**Screening Algorithm for FISH Annotations**

Notes 26 July: At each step where some sort of threshold value is required (blue font), I need to find automatically generated thresholds and justify these choices with evidence using the analysis methods I have been working with and new analysis methods I am writing using ideas we discussed during our meeting yesterday. Links to notebooks describing these (in the endnotes), are available for Steps 1 and 2; I am working on the analyses for Steps 3-5. Once I have written the code for the algorithm, I will test its performance on the data we have been using (the batches of spot images and the earlier Cy3.json).

To keep the algorithm modular, I limit one new threshold/pruning for each step so that it might be easier to work on each threshold/pruning separately.

**Algorithm**

**Input: raw annotations.**

1. Cluster only workers with lowa (= good) pairwise scores.

Assumes that workers with low pairwise scores have high specificity.[[1]](#footnote-1)

1. Look at clusters from Step 1. Sort clusters with few/manyb workers annotating (“putatively incorrect/correct”).

Assumes that clusters with a lot of workers annotating are correct.

1. Look at all[[2]](#footnote-2) workers. Sort workers who are in few/manyc "putatively correct" clusters.

Assumes workers who are in many putatively correct clusters “go with the good crowd.”[[3]](#footnote-3)

1. Keep “putatively incorrect” clusters which are mostlyd comprised of workers who are in many putatively correct clusters.

Assumes that “putatively incorrect” clusters with a large fraction of workers who go with the good crowd are probably correct.

1. Keep“putatively correct” clusters which are mostlyecomprised of workers who are in many putatively correct clusters.

Assumes that “putatively correct” clusters with a large fraction of workers who don’t click the spots everyone does are probably incorrect.

**Return: list of centroid coordinates for kept clusters.**

**Performance metrics**: (1) Fraction of spots detected. (2) Fraction of detections which are correct.

Endnotes:

1. Otsu’s seems to give the most aggressive thresholding for the positively skewed distribution of worker pairwise scores, compared with k-means (and compared with mean+1.5\*stdev, just out of curiosity). See [notebook](https://github.com/czbiohub/FISH-annotation/blob/BaseAnnotation/Annotation/batch_20180719_pairwise_scores_thresholding.ipynb).
2. K-means seems to give the most aggressive thresholding for the bimodal distribution of cluster membership, compared with Otsu’s. See notebooks: [SNR separately](https://github.com/czbiohub/FISH-annotation/blob/BaseAnnotation/Annotation/batch_20180719_cluster_sizes_and_snr.ipynb), [SNR overlaid](https://github.com/czbiohub/FISH-annotation/blob/BaseAnnotation/Annotation/batch_20180719_cluster_sizes_and_snr_together.ipynb).
3. K-means seems to give the best thresholding between the good group of workers with who find a lot of putatively correct clusters and the other outliers who find few putatively correct clusters. See [notebook](https://github.com/czbiohub/FISH-annotation/blob/BaseAnnotation/Annotation/batch_20180719_workers_pc_yield.ipynb).
4. Visually, okay separation between correct and incorrect clusters, sometimes. Perhaps separation would be improved if we found a better threshold to sort the Good Crowd of workers (workers who participate in a lot of big clusters)? Need to find a better threshold. K-means performs a lot better than Otsu’s but still keeps too many clusters with low fractions of workers from the good crowd. See notebook for histograms: number of “putatively incorrect” clusters vs. fraction of the cluster’s annotations which are from workers who are in many “putatively correct” clusters. The list should contain, for each “putatively incorrect” cluster, the fraction of the cluster’s annotations which are from workers who are in many “putatively correct” clusters. See [notebook](https://github.com/czbiohub/FISH-annotation/blob/BaseAnnotation/Annotation/batch_20180719_small_clusters_good_crowd.ipynb).
5. Need histograms: number of “putatively correct” clusters vs. fraction of the cluster’s annotations which are from workers who are in few “putatively correct” clusters.

1. Since when they detect something, usually a lot of other people detect it, too. [↑](#footnote-ref-1)
2. Looking at all workers because we want to take into account the visionaries (having bad pairwise scores, but in a lot of “putatively correct clusters”). [↑](#footnote-ref-2)
3. “Go with the good crowd” = click the spots everyone else does. [↑](#footnote-ref-3)