**Introduction: Exploring genetic diversity and behavioral variance through divergent strain sets**

1. Introduction

The realm of genetic studies has witnessed remarkable progress, especially when examining the correlation between genetics and specific phenotypes. Here we focus on Divergent Strain Sets to unravel the intricate relationship between genetic divergence and phenotypic behavior.

2. Background

Divergent Sets serve as a foundational block for this type of exploration. Out of 198 strains, 12 strains, distinguished by their genetic distinctiveness, make up the Divergent Set. These strains, when observed on a phylogenetic tree, reflect the nuances of genetic relatedness. Strains with greater distance from each other on the tree imply a more significant genetic distinction.

3. Objectives

To analyze the relationship between genetic diversity, as represented by the Divergent Set, and variations in collective behavior.

* Quantitative Analysis: To develop robust quantification measures for assessing the collective behavior of different nematode strains captured in videos.
* Genetic Correlation: To investigate the relationship between the genetic background of the nematode strains and their observed collective behaviors.
* Methodological Innovation: To establish a comprehensive methodology that integrates video analysis, feature extraction, and genetic data for studying collective behavior in nematodes.

4. Hypotheses

* H1: Nematode strains that are more closely related on the phylogenetic tree will exhibit more similar collective behaviors compared to those that are distantly related.
  + There will be a statistically significant correlation between the genetic composition of nematode strains and the phenotypic expression of their collective behavior.

5. Methods

* Phylogenetic Analysis: Employ the phylogenetic tree to ascertain genetic relatedness among the strains. Strains distant on the tree will be classified as genetically distinct.
* Behavioral Metrics Development: Formulate quantitative behavioral metrics to:
  + Catalog collective behavioral phenotypes.
  + Differentiate between distinct phenotypes.

6. Expected Outcomes

* Identification of patterns that link genetic diversity with behavioral variance.
* Development of reliable behavioral metrics that can effectively categorize and distinguish diverse phenotypes.
* Enhanced understanding of broader diversity patterns and genetic associations through the Mapping Set.

7. Future Scope

After deriving conclusive insights from the Divergent Set, and subsequently the Mapping Set, we anticipate the proposal of further specialized strain sets. This will cater to specific research needs, allowing for more targeted genetic and phenotypic studies.

To employ the Mapping Set, consisting of 48 strains, in broad diversity and genome-wide association studies, enhancing the scope after initial observations from the Divergent Set.

**Genral Steps:**

Step 1: Preprocessing

* Image Enhancement: Use techniques like histogram equalization to improve the contrast of your images.
* Noise Reduction: Apply filters to remove noise from the images.

Step 2: Blob Detection

* Thresholding: Convert the image to a binary format.
* Contour Detection: Use contour detection algorithms to identify the blobs.

Step 3: Feature Extraction

* Shape Descriptors: Calculate features like center of mass, area, perimeter, and circularity for each blob.

Step 4: Similarity Metrics

* Euclidean Distance: Calculate the distance between feature vectors of blobs from different datasets.
* Cosine Similarity: Measure the cosine of the angle between feature vectors.
* Statistical Tests: Use t-tests or ANOVA to compare the distributions of features.

Step 5: Data Visualization

* Heatmaps: Show similarity/dissimilarity matrices.
* Violin Plots: Plot features against each other to visualize similarities and differences as well as clusters.

Step 6: Optional Machine Learning

* Clustering Algorithms: Use k-means or hierarchical clustering to group similar blobs.
* Classification Algorithms: Train a classifier if you have labeled data.

**Plots:**

*Data set npr1, N2 and cb4856*

1. Frames images

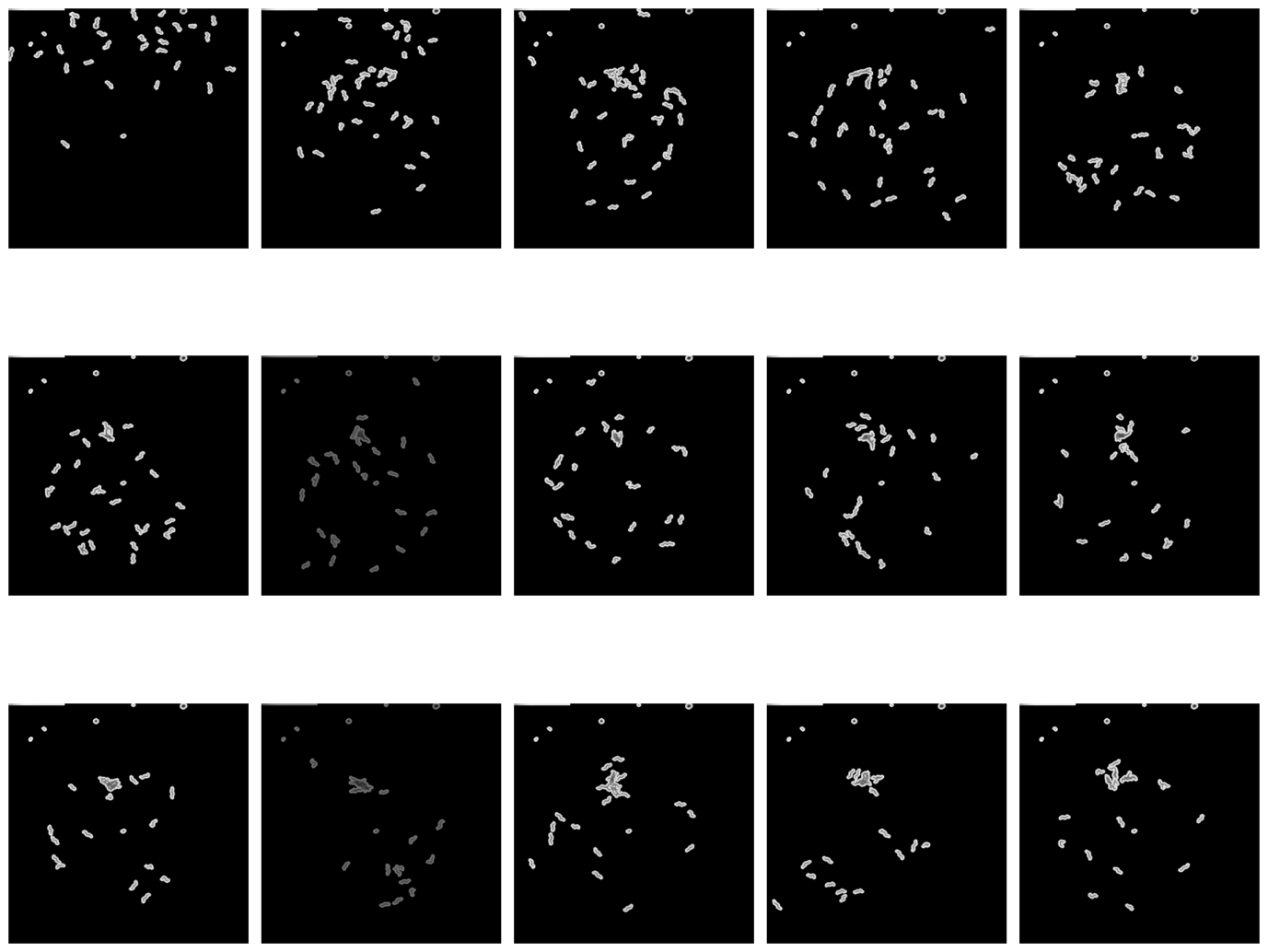


Figure : Frames used: I provide here one example of the extracted frames from the original hdf5 video file. This data set belongs to da609

The intensity of the frames does seem to vary and I am still unsure on why this is the case, and I am still trying to figure it out. This should not impact the quantification techniques, as I apply a binary mask to each frame so that only the blobs remain as white and the background black.

1. Blob distribution

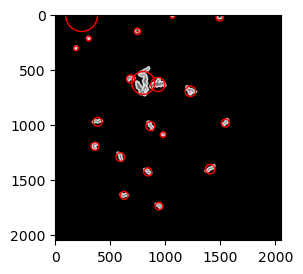


Figure : Detected center of mass for each blob for frame 15 of the da609 data set

After segmenting the blobs from the background I detected the center of mass of each blob and fitted a circle to the each blob. These values are then utilized to quantify the distribution of blobs in the frame, as well as their sizes.

* 1. Distribution KDE

The center of mass of the blobs was utilized to compute the Kernel Density estimation. This allows to visualize the distribution of the blobs centers on the plate. In the following image each subplot illustrates the KDE for the same frame number from the three different strains video. The darkest inner circle represents the highest density area, while the surrounding show the density decreasing. The skewness of the first subplot illustrate that the distribution is not symmetrical on the plat, while the second and third show a symmetrical distribution. It is also interesting to notes that all show one peak at the center indicating

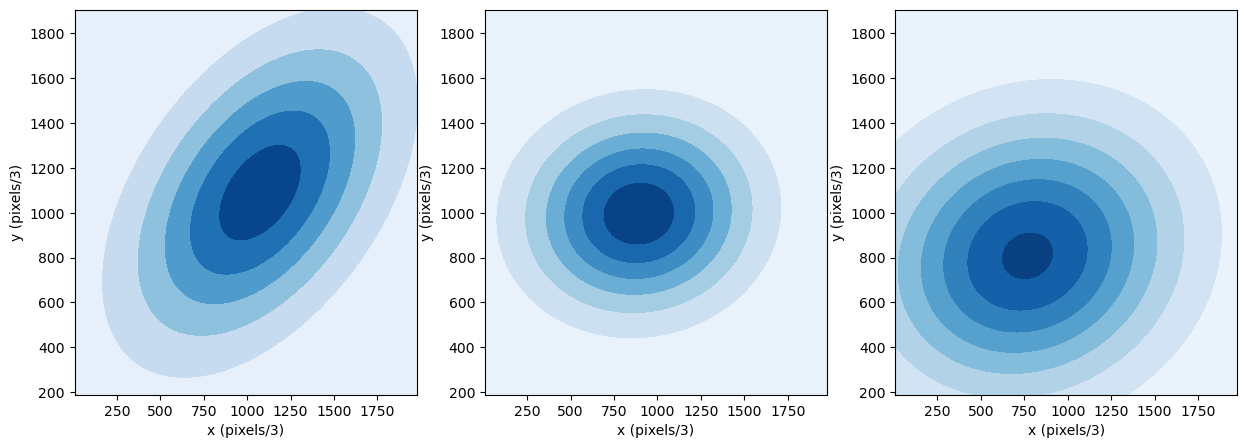


Figure : Kernel Density Estimation (KDE) on one frame for three data sets. This illustrates the distribution of the center of mass of the blobs on the plate. The first KDE corresponds to cb4856, the second to N2, and the third to da609.

* 1. Distance between blobs

In the following I have used the center of mass positions to estimate the pairwise distance between blobs for the same frame for all three different strains. The violin plot illustrates the distance for each different strain. All seem to indicate around the same pair wise distance as all median are included in the interquartile range of all three data sets. The problem with this measure is that it does not account for the rotational symmetries. A possible improvement to this is to look at the distance from the center of the image and look at the distance through the radial distance.

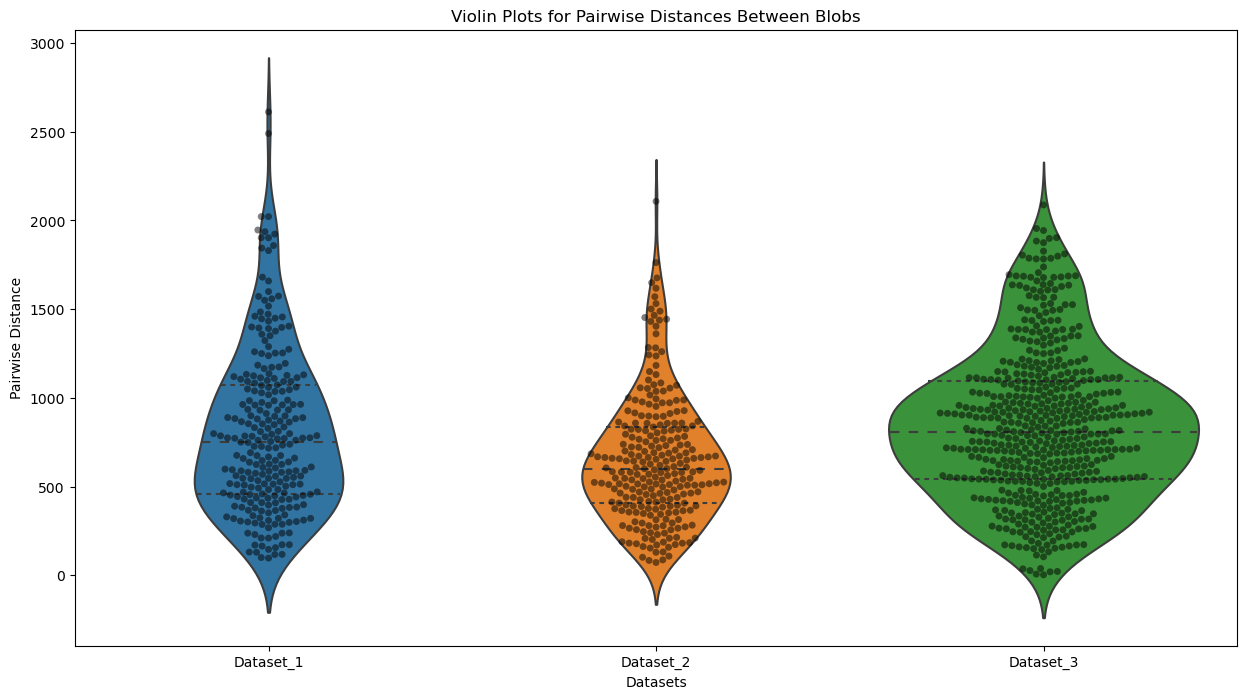


Figure 4: Violin plot for the pair wise distance between blobs center of mass in the same frame for the three strains. Dataset\_1 represents the data coming from cb4856, the second to N2, and the third to da609

Finally, I am still looking for a better visualization approach to explore these results. I have tried computing a heatmap for all distances, but I still have not found a way of summarizing the data to allow for easy comparison between different data sets or multiple frames. Any suggestion would be appreciated.

* 1. Distance from center

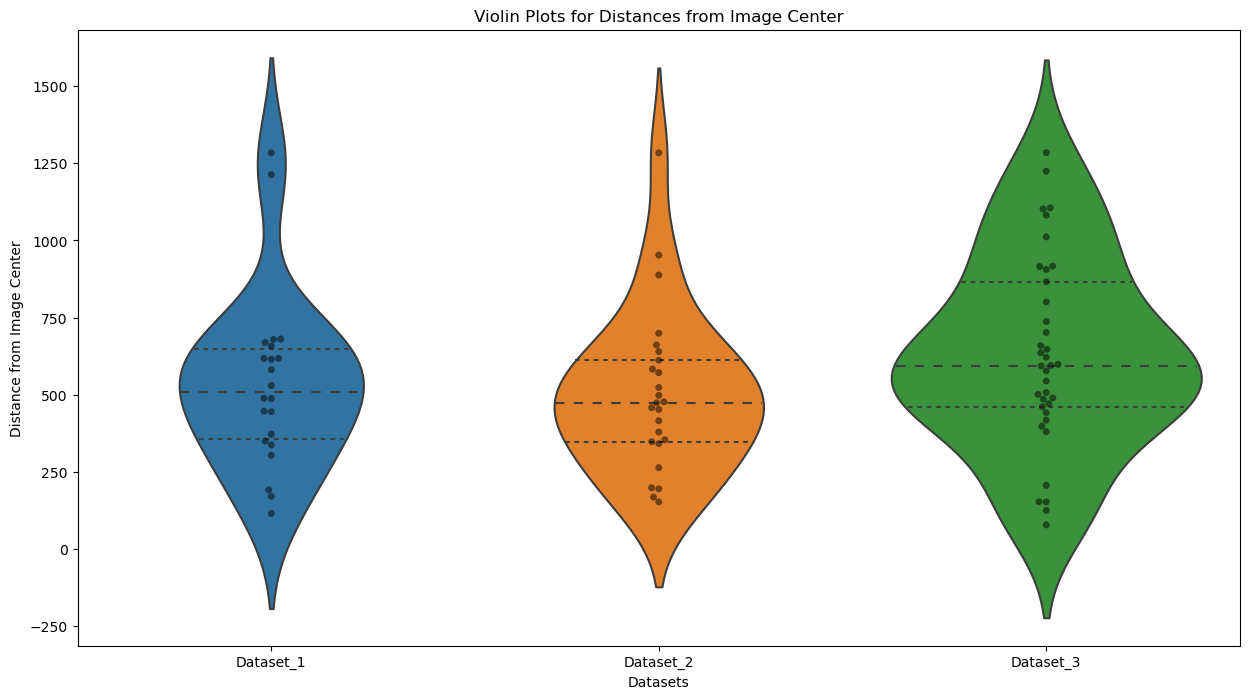


Figure 5: violin plot for distance between center of mass of each blob and center of the image. Dataset\_1 represents the data coming from cb4856, the second to N2, and the third to da609

1. Blob size: cluster density

In the following violin plot I am illustrating the area size in pixels for the three strains. All seem to present similar areas median, as all are inside the interquartile range of the other data sets.

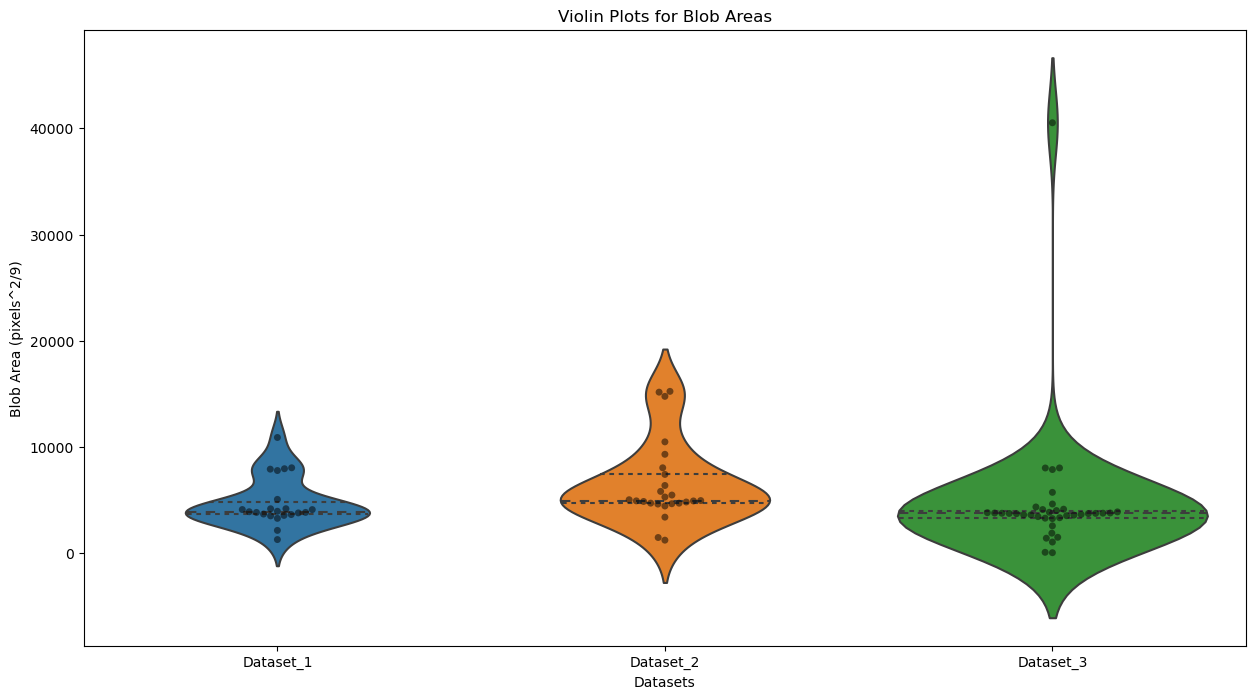


Figure 6: Violin plot for the area of the blobs detected in the same frame number for all different strains. Dataset\_1 corresponds to the strain cb4856, the second to N2, and the third to da609

1. Hu moments

In the following I am illustrating the 7 Hu moments for the same frame number for the different strains, as labeled in the legend.

The hu moments are a commonly utilized estimation of an image statistical pixel value properties. They are designed to capture essential characteristics of the shape of an object, such as its area, centroid, and orientation. Because they are invariant to certain transformations, they allow for the comparison of shapes in different images even if the shapes have been rotated, scaled, or translated.

* The first moment measures the area of the object.
* The second provides information about the spread of the object.
* The third illustrates how skew the object pixel distribution is.
* With the fourth we look at the elongation of the object in order to distinguish elongation and compactness.
* With the fifth we capture the information on the number of isolated regions in the object.
* With the sixth we look at the boundary of the object.
* And finally with the seventh we quantify the complexity og the object, meaning that the intricate details are identified.

This method of analysis has been developed for image analysis, but it’s interpretability is easier when applying the measurement to each blob independently more than the entire image. In fact this method is usually utilized in object recognition tasks. A possibility is to apply this individually to each blob and take the mean and standard deviation in order to represent one frame. I have to conduct more research on it.

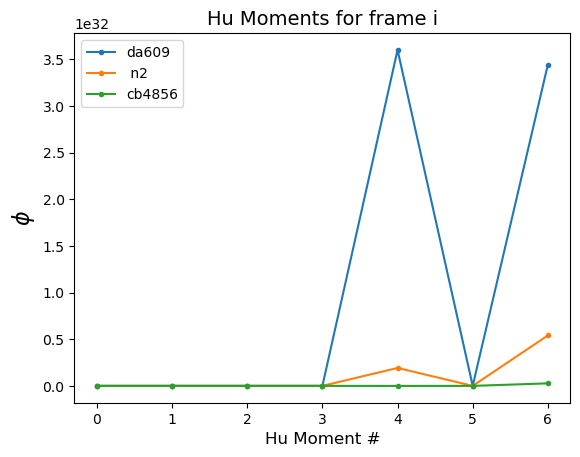


Figure : 7 Hu moments for the same frame for all three different strains considered. As indicated in the legend, cb4856 is shown in green, N2 in orange and da609 in blue.

The moments seem to indicate that for this frame the cb4856 has all moments close to zero, while da609 shows the larger values for the fourth and sixth moments.

1. Wavelet Coefficients

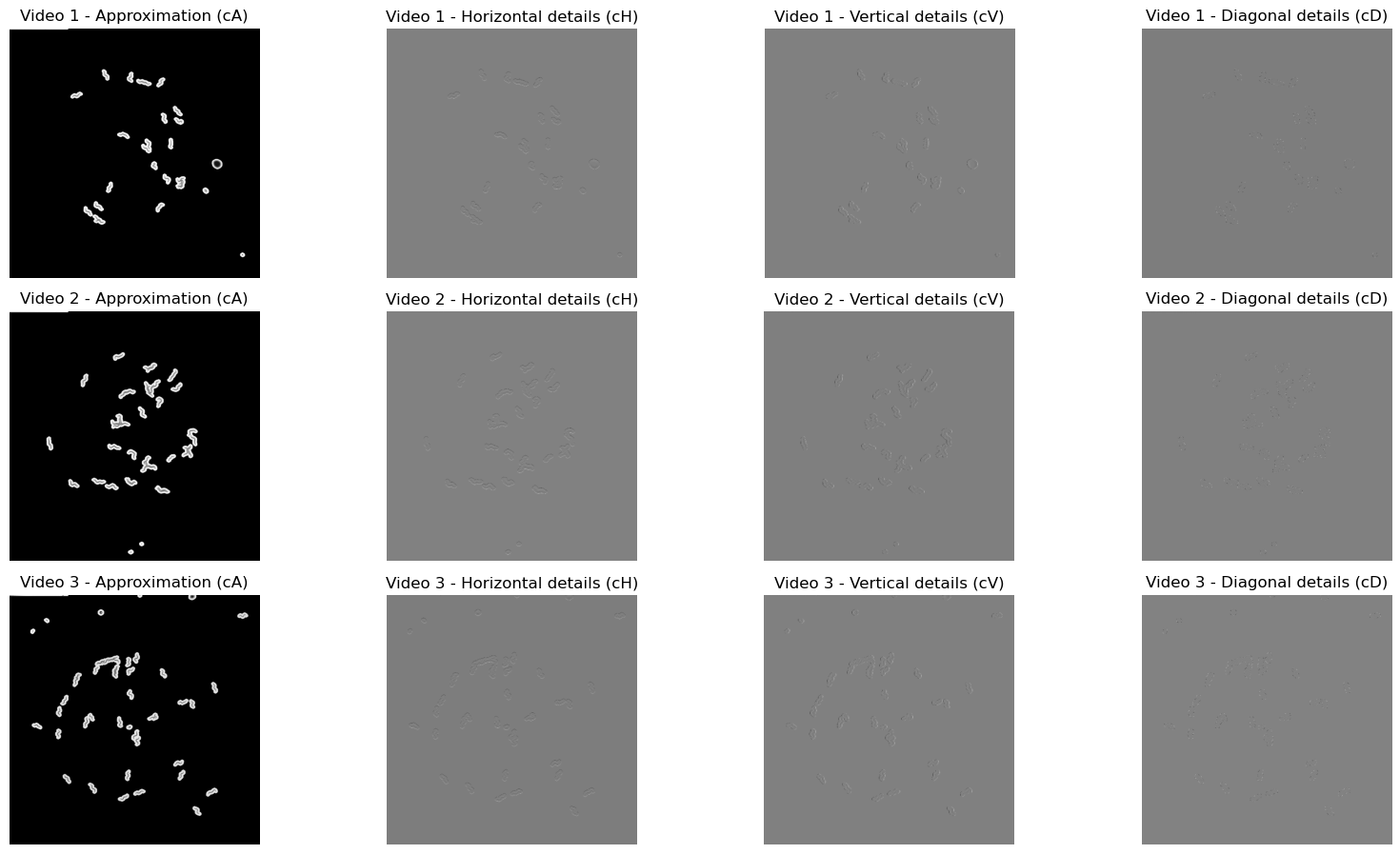


Figure : The figure provides an example of the original frame considered followed by the horizontal, vertical and diagonal details extracted through the wavelet transform. Again Video 1 corresponds to cb4856 is shown in green, N2 in orange and da609 in blue.

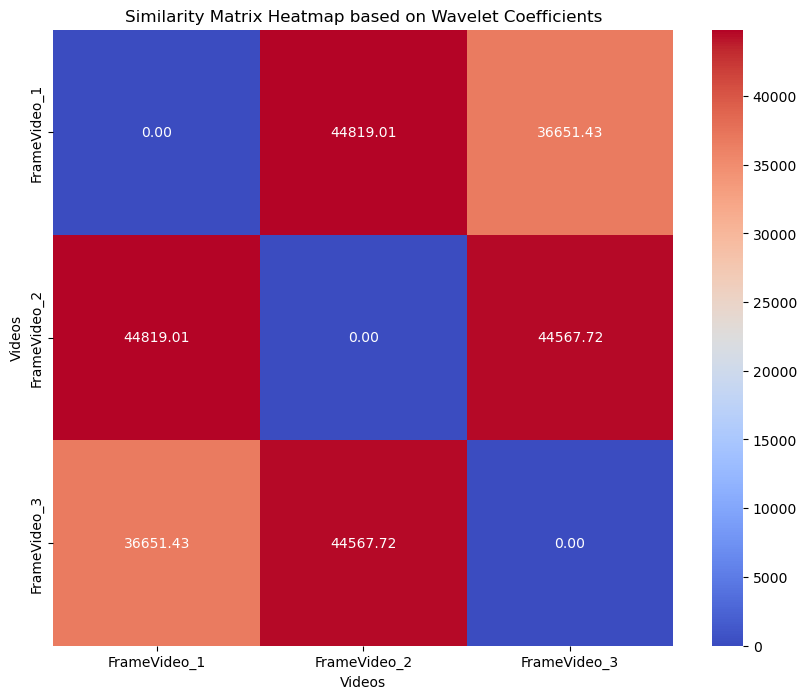


Figure : Heat map illustrating the distance between the three frames belonging to the three different strains. The image Index 0 corresponds to cb4856,1 to N2 in and da609 in blue.

Finally I have repeated the wavelet analysis on each video independently for each frame and obtained the distance matrix between all frames for each video independently. Here I provide the distance matrix for the first video.

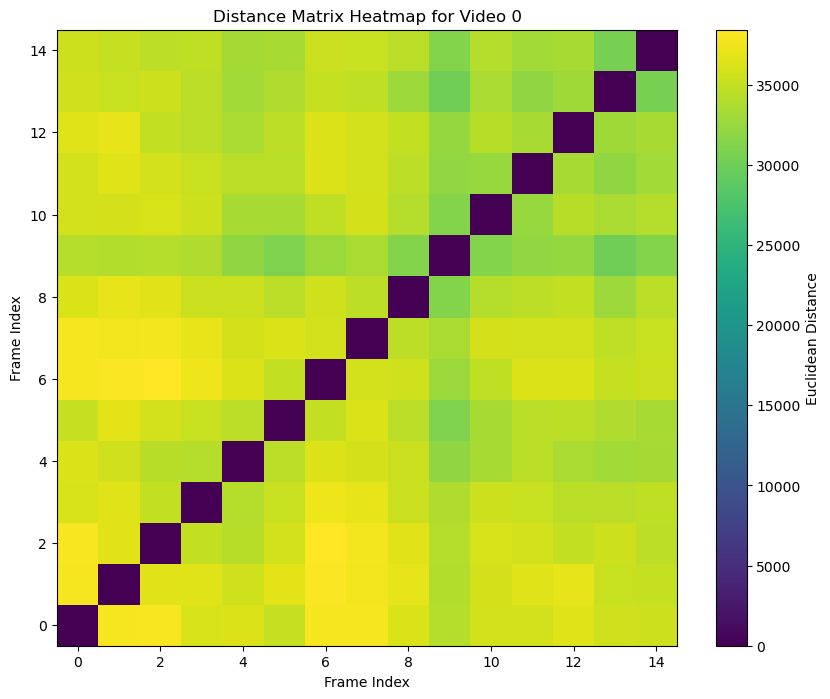
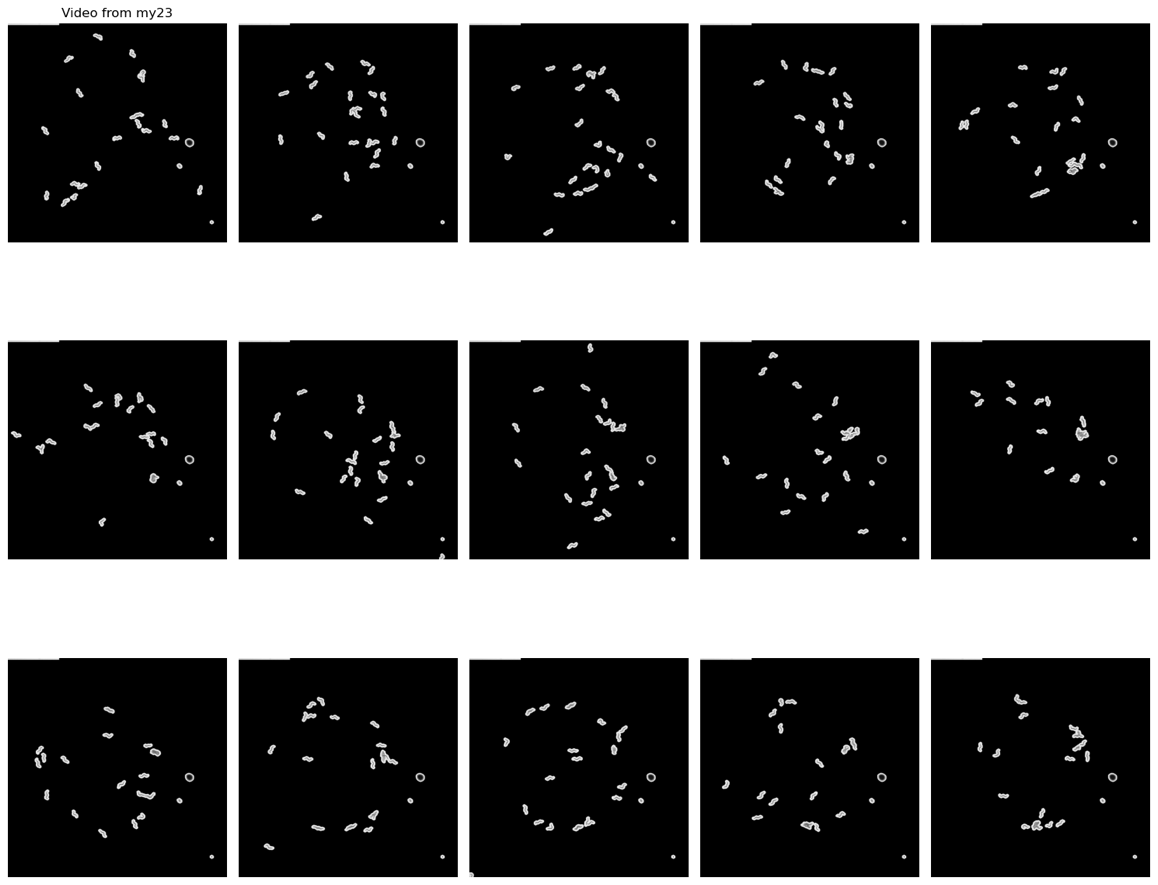


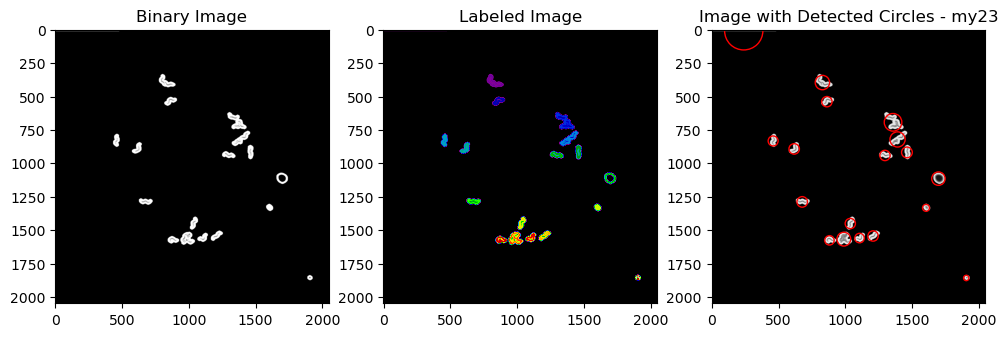
Figure : Distance matrix for all frames belonging to the video cb4856

*Divergent Data Set*

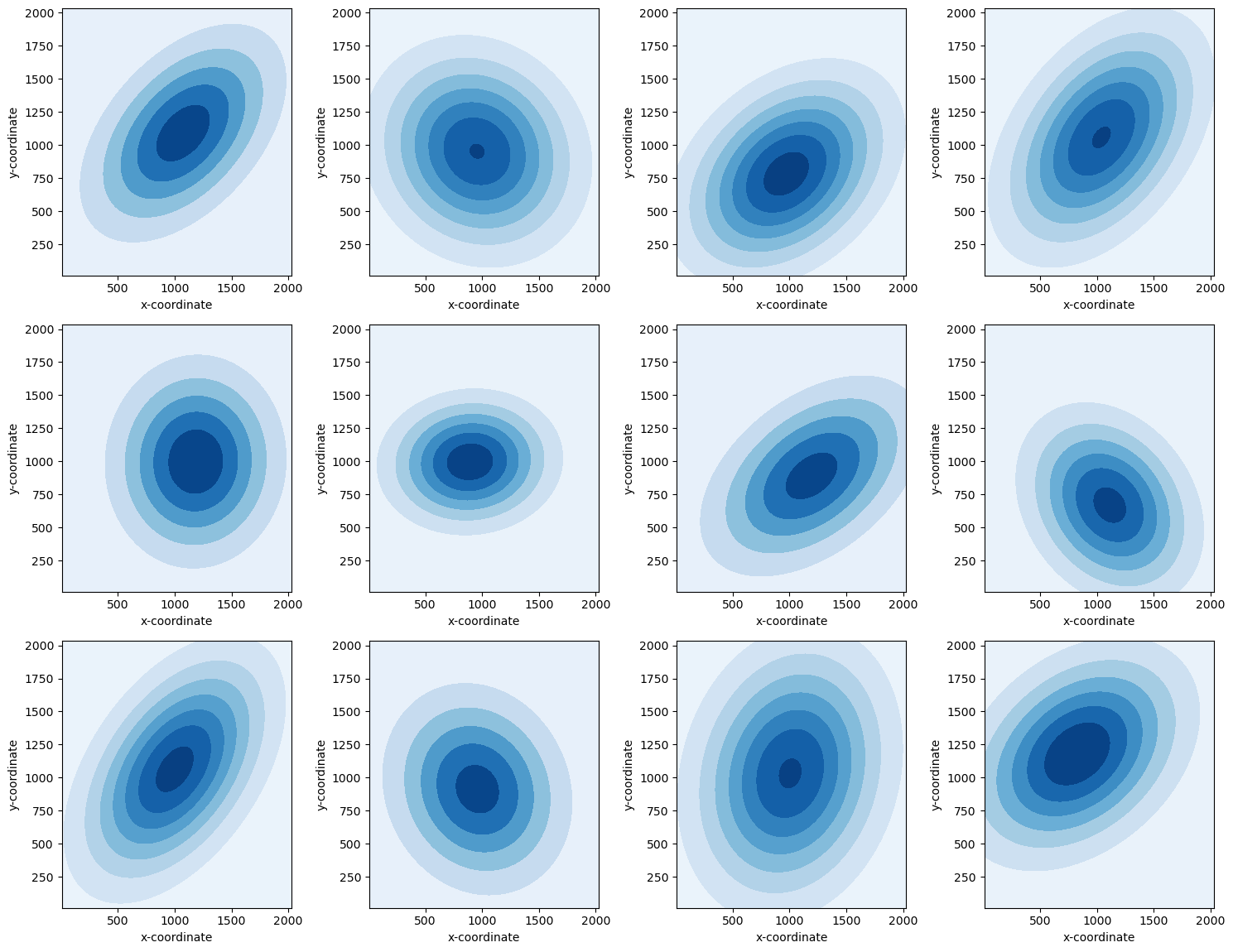
1. Frames images



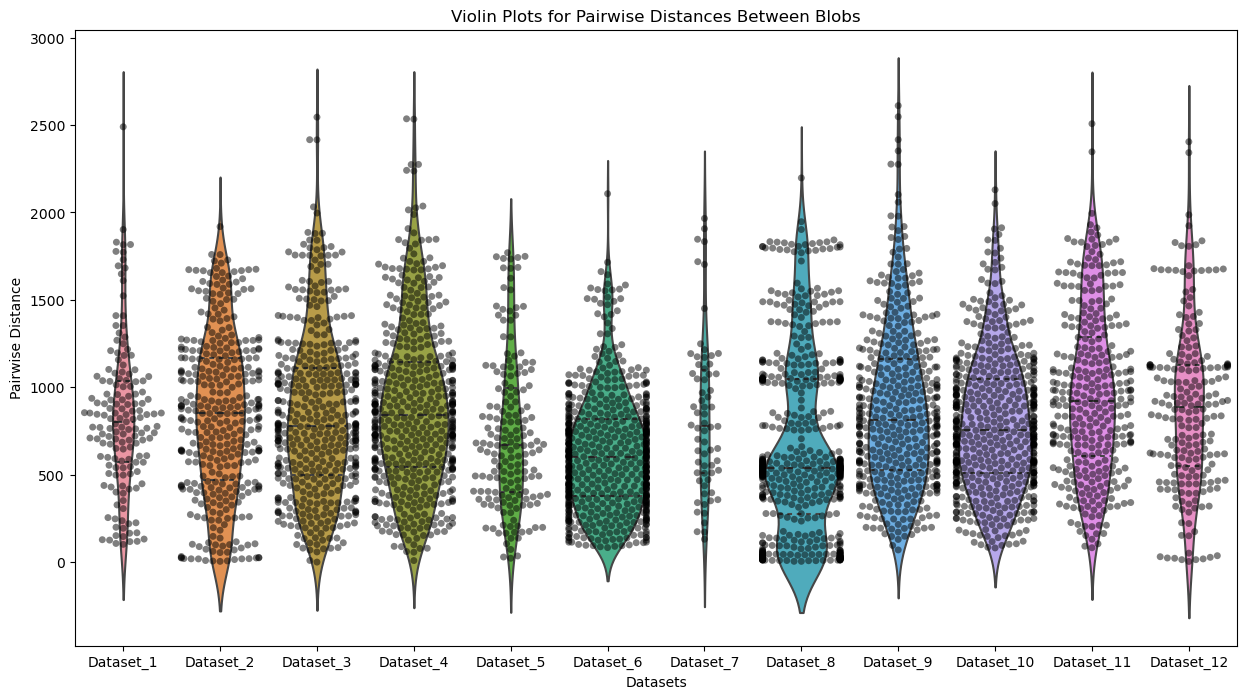
1. Blob distribution



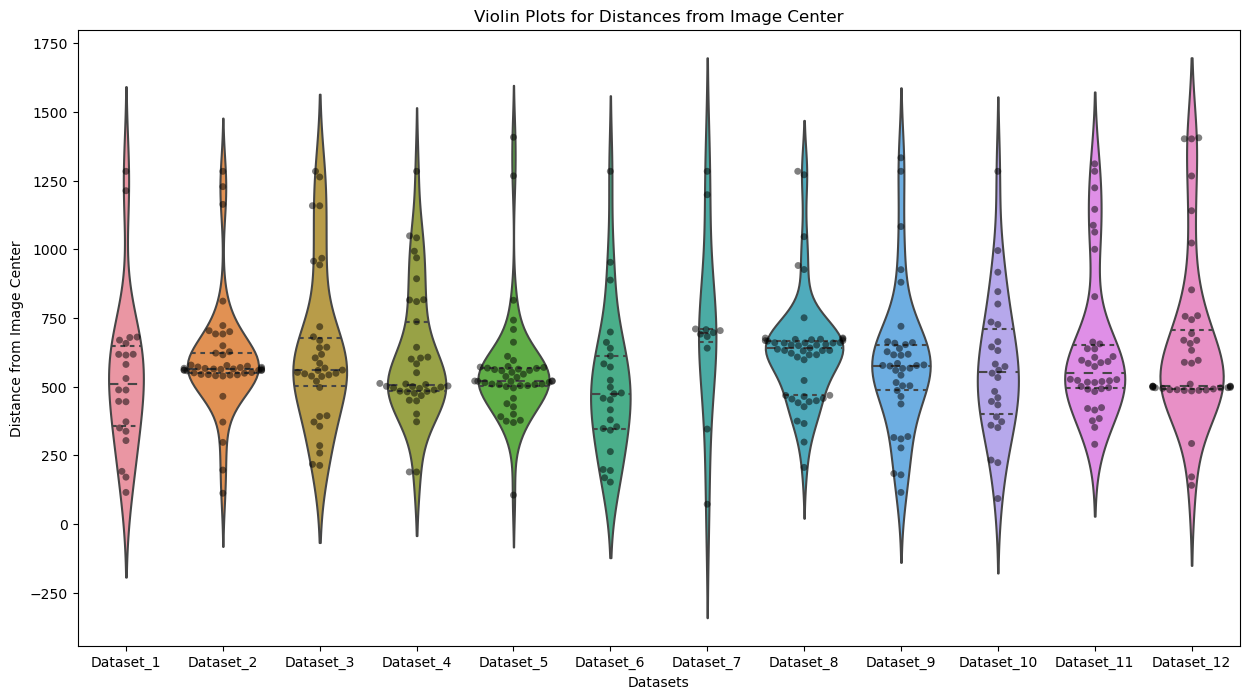
* 1. Distribution KDE



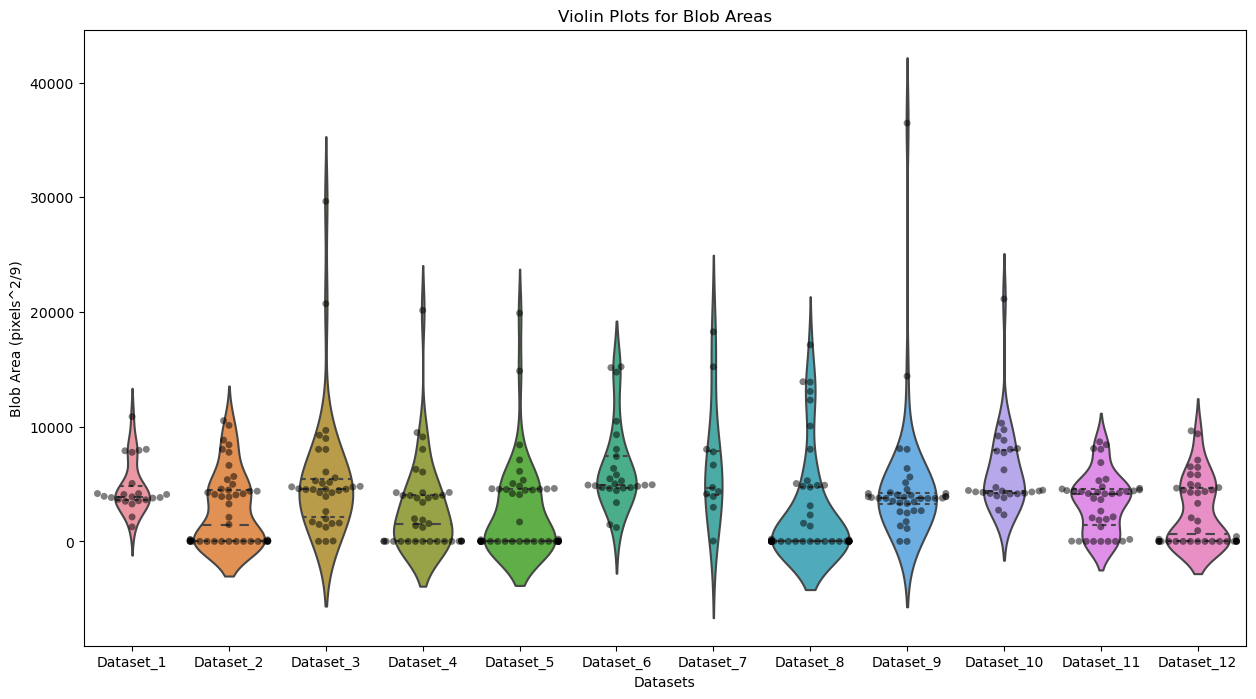
* 1. Distance between blobs



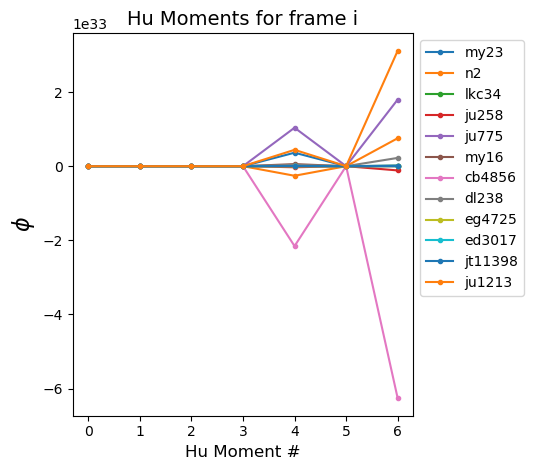
* 1. Distance from center



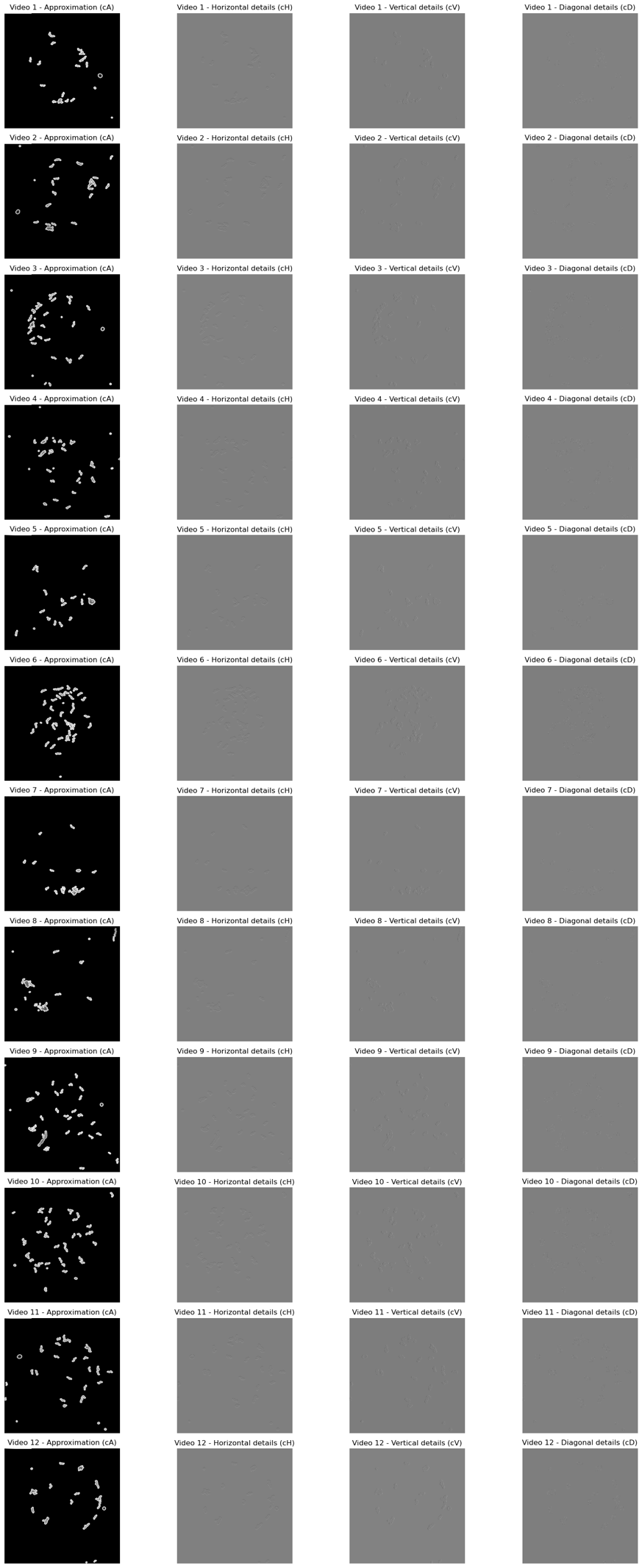
1. Blob size: cluster density

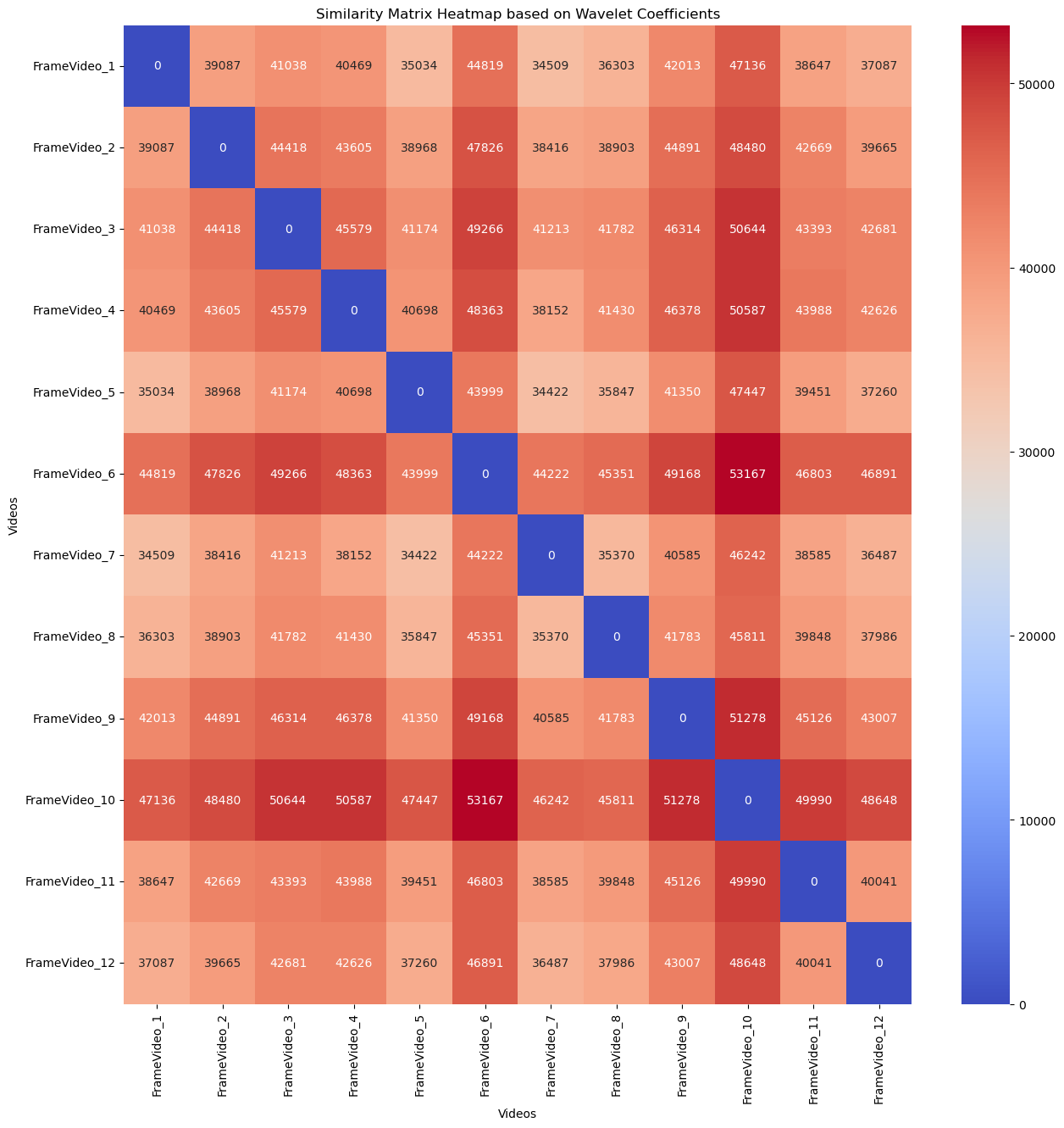


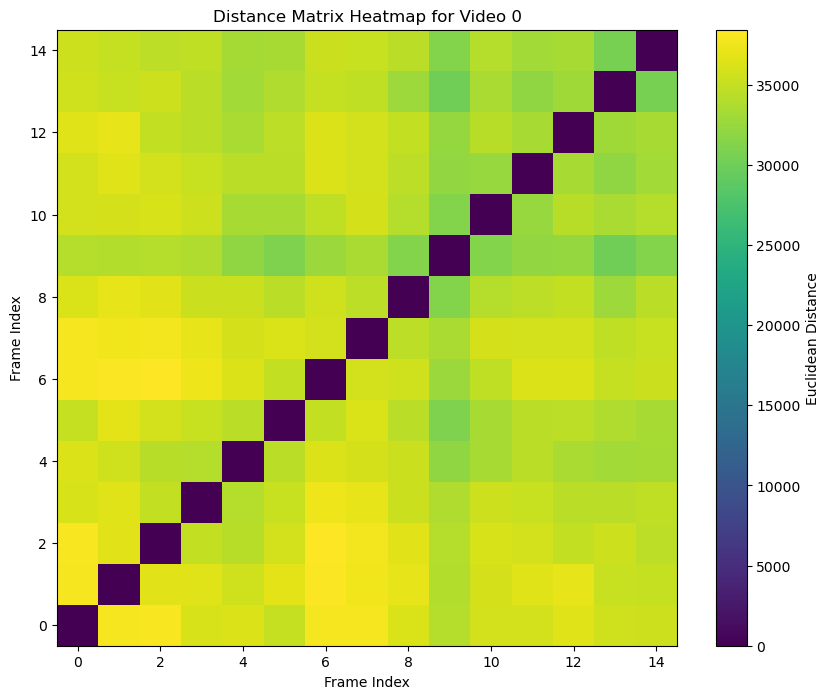
1. Hu moments



1. Wavelet Coefficients





****

**10. Conclusion**

Here I have provided a visualization of some of the potential quantifications for the collective behavior. Many other are still to investigate, but this was a first exploration of the easiest and more intuitive ones. I still have to decide on better ways to visualize the analysis. This includes also on thinking how this can be extended to the analysis of whole videos and not only single frame comparisons, as this adds another dimension not so simple to visualize at the moment.

**Next Steps:**

Thinking about possible steps forward. I will look into global and local features which can be useful in the characterization. The first stage will involve looking for possible useful measures. The choice will be made based on how good the measurements provide a quantification of similarity and difference between frames. All the characterization methods will then be combined in a feature vector.

* **Global Features**
  + Global features capture the overall characteristics of an image and are usually simple to compute. They are sensitive to rotation, scale, and translation.
  + Texture Descriptors: Use statistical measures to describe the texture.
  + Shape Descriptors: Use geometric properties like area, perimeter, and moments.

How to Compare:

* Euclidean Distance: Measure the distance between histograms or feature vectors.
* Chi-Square Test: Compare histograms using the chi-square statistic.

Local Features

* Local features capture intricate details by focusing on smaller regions of the image. They are usually robust to rotation, scale, and translation.
* SIFT (Scale-Invariant Feature Transform): Detects and describes local features in images.
* SURF (Speeded-Up Robust Features): An enhanced version of SIFT.
* ORB (Oriented FAST and Rotated BRIEF): A fast binary descriptor.

How to Compare:

* Feature Matching: Use algorithms like FLANN or Brute-Force to match local features between images.
* RANSAC (Random Sample Consensus): For robustly estimating the transformation parameters.

Combining Global and Local Features

* You can concatenate global and local features into a single feature vector for more robust similarity measures.
  + Fusion Methods: Combine feature vectors using weighted sums or other fusion techniques.
  + Machine Learning: Use the combined features as input to machine learning algorithms like SVM or Random Forest for classification tasks.