4EU+ Quantitative Microscopy Project

Name:	
Subject:	MB100P08 4EU+ Quantitative Microscopy
Semester:	Spring 2023/2024

The source code and other information is available here.

Definition of the Problem

Counting cells in a microscopic image may sound like a basic problem, but it is not as easy as it sounds. For a human, it is a very time-consuming task, and it is very easy to get lost in the counting, especially if there are many cells in the image.

For image analysis software, the problems remain but are obviously of a different kind. Counting cells in an image containing cells of various sizes and shapes is a challenging task, not made easier by the fact that it is not unusual for the cells to touch each other or even overlap when dealing with 3-dimensional images. All in all, segmentation and boundary identification of the cells are very difficult to do correctly.

As counting cells is the basis of almost all image analysis, the goal of this project is to compare two different segmentation workflows: Trainable Weka Segmentation (Arganda-Carreras et al., 2017) in Fiji (Schindelin et al., 2012) and Pixel Classification in Ilastik (Berg et al., 2019) with each other and also with manual cell counts as observed by two people. Both of these workflows are based on the gradual training of a classifier on selected samples, with the trained classifier then applied to the rest of the samples.

Data Background

The dataset used to compare the two workflows contained images of *Drosophila melanogaster* Kc167 cell line (Echalier & Ohanessian, 1969). Five samples were included, all stained with a specific DNA stain. Four samples, labelled *48*, *340*, *Anillin*, and *mad2*, had a particular gene knocked down. The last sample, labelled *nodsRNA*, contained wild-type cells (Carpenter et al., 2006).

Each of the five samples was imaged ten times, so the dataset contained 50 images with one DNA channel. The images were obtained with Zeiss Axiovert 200M microscope and provided as TIF files [1].

The ground truth, an essential part of the dataset, was provided in a tab-delimited text file containing the number of cells in each image, as counted by two different human observers [1]

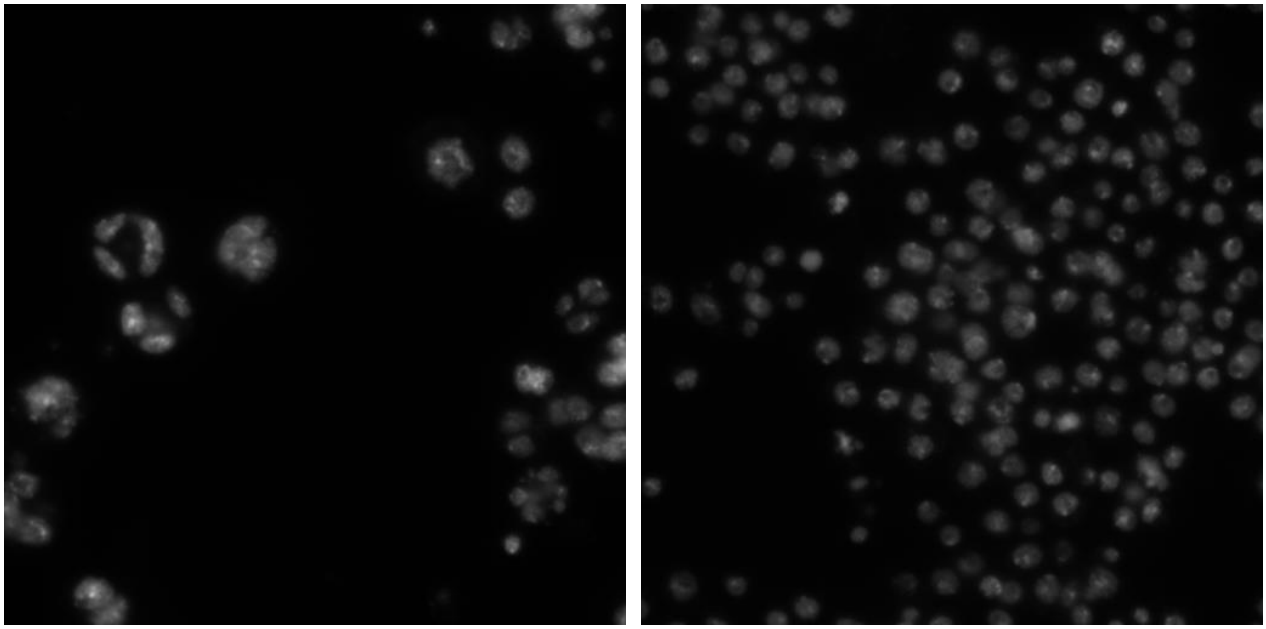

Figure 1: Examples of images from the dataset containing *Drosophila melanogaster* cells [1]

Process & Methods

As mentioned earlier, this project aimed to compare two different segmentation workflows. For both workflows, I trained five classifiers, one for each sample. The classifiers were always trained on one image of the sample and applied to the remaining images of the sample without any modification.

Ilastik

In Ilastik (Berg et al., 2019), I tried to use the Interactive Density Counting workflow, which estimates the density of objects in the image directly without performing segmentation. This approach allows counting objects even when they overlap [2]. However, this workflow is only appropriate when objects in the image have similar size, intensity, and shape, so it was unsuitable for the *Drosophila melanogaster* dataset, as cells appearing in the dataset's images did not meet this condition.

As my first choice of Ilastik workflow was inappropriate, I tried using Pixel Classification as the next option. This workflow assigns labels to pixels based on user annotations and pixel features. It uses a Random Forest Classifier trained on user annotations to provide a probability map of each class, and the training itself is based on direct interaction with the user [3].

To be able to load the images in TIF format into Ilastik, I saved the images with a greyscale look-up table using Fiji (Schindelin et al., 2012), as importing images with the default colour palette was impossible [4].

Then, I started the Pixel Classification workflow and uploaded one TIF-formatted image of a sample as input data. As I was unsure about the most helpful features, I decided to select all of them because this is the recommended approach when computation time is not a concern [3].

My goal was to separate the images into cell and background classes, so I used two labels, with Label 1 representing the background and Label 2 representing cells. After defining the labels, I trained the

classifier iteratively by choosing the label and drawing over pixels belonging to that label. The Live Update function allowed me to see the classifier's prediction based on its current training level.

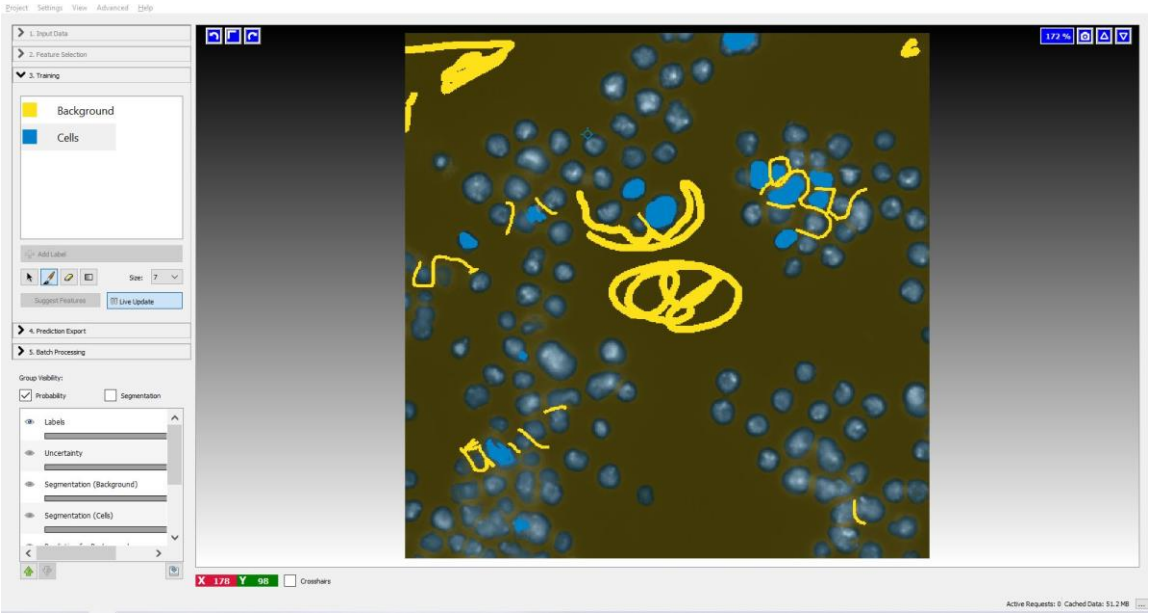

Figure 2: Example of Classifier Training in Ilastik, with yellow pixels representing the background and blue pixels representing cells

Rich colours represent the user's annotations, while muted colours are the classifier's predictions based on the user's annotations.

After training the classifier to a satisfactory degree, I imported the remaining images of the sample, applied the trained classifier to them without further modification, and exported the probability maps.

Subsequently, as the segmentation was finished, it was time to count the cells in Fiji. To be able to import images exported from Ilastik to Fiji, I first had to install the Ilastik Fiji plugin [5]. In Fiji, I batch-processed all of the probability map files from Ilastik with a macro that can be found here.

This macro smoothed and applied filters to the probability maps in order to remove noise and any remaining errors. It then counted the cells in all the images and showed the counts in the Summary window.

This whole process was repeated five times, once for every sample. One image was used to train the classifier, which was then applied to the nine remaining images. A more technical description of the process can be found here.

Fiji

The workflow in Fiji (Schindelin et al., 2012) had some similarities with the Ilastik one, but the details and names of the segmentation steps differed. Here, I used the Trainable Weka Segmentation plugin (Arganda-Carreras et al., 2017), which is a default part of Fiji.

As a first step, I opened one image of a sample, which I then used for training the classifier. I started the Trainable Weka Segmentation plugin and adjusted the settings. Then, I began training the classifier by picking parts of the image with the Freehand Line tool and clicking either on Add to Background or Add to Cells, depending on the part of the image chosen.

After adding at least one entry to both classes, clicking on the Train the Classifier button allowed me to see the classifier's predictions for the first time. I added information to both classes and updated the classifier until I was satisfied with the prediction result.

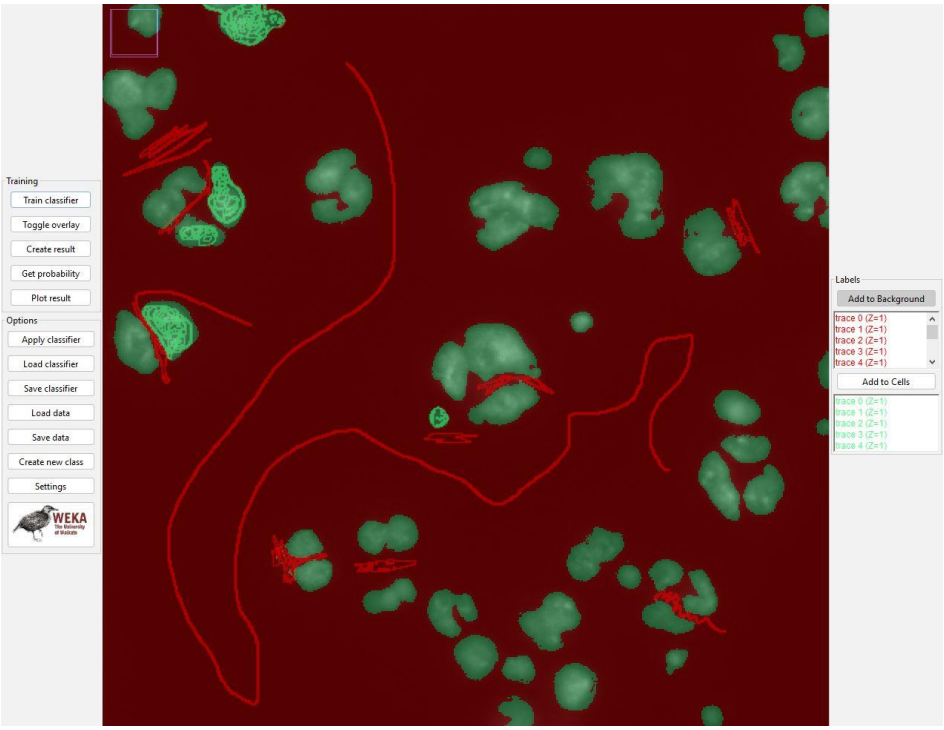

Figure 3: Example of Classifier Training in Fiji, with red parts representing the background and green parts representing cells

Rich colours represent the user's annotations, while muted colours are the classifier's predictions based on the user's annotations.

After that, I clicked on Apply Classifier and chose all images of the sample, including the one used for the training. At this point, it was time to count the cells. I batch-processed all of the segmentation images with a macro that can be found here. This macro smoothed and applied filters to the images to reduce the noise and remaining errors, counted the cells in all images in the folder and showed the counts in the Summary table.

This whole workflow was repeated five times, once for each sample. Each time, one image was used to train the classifier, which was then applied to the remaining nine images. A more technical description of the process can be found here.

Results

After acquiring the results from both the Ilastik and Fiji workflows, I visualised the results and ran several statistical tests in the R programming language (R Core Team, 2024).

Before starting the tests, I visualised the cell counts in each image according to two human counters and two segmentation workflows, as shown in *Figure 4*. From this simple visualisation, I could see that Human 1 usually counted fewer cells in each image, while the other three measurements obtained by Human 2, Ilastik and Fiji were, in most cases, more similar.

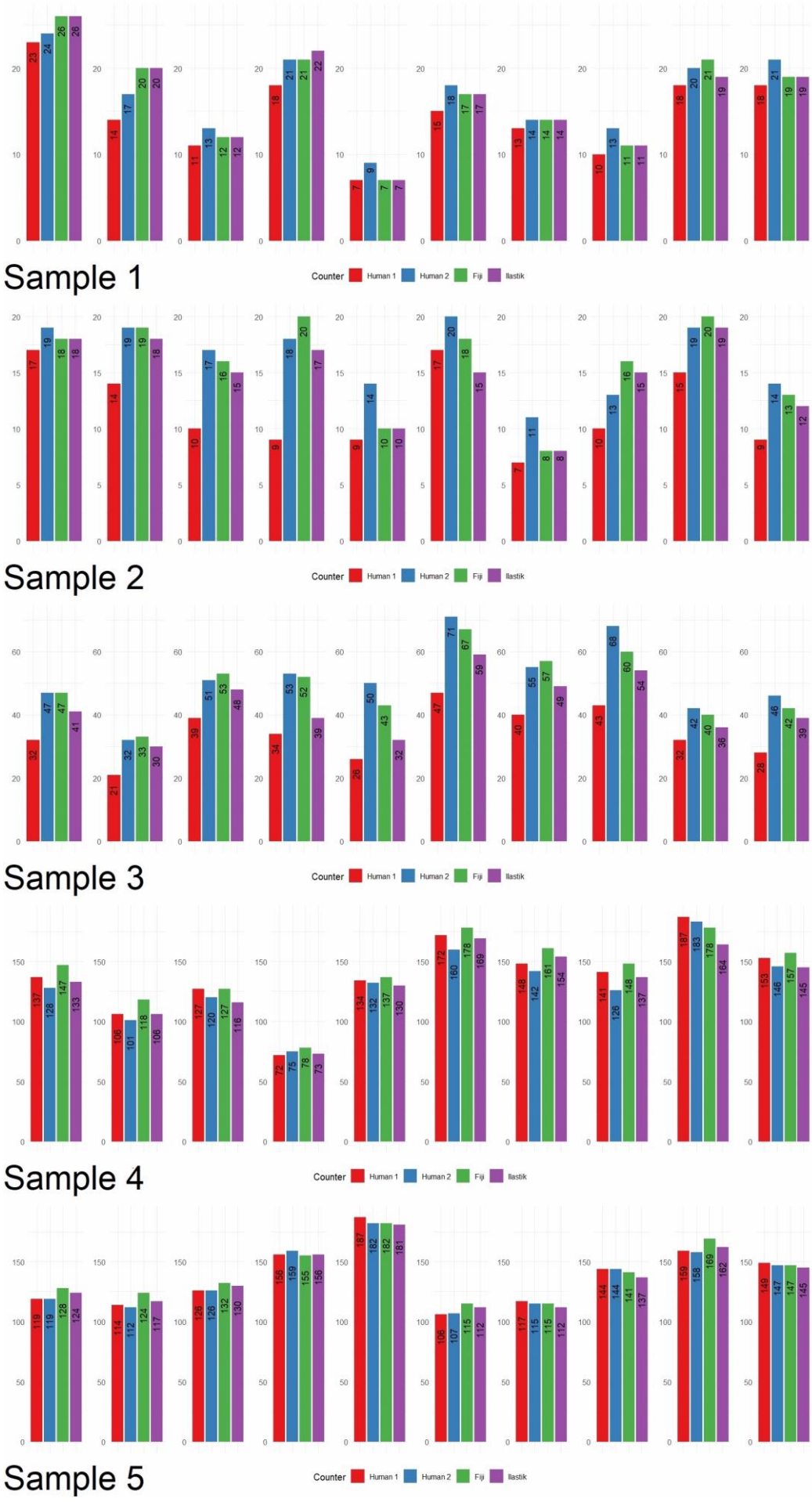

Figure 4: Plots for the five samples showing the cell count in each image according to two human counters and two segmentation workflows

To confirm this hypothesis and gain other information, I run several statistical tests in the R programming language (R Core Team, 2024) with the help of the following packages: *gridExtra* (Auguie, 2017), *patchwork* (Pedersen, 2024), *tidyverse* (Wickham et al., 2019), *reshape2* (Wickham, 2007) and *readxl* (Wickham & Bryan, 2023).

Firstly, as I was interested in the differences between human observers and segmentation workflows, I tested the normality of the differences between all pairs. I used the Shapiro-Wilk test (Shapiro & Wilk, 1965) for testing the data's normality, with all but one result showing that the differences between the data pairs were not normally distributed.

To better understand these results, I also visualised histograms and Q-Q plots of differences between all data pairs, as shown in *Figure 5*. Both confirmed that most of the differences between data pairs do not have a normal distribution.

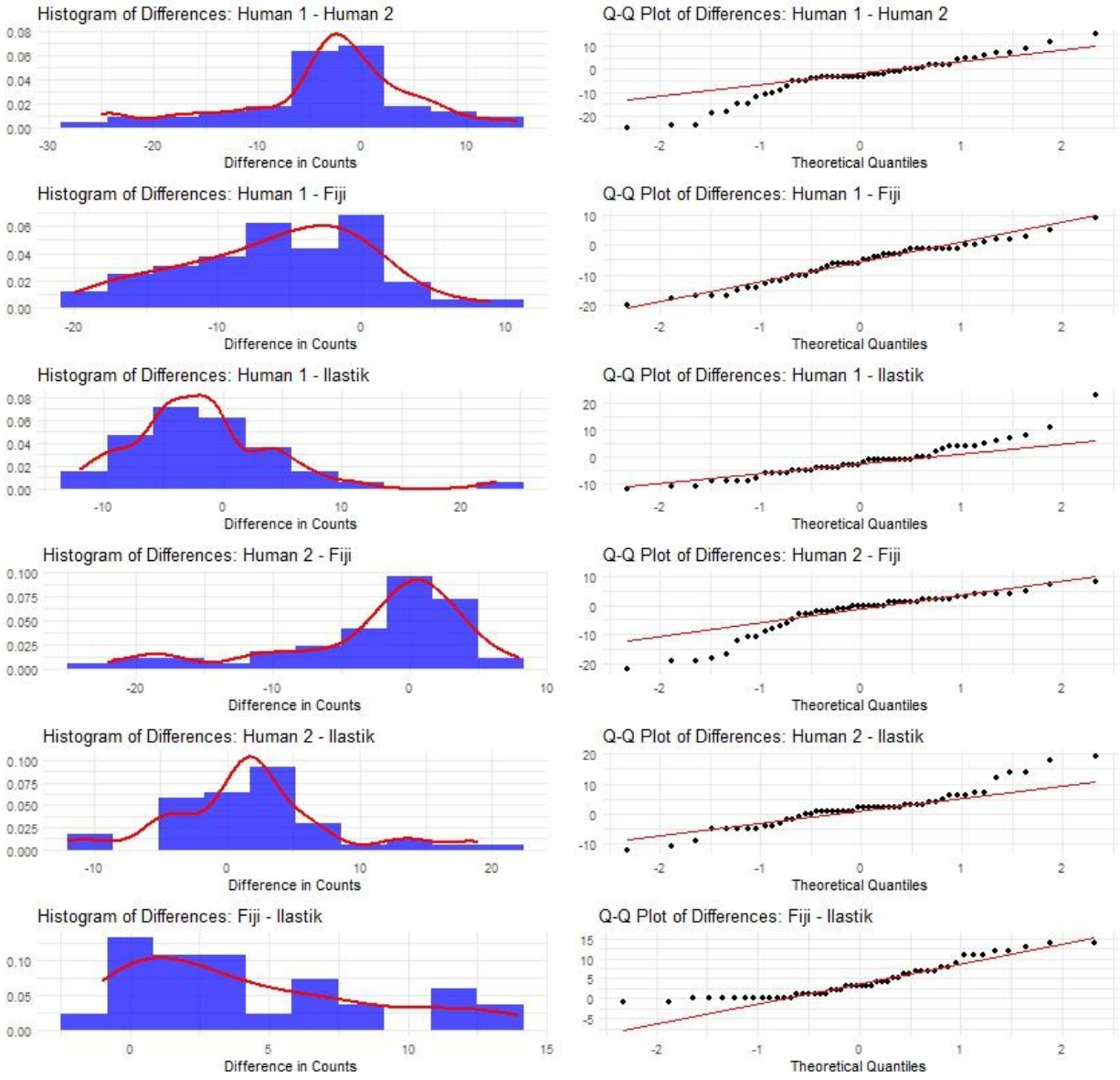

Figure 5: Visualisations of differences in cell counts between all data pairs

The left panel in each row shows a histogram of the differences with blue bars and overlaid red density curves. The x-axis represents the difference in cell counts, and the y-axis indicates density.

The right panel in each row displays a Q-Q plot, with the points representing observed quantiles against theoretical quantiles. The red line depicts the theoretical quantiles of a normal distribution.

Table 1: Results of the Shapiro-Wilk test for the differences between all data pairs

The p-values of all but one data pair are less than the significance level of 0.05; therefore, the differences between all but one data pair are not normally distributed.

Pairs	p-values
Human 1 – Human 2	0.00915
Human 1 – Fiji	0.330
Human 1 – Ilastik	0.00145
Human 2 – Fiji	0.0000325
Human 2 – Ilastik	0.00730
Fiji – Ilastik	0.000145

Because of that, all further analyses were conducted using non-parametric tests, which do not make assumptions about the data distribution. To find out if the population means between different measurements of each sample differ, I used the paired Wilcoxon signed-rank test (Wilcoxon, 1945).

Here, the differences between the test results for different pairs of measurements were significant. There were two pairs of measurements where the test results showed that their population means were not significantly different, with these pairs being Human 2 – Fiji and Human 2 – Ilastik. The most significant difference was found between Human 1 – Fiji and Fiji – Ilastik pairs, showing that the measurements between these pairs differed the most. The population means between pairs Human 1 – Human 2 and Human 1 – Ilastik were significantly different, but not to the level of the previously mentioned two pairs.

Table 2: Results of the Wilcoxon signed-rank test for the differences between all data pairs

The p-values of pairs Human 2 – Fiji and Human 2 – Ilastik are greater than the significance level of 0.05, showing that the difference between population means between these pairs is statistically insignificant. The other four pairs have p-values less than the significance level; therefore, the difference between population means is statistically significant.

Pairs	p-values
Human 1 – Human 2	0.0103
Human 1 – Fiji	0.000000424
Human 1 – Ilastik	0.0149
Human 2 – Fiji	0.102
Human 2 – Ilastik	0.056
Fiji – Ilastik	0.000000104

As a next step, I tried to use the Pearson correlation coefficient to ascertain the level of linear dependency between all data pairs. All of the results were over 0.99, showing a very strong level of linear

dependency. However, as the results did not differ much between different pairs, they could not be used to discuss the differences between measurement methods.

Next, I used the Friedman rank sum test (Friedman, 1937), a non-parametric test used to detect differences in measurements. The Friedman test's p-value was less than the significance level of 0.05, so there was a significant difference between the measurements.

However, the Friedman rank sum test alone cannot discern which pairs of measurements are the most different, so I used the Wilcoxon signed-rank test (Wilcoxon, 1945) with the Bonferroni multiple testing correction method. This method controls the increased risk of Type I errors that arise when multiple comparisons are needed [6].

In this case, six comparisons were made, so the significance level of 0.05 had to be adjusted by dividing the original significance level by the number of comparisons. Therefore, instead of the original significance level, I used the significance level of 0.0083 (gained as 0.05 / 6) to determine which differences between pairs were considered statistically significant.

Table 3: Results of the Wilcoxon signed-rank test with the Bonferroni correction for the differences between all data pairs

The p-values of pairs Human 1 – Human 2, Human 1 – Ilastik, Human 2 – Fiji and Human 2 – Ilastik are greater than the significance level of 0.0083, showing that the difference between population means between these pairs is statistically insignificant. The other two pairs have p-values lower than the significance level; therefore, the difference between population means is statistically significant.

Pairs	p-values
Human 1 – Human 2	0.062
Human 1 – Fiji	0.0000025
Human 1 – Ilastik	0.090
Human 2 – Fiji	0.613
Human 2 – Ilastik	0.339
Fiji – Ilastik	0.00000062

From the obtained results, I could discern that the most significant difference between the counts of the cells in each image was in Human 1 – Fiji and Fiji – Ilastik pairs. On the other hand, the most similar results were obtained by Human 2 – Fiji pair.

Conclusion

Counting cells in images is a challenging task, whether a human or a segmentation workflow is doing it. It is not made easier by the fact that there is no ground truth, and even the best results come only from the best guesses. In this project, I tried to compare two segmentation workflows with each other and, furthermore, with two human counters who did their best to count the number of cells in each of the fifty images.

From my results, I could tell that the results from both the Ilastik and Fiji segmentation workflows were more similar to those counted by Human 2, while Human 1 usually counted fewer cells in each image than the other three counters.

The most differences in counts were found between Human 1 – Fiji and Fiji – Ilastik pairs, where, in both cases, Fiji counted more cells in each image than Human 1 or Ilastik, respectively. It is hard to say precisely why that was the case, as it could be due to either Fiji's ability to recognise overlapping or clustered cells with high accuracy or Fiji's inability to learn to distinguish noise from real cells.

It is hard to define which workflow works better, but if I were forced to choose only one segmentation workflow for counting cells, I would most likely go with Ilastik. I found it easier and more intuitive to work with and, according to the Wilcoxon signed-rank test with the Bonferroni correction, the difference between population means between Ilastik and both Human 1 and Human 2 is not statistically significant, so it might not be as quick to classify any noise or errors as cells as Fiji.

References

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K. W., Schindelin, J., Cardona, A., & Sebastian Seung, H. (2017). Trainable Weka Segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics*, *33*(15), 2424–2426. https://doi.org/10.1093/bioinformatics/btx180

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics* (R package version 2.3). https://CRAN.R-project.org/package=gridExtra

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

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I., Friman, O., Guertin, D. A., Chang, J., Lindquist, R. A., Moffat, J., Golland, P., & Sabatini, D. M. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, *7*(10), R100. https://doi.org/10.1186/gb-2006-7-10-r100

Echalier, G., & Ohanessian, A. (1969). Isolement, en cultures in vitro, de lignées cellulaires diploïdes de Drosophila melanogaster [Isolation, in tissue culture, of Drosophila melangaster cell lines]. *Comptes Rendus Hebdomadaires Des Seances de l'Academie Des Sciences. Serie D: Sciences Naturelles*, *268*(13), 1771–1773.

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Pedersen, T. L. (2024). *patchwork: The Composer of Plots* (R package version 1.2.0). https://CRAN.R-project.org/package=patchwork

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. https://doi.org/10.1038/nmeth.2019

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Wickham, H. (2007). *Reshaping Data with the reshape Package* (R package version 1.4.4; Vol. 21, Issue 12, pp. 1–20). http://www.jstatsoft.org/v21/i12/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2023). *readxl: Read Excel Files* (R package version 1.4.3). https://CRAN.R-project.org/package=readxl

Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics*, *1*(6), 80–83.

Online Sources

References to online resources in the text are indicated by numbers in square brackets.

[1] Drosophila Kc167 cells. (n.d.). *Broad BioImage Benchmark collection*. Retrieved 18 May 2024, from https://bbbc.broadinstitute.org/BBBC002

[2] Interactive Density Counting. (n.d.). *Ilastik Documentation*. Retrieved 20 May 2024, from https://www.ilastik.org/documentation/counting/counting

[3] Pixel Classification. (n.d.). *Ilastik Documentation*. Retrieved 20 May 2024, from https://www.ilastik.org/documentation/pixelclassification/pixelclassification

[4] Unable to load images for Pixel Classification · Issue #2387. (n.d.). Ilastik, GitHub. Retrieved 20 May 2024, from https://github.com/ilastik/ilastik/issues/2387

[5] Ilastik Fiji plugin. (n.d.). *Ilastik Documentation*. Retrieved 21 May 2024, from https://www.ilastik.org/documentation/fiji_export/plugin

[6] Friedman Test in R. (n.d.). *Datanovia*. Retrieved 23 June 2024, from https://www.datanovia.com/en/lessons/friedman-test-in-r/