fisher's combined probability test [16, 17], which assumes that the original p-values are independent of each other and equally trust-worthy. On the other hand, for comparing the same fate population across stages, the pooled distributions are directly compared via a Mann-Whitney U-test.

Results

0.1 Establishment of a near-infrared nuclear reporter mouse line

Multicolor imaging is key to simultaneous recording of morphogenesis and cell fate specification. To enable visualization of cell nuclei in concert with various other molecular markers, which are typically tagged with green, red or far-red fluorescent proteins, we generated a novel spectrally distinct near-infrared nuclear mouse line expressing H2B-miRFP720 (Fig 1(A)-(B)). First, using 2C-HR-CRISPR [10] we targeted CAG H2B-miRFP720 to the TIGRE locus [11]. Early preimplantation embryos from this line showed uniform H2B-miRFP720 expression; however, by the mid blastocyst stage significant dimming of the fluorescent signal was noted, even in freshly isolated embryos (data not shown). A second mouse line harboring CAG ORF-2A-H2B-miRFP720 in the TIGRE locus however did not exhibit the same dimming issue, and rather showed a slight increase in H2B-miRFP720 intensity during preimplantation development. We therefore used two sgRNAs to delete the ORF-2A with Cas9 in this line, resulting in a CAG H2B-miRFP720 line (hereafter referred to as the H2B-miRFP720 mouse line) with bright reporter expression across all preimplantation stages (Fig 1(A)-(B)). This mouse line not only allows simultaneous imaging of up to four different reporters in mouse embryos (Fig 1(C)), but also results in reduced phototoxicity. Furthermore, the long wavelength used for its detection makes H2B-miRFP720 ideal for deep-tissue imaging.

0.2 A novel ground-truth dataset of preimplantation mouse embryos for comparing nuclear-segmentation methods.

Using selective plane illumination microscopy (SPIM) we acquired 3D live images of H2B-miRFP720-expressing preimplantation embryos at various developmental stages. By careful manual annotation, we created a new ground-truth dataset with full 3D nuclear instance segmentation. This dataset, which we call BlastoSPIM 1.0 (concatenation of blastocyst and SPIM), is one of the largest and most complete of its kind (S2 Table) with

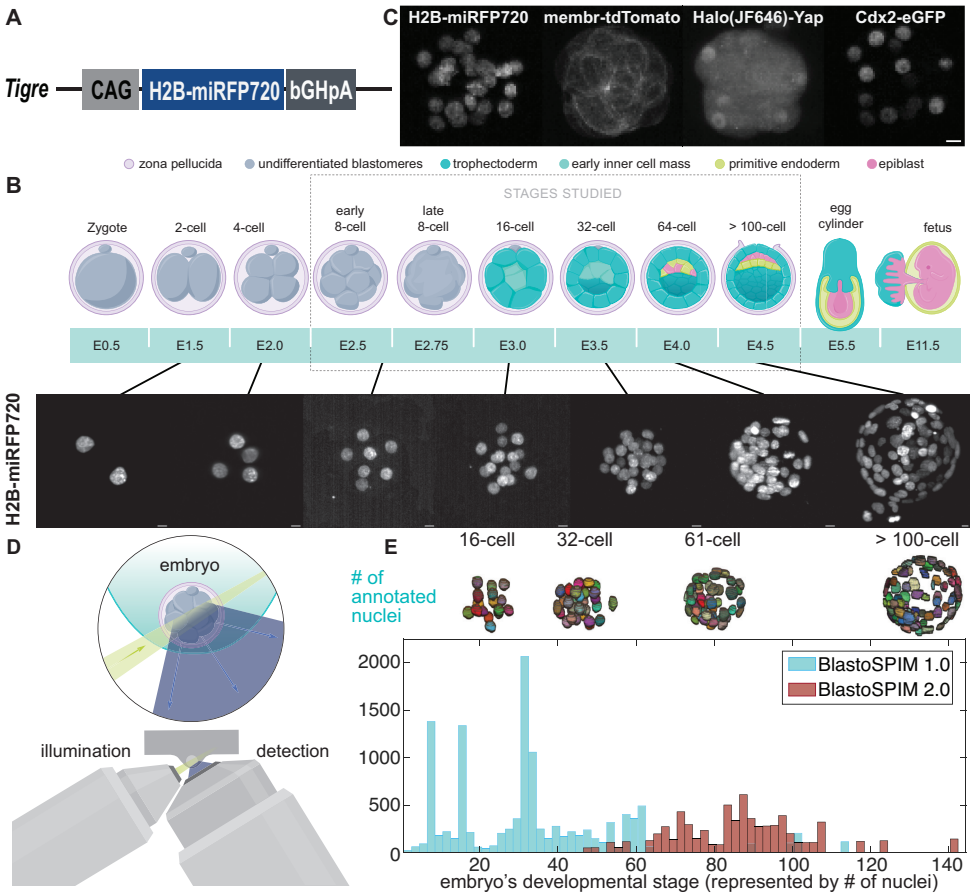

Fig 1. BlastoSPIM datasets, ground truth of nuclear instance segmentation for embryos expressing a novel near-infrared nuclear marker. (A) Schematic of targeted TIGRE locus with the CAG-H2B-miRFP720 insert. (B) Top: cartoon of preimplantation development in the mouse. After fertilization, the zygote undergoes rounds of division in the oviduct. At the 8-cell stage, compaction and polarization occur. By the 32-cell stage, a subset of cells called the trophectoderm (TE) form the embryo's surface; the remaining cells form the inner cell mass (ICM). The ICM cells begins to pattern into two fates, primitive endoderm (PE) and epiblast (EPI), by the 64-cell stage; by implantation, around the > 100-cell stage, the two inner fates are spatially segregated. Bottom: Maximum-intensity projected images – acquired with SPIM – of preimplantation embryos expressing H2B-miRFP720 at different developmental stages. Scale bar: 10 μm. (C) Preimplantation embryo expressing four spectrally distinct fluorescent reporters: CDX2-eGFP; membrane-tdTomato (mT/mG); Halo-YAP and H2B-miRFP720. Maximum intensity projections of images acquired with SPIM. Scale bar: 10 μm. (D) Live imaging preimplantation development. Green: light sheet used for illumination. Blue: emitted light is collected by the detection objective. (E) Histogram of number of nuclei per embryonic stage (represented by embryo cell number) for both BlastoSPIM 1.0 (blue, used for initial benchmarking of methods) and BlastoSPIM 2.0 (red, used for extending accurate segmentation to later stages). For four embryos from different stages, the ground truth of nuclear segmentation are shown.

more than 570 high-resolution, light-sheet images (Fig 1(E)). Fig 1(E) shows the number of annotated nuclei per developmental stage, from the 8-cell stage to beyond the 100-cell stage (see Dataset Characteristics for details). In total, across these images, approximately 12,000 nuclei are annotated. The quality, detail, and size of the BlastoSPIM dataset makes it unique relative to other currently available ground truth datasets for nuclear instance segmentation (S2 Table).

The BlastoSPIM 1.0 dataset, rather than focusing exclusively on early stages of development like the 16- cell stage, contains a wide range of developmental stages, particularly the 32-to-64 cell stages. To quantitatively illustrate the challenges posed by densely packed nuclei for instance segmentation, with BlastoSPIM 1.0 we calculated how nucleus-to-nucleus distances change from the 16-cell stage to the >100-cell stage. As a summary statistic for the density of nuclei, we computed the shortest distance from a nucleus's surface to another nucleus's surface (Fig 2). The median of this distance is typically 6.0 μm at the 16-cell stage, 2.9 μm at the 32-cell stage, 1.8 μm at the 64-cell stage, and $\approx$0.5 μ at the >100-cell stage. This drop in nearest-neighbor distance, with an increasing number of nuclei having a $< 1\mu m$ nearest-neighbor distance with successive developmental stages, is not accompanied by a comparable decrease in nuclear size (Fig 2); thus, the task of instance segmentation is expected to be considerably more difficult as development progresses.

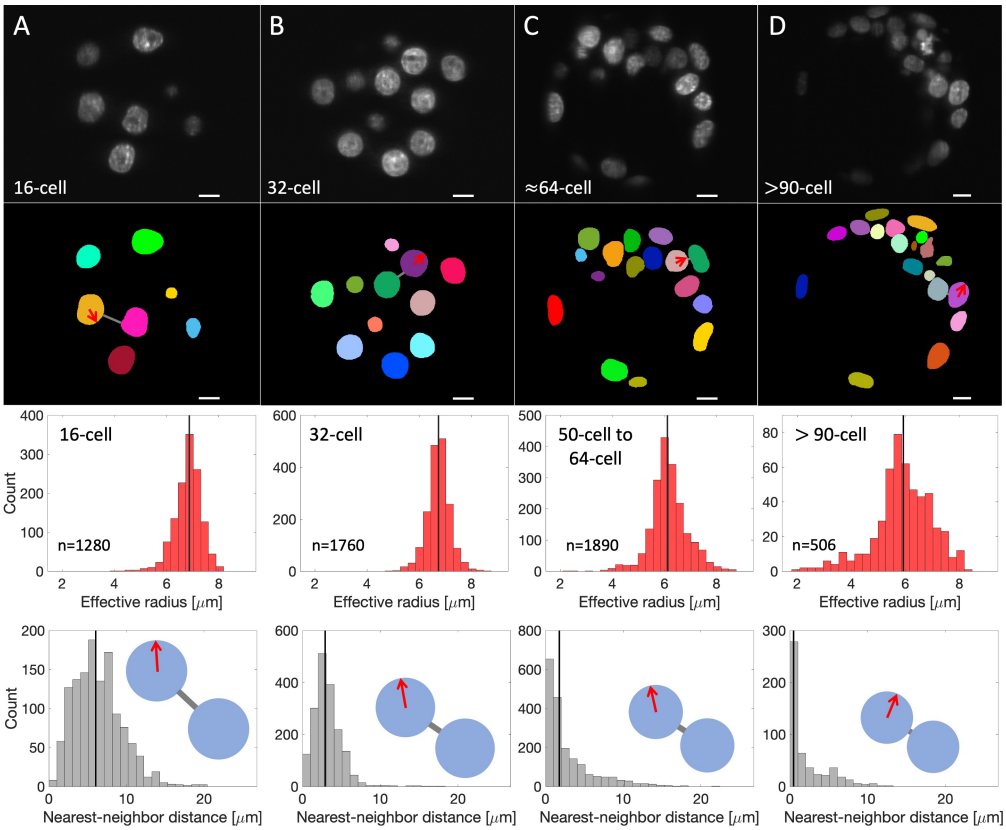

Fig 2. Nearest-neighbor distances between nuclei decrease dramatically during development. (A-D) Example z-slices and quantification for 16-cell (A), 32-cell (B), 50-to-64-cell (C), and >90-cell (D) embryos. The first two rows contain images and corresponding annotations. Each red arrow indicates the nucleus's effective radius, the radius of a sphere of equivalent volume. The gray lines indicate examples of shortest surface-to-surface distance. The third and fourth rows show that the effective radius and the shortest surface-to-surface distance decrease during development. Illustrations in the bottom histograms show that the latter decreases more than the former. Median of histogram in black. Scale bar: 10 μm.

The challenge for instance segmentation is due not only to nucleus-to-nucleus juxtaposition, but also to characteristics of image acquisition. For example, live images often have low SNRs because the exposure of embryos to light has to be limited to prevent phototoxicity [18]. In addition, the sample is imaged

along a single axis (the z-axis, by convention), resulting in voxel anisotropy – poorer z-resolution than xy-resolution. Our ground-truth dataset contains a range of SNR values (S1 Fig) and has a voxel anisotropy of approximately 10. In summary, because of its size as well as its diversity in both developmental stage and SNR, our ground-truth dataset of manually annotated 3D instances of nuclei is uniquely suited to interrogate the performance of any existing segmentation method – for achieving accurate nuclear instance segmentation in preimplantation mouse embryos.

0.3 Benchmarking of five instance segmentation methods on BlastoSPIM 1.0 reveals superior performance of Stardist-3D.

We used our ground-truth dataset to compare five instance segmentation networks (in S1 Table), including Cellpose [19], Stardist-3D [20], RDCNet [21], U3D-BCD [22], and UNETR-BCD [23]. These methods span a variety of network architectures, from those including recurrent blocks or transformers to more conventional U-Nets. They also represent the instances in different ways. For example, Stardist-3D computes a set of distances to the boundary, while Cellpose predicts gradients that are tracked to the instance center (S2 Intro Fig). Notably, we chose not to test QCANet [8] – the net currently reported to be the state-of-the-art in nuclear instance segmentation in early mouse embryos – because its two U-Nets are not jointly optimized, which unnecessarily increases the number of parameters, and because it cannot directly handle anisotropic images.

We trained each model with data from 482 3D images of embryos from BlastoSPIM 1.0 and then evaluated on two distinct test sets, one of low SNR data and one of moderate SNR data. We first evaluated the performance on the test set of 30 3D moderate SNR images. To interrogate stage-specific performance, we divided this test set into developmental stages such that it contained approximately 120 nuclei from each stage (*e.g.*, more images from earlier stages than later stages). To benchmark each method, we compared the ground-truth instances and model-inferred instances by computing matches based on the intersection-over-union (IoU). Based on this matching, we computed the precision, recall, and average precision, which are defined respectively as: $\frac{TP}{TP+FP}$, $\frac{TP}{TP+FN}$, $\frac{TP}{TP+FP+FN}$, where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. Whereas the precision and recall only penalize false positives and false negatives, respectively, the average precision penalizes both similarly. Table 1 shows the performance of each method at each developmental stage from the 8-cell stage to beyond the 100-cell stage.

We found that the Stardist-3D model outperformed all other methods as demonstrated by the average precision metric. This is significantly higher than the state-of-the-art results on similar (confocal) data from preimplantation mouse embryos, particularly in embryos with greater than 32 nuclei [8]. Of the 600 nuclei in all the embryos in the test set, Stardist-3D produced 14 false negatives and 17 false positives – thus achieving an overall precision and recall of 98% and 97%, respectively. We suspect that Stardist-3D's assumption of star-convexity, which approximates nuclear shapes in our dataset well (S3 Fig), facilitates Stardist-3D's learning of sufficient features to distinguish closely neighboring nuclei.

By comparison to Stardist-3D's strong performance across stages, the performance of the other methods depended strongly on developmental stage. For example, the U3D-BCD model also performed reasonably well at the 16- and 32-cell stages but was unable to detect several nuclei in later stages as they became more densely packed. The related UNETR-BCD method also performed well at the 16- and 32-cell stages, but its performance degraded more than that of U3D-BCD by the last developmental stage. By contrast, the performance of the other two methods slightly improved as development progressed. For example, Cellpose produced more false positives than true positives in early stages, but by the 64-cell stage, precision and recall surpassed 90%. Finally, RDCNet's precision increased with developmental stage as its recall decreased.

Figure 3 shows qualitative results of the five networks on an embryo with 62 nuclei based on two 2D image slices. U3D-BCD, UNETR-BCD, and RDCNet missed several nuclei due to under-segmentation – the merging of more than one nucleus into the same instance label. Stardist-3D missed the same nucleus in the xy- and xz-slices shown because two ground-truth instances (peach and purple) merged into one (purple). Cellpose produced no false negatives and one false positive (white) in this example. Despite Cellpose's relatively strong performance at the 64-cell stage, the segmentation masks are coarser than all other methods since it is trained at half the resolution in xy (see Network Implementation Details), and its tendency to produce false positives leads to poor generalization across developmental stages.

Despite Stardist-3D's strong performance in defining separate instances for closely packed nuclei (Fig 3),

Table 1. Performance Results on Moderate SNR Images Per Developmental Stage for Five Methods. Each network was trained on 482 3D images from BlastoSPIM 1.0, and was subsequently applied to a test set of moderate SNR images (30 images, 600 nuclei). For 8-cell to 64-cell stages, model hyperparameters were not independently adjusted. For the latest stage (> 100-cell), the probability threshold – used to define instances – was tuned to independently optimize each method's performance. Matching between ground-truth instances and inferred instances was based on the cutoff in intersection over union – as listed in "Dataset Splits and Evaluation Metric".

Stage	Method	Precision	Recall	Average Precision
≈ 8 Nuclei	Cellpose	0.44	0.95	0.43
(15 embryos)	RDC Net	0.33	0.92	0.32
(total nuclei 132)	U3D BCD	0.94	**0.98**	0.92
	UNETR	0.97	**0.98**	0.96
	Stardist-3D	**0.98**	**0.98**	**0.96**
≈ 16 Nuclei	Cellpose	0.44	0.86	0.41
(8 embryos)	RDC Net	0.78	0.83	0.67
(total nuclei 117)	U3D BCD	**1.00**	**0.99**	**0.99**
	UNETR	**1.00**	**0.99**	**0.99**
	Stardist-3D	**1.00**	**0.99**	**0.99**
≈ 32 Nuclei	Cellpose	0.74	0.98	0.73
(4 embryos)	RDC Net	0.80	0.74	0.63
(total nuclei 127)	U3D BCD	0.98	0.93	0.91
	UNETR	0.98	0.96	0.94
	Stardist-3D	**0.99**	**0.99**	**0.98**
≈ 64 Nuclei	Cellpose	0.92	**0.98**	0.91
(2 embryos)	RDC Net	0.90	0.85	0.78
(total nuclei 122)	U3D BCD	**1.00**	0.89	0.89
	UNETR	0.96	0.80	0.78
	Stardist-3D	**1.00**	0.97	**0.97**
> 100 Nuclei*	Cellpose	0.78	0.87	0.71
(1 embryo)	RDC Net	0.94	0.77	0.74
(total nuclei 102)	U3D BCD	0.95	0.78	0.75
	UNETR	**0.96**	0.71	0.69
	Stardist-3D	0.88	**0.94**	**0.83**

whether it could predict instances that overlap well with the ground truth remained unclear. By varying the IoU threshold used for matching (S3 Table,S4 Table), we found that the average precision on the original test set remained high even at a cutoff of 0.5, but decreased for an IoU threshold of 0.6 (S5 Table). This performance deterioration as the IoU threshold increases beyond ≈ 0.5 is common in 3D instance segmentation tasks [20, 24]; however, we expect than an IoU of ≈ 0.5 between ground-truth instances and Stardist-3D inferred instances is sufficient for nuclear tracking and measurement of fluorescent nuclear-localized factors.

Given the deterioration of model performance for late blastocysts, we set out to improve Stardist 3-D's accuracy by specifically training the net on late stage ground-truth data. We hand-annotated an additional 80 3D images of late stage embryos expressing H2B-miRFP720, containing more than 6600 nuclear instances – a data set we termed BlastoSPIM 2.0 (Fig 1(E)). We trained and validated a new Stardist 3-D model based on 72 late blastocysts from BlastoSPIM 2.0. This new model outperformed the previous Stardist-3D model from Table 1 on test images of late blastocysts, yet unsurprisingly underperformed it on test images of embryos with fewer than 64 nuclei (Table 2). For the remainder of this study, we term the Stardist-3D model trained on BlastoSPIM 1.0 from Table 1 the "early embryo model" (used for < 64 nuclei) and the subsequent Stardist-3D model trained on BlastoSPIM 2.0 as the "late blastocyst model" (used for > 64 nuclei).

We also evaluated the performance of both models on a more difficult test set – separate from the test set in Table 2 – comprised of images with a very low SNR ratio (S3 Table). The test set was binned by developmental stage. Though only a third of the images in the training set met our definition of low SNR (S1 Fig), the recall for the low SNR test set is at least 95 % for all stages from the 2-cell stage to the 32-cell stage

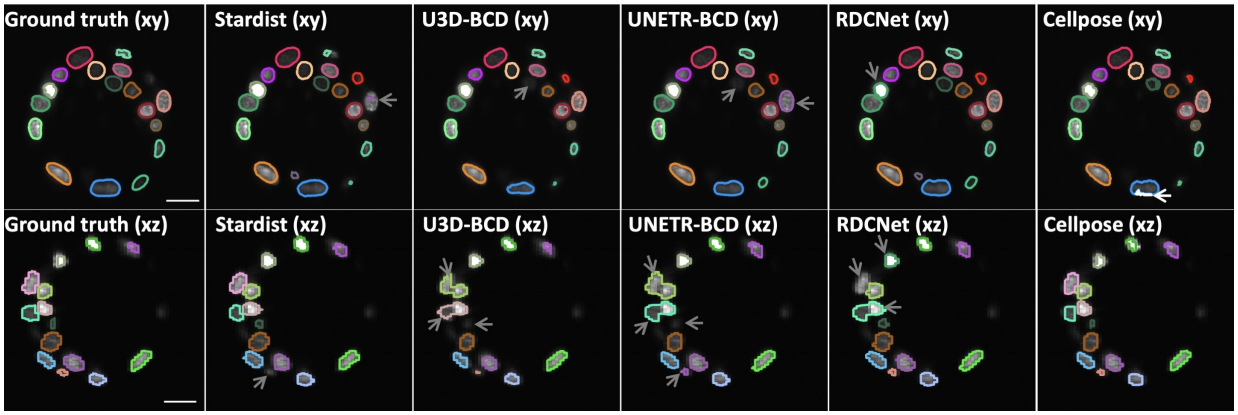

Fig 3. Qualitative evaluation of five instance-segmentation networks trained on BlastoSPIM and tested on a 62-cell embryo. Instance contours overlaid on a representative slice of the intensity image in xy (top panels) and in xz (bottom panels). Each panel is labelled as: Ground-truth, Stardist for Stardist-3D results, and similarly for other methods. Grey arrows indicate false negatives, including undersegmentation. White arrows denote false positives. Scale bars: 20 μm. Note that false positives and false negatives are defined by comparing the 3D instance segmentation results rather than the results shown in a single 2D slice.

(S6 Table, S7 Table, S8 Table). Although the performance deteriorated for low SNR images of $>$ 64-cell stage embryos, the late blastocyst model still achieved $\approx$ 90 % precision and recall (S6 Table). We note that the models' performance on low SNR images can be improved by tuning the probability hyperparameter used by Stardist (data not shown).

Overall, comparative analysis of five different methods trained on our BlastoSPIM dataset revealed not only state-of-the-art performance by Stardist-3D on a test set composed of moderate-to-high SNR, but also reliably high recall even for low SNR. Despite this strong performance, Stardist-3D's performance could be improved upon, particularly in inferring the exact shape of each nucleus; thus, the BlastoSPIM dataset – both BlastoSPIM 1.0 and BlastoSPIM 2.0 – can be used to test whether future neural network architectures can outperform Stardist-3D in accurately identifying the positions and shapes of nuclei in preimplantation embryos.

0.4 Using the BlastoSPIM-trained Stardist-3D models to characterize nuclear counts, shapes, and volumes in preimplantation embryo time series.

Next we used the two Stardist-3D models, one for early embryos and one for late blastocysts, to analyze time-lapse sequences of embryos for which no ground truth existed. First, we computed the time dynamics of the nuclear count. Supporting our conclusions from the test set (Table 2), the early model performs well up to the 32/64-cell transition (Fig 4(A)), while the late model performs better for the 64-cell and later stages (Fig 4(B)). Then, we sought to understand the relationship between developmental stage and nuclear volume (Fig 4). A previous study – based on fixed embryos – measured how nuclear volume depends on developmental stage, but did not measure how nuclear volume changes within a stage [25]. The QCANet study by Tokuoka *et al.* did measure the distribution of nuclear volumes over time in live images, but could not accurately segment nuclei after the 32-cell stage [8].

By analyzing time-lapse sequences of five different H2B-miRFP720-expressing preimplantation embryos developing from the 8-cell stage to the >100-cell stage, we found that the nuclear volumes peaked at the end of the 8-, 16-, 32-, and 64-cell stages and abruptly dropped after each round of division (for late Stardist model results, see (Fig 4(C); for results of both models, see S4 Fig). These trends are consistent with observations in unicellular eukaryotes, such as fission yeast [26, 27], where nuclear volume was shown to increase with cell cycle progression until an abrupt decrease at mitosis. Interestingly, in the embryo the rate of growth progressively slowed with each subsequent stage and peaked at lower and lower values, consistent with previously reported stage-specific scaling of nuclei [25] (Fig 4), although we cannot discount that part of

Table 2. Performance Results on Moderate SNR Images Per Developmental Stage for both the Stardist-3D early embryo model and the Stardist-3D late blastocyst model. This test set includes test images from both BlastoSPIM 1.0 and 2.0. Model hyperparameters were fixed for both models across all stages. *The decrease in the early model's recall on the ≈ 64-cell test set (relative to Table 1) is attributed to the incorporation of an embryo with 66 nuclei into the combined test set.

Stage	Method	Precision	Recall	Average Precision
≈ 8 Nuclei (15 embryos) (total nuclei 132)	early embryo	**0.98**	**0.98**	**0.96**
	late blastocyst	0.63	**0.98**	0.62
≈ 16 Nuclei (8 embryos) (total nuclei 117)	early embryo	**1.00**	**0.99**	**0.99**
	late blastocyst	0.71	**0.99**	0.71
≈ 32 Nuclei (4 embryos) (total nuclei 127)	early embryo	**0.99**	0.99	**0.98**
	late blastocyst	0.90	**1**	0.90
≈ 48 Nuclei (1 embryo) (total nuclei 48)	early embryo	**0.98**	0.96	**0.94**
	late blastocyst	0.96	**0.98**	**0.94**
≈ 64 Nuclei* (3 embryos) (total nuclei 188)	early embryo	**1**	0.88	0.88
	late blastocyst	0.96	**1**	**0.96**
≈ 80 Nuclei (2 embryos) (total nuclei 165)	early embryo	**0.98**	0.73	0.72
	late blastocyst	0.95	**0.98**	**0.93**
> 100 Nuclei (2 embryo) (total nuclei 208)	early embryo	**1**	0.73	0.73
	late blastocyst	0.96	**0.99**	**0.94**

this effect is due to increasing cell cycle asynchrony among cells in the embryo.

Because previous studies reported that at the 32- and 64-cell stages, the TE cells have larger aspect ratios than ICM cells [15, 28], we then asked whether the same statement holds true for the corresponding nuclei. To distinguish between ICM and TE cells, we used nuclear position as a proxy in 32-, 64-, and > 100-cell stage embryos, annotating nuclei close to the embryo surface as TE and nuclei deeper within the embryo as ICM. We found that in cavitated 32-cell stage embryos, the median TE nucleus's aspect ratio, given by its longest axis length over its shortest axis length, was already greater than that of the median ICM nucleus ($p < 0.001$) (Fig 4). At this stage, a TE nucleus's long axis tends to be 80 % longer than its shortest axis. Interestingly, the ICM nuclei are often not well approximated by a sphere, but instead often have long and medium axes that are both 5 % - 40 % longer than the shortest axis. By comparing our embryos at the 32-cell stage to those at the 64-cell stage, we found that the median aspect ratio of TE nuclei grows over that time ($p < 0.001$), while the median ICM nuclear aspect ratio remains unchanged ($p = 0.7$) (Fig 4). From the 64-cell stage to the > 100-cell stage, the nuclei of the ICM and TE both increase in aspect ratio ($p < 0.001$ and $p = 0.009$, respectively), yet at the > 100-cell stage, the TE nuclei continue to have higher aspect ratios than ICM nuclei ($p < 0.001$) (Fig 4). In summary, we found, as seen in studies of TE and ICM cell shapes, that TE nuclei flatten more during cavitation (from the 32-to-64 cell stages) than the ICM nuclei; however, both the TE and ICM nuclear aspect ratios increase by the >100-cell stage.

In addition to flattening nuclei, the formation and expansion of the cavity could plausibly affect the nuclear volumes of the two cell populations differently. Consistent with a recent publication [2], at the 32-cell stage, no statistically significant differences were found between the median TE nuclear volume and the median ICM nuclear volume ($p = 0.16$). However, by the early 64-cell stage, the median TE nucleus was larger than that of the ICM ($p < 0.001$) by about 15 % (Fig 4). The mechanism responsible for generating this difference in nuclear volumes remains unclear. Since nuclear volume depends on the time since the last division (Fig 4), we hypothesize that this difference could arise from cell cycle asynchrony, if the TE cells divided earlier [29] and thus had more time post-division to grow in volume relative to the ICM. It is

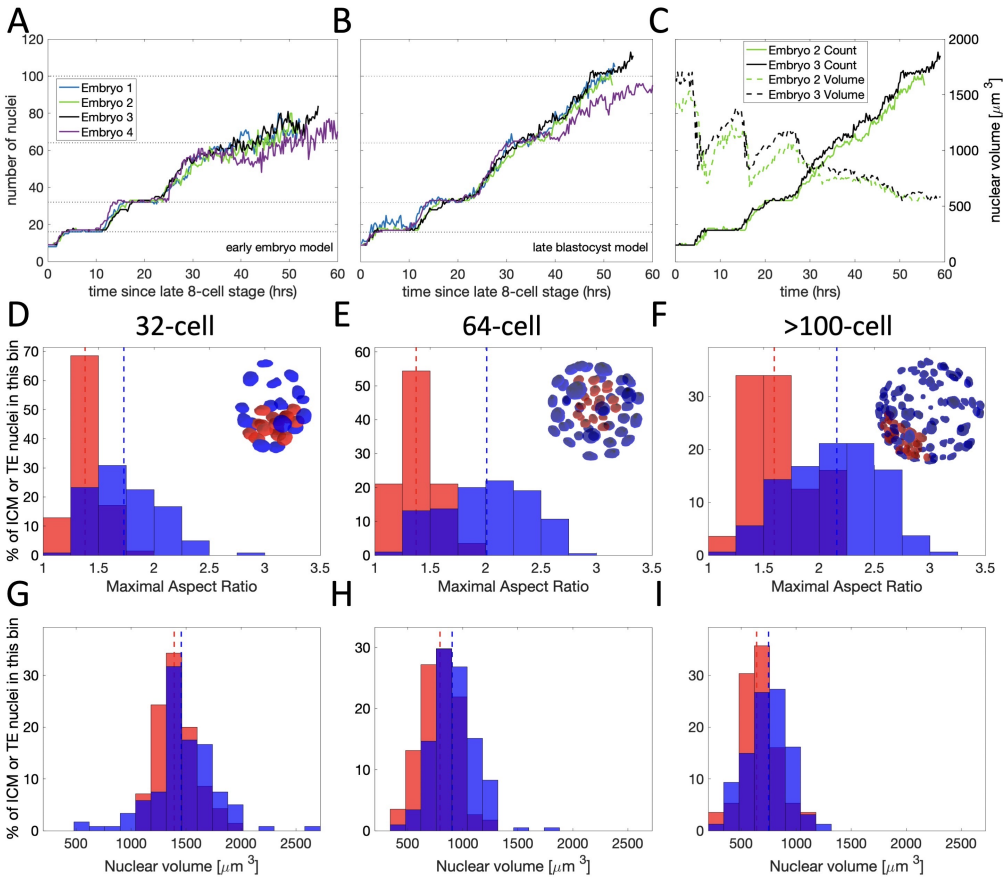

Fig 4. Measuring dependence of nuclear properties on developmental stage and cell fate via Stardist-3D inference. (A-B) Number of nuclei over time for 4 embryo sequences as inferred by the early embryo model and late blastocyst model, respectively. Dashed horizontal lines: 16, 32, 64, 100 nuclei respectively. (C) Number of nuclei and nuclear volume versus time for two representative embryo sequences from (B). (D-F) For 32-cell (left, pooled nuclei from 6 embryos) and 64-cell stage embryos (center, pooled nuclei from 5 embryos) and > 100-cell stage embryo (right, pooled nuclei from 2 embryos), the TE and ICM populations are compared based on their nuclear aspect ratios. Dashed vertical lines: median of ICM (red) and TE (blue) nuclei. Insets: renderings of an embryo at the 32-, 64-, and > 100-cell stage in which nuclei have been manually fate-assigned as ICM (red) and TE (blue) nuclei. (G-I) Same as (D-F) but comparisons are based on nuclear volumes. See text for p-values comparing each pair of distributions in (D-I).

also possible that forces from the cavity onto the ICM affect nuclear volume regulation. By analysis of two >100-cell stage embryos, we found that the median TE nuclear volume remains larger than the median ICM nuclear volume for embryos with >100 nuclei ($p = 0.003$) (Fig 4). Just as the model's segmentations detect differences between ICM and TE, our model will likely play an important role in quantifying the changes in nuclear properties that accompany the next fate decision, specifying the embryo proper and extra-embryonic tissue within the ICM [30].

0.5 Generalization of our trained Stardist-3D models to different model organisms.

Finally, we wished to address whether our advances in instance segmentation for the mouse embryo generalize to other systems. In principle, our model's performance on a dataset not used for training could depend on the organism, the method for nuclear labeling, and the imaging technique (*i.e.*, confocal or light-sheet). We evaluated our Stardist-3D model on a ground-truth set composed of 9 live light-sheet images of *Platynereis dumerilli* embryos from the 38- to the 392-cell stage [24] (Fig 5(A)-(D)), in which nuclei were labeled by

microinjection of a fluorescent tracer.

We first optimized our Stardist-3D early embryo model ("early mouse only" model) – without retraining – by using four of the ground-truth *Platynereis* images to tune a single Stardist-3D hyperparameter, the probability threshold. Applying that "early mouse only" model to the five other images, we found that it performed well, at greater than 95% precision and recall, on early *Platynereis* embryos, from the 76- to 198- cell stages . With a precision and recall of 98% and 87%, respectively, for an image with $\approx$ 400 nuclei, the "early mouse only"'s performance deteriorated slightly at the latest developmental stages (Table 3). Performing the same exercise with the "late mouse only" model revealed that its performance on the test set was weaker than the "early mouse only" model for early developmental stages (Fig 5), but for the latest developmental stages (with 281 and 392 nuclei), the "late mouse only" model's precision and recall remained above 90 %.

Table 3. Evaluation of our Stardist-3D model ("mouse only") on unseen data from *Platynereis* at an IoU threshold of 0.1.

Nuclear Count	Precision (Early)	Recall (Early)	Average Precision (Early)	Precision (Late)	Recall (Late)	Average Precision (Late)
76	0.96	1	0.96	0.82	0.97	0.80
162	0.99	0.95	0.94	0.92	0.96	0.89
198	0.96	0.96	0.93	0.89	0.98	0.87
281	0.97	0.90	0.88	0.92	0.94	0.87
392	0.98	0.87	0.85	0.94	0.91	0.85

Our analysis with no model retraining (Table 3, S9 Table, and S10 Table) demonstrated that our models, especially the "early mouse only" model, performed well for embryos with even more nuclei than late blastocysts. Nonetheless, the model performance deteriorated for higher IoU thresholds (S9 Table and S10 Table), meaning that our model-predicted instances did not precisely overlap with corresponding ground-truth instances. This observation motivated us to retrain and revalidate our Stardist-3D models, originally used for the mouse, based on seven of the *Platynereis* ground-truth images embryo. We refer to these two retrained models as the "early-mouse-then-worm" model and "late-mouse-then-worm" model. We compared each retrained model's performance to the performance of a model trained on the same *Platynereis* training data, but without the mouse model as an initial condition ("worm only" model). Relative to the "worm only" model, the "early-mouse-then-worm" model achieved higher precision and recall on the 162-cell test embryo at an IoU cutoff of 0.5 (S13 Table and Fig 5). Similarly, the "late-mouse-then-worm" model achieved higher precision and recall on the 198-cell test embryo than the "worm only" model at an IoU cutoff of 0.5. Thus, our models without further training, particularly the "early mouse only" model, performed well up to the 198-cell stage for low IoU cutoffs, and the retraining of our mouse models on *Platynereis* data enabled them to outperform a "worm only" model in accurately predicting nuclear shapes and positions for higher IoU cutoffs (S11 Table, S12 Table, and S13 Table).

Conclusion

To understand how individual cells' behaviors contribute to morphogenetic events, biologists acquire staggering amounts of time-lapse images of these processes. Quantifying the properties and behaviors of individual cells in such image series requires instance segmentation: identifying which voxels belong to which object. Although many measurements require segmentation of entire cells, instance segmentation of nuclei is useful for estimating the relative positions of cells, classifying by mitotic stage, and measuring the expression of nuclear-localized factors. Nuclear instance segmentation is challenging for several reasons including nucleus-to-nucleus proximity, variations in nuclear shape, voxel anisotropy, and low SNR. By comparative analysis of five different neural networks on a newly collected and publicly available ground-truth dataset, BlastoSPIM, we have shown which of these networks best addresses these challenges in the preimplantation mouse embryo.

Our comparative analysis revealed state-of-the-art performance by Stardist-3D (early embryo model) across developmental stages. Both precision and recall remained above 95 %, even at the 64-cell stage (Table 1). Similar performance was achieved even on a separate test set with low SNR. In contrast, the performance of other methods varied, with Cellpose and RDCNet producing many false positives, particularly at early developmental stages, and U3D-BCD and UNETR missing several nuclei for the 64-cell stage and

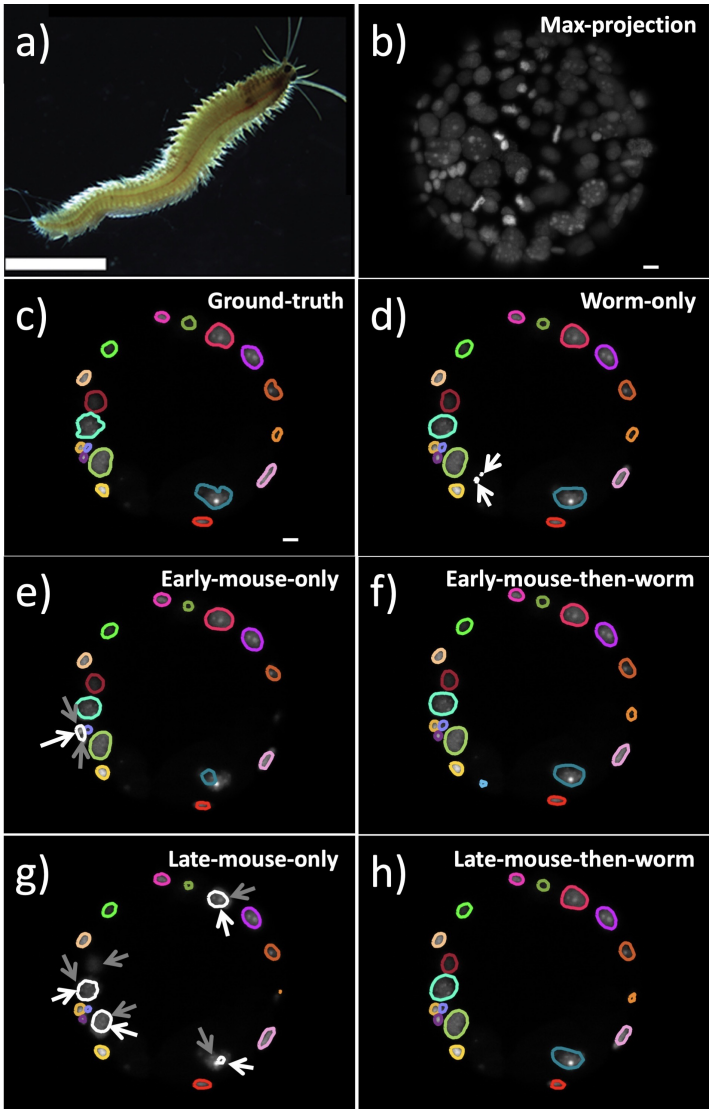

Fig 5. Usefulness of the mouse embryo datasets (BlastoSPIM 1.0 and 2.0) for the analysis of images of the worm *Platynereis* (A) Mature female *P. dumerilli*, image from: [31]. Scale bar: 4 *mm*. (B) Maximum intensity projection of image of *Platynereis* embryo containing 162 nuclei. Adapted from [24]. Scale bar: 10 μm. (C-H) Instance contours overlaid over intensity image for a single slice of the 162-cell test embryo. Comparison of ground-truth (C) to a Stardist-3D model trained solely on *Platynereis* data ("worm only") (D), our early embryo Stardist-3D model without further training ("early mouse only") (E), our early embryo Stardist-3D model with further training ("early-mouse-then-worm") (F), our late blastocyst Stardist-3D model without further training ("late mouse only") (G), and our late blastocyst Stardist-3D model with further training ("late-mouse-then-worm") (H) . Scale bar: 10 μm. White (gray) arrows denote false positive (negatives). Note that each white-grey arrow pair in (G) results from the relatively high IoU cutoff (0.5); at this cutoff, the model-inferred instance is not matched with a ground-truth instance, which leads to a false negative paired with a false positive.

later stages (Table 1). To further improve segmentation performance at later stages of preimplantation development, we hand-annotated a second ground truth dataset of nuclei in late blastocyst embryos and trained a second Stardist model (late blastocyst model), with which precision and recall also remained ≥ 95 % even in $>$ 100-cell stage embryos. Therefore we not only present trained Stardist-3D models with superior performance for nuclear instance segmentation in time-lapse images of early mouse embryos, but share large ground truth datasets (BlastoSPIM 1.0 and 2.0), which will be an important resource for evaluating the

performance of future CNN architectures because of the dataset's size and quality and nuclear density relative to other currently available datasets (S2 Table).

We used our trained Stardist-3D models to segment nuclei in time series for which no ground truth existed. These segmentations revealed oscillations of nuclear volume with the cell cycle: volumes gradually increased throughout interphase and peaked just before mitosis, resulting in a sudden volume drop. We also found that the growth rate of nuclei slowed progressively from the 8-cell stage to the 16-cell stage to the 32-cell stage to the 64-cell stage. We extended these analyses to test whether nuclear geometries are correlated with fate. First, we confirmed that TE nuclei have significantly higher aspect ratio – with their long axis typically $\approx$ 80% longer than their shortest axis – than ICM nuclei, both at 32-cell stage and the 64-cell stage. Second, we showed that though TE and ICM nuclei do not differ in volume at the 32-cell stage, TE nuclei become larger than ICM nuclei by the 64-cell stage (Fig 4). The TE-ICM difference in nuclear volume detected at the 64-cell stage was also detected at the >100-cell stage, and the aspect ratios of both TE and ICM nuclei increased at the >100-cell stage relative to the 64-cell stage. We expect that this instance segmentation model will enable many more insights into mouse development, including into the fate decision occurring within the ICM.

We next tested whether our model generalizes well to other imaging data. We took advantage of a publicly available annotated dataset of developing annelid *Platynereis dumerilli* embryos, injected with a nuclear fluorescent tracer and live-imaged with a light-sheet microscope (Fig 5) [24]. We found that our models – without further retraining – performed well on the data up to the 198-cell stage. When we trained our model further on *Platynereis* ground truth data from *Platynereis* images, it outperformed – at higher IoU thresholds, like 0.5 – a model trained solely trained on *Platynereis* data. Thus, our models without retraining, particularly the "early-mouse-only" model, performed well at low IoU thresholds on embryos with many nuclei, up to $\approx$ 200. When retrained on ground truth data from the *Platynereis* model systems, our model achieved more accurate nuclear segmentation at high IoU thresholds when compared to a model trained on *Platynereis* data alone.

The generalizability of our model fills a clear need since many publicly available models work only in 2D, segment only cell boundaries, or are trained only on high SNR images [32]. Given our model's performance even without fine-tuning, small hand-corrections of our model's predictions on a different biological system could be used to generate training data, as long as that system's nuclei satisfy the star-convexity assumption of Stardist. We expect that BlastoSPIM and our Stardist-3D model, in conjunction with other publicly available datasets and pre-trained models [33], will play a key role in the development of truly generalist models. BlastoSPIM 1.0 and 2.0 and the Stardist-3D models trained on them are, furthermore, the first crucial steps towards automated (3+t)-D analysis of early mouse development, which, by incorporating the construction of lineage trees, can reveal the temporal dynamics of individual nuclei as fate decisions transpire.

Supplementary Information

S1 Table. Tools for 3D nucleus segmentation. Description of base network architecture and modifications, network output, loss functions, and post-processing methods.

Tool name	Base network	Network output	Loss metrics	Post-processing
Cellpose [19]	2D U-Net with residual blocks and style transfer	horizontal/vertical gradient maps, cell probability map	L2 loss (gradients), cross-entropy loss (cell probability)	probability threshold and test-time enhancements
QCANet [8]	Two 3D U-Net's, hyperparameter tuning by Bayesian optimization	Semantic segmentation, nucleus center detection	Dice loss	Reinterpolation and marker-based watershed
NuSeT [34]	2D U-Net integrated with Region Proposal Network (RPN)	Semantic segmentation, bounding box with score	cross-entropy loss + Dice loss (segmentation), class loss and regression loss (detection)	watershed and 3D reconstitution from 2D slices
Stardist [20]	3D ResNet or 3D U-Net	radial distances to object boundary, object probability (OP) with distance transform	cross-entropy loss (OP), OP-weighted mean absolute error with regularization (radial distances)	OP threshold and non-maximum suppression
RDCNet [21]	3D recurrent block with stacked dilated convolutions	semantic classes, semi-convolutional embeddings	Embedding soft jaccard (ESJ) loss	Margin thresholds, Hough voting
EmbedSeg [24]	3D Branched ERF-Net	pixel embeddings, clustering bandwidth, seed probability	Lovász-Softmax loss + seed loss + smoothness loss	seed probability threshold, cluster bandwidth threshold
U3D-BCD [22]	3D U-Net with residual blocks substituted for convolutional layers	foregrounds masks, instance contours, signed-distance-transform map	Weighted sum of cross-entropy loss and dice loss for foreground and contour; mean-squared error for signed distance	seed detection via threshold on foreground probability and distance value, marker-controlled watershed
UNETR-BCD [23]	Stack of transformers connected to 3D CNN-based decoder	foregrounds masks, instance contours, signed-distance-transform map	Weighted sum of cross-entropy loss and dice loss for foreground and contour; mean-squared error for signed distance	seed detection via threshold on foreground probability and distance value, marker-controlled watershed

S2 Table. Ground-truth, three-dimensional annotations of nuclei. Other examples of publicly available ground-truth data sets for instance segmentation of nuclei. *The entire ground-truth dataset used in [8] contains more than 6000 time-series of early mouse embryos, of which only 165 have been made publicly available.

Name	Microscopy	Nuclear Labeling	Sample	Image Count	Network results
NucMM-Z [22]	Serial-section electron microscopy	N/A	Zebrafish brain	1	Cellpose3D, Stardist-3D, U3D-BCD
NucMM-M [22]	Micro-CT	N/A	Mouse visual cortex	1	Cellpose3D, Stardist-3D, U3D-BCD
BBBC050* [8]	Confocal microscopy	H2B-mRFP1, H2B-mCherry	Pre-implantation mouse embryo from the pro-nuclear stage to 53-cell stage	165	QCANet, 3D U-Net, 3D Mask R-CNN
C. elegans developing embryo [35]	Confocal microscopy	histone-GFP	*C. elegans* embryo between 2-cell stage and ≈300-cell stage	9	QCANet, 3D U-Net, 3D Mask R-CNN
Platynereis-Nuclei-CBG [24]	Light-Sheet Microscopy	Fluorescent nuclear tracer injected	*Platynereis dumerilii* embryo between 0 and 16 hours post-fertilization	9	Cellpose3D, Stardist-3D, EmbedSeg
Platynereis-ISH-Nuclei-CBG [24]	Confocal Microscopy	DAPI	*Platynereis dumerilii* specimins 16 hours post-fertilization	2	Cellpose3D, Stardist-3D, EmbedSeg
Parhyale hawaiensis-Nuclei [20]	Confocal Microscopy	H2B-eGFP	*Parhyale hawaiensis* embryo between 46 hours post-amputation (hpa) and 110 hpa	6	U-Net, Stardist-3D, Cellpose3D, EmbedSeg
C. elegans-Nuclei [20]	Confocal Microscopy	DAPI	*C. elegans* embryo at the 558-cell stage	28	U-Net, Stardist-3D
Mouse-Skull-Nuclei-CBG [24]	Confocal Microscopy	DAPI	Nuclei from the skull of developing mouse embryos	2	Cellpose3D, Stardist-3D, EmbedSeg
Peri-implantation mouse embryos [36]	Confocal Microscopy	Antibody staining	Peri-implantation mouse embryos	35	3D U-Net

S3 Table. High SNR test set: Performance Results Per Developmental Stage for Stardist-3D at an IOU Threshold of 0.3. This table is analogous to Table 2, which used an IoU cutoff of 0.1. Model hyperparameters were fixed for both models across all stages. *The decrease in the early model's recall on the ≈ 64-cell test set (relative to Table 1) is attributed to the incorporation of an embryo with 66 nuclei into the combined test set.

Stage	Method	Precision	Recall	Average Precision
≈ 8 Nuclei (15 embryos) (total nuclei 132)	early embryo	**0.96**	**0.97**	**0.93**
	late blastocyst	0.62	**0.97**	0.60
≈ 16 Nuclei (8 embryos) (total nuclei 117)	early embryo	**1**	**0.99**	**0.99**
	late blastocyst	0.71	0.98	0.70
≈ 32 Nuclei (4 embryos) (total nuclei 127)	early embryo	**0.99**	0.99	**0.98**
	late blastocyst	0.90	**1**	0.90
≈ 48 Nuclei (1 embryo) (total nuclei 48)	early embryo	**0.98**	0.96	**0.94**
	late blastocyst	0.96	**0.98**	**0.94**
≈ 64 Nuclei* (3 embryos) (total nuclei 188)	early embryo	**1**	0.88	0.88
	late blastocyst	0.95	**0.99**	**0.95**
≈ 80 Nuclei (2 embryos) (total nuclei 165)	early embryo	**0.98**	0.72	0.71
	late blastocyst	0.95	**0.97**	**0.92**
> 100 Nuclei (2 embryo) (total nuclei 208)	early embryo	**1**	0.73	0.73
	late blastocyst	0.96	**0.99**	**0.94**

S4 Table. High SNR test set: Performance Results Per Developmental Stage for Stardist-3D at an IOU Threshold of 0.5. This table is analogous to Table 2, which used an IoU cutoff of 0.1. Model hyperparameters were fixed for both models across all stages. *The decrease in the early model's recall on the ≈ 64-cell test set (relative to Table 1) is attributed to the incorporation of an embryo with 66 nuclei into the combined test set.

Stage	Method	Precision	Recall	Average Precision
≈ 8 Nuclei (15 embryos) (total nuclei 132)	early embryo	**0.95**	**0.95**	**0.91**
	late blastocyst	0.49	0.77	0.42
≈ 16 Nuclei (8 embryos) (total nuclei 117)	early embryo	**0.98**	**0.97**	**0.96**
	late blastocyst	0.66	0.91	0.62
≈ 32 Nuclei (4 embryos) (total nuclei 127)	early embryo	**0.98**	0.98	**0.97**
	late blastocyst	0.89	**0.99**	0.89
≈ 48 Nuclei (1 embryo) (total nuclei 48)	early embryo	**0.96**	0.94	**0.90**
	late blastocyst	0.94	**0.96**	**0.90**
≈ 64 Nuclei* (3 embryos) (total nuclei 188)	early embryo	**0.94**	0.83	0.79
	late blastocyst	0.92	**0.96**	**0.89**
≈ 80 Nuclei (2 embryos) (total nuclei 165)	early embryo	0.85	0.63	0.57
	late blastocyst	**0.91**	**0.93**	**0.85**
> 100 Nuclei (2 embryo) (total nuclei 208)	early embryo	0.87	0.63	0.58
	late blastocyst	**0.90**	**0.93**	**0.84**

S5 Table. High SNR test set: Performance Results Per Developmental Stage for Stardist-3D at an IOU Threshold of 0.6. This table is analogous to Table 2, which used an IoU cutoff of 0.1. Model hyperparameters were fixed for both models across all stages. *The decrease in the early model's recall on the ≈ 64-cell test set (relative to Table 1) is attributed to the incorporation of an embryo with 66 nuclei into the combined test set.

Stage	Method	Precision	Recall	Average Precision
≈ 8 Nuclei (15 embryos) (total nuclei 132)	early embryo	0.86	0.86	0.76
	late blastocyst	0.44	0.69	0.37
≈ 16 Nuclei (8 embryos) (total nuclei 117)	early embryo	0.92	0.91	0.85
	late blastocyst	0.62	0.86	0.56
≈ 32 Nuclei (4 embryos) (total nuclei 127)	early embryo	0.98	0.98	0.97
	late blastocyst	0.83	0.92	0.77
≈ 48 Nuclei (1 embryo) (total nuclei 48)	early embryo	0.87	0.85	0.76
	late blastocyst	0.86	0.88	0.76
≈ 64 Nuclei* (3 embryos) (total nuclei 188)	early embryo	0.90	0.79	0.73
	late blastocyst	0.88	0.91	0.81
≈ 80 Nuclei (2 embryos) (total nuclei 165)	early embryo	0.74	0.55	0.46
	late blastocyst	0.80	0.82	0.69
> 100 Nuclei (2 embryo) (total nuclei 208)	early embryo	0.75	0.54	0.46
	late blastocyst	0.69	0.71	0.54

S6 Table. Performance Results of Stardist-3D on Low SNR Images Per Developmental Stage at an IoU cutoff of 0.1. *Note that model performance for images (with low SNR) of late blastocysts can be improved by lowering the probabililty hyperparameter used by Stardist-3D (data not shown).

Stage	Method	Precision	Recall	Average Precision
≈ 2 Nuclei	early embryo	**1**	**1**	**1**
(2 embryos)	late blastocyst	0.27	0.75	0.25
(total nuclei 4)				
≈ 4 Nuclei	early embryo	**1**	**1**	**1**
(2 embryos)	late blastocyst	0.47	**1**	0.47
(total nuclei 8)				
≈ 8 Nuclei	early embryo	**1**	0.99	**0.99**
(33 embryos)	late blastocyst	0.99	**1.0**	**0.99**
(total nuclei 270)				
≈ 16 Nuclei	early embryo	**0.96**	0.98	**0.94**
(19 embryos)	late blastocyst	0.88	**1**	0.88
(total nuclei 268)				
≈ 32 Nuclei	early embryo	**1**	**1**	**1**
(2 embryos)	late blastocyst	0.97	**1**	0.97
(total nuclei 66)				
≈ 64 Nuclei *	early embryo	**0.97**	0.67	0.66
(2 embryos)	late blastocyst	0.89	**0.93**	**0.83**
(total nuclei 123)				
≈ 95 Nuclei	early embryo	**0.99**	0.73	0.72
(3 embryos)	late blastocys	0.94	**0.98**	**0.92**
(total nuclei 280)				

S7 Table. Performance Results of Stardist-3D on Low SNR Images Per Developmental Stage at an IoU cutoff of 0.3. *Note that model performance for images (with low SNR) of late blastocysts can be improved by lowering the probabililty hyperparameter used by Stardist-3D (data not shown).

Stage	Method	Precision	Recall	Average Precision
≈ 2 Nuclei (2 embryos) (total nuclei 4)	early embryo late blastocyst	**1** 0	**1** 0	**1** 0
≈ 4 Nuclei (2 embryos) (total nuclei 8)	early embryo late blastocyst	**1** 0.47	**1** **1**	**1** 0.47
≈ 8 Nuclei (33 embryos) (total nuclei 270)	early embryo late blastocyst	**1** 0.98	**0.99** **0.99**	**0.99** 0.96
≈ 16 Nuclei (19 embryos) (total nuclei 268)	early embryo late blastocyst	**0.96** 0.86	**0.98** **0.98**	**0.94** 0.85
≈ 32 Nuclei (2 embryos) (total nuclei 66)	early embryo late blastocyst	**1** 0.97	**1** **1**	**1** 0.97
≈ 64 Nuclei * (2 embryos) (total nuclei 123)	early embryo late blastocyst	**0.97** 0.84	0.67 **0.87**	0.66 **0.74**
≈ 95 Nuclei (3 embryos) (total nuclei 280)	early embryo late blastocyst	**0.98** 0.93	0.72 **0.98**	0.71 **0.91**

S8 Table. Performance Results of Stardist-3D on Low SNR Images Per Developmental Stage at an IoU cutoff of 0.5. *Note that model performance for images (with low SNR) of late blastocysts can be improved by lowering the probabililty hyperparameter used by Stardist-3D (data not shown).

Stage	Method	Precision	Recall	Average Precision
≈ 2 Nuclei (2 embryos) (total nuclei 4)	early embryo	**1**	**1**	**1**
	late blastocyst	0	0	0
≈ 4 Nuclei (2 embryos) (total nuclei 8)	early embryo	**1**	**1**	**1**
	late blastocyst	0.24	0.5	0.19
≈ 8 Nuclei (33 embryos) (total nuclei 270)	early embryo	**1**	**0.99**	**0.99**
	late blastocyst	0.88	0.89	0.79
≈ 16 Nuclei (19 embryos) (total nuclei 268)	early embryo	**0.94**	**0.96**	**0.90**
	late blastocyst	0.70	0.79	0.59
≈ 32 Nuclei (2 embryos) (total nuclei 66)	early embryo	**0.98**	**0.98**	**0.97**
	late blastocyst	0.91	0.94	0.86
≈ 64 Nuclei * (2 embryos) (total nuclei 123)	early embryo	**0.92**	0.64	0.61
	late blastocyst	0.66	**0.69**	0.51
≈ 95 Nuclei (3 embryos) (total nuclei 280)	early embryo	**0.86**	0.63	0.63
	late blastocyst	0.84	**0.88**	**0.75**

S9 Table. Stardist-3D Results on *Platynereis* set at an IoU Cutoff of 0.3. Analogous to Table 3.

Nuclear Count	Precision (Early)	Recall (Early)	Average Precision (Early)	Precision (Late)	Recall (Late)	Average Precision (Late)
76	0.96	1	0.96	0.78	0.92	0.73
162	0.98	0.94	0.92	0.89	0.93	0.83
198	0.96	0.96	0.92	0.87	0.95	0.83
281	0.96	0.89	0.86	0.91	0.93	0.85
392	0.96	0.85	0.82	0.93	0.90	0.84

S10 Table. Stardist-3D Results on *Platynereis* set at an IoU Cutoff of 0.5. Analogous to Table 3.

Nuclear Count	Precision (Early)	Recall (Early)	Average Precision (Early)	Precision (Late)	Recall (Late)	Average Precision (Late)
76	0.95	0.99	0.94	0.71	0.84	0.63
162	0.92	0.88	0.81	0.80	0.83	0.68
198	0.91	0.91	0.83	0.82	0.90	0.75
281	0.84	0.78	0.67	0.83	0.84	0.72
392	0.84	0.74	0.65	0.87	0.84	0.75

S11 Table. We retrained our models on *Platynereis* ground truth and quantified performance (at an IoU cutoff of 0.1) on test *Platynereis* embryos.

Stage	Model	TP	FN	FP	Precision	Recall
162 Nuclei	early "mouse-then-worm"	158	4	1	0.99	**0.98**
	late "mouse-then-worm"	157	5	0	**1**	0.97
	"worm only"	158	4	2	0.99	**0.98**
198 Nuclei	early "mouse-then-worm"	194	4	5	0.97	**0.98**
	late "mouse-then-worm"	194	4	2	**0.99**	**0.98**
	"worm only"	194	4	6	0.97	**0.98**

S12 Table. We retrained our models on *Platynereis* ground truth and quantified performance (at an IoU cutoff of 0.3) on test *Platynereis* embryos.

Stage	Model	TP	FN	FP	Precision	Recall
162 Nuclei	early "mouse-then-worm"	158	4	1	**0.99**	**0.98**
	late "mouse-then-worm"	156	6	1	0.99	0.96
	"worm only"	158	4	2	0.99	**0.98**
198 Nuclei	early "mouse-then-worm"	194	4	5	0.97	**0.98**
	late "mouse-then-worm"	194	4	2	**0.99**	**0.98**
	"worm only"	193	5	7	0.97	0.97

S13 Table. We retrained our models on *Platynereis* ground truth and quantified performance (at an IoU cutoff of 0.5) on test *Platynereis* embryos.

Stage	Model	TP	FN	FP	Precision	Recall
162 Nuclei	early "mouse-then-worm"	155	7	4	**0.97**	**0.96**
	late "mouse-then-worm"	152	10	5	0.97	0.94
	"worm only"	153	9	7	0.956	0.944
198 Nuclei	early "mouse-then-worm"	189	9	10	0.95	0.95
	late "mouse-then-worm"	192	6	4	**0.98**	**0.97**
	"worm only"	190	8	10	0.950	0.960

S1 Intro Fig. Segmentation tasks applied to images of a pastoral scene [37] and of a mouse embryo. (A) Raw image to be segmented. From top to bottom, an image of cows in a pasture, maximum intensity projections of 3D image of 16-cell mouse embryo, a z-slice from the 3D image. 3D image has dimensions (83.4 μm, 83.4 μm, 68 μm). Scale bar: 10 μm. (B) Semantic segmentation for images in (A). (C) Object detection for images in (A). (D) Instance segmentation for images in (A).

S2 Intro Fig. Comparison of five deep-learning-based methods for nuclear instance segmentation. Each column depicts a different method. From top to bottom, the rows illustrate how images are inputted into the network, the network's architecture, network outputs for a single nucleus, the post-processing steps, and the 3D instance segmentation, respectively.

S1 Fig. SNR of each image in the BlastoSPIM dataset. (A-B) Histogram, for annotated images in the original BlastoSPIM set and in the corrected late blastocyst segmentations, respectively, of the difference between mean foreground intensity and the mean background intensity. Black line: Cutoff for separating low SNR from moderate-to-high SNR in the original BlastoSPIM dataset.

S2 Fig. Number of nuclei annotated per developmental stage for corrected late blastocyst segmentations. This plot is analogous to Fig 1(E).

S3 Fig. Quantifying how well star-convex approximation applies to nuclear shapes in ground-truth time series data. We fit each nucleus to a star-convex shape, using 128 rays. For a single embryo, for which we have annotated ground truth for 89 consecutive timepoints (time points acquired every 15 minutes), we plot a box for each time to illustrate how well this fit performs, in terms of IoU. When all nuclei are in interphase, the star-convex fit performs quite well, at more than 90 percent IoU between the ground truth and the model-generated instance. During the transition from the 16-cell stage to the 32-cell stage and from the 32-cell stage to the 64-cell stage, the fit quality degrades. A small number of nuclei, about five in this time series cannot be fit by a star-convex shape, resulting in an IoU of less than 40 percent. We expect that the outlier nuclei (red) – which are not well fit by a star-convex shape – are likely mitotic, most likely in either metaphase or anaphase when the shape of the condensed chromatin is often complex. Black dashed line: the number of nuclei versus time.

S4 Fig. Nuclear volumes versus developmental time in live images, as inferred by both early embryo and late blastocyst Stardist-3D models. (A-D) For embryos 1-4 (indexed as in Fig 4), we plot both the median nuclear volume (blue) and the sum of nuclear volumes (orange) over time. Solid lines: inference based on early embryo model. Dashed line: inference based on late blastocyst model.

S5 Fig. Precision-recall curve as a function of Stardist-3D probability threshold for "early mouse-then-worm" model and "worm only" model. For all panels, the Stardist-3D probability threshold varies from 0.4 to 0.7. Lower precision (recall) values tend to occur for lower (higher) probability thresholds, respectively. At high IoU, the "early mouse-then-worm" model outperformed "worm only" model across these values of Stardist-3D probability threshold. (A-B) IoU threshold of 0.1 applied to 162-nuclei and 198-nuclei embryo, respectively. (C-D) IoU threshold of 0.3 applied to 162-nuclei and 198-nuclei embryo, respectively. (E-F) IoU threshold of 0.5 applied to 162-nuclei and 198-nuclei embryo, respectively. (G-H) IoU threshold of 0.6 applied to 162-nuclei and 198-nuclei embryo, respectively.

S6 Fig. Precision-recall curve as a function of Stardist-3D probability threshold for "late mouse-then-worm" model and "worm only" model. For all panels, the Stardist-3D probability threshold varies from 0.4 to 0.7. Lower precision (recall) values tend to occur for lower (higher) probability thresholds, respectively. At high IoU, the "late mouse-then-worm" model outperformed "worm only" model across these values of Stardist-3D probability threshold. (A-B) IoU threshold of 0.1 applied to 162-nuclei and 198-nuclei embryo, respectively. (C-D) IoU threshold of 0.3 applied to 162-nuclei and 198-nuclei embryo, respectively. (E-F) IoU threshold of 0.5 applied to 162-nuclei and 198-nuclei embryo, respectively. (G-H) IoU threshold of 0.6 applied to 162-nuclei and 198-nuclei embryo, respectively.

0.6 Description of Segmentation Methods

An important factor influencing model performance is the base-network used to learn the relationship between input images and the output instance representation. Many relevant methods (Fig 6, Fig 7, S1 Table), including QCANet [8], NuSeT [34], Cellpose [19], U3D-BCD [22], and EmbedSeg [24], are adaptations of the U-Net [38]. The current state-of-the-art model for nuclear segmentation in mouse embryos, QCANet uses two independent 3D U-Nets, one for a semantic map of nuclei and one for detecting nuclear centroids. On the other hand, NuSeT uses a 2D U-Net for semantic segmentation and a region proposal network (RPN) that shares the encoder for predicting boxes (S1 Intro Fig). Two methods, Cellpose and U3D-BCD, modify the U-Net by replacing the standard convolutional blocks with residual blocks, which incorporate identity-mapping short-cut connections to prevent high training error for deep networks [39]. EmbedSeg uses a 3D branched ERFNet (Efficient Residual Factorized Network), that also employs residual blocks with 1D factorized convolutions to reduce computational costs. These blocks are combined with downsampling and upsampling blocks to form the sequential encoder-decoder structure. Thus, these first five methods rely on convolutional maps generated by a U-shaped structure.

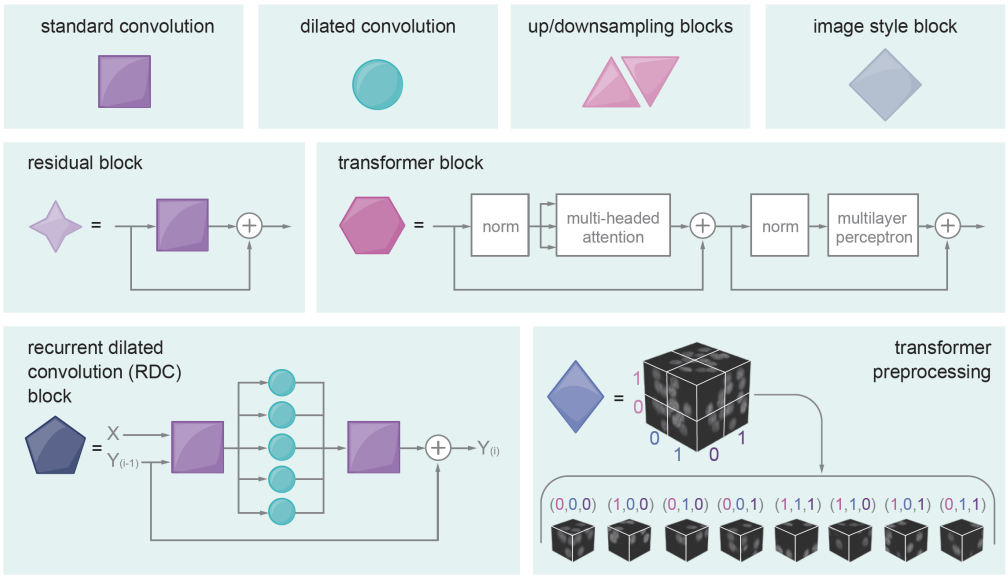

Fig 6. Components of the network architectures tested on our dataset. Dark purple, standard convolutional layer; teal circle, dilated convolutional layer; pink arrow, upsampling and downsampling block for U-Net-derived networks; gray square, global average pooling operation for computing image style in Cellpose; purple four-sided star, residual block, where the input to the convolutional layer is added to its output via a skip connection (the number of convolutional layers within the block may differ); blue pentagon, recurrent block in RDCNet *; magenta hexagon, transformer block; light blue diamond, linear projection and positional embedding of image patches in UNETR-BCD. *The recurrent block's previous output ($Y_{(i-1)}$) is combined with the original input (X) to the recurrent block through concatenation and convolution. Then, a stacked, dilated convolution block with shared weights is applied.

The remaining networks, including Stardist-3D, RDCNet, and UNETR-BCD, differ significantly from the previously discussed networks. Stardist-3D is the most similar to the previous methods because it relies on a series of residual blocks [39], containing convolutional layers; its distinguishing feature is its lack of down-sampling and, thus, of a U-shape. In contrast to Stardist-3D which uses different blocks connected successively, RDCNet uses the same block – a stacked, dilated convolution block – iteratively to refine the network outputs by operating on the initial input and the latest iteration's output (Fig 6 and Fig 7). Unlike all other methods, RDCNet also uses a semi-convolutional operation, one that explicitly includes each voxel's positional information. The final network, UNETR, the base network of UNETR-BCD, departs radically from other included methods by its use of transformers for the contracting path of the U-Net (Fig 7). Transformers, widely used in natural language processing, learn potentially long-ranged interactions between

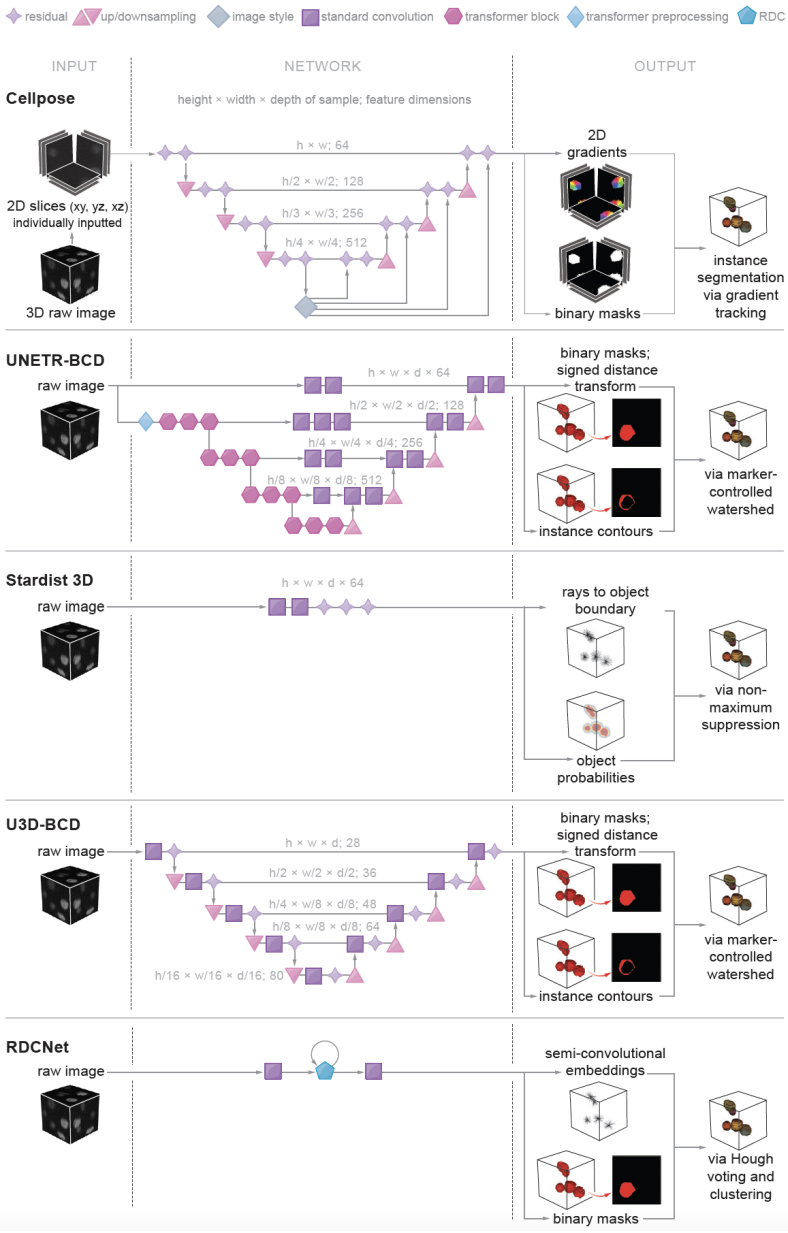

Fig 7. The network architectures tested on our dataset. (A) Cellpose preprocesses a 3D image into XY, XZ, and YZ slices. For each 2D slice, a modified U-Net is trained to predict two gradient maps (e.g. XY and XZ for X direction); these maps are combined to calculate components of a 3D gradient vector field. (B) U3D-BCD predicts a signed-distance map and instance contours via a modified 3D U-Net, where for each spatial resolution a convolutional layer is followed by a residual block containing two convolutional layers. (C) Stardist uses two convolutional layers with kernel sizes 7x7x7 and 3x3x3 followed by three residual blocks that each contain three convolutional layers with kernel size 3x3x3 to predict star-convex object boundaries. (D) UNETR-BCD divides an image into patches, linearly projects each patch into a vector and applies positional embedding to preserve the patches' relative spatial arrangement. The embedded patches are then passed into a sequence of 12 transformer blocks followed by a U-Net decoder. (E) RDCNet consists of a recurrent block between two convolutional blocks. Outputs are iteratively refined through the recurrent block.

image patches, similar to the relationships between words in a sentence. Just as in a U-net, UNETR connects the contracting path, here occupied by transformers, with an expanding path, a collection of convolutional and deconvolutional operators, via skip connections. Thus, Stardist-3D and RDCNet differ from other methods by lacking U-net-shaped structures, and UNETR-BCD has the shape of a U-Net but encodes via transformers.

The key remaining difference between the methods is their output instance representations and the necessary post-processing steps. Although a common output for all the networks is a voxel-wise score related to whether a voxel belongs to a nucleus, each method has a different way of combining this score with its other outputs to represent individual instances. Both QCANet and NuSeT predict a binary semantic segmentation map, which is combined with either nucleus center detection (QCANet) or object bounding boxes (NuSeT) through marker-based watershed in post-processing. However, since NuSeT uses a 2D network, it requires an additional post-processing step of combining 2D instances into 3D masks. Cellpose predicts a cell probability, thresholded to distinguish between foreground and background, as well as a gradient vector for each pixel. Gradient tracking and clustering are then performed to determine instances. While Cellpose also uses a 2D network, it is able to segment 3D images by making 2D predictions in each of the three spatial directions and estimating 3D gradients from the different 2D slices. Stardist outputs an object probability map and distances to an object boundary, represented as a star-convex polyhedron. In post-processing, non-maximum suppression is used to remove duplicate instance predictions, and the object probability threshold can be tuned to reduce false positives or negatives. RDCNet outputs a semantic segmentation map, then predicts semi-convolutional embeddings – a vector pointing from each foreground voxel to the center of the corresponding instance. The instance centers are determined through a Hough voting scheme, and voxel embeddings are clustered with a tunable margin during post-processing. Analogously, EmbedSeg predicts the probability of a voxel being an instance center, voxel embeddings, and clustering bandwidth or margin. In contrast to all other methods, U3D-BCD and UNETR-BCD predict multiple representations, including a semantic foreground mask, instance contours, and a signed distance map. Marker-controlled watershed on the predicted distance map is used for post-processing in these two methods.

0.7 Network Implementation Details

Each model was trained with data from 482 3D images of whole embryos. Each embryo was cropped into 8 to 16 patches depending on the size for a total of 4363 patches. Each patch had a resolution of 64x256x256. The patches overlap such that all voxels of a nucleus were fully contained in at least one patch. The raw intensity images were bit-shifted by four bits to the right, so that all voxel intensities are in the range between 0 and 255. Any value still above 255 was capped at 255.

0.7.1 RDC Net

For all hyperparameter combinations sampled for training, a few were held constant. The down sampling factors were chosen to be 1, 10, and 10, for the z, x, and y directions, respectively, to account for anisotropy. Spatial dropout was chosen to be 0.1, following the original paper. All networks were trained for a maximum of 200 epochs, batch size of 2, with the Adam optimizer and Cosine Decay Restarts scheduler with learning rates from 10-3 to 0. The set of model weights that resulted in the lowest validation loss across all epochs was saved. The patch size was either the original crop (64x256x256) or 32x256x256 (32 random, consecutive Z slices from the original crop). The number of groups (parallel stacked, dilated convolution blocks with shared weights), dilation rates, number of channels per group, number of iterations, and the margin parameter were also adjusted to observe their effects on network performance.

During inference on test images, each raw image was broken into patches with the same size as those the model was trained on. The test patches were passed through the model and the resulting label patches were stitched together by discarding redundant masks and any masks touching the patch boundaries, assuming each nucleus is located at the center of at least one patch.

0.7.2 Cellpose

Cellpose is called a generalist method for cell and nuclei instance segmentation. It is based on a 2D U-Net with residual blocks and style transfer. The objects are modeled as a diffusion gradient. The output is composed of horizontal and vertical gradient maps and a segmentation probability map. Since the original

cellpose model is 2D, the 3D patches were broken into 2D slices for training. The source code was modified to include a data loader, since the size of the 2D training set is orders of magnitude larger than the original Cellpose dataset. Models were trained for a maximum of 1000 epochs, either from scratch or from a pretrained Cellpose model. Test images were down-sampled by a factor of 0.5 in X and Y to improve performance since Cellpose is prone to over-segmentation for our full-resolution images in 3D. The patching and stitching method was the same as for RDCNet.

0.7.3 Stardist

3D Stardist was trained with patches of 32x256x256 sampled from the full size patches. Input intensity was normalized capping values below 1% and above 99.8%. For sampling the star convex in 3D, we used 96 rays with a grid of 1x4x4 to compensate for the anisotropy. Data augmentation included 2D flips and grid warping. For late stage embryos, more than 2.5 days, we used the optimal threshold based on the subset of training data in this category.

0.7.4 U3D BCD and UNETR

In U3D BCD, a 3D Residual U-Net, and UNETR, which uses a transformer as an encoder, the instance segmentation problem is broken down into learning hybrid representations i.e., semantic, contour and signed distance transform maps with the help of neural networks, and using watershed algorithm to separate instances.

UNETR encoder's transformer uses an embedding dimension of 768, the input volume is patched into volumetric tokens of dimensions 16 $\times$16$\times$16, and multi-head self-attention is performed with 12 heads. Augmentations, in the form of randomized brightness and contrast, flips, rotations and elastic deformations, were used. Finally, the input volumes were randomly cropped to 16$\times$128$\times$128, before passing them through the network. Adam optimizer with decaying learning rate was chosen for training. Weighted sum of Binary Cross Entropy (BCE) Loss and Dice Loss is taken for foreground and contour masks, while Mean Squared Error (MSE) was utilized for signed distance transform map predictions. Inference is performed by processing overlapping sliding windows across the large volumes of testing set. During post-processing, the multi-channel outputs from networks are combined by thresholding them appropriately to find instance seeds (or markers). Similarly, a more relaxed threshold on the outputs is used to obtain the foreground mask. Thereafter, marker-controlled watershed algorithm can be used with the help of seeds and predicted distance map to find instances.

0.7.5 Training with Synthetic Data

To counter the limited number of samples with densely-packed nuclei, we generate artificial samples to learn generalized features. This is made possible by modeling nuclei as 3-dimensional Gaussian kernels, of dimensions x, y, z where x, y $\in$ [100, 150] and z $\in$ [3, 6]. Elastic deformations, randomized lighting, and addition of noise are done to match SNR ratios with that of actual data-set. The models are pre-trained with this simulated data, allowing the network to fine-tune its predictions on the actual data-set.

Acknowledgments

We thank Lucy Reading-Ikkanda (Flatiron Institute) for figure artwork and Marta Baziuk (Princeton), Alana Bernys (Princeton) and Christopher Catalano (Princeton) for work on ground-truth annotations. This publication was made possible by grant number 1R01HD107026-01 (Posfai) from the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIH.

References

1. X. Lou, M. Kang, P. Xenopoulos, S. Muñoz-Descalzo, A.-K. Hadjantonakis, A rapid and efficient 2D/3D nuclear segmentation method for analysis of early mouse embryo and stem cell image data, Stem Cell Reports 2 (3) (2014) 382–397. doi:10.1016/j.stemcr.2014.01.010.

2. D. J. Barry, C. Gerri, D. M. Bell, R. D'Antuono, K. K. Niakan, GIANI – open-source software for automated analysis of 3D microscopy images, Journal of Cell Science 135 (10) (2022) jcs259511. doi:10.1242/jcs.259511.
URL https://journals.biologists.com/jcs/article/135/10/jcs259511/275435/GIANI-open-source-software-for-automated-analysis

3. G. Blin, D. Sadurska, R. Portero Migueles, N. Chen, J. A. Watson, S. Lowell, Nessys: A new set of tools for the automated detection of nuclei within intact tissues and dense 3D cultures, PLOS Biology 17 (8) (2019) e3000388. doi:10.1371/journal.pbio.3000388.
URL https://dx.plos.org/10.1371/journal.pbio.3000388

4. S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, A. Kreshuk, ilastik: interactive machine learning for (bio)image analysis, Nature Methods 16 (12) (2019) 1226–1232. doi:10.1038/s41592-019-0582-9.
URL http://www.nature.com/articles/s41592-019-0582-9

5. J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, A. E. Carpenter, Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl, Nature Methods 16 (12) (2019) 1247–1253. doi:10.1038/s41592-019-0612-7.
URL http://www.nature.com/articles/s41592-019-0612-7

6. O. Cicek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in: S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, W. Wells (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Vol. 9901, Springer International Publishing, Cham, 2016, pp. 424–432. doi:10.1007/978-3-319-46723-8_49.
URL https://link.springer.com/10.1007/978-3-319-46723-8_49

7. P. F. Jaeger, S. A. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, K. H. Maier-Hein, Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection, in: A. V. Dalca, M. B. McDermott, E. Alsentzer, S. G. Finlayson, M. Oberst, F. Falck, B. Beaulieu-Jones (Eds.), Proceedings of the Machine Learning for Health NeurIPS Workshop, Vol. 116 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 171–183.
URL https://proceedings.mlr.press/v116/jaeger20a.html

8. Y. Tokuoka, T. G. Yamada, D. Mashiko, Z. Ikeda, N. F. Hiroi, T. J. Kobayashi, K. Yamagata, A. Funahashi, 3D convolutional neural networks-based segmentation to acquire quantitative criteria of the nucleus during mouse embryogenesis, npj Systems Biology and Applications 6 (1) (2020) 32. doi:10.1038/s41540-020-00152-8.
URL https://www.nature.com/articles/s41540-020-00152-8

9. E. Posfai, S. Petropoulos, F. R. O. de Barros, J. P. Schell, I. Jurisica, R. Sandberg, F. Lanner, J. Rossant, Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo, eLife 6 (2017) e22906. doi:10.7554/eLife.22906.
URL https://elifesciences.org/articles/22906

10. B. Gu, E. Posfai, J. Rossant, Efficient generation of targeted large insertions by microinjection into two-cell-stage mouse embryos, Nature Biotechnology 36 (7) (2018) 632–637. doi:10.1038/nbt.4166.
URL http://www.nature.com/articles/nbt.4166

11. J.-P. Concordet, M. Haeussler, CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens, Nucleic Acids Research 46 (W1) (2018) W242–W245. arXiv:https://academic.oup.com/nar/article-pdf/46/W1/W242/25110393/gky354.pdf, doi:10.1093/nar/gky354.
URL https://doi.org/10.1093/nar/gky354

12. B. Gu, E. Posfai, M. Gertsenstein, J. Rossant, Efficient generation of large-fragment knock-in mouse models using 2-cell (2c)-homologous recombination (hr)-crispr, Current Protocols in Mouse Biology 10 (1) (2020) e67. arXiv:https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpmo.67, doi:https://doi.org/10.1002/cpmo.67.
URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpmo.67

13. K. McDole, Y. Zheng, Generation and live imaging of an endogenous Cdx2 reporter mouse line, genesis 50 (10) (2012) 775–782. doi:10.1002/dvg.22049.
URL https://onlinelibrary.wiley.com/doi/10.1002/dvg.22049

14. M. D. Muzumdar, B. Tasic, K. Miyamichi, L. Li, L. Luo, A global double-fluorescent Cre reporter mouse, Genesis (New York, N.Y.: 2000) 45 (9) (2007) 593–605. doi:10.1002/dvg.20335.

15. R. Niwayama, P. Moghe, Y.-J. Liu, D. Fabrèges, F. Buchholz, M. Piel, T. Hiiragi, A Tug-of-War between Cell Shape and Polarity Controls Division Orientation to Ensure Robust Patterning in the Mouse Blastocyst, Developmental Cell 51 (5) (2019) 564–574.e6. doi:10.1016/j.devcel.2019.10.012.
URL https://linkinghub.elsevier.com/retrieve/pii/S1534580719308561

16. R. A. Fisher, Statistical Methods for Research Workers, in: S. Kotz, N. L. Johnson (Eds.), Breakthroughs in Statistics, Springer New York, New York, NY, 1992, pp. 66–70. doi:10.1007/978-1-4612-4380-9_6.
URL http://link.springer.com/10.1007/978-1-4612-4380-9_6

17. J. Lasky-Su, Statistical Techniques for Genetic Analysis, in: Clinical and Translational Science, Elsevier, 2017, pp. 347–362. doi:10.1016/B978-0-12-802101-9.00019-3.
URL https://linkinghub.elsevier.com/retrieve/pii/B9780128021019000193

18. P. P. Laissue, R. A. Alghamdi, P. Tomancak, E. G. Reynaud, H. Shroff, Assessing phototoxicity in live fluorescence imaging, Nature Methods 14 (7) (2017) 657–661. doi:10.1038/nmeth.4344.
URL http://www.nature.com/articles/nmeth.4344

19. C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, Cellpose: a generalist algorithm for cellular segmentation, Nature Methods 18 (1) (2021) 100–106. doi:10.1038/s41592-020-01018-x.
URL http://www.nature.com/articles/s41592-020-01018-x

20. M. Weigert, U. Schmidt, R. Haase, K. Sugawara, G. Myers, Star-convex polyhedra for 3d object detection and segmentation in microscopy, in: The IEEE Winter Conference on Applications of Computer Vision (WACV), 2020. doi:10.1109/WACV45572.2020.9093435.

21. R. Ortiz, G. de Medeiros, A. H. F. M. Peters, P. Liberali, M. Rempfler, RDCNet: Instance Segmentation with a Minimalist Recurrent Residual Network, in: M. Liu, P. Yan, C. Lian, X. Cao (Eds.), Machine Learning in Medical Imaging, Vol. 12436, Springer International Publishing, Cham, 2020, pp. 434–443. doi:10.1007/978-3-030-59861-7_44.
URL https://link.springer.com/10.1007/978-3-030-59861-7_44

22. Z. Lin, D. Wei, M. D. Petkova, Y. Wu, Z. Ahmed, K. S. K, S. Zou, N. Wendt, J. Boulanger-Weill, X. Wang, N. Dhanyasi, I. Arganda-Carreras, F. Engert, J. Lichtman, H. Pfister, NucMM Dataset: 3D Neuronal Nuclei Instance Segmentation at Sub-Cubic Millimeter Scale, in: M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Vol. 12901, Springer International Publishing, Cham, 2021, pp. 164–174. doi:10.1007/978-3-030-87193-2_16.
URL https://link.springer.com/10.1007/978-3-030-87193-2_16

23. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, D. Xu, UNETR: Transformers for 3D Medical Image Segmentation, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2022, pp. 1748–1758. doi:10.1109/WACV51458.2022.00181.
URL https://ieeexplore.ieee.org/document/9706678/

24. M. Lalit, P. Tomancak, F. Jug, Embedseg: Embedding-based instance segmentation for biomedical microscopy data, Medical Image Analysis 81 (2022) 102523. doi:https://doi.org/10.1016/j.media.2022.102523.
URL https://www.sciencedirect.com/science/article/pii/S1361841522001700

25. E. Tsichlaki, G. FitzHarris, Nucleus downscaling in mouse embryos is regulated by cooperative developmental and geometric programs, Scientific Reports 6 (1) (2016) 28040. doi:10.1038/srep28040.
URL http://www.nature.com/articles/srep28040

26. F. R. Neumann, P. Nurse, Nuclear size control in fission yeast, Journal of Cell Biology 179 (4) (2007) 593–600. doi:10.1083/jcb.200708054.
URL https://rupress.org/jcb/article/179/4/593/45063/Nuclear-size-control-in-fission-yeast

27. P. Jorgensen, N. P. Edgington, B. L. Schneider, I. Rupeš, M. Tyers, B. Futcher, The Size of the Nucleus Increases as Yeast Cells Grow, Molecular Biology of the Cell 18 (9) (2007) 3523–3532. doi:10.1091/mbc.e06-10-0973.
URL https://www.molbiolcell.org/doi/10.1091/mbc.e06-10-0973

28. C. J. Chan, M. Costanzo, T. Ruiz-Herrero, G. Mönke, R. J. Petrie, M. Bergert, A. Diz-Muñoz, L. Mahadevan, T. Hiiragi, Hydraulic control of mammalian embryo size and cell fate, Nature 571 (7763) (2019) 112–116. doi:10.1038/s41586-019-1309-x.

29. K. McDole, Y. Xiong, P. A. Iglesias, Y. Zheng, Lineage mapping the pre-implantation mouse embryo by two-photon microscopy, new insights into the segregation of cell fates, Developmental Biology 355 (2) (2011) 239–249. doi:10.1016/j.ydbio.2011.04.024.
URL https://linkinghub.elsevier.com/retrieve/pii/S0012160611002569

30. Y. Yamanaka, F. Lanner, J. Rossant, FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst, Development 137 (5) (2010) 715–724. doi:10.1242/dev.043471.
URL https://journals.biologists.com/dev/article/137/5/715/44186/FGF-signal-dependent-segregation-of-primitive

31. E. Kuehn, A. W. Stockinger, J. Girard, F. Raible, B. D. Özpolat, A scalable culturing system for the marine annelid Platynereis dumerilii, PLOS ONE 14 (12) (2019) e0226156. doi:10.1371/journal.pone.0226156.
URL https://dx.plos.org/10.1371/journal.pone.0226156

32. N. Koushki, A. Ghagre, L. K. Srivastava, C. Sitaras, H. Yoshie, C. Molter, A. J. Ehrlicher, Lamin a redistribution mediated by nuclear deformation determines dynamic localization of yap, bioRxiv (2020). arXiv:https://www.biorxiv.org/content/early/2020/03/20/2020.03.19.998708.full.pdf, doi:10.1101/2020.03.19.998708.
URL https://www.biorxiv.org/content/early/2020/03/20/2020.03.19.998708

33. W. Ouyang, F. Beuttenmueller, E. Gómez-de Mariscal, C. Pape, T. Burke, C. Garcia-López-de Haro, C. Russell, L. Moya-Sans, C. de-la Torre-Gutiérrez, D. Schmidt, D. Kutra, M. Novikov, M. Weigert, U. Schmidt, P. Bankhead, G. Jacquemet, D. Sage, R. Henriques, A. Muñoz-Barrutia, E. Lundberg, F. Jug, A. Kreshuk, Bioimage model zoo: A community-driven resource for accessible deep learning in bioimage analysis, bioRxiv (2022). arXiv:https://www.biorxiv.org/content/early/2022/06/08/2022.06.07.495102.full.pdf, doi:10.1101/2022.06.07.495102.
URL https://www.biorxiv.org/content/early/2022/06/08/2022.06.07.495102

34. L. Yang, R. P. Ghosh, J. M. Franklin, S. Chen, C. You, R. R. Narayan, M. L. Melcher, J. T. Liphardt, NuSeT: A deep learning tool for reliably separating and analyzing crowded cells, PLOS Computational Biology 16 (9) (2020) e1008193. doi:10.1371/journal.pcbi.1008193.
URL https://dx.plos.org/10.1371/journal.pcbi.1008193

35. V. Ulman, M. Maska, K. E. G. Magnusson, O. Ronneberger, C. Haubold, N. Harder, P. Matula, P. Matula, D. Svoboda, M. Radojevic, I. Smal, K. Rohr, J. Jalden, H. M. Blau, O. Dzyubachyk, B. Lelieveldt, P. Xiao, Y. Li, S.-Y. Cho, A. C. Dufour, J.-C. Olivo-Marin, C. C. Reyes-Aldasoro, J. A. Solis-Lemus, R. Bensch, T. Brox, J. Stegmaier, R. Mikut, S. Wolf, F. A. Hamprecht, T. Esteves, P. Quelhas, O. Demirel, L. Malmström, F. Jug, P. Tomancak, E. Meijering, A. Munoz-Barrutia, M. Kozubek, C. Ortiz-de Solorzano, An objective comparison of cell-tracking algorithms, Nature Methods 14 (12) (2017) 1141–1152. doi:10.1038/nmeth.4473.
URL http://www.nature.com/articles/nmeth.4473

36. V. Bondarenko, M. Nikolaev, D. Kromm, R. Belousov, A. Wolny, S. Rezakhani, J. Hugger, V. Uhlmann, L. Hufnagel, A. Kreshuk, J. Ellenberg, A. Erzberger, M. Lutolf, T. Hiiragi, Coordination between embryo growth and trophoblast migration upon implantation delineates mouse embryogenesis, bioRxiv (2022). arXiv:https://www.biorxiv.org/content/early/2022/06/16/2022.06.13.495767.full.pdf, doi:10.1101/2022.06.13.495767.
URL https://www.biorxiv.org/content/early/2022/06/16/2022.06.13.495767

37. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Vol. 8693, Springer International Publishing, Cham, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.
URL http://link.springer.com/10.1007/978-3-319-10602-1_48

38. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Vol. 9351, Springer International Publishing, Cham, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28.
URL http://link.springer.com/10.1007/978-3-319-24574-4_28

39. K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
URL http://ieeexplore.ieee.org/document/7780459/