Validation and tuning of *in situ* transcriptomics image processing workflows with crowdsourced annotations

Jenny Vo-Phamhi, Kevin Yamauchi, and Rafael Gómez-Sjöberg

Chan Zuckerberg Biohub, San Francisco, California

Keywords: in situ *transcriptomics, image processing, crowdsourced annotation, ground truth*

Abstract

Recent advancements in *in situ* methods, such as multiplexed *in situ* RNA hybridization and *in situ* RNA sequencing, have deepened our understanding of the way biological processes are spatially organized in tissues. Automated image processing and spot-calling algorithms for analyzing *in situ* transcriptomics images have many parameters which need to be tuned for optimal detection. Having ground truth datasets (images where there is very high confidence on the accuracy of the detected spots) is essential for evaluating these algorithms and tuning their parameters.

We present a first-in-kind open-source toolkit and framework for *in situ* transcriptomics image analysis that incorporates crowdsourced annotations, alongside expert annotations, as a source of ground truth for the analysis of *in situ* transcriptomics images. The kit includes tools for preparing images for crowdsourcing annotation to optimize crowdsourced workers' ability to annotate these images reliably, performing QC on worker annotations, extracting candidate parameters for spot-calling algorithms from sample images, tuning parameters for spot-calling algorithms, and evaluating spot-calling algorithms and worker performance. These tools are wrapped in a modular pipeline with a flexible structure that allows users to take advantage of crowdsourced annotations from any source of their choice. We tested the pipeline using real and synthetic *in situ* transcriptomics images and annotations from the Amazon Mechanical Turk system obtained via Quanti.us. Using real images from *in situ* experiments and simulated images produced by one of the tools in the kit, we studied worker sensitivity to spot characteristics and established rules for annotation quality control (QC). We explored and demonstrated the use of ground truth generated in this way for validating spot-calling algorithms and tuning their

parameters, and confirmed that consensus crowdsourced annotations are a viable substitute for expert-generated ground truth for these purposes.

Data Availability

The *In Situ* Transcriptomics Annotation (INSTA) pipeline software is available from https://github.com/czbiohub/instapipeline. The SpotImage software is available from https://github.com/czbiohub/spotimage. The figures and data for this project are available from https://github.com/czbiohub/instapaper.

Introduction

Diversity of form follows diversity of function in biological tissues. The anatomy and cellular properties of each tissue come from cell-specific gene expression patterns.(1) To understand important biological processes, such as development, wound healing, and disease, it is necessary to study the 3-dimensional spatial architecture of biological tissues and their gene expression patterns at the cellular (or even subcellular) level. Recent advancements in *in situ* methods(2–9) (e.g., DNA(10–13), RNA(10,14,15), and protein(10,16) measurements in tissue sections) have deepened our understanding of the way biological processes are spatially organized in tissues. In particular, recent *in situ* transcriptomics tools, such as multiplexed *in situ* RNA hybridization and *in situ* RNA sequencing, have facilitated the spatial mapping of gene expression with subcellular resolution.(1)

In situ transcriptomics methods utilize the binding of fluorescent probes to specific RNA target molecules with high complementarity within cultured cells and tissue sections. Extracting

the positions of the fluorescent probes, which appear in microscopy images as bright spots, presents a key image processing challenge. Automated spot detection is not trivial due to noise arising from light scattering and background autofluorescence.(17) Although automated image processing and spot-calling algorithms exist (for brevity, from now on, we will use the term "spot-calling algorithm" to refer to the whole image processing and spot-finding pipeline), they have many parameters which need to be tuned for optimal detection.(18–22) Having ground truth datasets (images where there is very high confidence on the accuracy of the detected spots) is essential for evaluating these algorithms and tuning their parameters.

Studies typically use synthetic images to evaluate or test the performance of any spot detector because ground truth does not inherently exist for real images.(17) The typical way of generating ground truth datasets for real images is having an expert inspect the images and annotate the valid spot locations by hand.(17) In cases where manual annotation of a large *in situ* transcriptomics image dataset by an expert is unfeasible, it is necessary to have alternative sources of ground truth. One proposed solution to this problem is iterative human-in-the-loop deep learning workflows, where ground truth generated by spot-calling algorithms can be manually refined.(23) Since valid spot locations can often be apparent even to minimally-trained, non-expert human eyes, we propose that crowdsourcing is a feasible way to generate large, high quality ground truth datasets.

Crowdsourcing refers to the use of web-based systems to recruit random volunteers or paid workers to perform tasks remotely. Recent work has indicated that carefully crowdsourced annotations can expedite data processing tasks that have a visual component. Volunteer-based

citizen science has made substantial contributions to areas of biology from proteomics(24–26) to ecology(27). When tasks are less intrinsically interesting to volunteers, minimally-trained workers can complete tasks for small payments through crowdsourcing platforms such as Amazon's Mechanical Turk (MTurk), and the consensus annotations (across multiple workers or "turkers") can be highly comparable with expert annotations, and sufficiently reliable for use as training data for detection algorithms.(27–29) Therefore, we hypothesized that consensus from crowdsourced annotations can be used as a substitute for ground truth to tune and benchmark spot-calling algorithms. However, there are no published *in situ* transcriptomics pipelines that can incorporate ground truth from crowdsourced annotations. Such pipelines should have mechanisms to prepare images for annotation, process annotations, establish consensus from the annotations, and generate annotation performance metrics.

In this paper, we present INSTA (IN situ Sequencing and Transcriptomics Annotation), an open-source toolkit and framework that incorporates crowdsourced annotations alongside expert annotations as a source of ground truth for the analysis of *in situ* transcriptomics images. Using real images from *in situ* experiments, and simulated images produced by a tool we developed to generate synthetic, customizable *in situ* transcriptomics images, we explored worker sensitivity to the size, quantity, and density of spots, and we established rules for annotation quality control. Based on these rules, we developed tools for preparing images to optimize workers' ability to annotate these images reliably, performing quality control (QC) on worker annotations to get maximum value from them, extracting candidate parameters for spot-calling algorithms from sample images, tuning parameters for spot-calling algorithms, and evaluating spot-calling algorithms and worker performance. We wrapped these tools in a

modular pipeline with a flexible structure that allows users to take advantage of crowdsourced annotations. The toolkit includes an annotation ingestion class designed to work with Quanti.us(28), and it can be easily adapted to work with any crowdsourcing system by creating custom annotation ingestion classes. We tested the pipeline using images from the *in situ* transcriptomics dataset from starfish(30,31), a Python library for analysis of image-based transcriptomics data developed by the Chan Zuckerberg Initiative, and annotations from MTurk via Quanti.us. We explored and demonstrated the use of ground truth generated in this way for validating spot-calling algorithms and tuning their parameters, and confirmed that consensus crowdsourced annotations are a viable substitute for expert-generated ground truth for these purposes.

In addition to this pipeline, we created a tool to generate synthetic *in situ* images, which we call SpotImage. This tool receives background images from real *in situ* experiments and adds simulated spots to them. The user can vary spot characteristics including size, shape, location, distribution throughout the image, and signal to noise ratio (SNR). These parametrized synthetic images are very useful for testing crowdsourced annotations and spot calling algorithms. More details in Supplementary Text 1.

Materials and Methods

Structure of a modular pipeline for tuning *in situ* transcriptomics image processing with crowdsourced annotations

This section provides a high-level overview of INSTA, a pipeline for crowdsourcing annotations and for tuning and evaluating spot-calling methods (Figure 1). Greater detail will be

provided in Sections III and IV, and examples with two different *in situ* transcriptomics chemistries are provided in Sections V and VI.

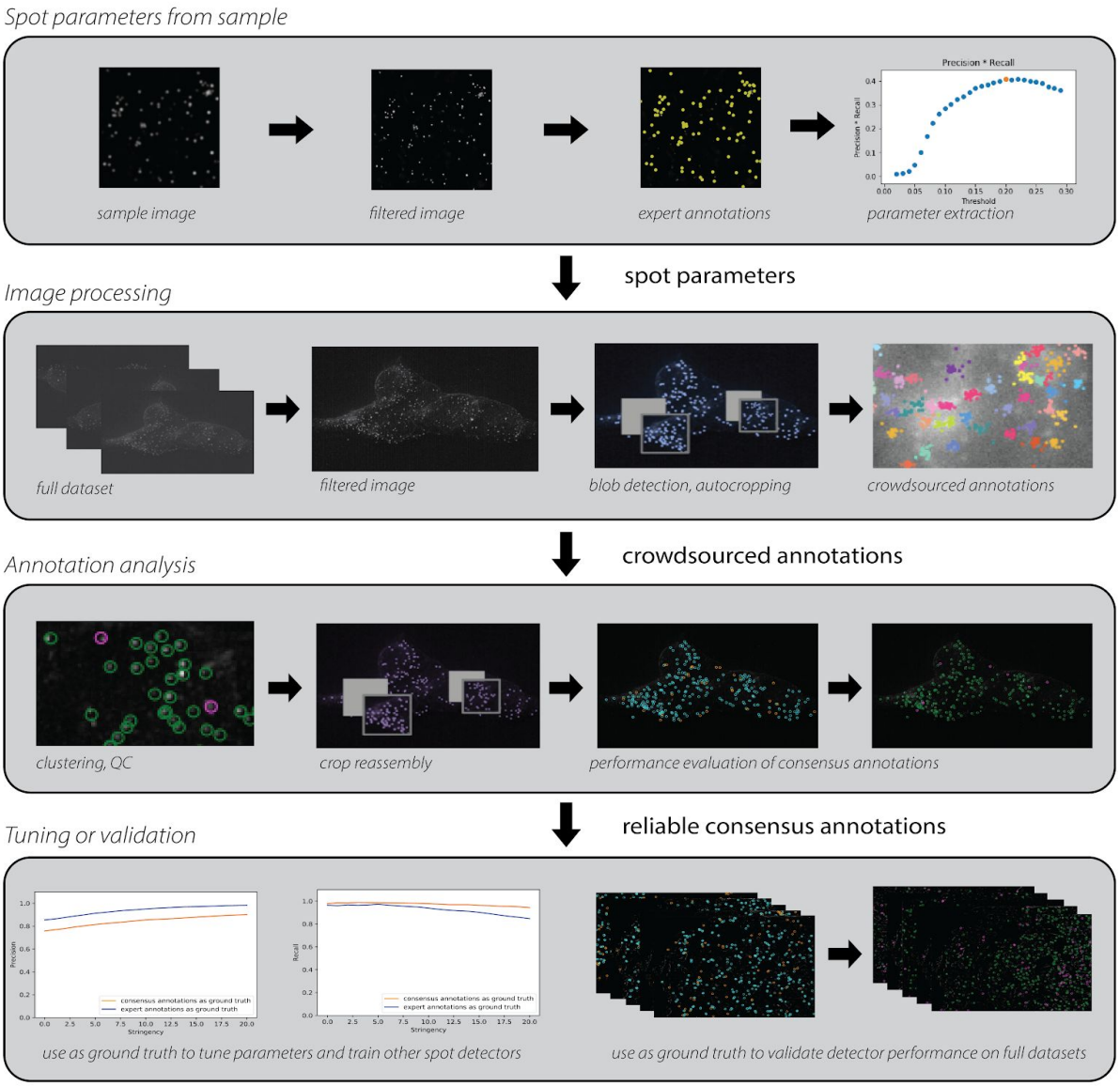

Figure 1: INSTA (IN situ Sequencing and Transcriptomics Annotation) is a pipeline for tuning and validating spot detection methods using crowdsourced annotations.

The input to the pipeline consists of images from a particular *in situ* transcriptomics chemistry and a spot detection algorithm to be optimized for that chemistry. An expert

designates one image as a representative image for the *in situ* chemistry used, and annotates it. The remaining images are assigned to the test dataset..

From this small amount of expert input, the tool learns approximately what a spot in this chemistry should look like. That is, a script extracts parameters which characterize the brightness (intensity) and size (sigma of a 2D Gaussian approximation of a spot) profiles of the spots of that chemistry. These parameters are passed to a blob detector that uses scikit-image's implementation of the Laplacian of Gaussian algorithm.(18)

The pipeline then processes each test image separately. For each test image:

- The blob detector uses the spot parameters it learned from parameter extraction to do rough, first-pass spot-calling.
- A script detects the crowded regions and recursively crops the images until the sub-images are sufficiently uncrowded that a human worker should easily be able to click on all the spots without getting frustrated or tired.
- All the pieces of the image – the crops and the parent images – are sent to Quanti.us, which is a platform for crowdsourced image annotation through Amazon's Mechanical Turk platform. Custom crowdsourced data ingestion classes can be written to allow the pipeline to work with any crowdsourced annotation system.
- Each image is annotated by a user-defined number of workers (typically 20 to 30). In the annotation analysis stage, all the crowdsourced annotations are clustered and QC is performed based on characteristics of the spots and clusters to produce

consensus annotations, which are then reassembled to produce an original image that has been annotated with high precision and recall.

- These annotations can be used to tune and train other spot detectors. They can also be used to validate or quantify the performance of the spot detector that this run of the pipeline attempted to optimize, or of another detector.
- If the optimized spot detector's performance is satisfactory, the detector may be useful for other images. If the detector's performance is unsatisfactory, the parameters can be modified and the detector can be reevaluated against the worker annotations.

Two key aspects of the pipeline should be highlighted: First, individual segments of the pipeline may be used separately for assorted purposes. For example, to simply get annotations for images without optimizing any spot detectors, the latter portion of the pipeline can be used to crop images and QC the crowdsourced annotations, with cropping based on spots detected by some detector that has been deemed sufficiently good for the purpose of preliminary detection. Second, workers tend to perform poorly on full-size raw images because there are too many spots and the spots are too close together. The pipeline includes a recursive cropping functionality that automatically breaks up each image into sections that the workers can handle effectively.

Section III will discuss the limits of what workers can accurately annotate with regard to brightness, density, number of spots, etc. Section IV will further discuss ways to prepare images to optimize worker performance.

Results

Worker performance is limited by spot crowding, visibility, and quantity

We crowdsource annotations using Quanti.us (30) and perform clustering and quality control on the annotations to arrive at consensus coordinates for spot locations. Each image sent to Quanti.us is annotated by 25 workers (Supplementary Text 2). The resulting annotations are then clustered via Affinity Propagation to find the initial set of annotation clusters - this algorithm does not require *a priori* knowledge of the number of clusters.(32,33) Given that some of the annotations do not correspond to spot locations and some of the annotations cover adjacent spots (Supplementary Figure 3), we next perform quality control to identify false positives and unmix adjacent clusters.

To determine if a cluster is a false positive, we threshold the clusters by the number of annotations in the cluster (Fig. 2a). In annotations of synthetic images we observed that clusters are distributed bimodally (Supplementary Text 3) by number of annotations. Clusters with few annotations tend to be incorrect (that is, the cluster centroid is not within a given threshold pixel radius of the closest actual spot location). So for the first QC step, clusters are sorted by number of annotations and one-dimensional k-means with $k = 2$ is applied to find the threshold number of annotations. All clusters with fewer annotations than this threshold are removed. This thresholding method tends to be aggressive; we would rather miss spots than “detect” incorrect spots, since in reality it is inevitable that some spots will be missed anyway when signals are too faint or overlap. In an experiment with synthetic images, this QC step yielded 100%, 100%, and 100% precision (40%, 13%, and 7% increase compared with the precision obtained without thresholding clusters by number of annotations) and 51%, 78%, and 99% recall (19%, 12%, and

0% decrease) for images with mean SNR = 5, 10, and 15 respectively (Supplementary Figure 4a).

To detect whether a cluster corresponds to multiple spots that are very close together, we threshold the clusters by the fraction of unique workers who contribute multiple times to the cluster. We observed that when spots are very close together, the clusters associated with multiple spots may clump together into one annotation cluster, but some workers do detect that the spots are supposed to be separate and those workers contribute more than one click within the region that the clustering algorithm detects as one cluster (See pink clusters in Supplementary Figure 5). The threshold fraction is found between the main mode of the distribution and the tail by identifying the point of steepest increase in histogram values. All clusters with a greater fraction of multiple-clicking workers than this threshold fraction are removed. Therefore, the fraction of workers who contribute only once can predict whether a cluster is actually clumpy, even if sometimes the actual spots are so close that most of the workers looking at them interpret them as one spot and it is not possible to identify that cluster as clumpy (Figure 2b). We declump each clumpy cluster using two-dimensional *k*-means (Figure 2b). In the same experiment with synthetic images, this QC step yielded 64%, 88%, and 94% precision (4%, 1%, and 0% increase over results without QC) and 67%, 87%, and 96% recall (3%, 3%, and 3% decrease) for images with mean SNR = 5, 10, and 15 respectively (Supplementary Figure 4b).

Performing declumping after false positive detection boosts recall. In the same experiment with synthetic images, removing small clusters and separating clumpy clusters yielded 100%, 100%, and 100% precision (40%, 13%, and 7% increase in precision) and 51%,

78%, and 98% recall (19%, 12%, and 2% decrease in recall) for images with mean SNR = 5, 10, and 15 respectively (Figure 2c).

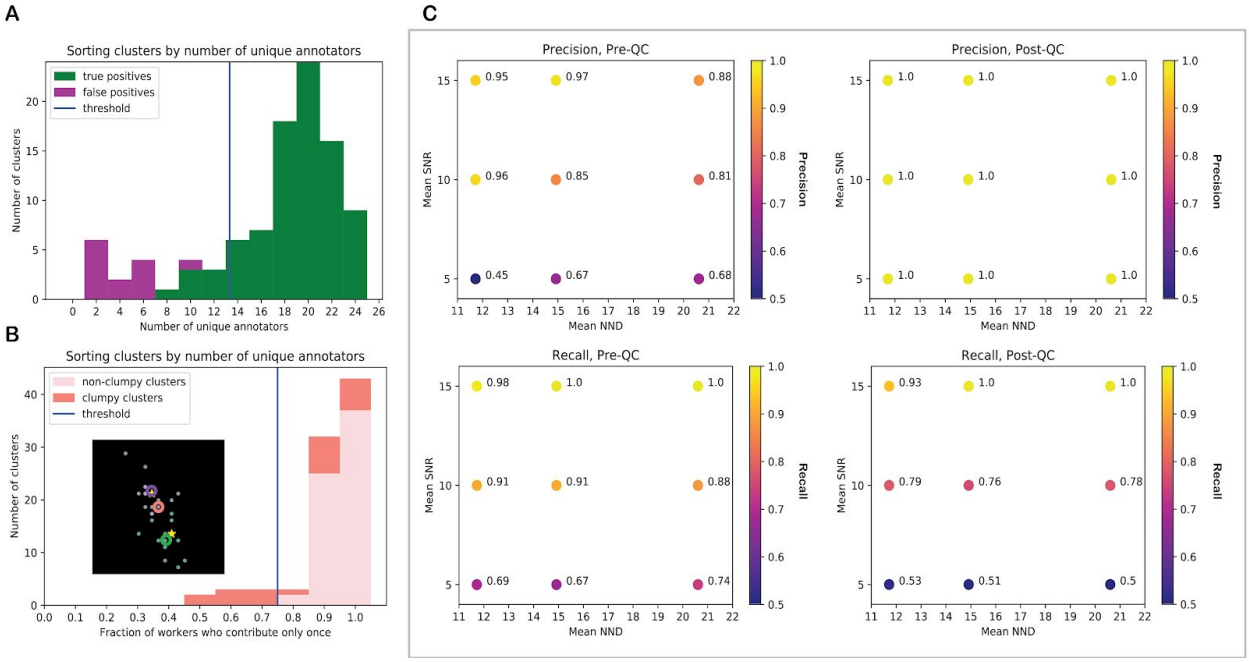

Figure 2: QC, including cluster size thresholding and declumping, improves precision, sometimes at the expense of recall, for images with lower SNR values. (A) *Clusters with fewer workers tend to be incorrect. Sort clusters by number of unique workers annotating them. The fraction of workers who contribute once can predict whether a cluster is clumpy (it corresponds to multiple image spots that are close together).* ***(B)*** *Sort clusters by fraction of unique workers contributing. Isolate and declump the clusters where many workers contribute more than once, as the inset demonstrates using real data.* (*Inset – orange circle: original centroid, green and purple circles: new centroids found by declumping, green and purple dots: worker annotations assigned to new centroids by declumping, stars: actual spot locations.) Clumps are declumped using 2D k-means. For more examples of declumped clusters, see Supplementary Figure 1.* ***(C)*** *Thresholding clusters by the number of annotations in the cluster and by the fraction of unique workers who contribute multiple times to the cluster improved recall by 17%, while decreasing precision by 11% on average, in an experiment with images of mean SNR = 5, 10, and 15 and average nearest neighbor distance (NND) ≈ 11.5, 15, and 20.5.*

We observed some limits on the utility of QC. When spot visibility was poor, nearest neighbor distance was very small, or an image had too many spots, QC was unable to improve

precision and recall to acceptable levels since the quality of the raw annotations was so low. Using the NND-varying and SNR-varying features of our SpotImage tool, we tested the limits of spot visibility and crowdedness that workers can accurately annotate. In an experiment with synthetic images of mean SNR = 3, 5, 7, 9, and 11, and spot size = 0.5, 1.0, and 1.75 pixels (Supplementary Figure 6a), small spot sizes required larger mean SNR to achieve recall greater than 50%. For example, when the spot size (sigma of a gaussian fit to the intensity) was half a pixel (about 5 microns), recall was zero until SNR > 9 (Supplementary Figure 6b). At lower SNR values, even spots with large nearest neighbor distances tended to be missed, and as spot SNR increased, the mean NND of undetected spots decreased (Supplementary Figure 6c). We observed that the radius of the symbol that Quanti.us uses to mark spots identified by workers limits the minimum nearest neighbor distance between spots which we can reasonably expect workers to discern to 4% of the image's width. When spots are closer than this quantity, the mark on one spot obscures neighboring spots. We tested inverting the images (dark spots on a light background) before submitting them to Quanti.us to see if this would improve worker performance. A linear regression between recall with inversion and recall without inversion resulted in a slope of 1.004, with a Pearson's correlation coefficient of $r = 0.985$ (Supplementary Figure 6d), making it clear that inversion is not helpful.

We also sought to understand the total number of spots that workers can annotate in one image. In an experiment with SNR = 3, 5, 7 and number of spots per image between 50 and 225, the number of clicks per worker per image increased as the number of spots in the image increased, until it leveled off at around 120 on average, suggesting that 120 was the upper bound on the total number of spots workers were willing to click for the payment offered in these

experiments ($0.05 per image) (Supplementary Figure 7a). As the number of spots increased beyond 50, the fraction of spots that workers were willing to click decreased. On average, workers annotated almost all spots for images with 50 spots but only about 60% of all spots for images with 200 spots (Supplementary Figure 7b). However, even though each worker annotated a smaller fraction of the spots as the number of spots in the image increased, most spots were still getting annotations from at least half the workers (Supplementary Figure 7c). In other words, some workers detected spots which other workers missed.

These effects of spot visibility, crowding, and quantity on worker performance are intuitive, but these results provide quantitative metrics to understand worker performance as a function of image quality, and objective guidelines for image pre-processing that allow us to optimize the performance for a given image dataset. In the next section, we present the methods we developed to prepare images which workers are more likely to annotate reliably.

Image pre-processing improves workers' ability to annotate images reliably

Given the guidelines we identified for the spot density, total number of spots and number of spots that can be accurately annotated in a given image, raw *in situ* transcriptomics images would need to be pre-processed in order to meet the guidelines.

In the first preparation step, the raw images are pre-processed with filters to enhance contrast between spots and the rest of the image (Figure 3a). This process is described in greater detail in Supplementary Text 4. In the second step, the filtered images are automatically subdivided ("autocropped") to produce child images with sufficiently small spot densities and

large nearest neighbor distances to be effectively annotated by workers (Figure 3b).[1] In an experiment with a real single molecule fluorescence *in situ* hybridization (smFISH) image with 268 spots of typical contrast and density, annotated by 25 workers, applying cropping resulted in precision and recall of 97% and 87%, improvements of 50% and 64% respectively, over the un-cropped image (Figure 3c).

[1] More details in Supplementary Information.

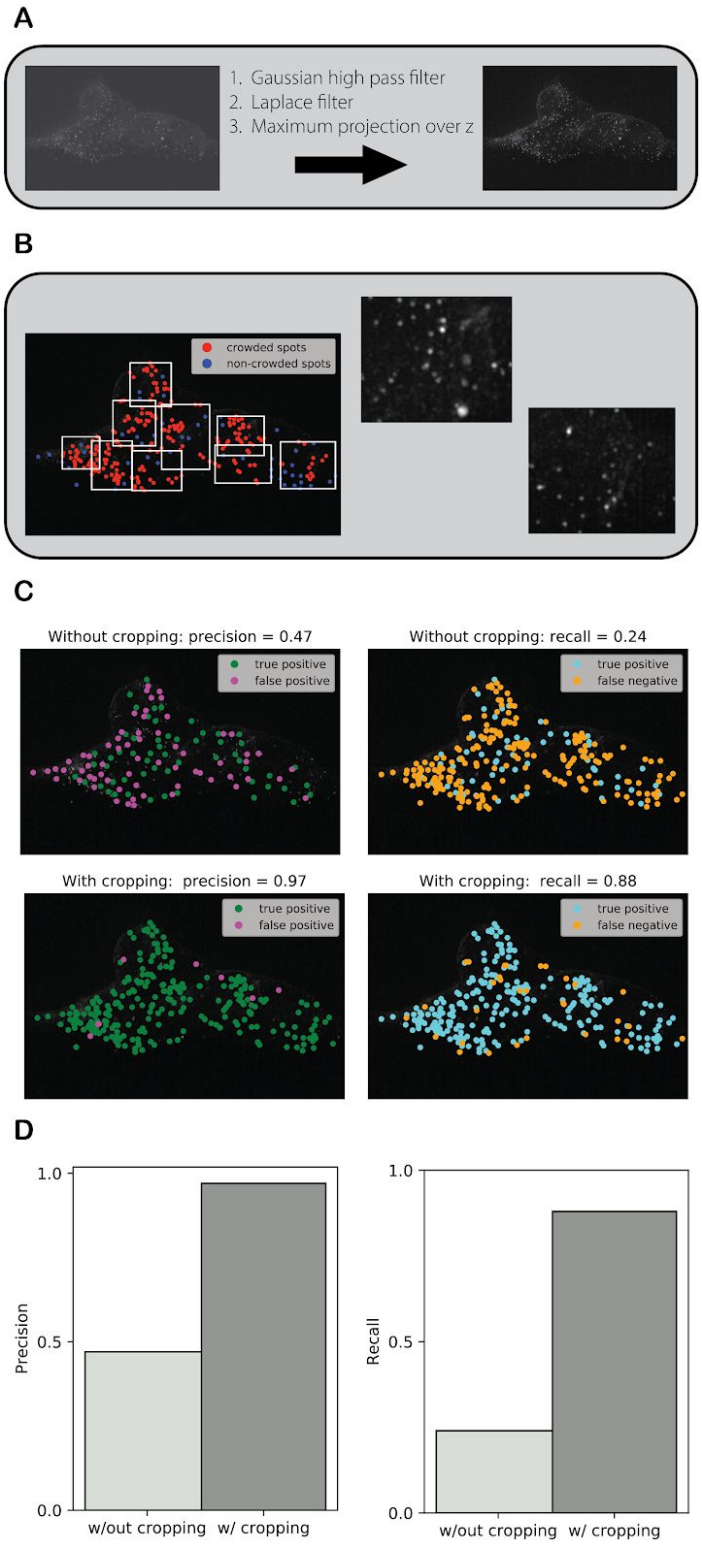

Figure 3: Images are filtered and autocropped so that they are easier for workers to annotate. ***(A)*** *Raw images are pre-processed with a gaussian high pass filter, a Laplace filter, and a maximum intensity projection over z.* ***(B)*** *Crowded spots detected and bound. Rough, first-pass*

spot-calling enables the detection of crowded spots and subsequent autocropping. ***(C)*** *True positive = consensus in concordance with expert annotation, false positive = consensus not in concordance with expert annotation, and false negative = no consensus found for an expert annotation. The distance between a correct consensus annotation and the nearest expert annotation is no more than the user-defined correctness threshold. The distance between a detected expert annotation and the nearest consensus annotation is also no more than the user-defined correctness threshold.* ***(D)*** *Applying cropping resulted in precision and recall of 97% and 87%, improvements of 50% and 64%, respectively, over the un-cropped image.*

There are limits on how much an image can be cropped to increase recall. Somes spots are so close together that the spots themselves appear to clump in the parent image, so no amount of zooming in would enable them to be separated. The user should also keep in mind that in the Quanti.us website cropped images are upscaled by interpolation to a size which workers can annotate, so the pixels in the upscaled image which the workers are annotating are not a perfect representation of the original data. The user should also consider the relationship between the number of crops and the overall cost of the experiment. Annotations from Quanti.us cost only five cents per worker per image (in 2020), but many workers may be needed to ensure good coverage and proper clustering and de-clumping of clusters.

Helper images

Because images from experiments using RCA (Rolling Circle Amplification, an *in situ* transcriptomics chemistry which is described in Supplementary Text 5) have a lot of variation in spot sizes, we sought to investigate whether we could improve the quality of the worker annotations by providing better guidance at the crowdsourcing interface, and whether including the parent image with the crops removed in the stack (e.g. second image from left, Supplementary Figure 8B) affects worker accuracy. The parent image appears to the workers on

the Quanti.us interface at the same size as the crops, so the spots in the parent image appear much smaller to the workers than the spots in the crops. Additionally, we designed four different helper images – two variants, with or without circles drawn around the correct spots – which illustrate what the workers should click (Supplementary Figure 10).

On average, the inclusion of helper images increased precision by 14% (95% with helper images and 81% without) and decreased recall by 16% (59% with helper images and 76% without) (Supplementary Figure 11). When the spots in the helper image were circled, precision was 0.4% higher and recall was 3.4% higher. Workers expressed little preference between the two variants of helper images. On average, precision and recall with images of the first variant were only 0.5% greater and 4% less than precision and recall of the second variant, respectively. However, including the parent image in the stack, which workers would also annotate, decreased both precision and recall by 6% and 4% respectively.

These results suggest that the workers were less likely to click spots they felt unsure about when helper images were provided. These results also demonstrate that for some images it is disadvantageous to show the workers the parent image, as it confuses them because of the drastic difference in scale between the parent image and the crops.

These image preparation steps complete the pipeline described in Section II (Figure 1). The performance of the pipeline will now be demonstrated with a vignette, using the rolling circle amplification (RCA) chemistry.

The results of validating spot calling algorithms using crowdsourced ground truth are comparable to the results using expert annotations

We tested the usage of crowdsourced annotations resulting from the pipeline as ground truth to validate spot calling algorithms. Using available images from RCA experiments, we sought to evaluate how well crowdsourced annotations agree with expert annotations to assess the generalizability of the tuned spot-calling parameters.

To produce the crowdsourced annotations, the inputs to the annotation generation pipeline were: One sample image with the RCA chemistry, expert spot location annotations for that image, and three test images without annotations. All images were downloaded from an *in situ* sequencing (ISS) experiment in the starfish database.(34) The spots which an expert had annotated in the sample image were analyzed to extract spot detection parameters intaken by starfish's BlobDetector method, which implements the Laplacian of Gaussian spot detection approach. These parameters were used for first-pass blob detection on the test images, and the resulting blob coordinates were used to autocrop the test images. All crops were then sent to Quanti.us to be annotated by 25 workers each. Further details are explained in Supplementary Text 6.

We sought to evaluate how well the resulting crowdsourced annotations agreed with expert annotations. Precision and recall for the consensus annotations, based on an expert's evaluations of the test images, were 95% and 70%, 92% and 89%, and 81% and 76% for images ISS_rnd0_ch1_z0, ISS_rnd0_ch3_z0, and ISS_rnd1_ch1_z0 respectively (Figure 4). The Jaccard similarity indices (intersection over union) between the consensus annotations and the expert annotations were 71%, 85%, and 68%, respectively. We then sought to understand how well the consensus annotations and the expert annotations agreed with each other by evaluating the level of concurrence between different experts and between the same expert annotating the same

image twice six months apart (at times t_0 and $t_1 = t_0 + 6$ mos.)(Supplementary Text 7). The Jaccard indices (intersection over union) for Expert #1 at t_0 and Expert #1 at t_1, Expert #1 at t_0 and Expert #2 at t_1, and Expert #1 at t_1 and Expert #2 at t_1 were 73%, 78%, and 82% respectively (Supplementary Figure 12). Thus, the agreement we saw between the consensus and expert annotations for the RCA images are in the same range as the intra- and inter-expert Jaccard indices.

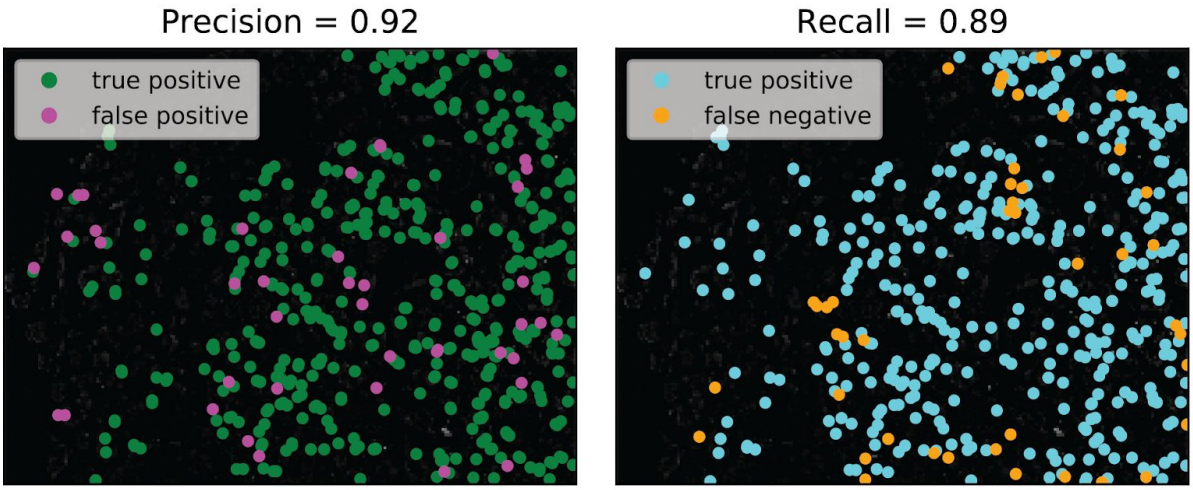

Figure 4: The* in situ *transcriptomics annotation pipeline was tested using RCA (Rolling Circle Amplification) images from an* in situ *sequencing (ISS) experiment in the starfish database. *Worker consensus annotations for RCA test image ISS_rnd0_ch3_z0 achieved 92% precision and 89% recall based on expert consensus annotations. The Jaccard similarity index (intersection over union) between the consensus annotations and the expert annotations was 0.85.*

We also sought to assess the generalizability of the tuned spot-calling parameters and the potential utility of using crowd-sourced annotations to generate ground truth for a large number of images. To do this, we ran starfish's BlobDetector method using the spot parameters which had been extracted in this vignette on thirteen other images from starfish's RCA dataset which

had not been annotated by experts. By visual inspection, these images had a greater variation in spot size, brightness, and quantity than the other images. As ground truth, we used consensus annotations for these images. The precision of the BlobDetector method was 82 ± 10%, and the recall was 68 ± 17% (mean ± standard deviation, Supplementary Figure 14). These results suggest that when a set of spot parameters tuned to a particular channel and field of view for a chemistry are used for other channels and fields of view for the same chemistry, the spots detected are likely to be correct but fewer spots may be detected.

The finding that the Jaccard indices for the consensus and expert annotations for the RCA images were similar to the intra- and inter-expert Jaccard indices strongly suggests that consensus annotations can be used in place of expert annotations. We also found that spot parameters found for a given chemistry using consensus annotations as ground truth can be used to automatically find spots with precision (82 ± 10%, mean ± standard deviation) comparable to the agreement between two experts annotating the same image (82%). Together, these findings are especially useful for the processing of large datasets that would be infeasible for an expert to annotate by hand.

Crowdsourced ground truth is useful for tuning and validating spot calling parameters

This section tests the usage of crowdsourced annotations resulting from the pipeline as ground truth to tune and validate their parameters. Ground truth is essential for testing how well a spot-calling algorithm generalizes to other *in situ* transcriptomics chemistries and tuning parameters (Supplementary Text 8). We tested whether consensus and expert annotations

function similarly well as ground truth to tune parameters for spot-calling algorithms, using the RCA dataset from starfish.(31) We also bootstrapped our consensus and expert annotations for the RCA dataset to explore the minimum number of ground truth annotations needed to effectively tune a spot-calling algorithm.

Usage of consensus annotations and expert annotations as ground truth to tune the LMP (starfish's LocalMaxPeakfinder) stringency parameter, which controls the intensity cutoff to detect a peak, resulted in similar precision vs. stringency trends, as well as recall vs. stringency trends (Figure 5). The optimal stringency parameter found using consensus annotations as ground truth resulted in a lower precision and slightly higher recall (89.4% and 95.4% respectively, compared with 94.3% and 94.8% from using expert annotations as ground truth – 5.3% and 0.63% difference respectively). These results reflect the fact that some of the spots annotated by the worker consensus were not annotated by the more conservative expert.

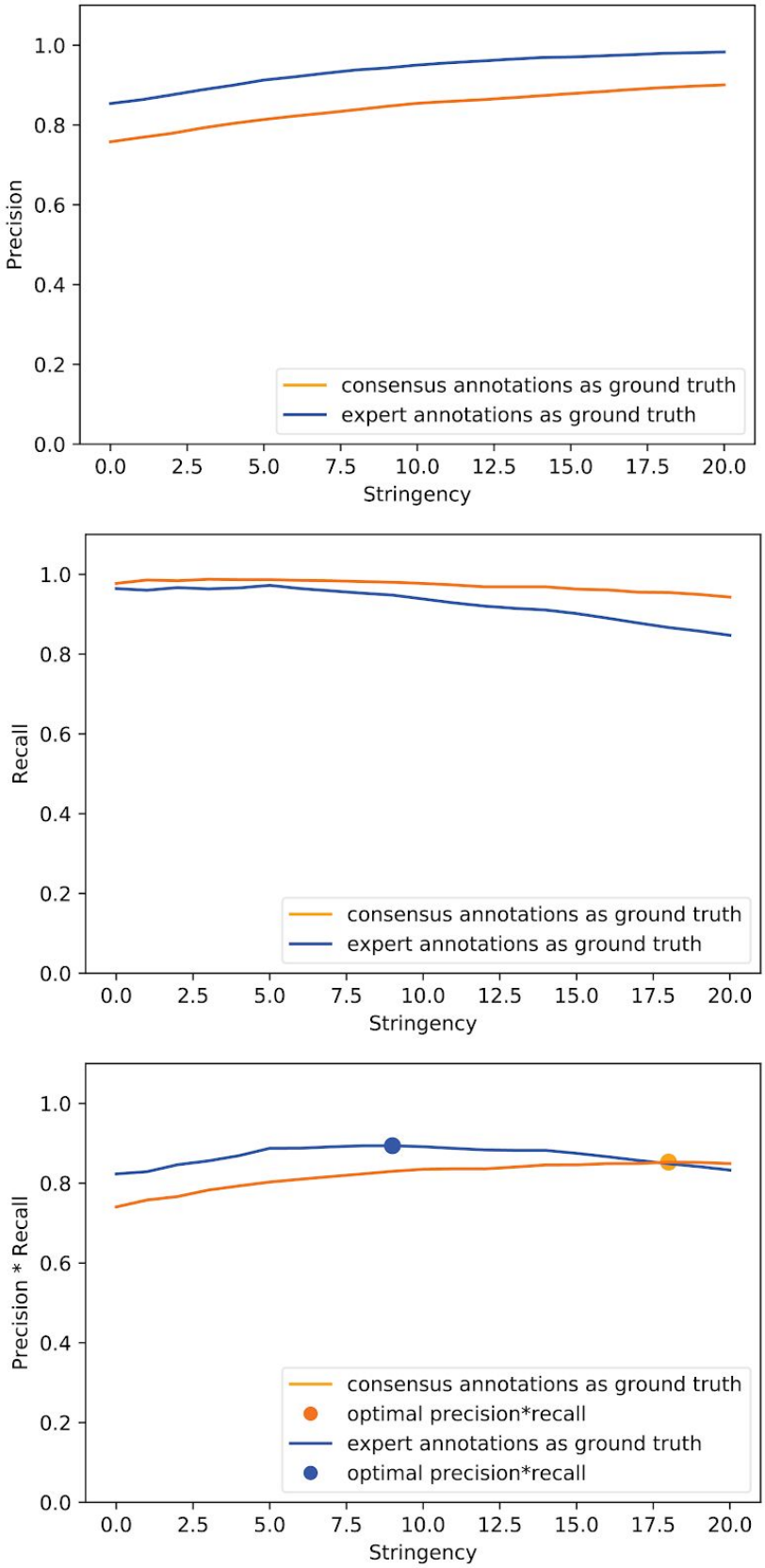
Precision
Stringency
consensus annotations as ground truth
expert annotations as ground truth
Recall
Stringency
consensus annotations as ground truth
expert annotations as ground truth
Precision * Recall
Stringency
consensus annotations as ground truth
optimal precision*recall
expert annotations as ground truth
optimal precision*recall

Figure 5. The optimal "stringency" parameter for starfish's LocalMaxPeakfinder (LMP) with Rolling Circle Amplification (RCA) images from the starfish database resulted in lower precision and slightly higher recall (89.4% and 95.4% respectively) when consensus annotations were used as ground truth for parameter tuning, compared with 94.3% and 94.8% precision and recall which were achieved when expert annotations were used as ground truth for parameter tuning.

Supplementary Text 9 shows that the training behavior of the expert and crowdsourced annotations is very similar. That is, performance converges to roughly the same level and at roughly the same rate. Fifteen ground truth annotations were enough to get 99.1% and 98.1% of the maximum precision performance when the annotations were produced by experts and worker consensus, respectively. With the same number of annotations, 97.6% and 96.6% of the maximum recall performance was achieved with annotations produced by experts and worker consensus, respectively. Together with the result that consensus annotations and expert annotations resulted in similar precision vs. stringency and recall vs. stringency trends when used as ground truth to tune the LMP stringency parameter (Figure 5), this suggests that consensus annotations are a viable substitute for expert-generated ground truth for both parameter tuning and algorithm validation.

Discussion

We developed INSTA, a pipeline to prepare *in situ* transcriptomics images for crowdsourced annotation and integrated these pipeline components into an open-source toolkit which takes an *in situ* image dataset or a spot detector as input, prepares the images for crowdsourcing annotation, receives the annotations, and outputs consensus locations for the spots and/or optimized parameters for the detector. The pipeline was designed to be flexible

enough to easily include components of the user's choice (e.g. custom crowdsourced annotation ingestion classes) and to accommodate manual user intervention at different points in the pipeline. We also created a tool (SpotImage) to generate simulated *in situ* images with adjustable parameters, such as spot density and SNR. Using simulated and real *in situ* transcriptomics images, we developed QC rules for crowdsourced annotations based on observable aspects of the annotation data, such as cluster characteristics, and used these rules to develop QC methods to optimize consensus precision and recall. We also gained insight into critical aspects of how the quantity, size, SNR, and crowdedness of spots in images all influence worker behavior and thus annotation quality.

We demonstrated the pipeline using RCA images, resulting in consensus annotations with high precision and recall compared to expert annotations. We also demonstrated that consensus and expert annotations are equally suitable as ground truth. While consensus annotations are useful when large amounts of ground truth are needed to check or validate the performance of spot-calling algorithms, a few dozen expert annotations alone may be sufficient to tune a spot-calling algorithm such as Starfish's BlobDetector, even if that would not be enough to properly validate the algorithm (obtain statistically significant measures of precision and recall). Crowdsourced ground truth is vital for validating algorithm performance on a large dataset with many images, or even just one image with thousands of spots.

The pipeline only requires user intervention for the reviewing and editing of autocrops. We believe it is particularly important for the user to have the option to intervene at this stage rather than leave the pipeline as an end-to-end black box. While it would be theoretically ideal to automatically learn cropping parameters which maximize accuracy, minimize the number of

crops used (and therefore cost incurred), and generalize perfectly to all images of a given chemistry, spot distribution characteristics vary too much between the individual *in situ* transcriptomics images of most chemistries for this to be reasonable. The researcher needs the opportunity to balance the tradeoff between crop detail and crowdsourcing cost. If a cheap experiment yields a very large dataset with many images, a user may be less concerned with maximizing data extracted from each image, but if each image costs more to produce, the researcher might wish to be more detailed with cropping (Supplementary Text 10).

Even without budgetary constraints, auto-cropping is only useful to a certain extent because zooming and scaling crops up to the size where they can be displayed for annotation inherently involves interpolation; at a certain point the workers may be annotating a crop that is not faithful to the original image. The toolkit user should be able to intervene before this point.

INSTA can be used to annotate publicly-available image datasets, especially for researchers who do not have access to wet labs. The sample data available through starfish would be a good starting point. We will update pipeline usage and guidelines if we experiment with processing images of other, more challenging chemistries, and strive to make the pipeline usage as generalizable to other chemistries as possible.

Acknowledgements

We would like to thank all members of the Chan Zuckerberg Biohub and especially the Bioengineering Platform for their useful discussions and feedback. We also thank Quanti.us for their support, the Chan Zuckerberg Initiative starfish team for helpful discussions, and all the crowd workers for their diligent contributions to the development and testing of the INSTA pipeline.

References

1. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018 Jul 27;361(6400):eaat5691.
2. Yuste R. Fluorescence microscopy today. Nat Methods. 2005 Dec;2(12):902–4.
3. Agard DA, Hiraoka Y, Shaw P, Sedat JW. Fluorescence microscopy in three dimensions. Methods Cell Biol. 1989;30:353–77.
4. Tsien RY. The green fluorescent protein. Annu Rev Biochem. 1998;67:509–44.
5. Langer-Safer PR, Levine M, Ward DC. Immunological method for mapping genes on Drosophila polytene chromosomes. Proc Natl Acad Sci U S A. 1982 Jul;79(14):4381–5.
6. Boyle S, Rodesch MJ, Halvensleben HA, Jeddeloh JA, Bickmore WA. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 2011 Oct;19(7):901–9.
7. Bienko M, Crosetto N, Teytelman L, Klemm S, Itzkovitz S, van Oudenaarden A. A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. Nat Methods. 2013 Feb;10(2):122–4.
8. Beliveau BJ, Joyce EF, Apostolopoulos N, Yilmaz F, Fonseka CY, McCole RB, et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc Natl Acad Sci U S A. 2012 Dec 26;109(52):21301–6.
9. Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of Single RNA Transcripts in Situ. Science. 1998 Apr 24;280(5363):585–90.
10. Crosetto N, Bienko M, van Oudenaarden A. Spatially resolved transcriptomics and beyond. Nat Rev Genet. 2015 Jan;16(1):57–66.
11. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011 Apr 7;472(7341):90–4.
12. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science. 2012 Dec 21;338(6114):1622–6.
13. de Bourcy CFA, De Vlaminck I, Kanbar JN, Wang J, Gawad C, Quake SR. A quantitative comparison of single-cell whole genome amplification methods. PloS One. 2014;9(8):e105585.
14. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011 Jul;21(7):1160–7.
15. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013 Nov;10(11):1096–8.
16. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012 Sep 27;2(3):666–73.
17. Mabaso MA, Withey DJ, Twala B. Spot detection methods in fluorescence microscopy imaging: A review. 2018 [cited 2019 Sep 9]; Available from: https://researchspace.csir.co.za/dspace/handle/10204/10606
18. Blob Detection — skimage v0.16.dev0 docs [Internet]. [cited 2019 Sep 9]. Available from: https://scikit-image.org/docs/dev/auto_examples/features_detection/plot_blob.html
19. ISS Processing Workflow — starfish documentation [Internet]. [cited 2019 Sep 9]. Available from: https://spacetx-starfish.readthedocs.io/en/stable/usage/data_processing_examples/iss_pipeline.html
20. Codeluppi S, Borm LE, Zeisel A, Manno GL, Lunteren JA van, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat Methods. 2018

Nov;15(11):932–5.
21. Lindeberg T. Feature Detection with Automatic Scale Selection. Int J Comput Vis. 1998 Nov 1;30(2):79–116.
22. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008 Oct;5(10):877–9.
23. Chen J, Ding L, Viana MP, Hendershott MC, Yang R, Mueller IA, et al. The Allen Cell Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. bioRxiv. 2018 Dec 8;491035.
24. Sullivan DP, Winsnes CF, Åkesson L, Hjelmare M, Wiking M, Schutten R, et al. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. Nat Biotechnol. 2018 Sep;36(9):820–8.
25. Horowitz S, Koepnick B, Martin R, Tymieniecki A, Winburn AA, Cooper S, et al. Determining crystal structures through crowdsourcing and coursework. Nat Commun. 2016 16;7:12549.
26. Koepnick B, Flatten J, Husain T, Ford A, Silva D-A, Bick MJ, et al. De novo protein design by citizen scientists. Nature. 2019 Jun;570(7761):390–4.
27. Zhou N, Siegel ZD, Zarecor S, Lee N, Campbell DA, Andorf CM, et al. Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning. PLoS Comput Biol. 2018;14(7):e1006337.
28. Hughes AJ, Mornin JD, Biswas SK, Beck LE, Bauer DP, Raj A, et al. Quanti.us: a tool for rapid, flexible, crowd-based annotation of images. Nat Methods. 2018 Aug;15(8):587.
29. Mitry D, Zutis K, Dhillon B, Peto T, Hayat S, Khaw K-T, et al. The Accuracy and Reliability of Crowdsource Annotations of Digital Retinal Images. Transl Vis Sci Technol. 2016 Sep 1;5(5):6–6.
30. spacetx/starfish [Internet]. GitHub. [cited 2019 Sep 9]. Available from: https://github.com/spacetx/starfish
31. starfish: scalable pipelines for image-based transcriptomics — starfish documentation [Internet]. [cited 2019 Sep 9]. Available from: https://spacetx-starfish.readthedocs.io/en/latest/
32. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007 Feb 16;315(5814):972–6.
33. Dueck D, Frey BJ. Non-metric affinity propagation for unsupervised image categorization. In: 2007 IEEE 11th International Conference on Computer Vision. 2007. p. 1–8.
34. ISS notebook [Internet]. spacetx; 2020 [cited 2020 Jun 26]. Available from: https://github.com/spacetx/starfish/blob/master/notebooks/ISS.ipynb