



Universiteit
Leiden
The Netherlands

Pattern Recognition in High-Throughput Zebrafish Imaging

Nezhinsky, A.E.

Citation

Nezhinsky, A. E. (2013, November 21). *Pattern Recognition in High-Throughput Zebrafish Imaging*. Retrieved from <https://hdl.handle.net/1887/22286>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/22286>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/22286> holds various files of this Leiden University dissertation

Author: Nezhinsky, A.E.

Title: Pattern recognition in high-throughput zebrafish imaging

Issue Date: 2013-11-21

5 Data and Pattern Analysis in Infection Studies in Zebrafish

Based on:

A. Nezhinsky, E.J. Stoop, A.M. van der Sar & F.J. Verbeek: Numerical Analysis of Image Based High Throughput Zebrafish Infection Screens: /Matching Meaning with Data/. In: Bioinformatics 2012, Proceedings of the Int. Conf. on Bioinformatics Models, Methods and Algorithms: 257–262 (2012)

and

A. Nezhinsky, E.J. Stoop, A.A. Vasylevska, A.M. van der Sar & F.J. Verbeek: Spatial Analysis of Bacterial Infection Patterns in Zebrafish. In: Proceedings 21th Annual Belgian-Dutch Conference on Machine Learning (2012)

and

A. Nezhinsky, A.M. van der Sar & F.J. Verbeek: In Depth Analysis of High Throughput Zebrafish Infection Screens. In preparation (2013)

5.1 Introduction

Tuberculosis is a serious disease and a significant part of the world population is infected. Unfortunately, effective treatment is still difficult due to bacteria resistance. In order to elucidate which genes are responsible for infection, the behavior of the tuberculosis bacteria *Mycobacterium tuberculosis* needs to be analyzed. In our study this behavior is modeled by a close relative — the *Mycobacterium marinum* (*MM*). The *MM* is hosted in cold blooded animals. For our study the zebrafish is used as a host. Zebrafish makes a good model for analysis as its immune system is in many ways comparable to human. The zebrafish larvae can be obtained in large numbers and studied by HT imaging (cf. Chapters 2, 3).

Infection of the zebrafish with *MM* is characterized by the presence of granulomas. Granulomas are clusters of immune cells and bacteria indicating infection. They can be visualized with fluorescent agents.

In order to determine which genes of *MM* are involved in formation of granulomas we obtained a dataset of 1000 random mutants of the *MM* bacteria and screened for those mutants that were not able to efficiently infect zebrafish larvae [54]. In this manner 30 mutants have been identified that were unable to infect larvae (cf. [55]).

In order to gain more insight in the progression of *MM* infection it is required to analyze infection spread in the host, c.q. the zebrafish, over a certain period of time. This requires the following questions to be answered:

- (1) Is there a pattern in the organization of granuloma clusters in certain tissues?
- (2) Does appearance differ for certain bacterial mutants?

This analysis is accomplished with the HT imaging as described in Chapters 4. For each zebrafish a brightfield and fluorescence microscopy image is acquired. The analysis included localization of the zebrafish shape and qualitative estimation of the granuloma cluster size and spread. Consequently, no quantitative data could be retrieved. The results are presented by means of different types of custom designed infographics.

We have designed and implemented an automated framework for shape retrieval and cluster analysis [43]. This framework has been applied to large scale HT applications [54].

The basis of the software framework is an algorithm for shape retrieval to automatically find the zebrafish shape(s) in the image. The algorithm uses deformable template matching [23] and labels the regions for further analysis. This approach made it possible to analyze the infection amount per fish in an automated fashion (cf. Chapter 4).

Proof of Principle: Wild-type versus Mutant 714

As a proof of principle, a study for the detailed analysis was performed to find a strategy for analysis. As an initial test case mutant 714 was chosen; this is one of the 30 mutants which is not successful in infecting the fish. We are focusing on the question: can the size of the granuloma clusters be analyzed from the infected fish? In addition, we compare similarity in larvae infected with the wild-type *MM* and mutant 714.



Figure 5.1: A brightfield image may contain up to 3 shapes of the zebrafish larva.

In depth Analysis: Wild-type versus a List of Mutants

After the proof of principle from the first assessment we apply our algorithm for in depth analysis of a larger set of mutants. We compare the results of infection to the readout for the wild-type. The following mutants indicated by a number were part of our analysis: 262, 308, 414, 415, 421, 423, 431, 730, 748, 801, 817, 885, 941 and 943. In order to process this amount of data an objective test was needed.

Microscopy

Images were taken in batches of 30 wells. Each batch contained sibling zebrafish larvae. Images were acquired with a Leica DC500 microscope. The general layout of an experiment takes 3 zebrafish larvae per well. As a result a single image can contain up to 3 individuals.

For each well, both brightfield and fluorescent images were acquired. The brightfield image contains the zebrafish shape, while the fluorescent image contains the signal at granuloma locations. An example of such images for a single well is presented in Figure 5.1 and 5.2 (cf. Chapters 2, 3). As can be seen the fluorescent images have very low and different intensity values. By inspecting the images, it is not difficult to conclude that consistent manual analysis therefore will be very difficult and imprecise.

5.2 Analysis of High Throughput Zebrafish Infections Screens Based on Approach 2

The images are the input for the analysis framework *ZFA* (cf. Chapter 4). As algorithm for the pattern recognition we used Approach 2 (cf. Section 2.3). First, the zebrafish are localized and the result is used as a mask for the fluorescent image. Within the mask

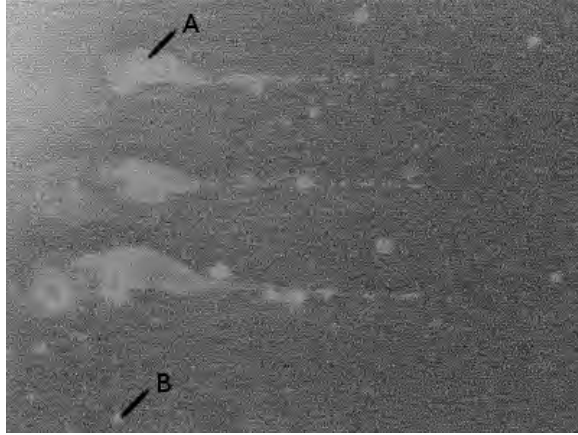


Figure 5.2: A fluorescent image containing the signal of granuloma spread. For visualization purposes the contrast is enhanced. The image contains specific (A) and non specific (B) staining as well as systemic noise.

a threshold value is determined. Finally the data is analyzed and written to a comma separated file.

The data set we have used for the analysis consisted of 189 infected zebrafish larvae. The larvae were divided into 3 groups: not infected larvae *NI* (5), infected with *Mycobacterium marinum* wild-type *MM* (67) and those infected with the 714 mutant *714M* (117).

For the infection approximately the same amount of bacteria was used; the volume was plated on 7H10 plates. At injection the zebrafish were 6 days old. The infection has progressed for 5 days, after this the imaging was performed.

5.2.1 Brightfield Imaging: Shape Localization and Annotation

The ZFA determines a region of interest (ROI) and provides annotation of the relevant areas. A deformable template is used for the retrieval of the zebrafish shapes. In order to be able to detect different regions of the fish, the template was divided into 11 regions (or *slices* in Section 2.3) counting from head to tail (numbered from 0 to 10). We have chosen for division into 11 parts, as it is empirically established that this amount of parts allows best for annotating the shape as well as doing spatial analysis. In Figure 5.3 the graphical representation of a prototype template used for our experiments is shown. Enumerations (i) with values 0 and 1 are considered as the head regions, 2 till 4 as the trunk regions and the remaining parts as the tail regions. This division of the larva in regions has been successfully used in other applications [63]. The injection point for infection is located at approximately region 5 [11] and this can be used for the analysis.

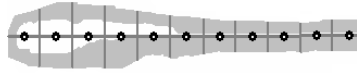


Figure 5.3: Graphical representation of the prototype template used for our experiments. The vertical bars divide the zebrafish into regions. The template is created from averaging a test set of 20 training shapes

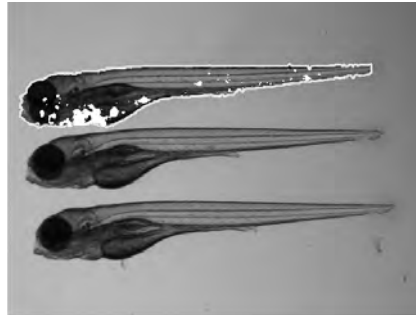


Figure 5.4: Graphical output of the framework overlaid on the original image for the top fish in the image. The light gray line denotes the shape mask contour and the white regions indicate the presence of granuloma formation. This image is created in an automated fashion.

5.2.2 Fluorescent Imaging: Analysis

For the actual measurement of clusters fluorescent images were used and related to the mask size, obtained from the brightfield images. The NI group is expected to have no infection at all and thus the level of their maximal fluorescent signal is considered as noise level n . No infection / granuloma formation is present in NI, therefore this group was only used to obtain reference for a noise level. In the other groups all signal below n is considered noise, while all signal above n represents granuloma presence. This signal is analyzed per fish and written to a *csv* file as shown in Table 5.1. The notions will be explained in the sequel.

5.2.3 Output and Dataset Creation

Output of the analysis software is created in the form of an overlay image containing the detected zebrafish and the infection as shown in Figure 5.4, together with a *csv* file (cf. Chapter 4).

Clusters of bacteria are labeled and for each of the clusters the surface area is determined. The infection is present in clusters:

Definition 1. *ClusterCount* is the amount of clusters present throughout the entire zebrafish larva, denoted by CC . A single cluster is defined as a connected collection of

Field name	Explanation
TotalArea	Total shape area
ClusterCount CC	Amount of clusters in the larva
ClusterSize CS	Area covered by all clusters
ClusterCountAt[i] $CC[i]$	Like CC but for a single template region
ClusterSize[i] $CS[i]$	Like CS but for a single template region

Table 5.1: Fields contained in the output csv per larva, i to the template region number (0 to 10) as described in Section 5.2.1.

pixels of which every pixel is covering an area that is classified as infection. The image pixels are classified as *infection* in an automated fashion as described in Section 4.3.2. Background pixels are automatically excluded from the analysis.

The total area of the spread is the sum over all clusters:

Definition 2. *ClusterSize* is the total area covered by the infection throughout the entire zebrafish larva, denoted by CS . Area is expressed in units that represent the amount of pixels that are classified as *infection*. This classification was performed in an automated fashion as described in Section 4.3.2. Background pixels are automatically excluded from the analysis.

From the template we define 11 regions and a cluster is always assigned to the region center with closest geometrical distance [43].

5.2.4 Analysis and Results

In this section an analysis of the results is presented. First the distribution of the clusters is discussed and second a relation to the amount of infection is derived from the data.

Distribution of Cluster Amount

In our study we set out to analyze the relationship between mutant and wild-type with respect to the amount of clusters and spatial distribution. To that end we compare the average number of clusters (feature Cluster Count, $CC[i]$) between MM and $714M$ (cf. Figure 5.5).

Definition 3. *Cluster Count[region]* is the amount of clusters present in a certain region (or slice), denoted by $CC[i]$. A cluster is defined as a connected collection of pixels of which every pixel is covering an area that is classified as infection.

The image pixels are classified as *infection* in an automated fashion as described in Section 4.3.2. Background pixels are automatically excluded from the analysis.

From the initial measurements this distribution does not seem to be conclusive. This is due to the fact that the mean was taken from a dataset with a certain scatter. This scatter is shown in Figure 5.6.

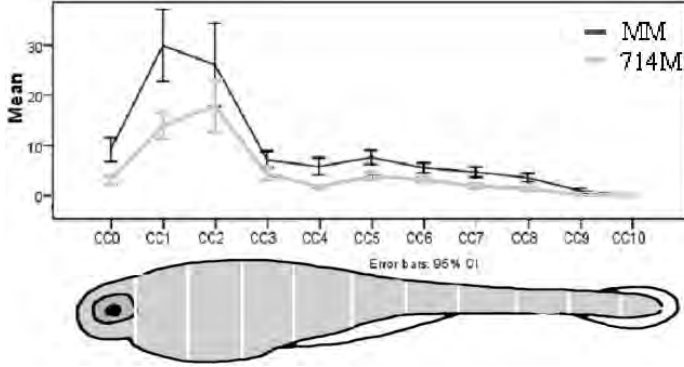


Figure 5.5: Spatial comparison of the average amount of clusters for MM and 714M in relation to the zebrafish template with a 95% confidence interval.

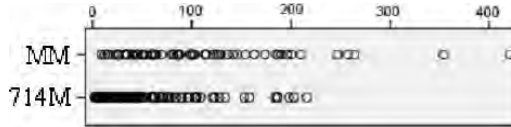


Figure 5.6: Scatter plot of the amount of clusters in each case of the test set used.

We are interested in the distribution of the granuloma clusters and in order to analyze the different batches in the same way we normalize the $CC[\#]$ over the total CC . The normalization is performed for each individual case and subsequently the mean is calculated. In Figure 5.7 the results are depicted.

Definition 4. *Normalized Cluster Count [region]* is the normalized $CC[i]$ feature, denoted by $CCn[i]$. Normalization is done over the total amount of clusters throughout the entire zebrafish larva region: $CCn[i] = CC[i]/CC$

From the graph we can observe that MM and $714M$ have the same behavior. The mean and the 95% confidence interval suggest that the two distributions can be considered as similar. Additionally we assume, after inspection of a Q-Q plot, that the data for each variable has a normal distribution. Our null hypothesis H_0 states that, under assumption that the two groups are independent, their variances are equal. We therefore apply Levene’s Test for Equality of Variances to the $CCn[i]$, $0 < i < 10$. The results are shown in Table 5.2.

From the Levene’s test we obtain the value of p , that stands for significance. If $p > 0.05$ then the null hypothesis is accepted. For zebrafish regions 1, 3, 4, 5, 8, 10 the hypothesis is accepted, the corresponding variances are equal (marked with * in Table 5.2). For regions 0, 2, 6, 7 and 9 the variances significantly differ. Finally, we performed the independent samples t-test. Based on the results from Levene’s test we know which variances significantly differ; in Table 5.3 only the correct assumptions are listed.

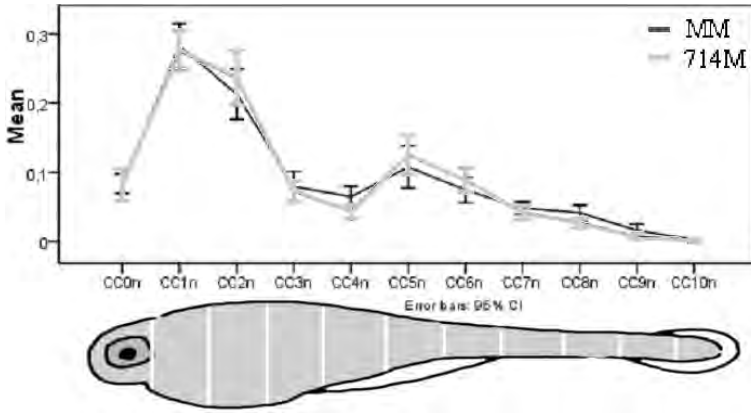


Figure 5.7: Spatial comparison of the normalized amount of clusters for MM and 714M with a 95% confidence interval as compared to the zebrafish template.

Region	H_0 Var	p value	
$CCn[0]$	=	0.017	
$CCn[1]$	=	0.218	*
$CCn[2]$	=	0.010	
$CCn[3]$	=	0.526	*
$CCn[4]$	=	0.512	*
$CCn[5]$	=	0.221	*
$CCn[6]$	=	0.011	
$CCn[7]$	=	0.042	
$CCn[8]$	=	0.488	*
$CCn[9]$	=	0.046	
$CCn[10]$	=	0.662	*

Table 5.2: Levene’s Test for Equality of Variances for Cluster Count $CCn[i]$. Equal variances are marked with *.

Region	Var	t	Sig. (2t)	Mean Diff	Std. Err Diff	
$CCn[0]$	\neg	0.014	0.989	0.000	0.014	
$CCn[1]$	$=$	0.244	0.808	0.006	0.023	
$CCn[2]$	\neg	-0.909	0.365	-0.024	0.027	
$CCn[3]$	$=$	0.417	0.677	0.005	0.013	
$CCn[4]$	$=$	2.216	0.028	0.020	0.009	*
$CCn[5]$	$=$	-0.844	0.400	-0.018	0.022	
$CCn[6]$	\neg	-1.225	0.222	-0.015	0.013	
$CCn[7]$	\neg	0.815	0.416	0.005	0.007	
$CCn[8]$	$=$	2.093	0.038	0.013	0.006	*
$CCn[9]$	\neg	1.603	0.112	0.008	0.005	
$CCn[10]$	$=$	0.270	0.787	0.000	0.001	

Table 5.3: t-test for Equality of Means. Significant difference in the mean value is marked with *.

We observe that there is a significant difference (meaning significance < 0.05) in the mean value for regions 4 and 8 (marked with * in Table 5.3). Preliminary conclusions are as follows. In the head the highest proportion of clusters is found, followed by the injection site. Globally the distribution is the same for both MM and $714M$. Exceptions are region 4, adjacent to the injection site, and region 8 where a significant larger distribution of the clusters of MM bacteria is found compared to $714M$.

Distribution of the Amount of Infection

Next, we analyze the relation between mutant and wild-type in the area covered by granulomas:

Definition 5. *Cluster Size [region]* is the total area covered by the infection in region i , denoted by $CS[i]$. The area is expressed in units that represent the amount of pixels covering the region of interest within in the input images.

The image pixels are classified as *infection* in an automated fashion as described in Section 4.3.2. Background pixels are automatically excluded from the analysis.

In Figure 5.8 a graph, with 95% confidence, is depicted of the comparison of the average area of infection between MM and $714M$.

Again, as with the cluster count, we normalize the $CS[i]$ over the total CS for both MM and $714M$ and compare the results in Figure 5.9.

Definition 6. *Normalized Cluster Size [region]* is the normalized $CS[i]$ feature, denoted by $CSn[i]$. Normalization is done over the total area covered by the infection throughout the entire zebrafish larva region: $CSn[i] = CS[i]/CS$.

From the graph in Figure 5.9 it seems that the mean and the distribution are similar for some regions and different for others; i.e., 1, 4, 5 seem to have a very different mean.

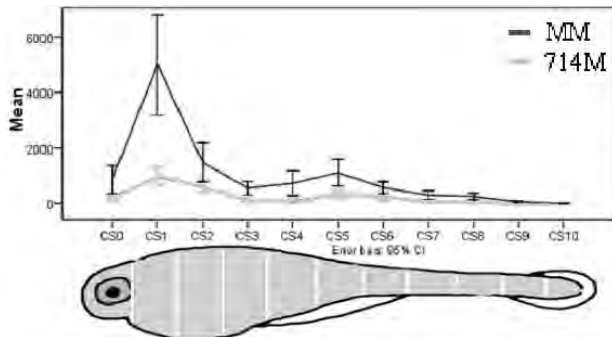


Figure 5.8: Spatial comparison of the average $CS[i]$ for MM and $714M$ in comparison to the zebrafish template.

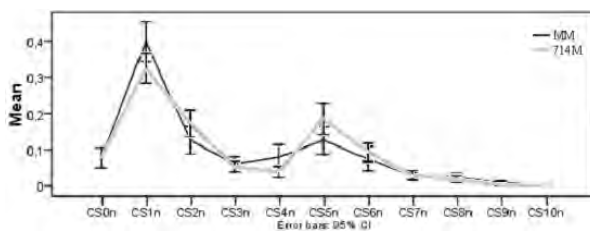


Figure 5.9: Spatial comparison of the normalized average total cluster size for MM and $714M$.

Region	H_0 Var	p value	
$CSn[0]$	=	0.266	*
$CSn[1]$	=	0.807	*
$CSn[2]$	=	0.025	
$CSn[3]$	=	0.690	*
$CSn[4]$	=	0.001	
$CSn[5]$	=	0.002	
$CSn[6]$	=	0.388	*
$CSn[7]$	=	0.255	*
$CSn[8]$	=	0.907	*
$CSn[9]$	=	0.036	
$CSn[10]$	=	0.274	*

Table 5.4: Levene’s test for equality of variances for total normalized Cluster Size $CSn[i]$, i is the region id.

Again, our null hypothesis states that, under the assumption that the two groups are independent, the variances are equal. We apply Levene’s Test for Equality of Variances to the $CS[i]$, $0 < i < 10$. The results are shown in Table 5.4.

For regions 0, 1, 3, 6, 7, 8 and 10 the significance is > 0.05 and thus our hypothesis is accepted. In Figure 5.9, it can be seen that the differences in the mean are considerable though the variances for MM and $714M$ remain the same.

To determine the difference of the mean values we use the knowledge gained from the Levene’s test to do the independent samples t-test. The results are shown in Table 5.5.

We observe there is a significant difference in the mean value for regions 1, 4 and 5 (marked with * in Table 5.5). This result is in correspondence with the initial observation of Figure 5.9 in which these means seemed rather different. The preliminary conclusions from these findings are the following. The larger part of the infection migrates towards the head of the zebrafish. The second largest part of infection, however, remains at the injection site. In wild-type infected fish (MM), a larger proportion of clusters is located in the head compared to the $714M$ mutant; i.e., significant difference of the mean while the variance is equal.

5.2.5 Conclusions

We have used a novel framework for automated granuloma cluster recognition in order to analyze the spatial distribution in zebrafish larvae. As a proof of concept we have analyzed the data for the zebrafish larva infected with the wild-type *Mycobacterium marinum* and a MM mutant $714M$.

From a statistical analysis of the data we can derive information on the spread of granulomas. In fish infected with the wild-type *Mycobacterium marinum* a higher amount of granuloma clusters is found. However, if we look at the normalized spread of infection it behaves approximately the same; it either stays at the site of the injection or it moves

Region	Var	t	Sig. (2t)	Mean Diff	Std. Err Diff	
<i>CSn</i> [0]	=	0.016	0.987	0.000	0.021	
<i>CSn</i> [1]	=	2.065	0.040	0.073	0.035	*
<i>CSn</i> [2]	∩	-1.739	0.084	-0.046	0.027	
<i>CSn</i> [3]	=	0.517	0.606	0.006	0.012	
<i>CSn</i> [4]	∩	2.176	0.032	0.041	0.019	*
<i>CSn</i> [5]	∩	-2.102	0.037	-0.060	0.029	*
<i>CSn</i> [6]	=	-0.850	0.397	-0.018	0.022	
<i>CSn</i> [7]	=	-0.305	0.761	-0.003	0.009	
<i>CSn</i> [8]	=	0.558	0.578	0.004	0.007	
<i>CSn</i> [9]	∩	1.443	0.151	0.004	0.003	
<i>CSn</i> [10]	=	-0.528	0.598	0.000	0.001	

Table 5.5: t-test for Equality of Means for the normalized Cluster Size feature. Significant difference in the mean value is marked with *.

towards the head of the larva. For the wild-type *Mycobacterium marinum* it seems that the infection is more likely to migrate towards the head compared to the 714 mutant; in the 714 mutant it is established that the majority of the infection remains located at the injection site. The percentage of the amount of clusters per region is distributed approximately in the same way for both groups in this test.

In the following sections this approach is further elaborated with more mutants and a larger dataset. Moreover, other measurement parameters are to be considered in the analysis. Large volumes of data will allow to do predictions from the measurements using machine learning approaches.

5.3 Analysis of High Throughput Zebrafish Infections Screens Based on Approach 3

In the previous section we have analyzed the infection spread within the zebrafish larvae. In this section a pilot analysis is shown for Approach 3 that is described in Chapter 3. We repeated the experiment of Section 5.2. Again, we have chosen mutant 714 (*714M*), as it is reported as one of the 30 mutants which does not make the fish ill. As a result of this study we could, indeed, find certain characteristic patterns in granuloma cluster spread and size for both *MM* and *714M*. The acquisition is accomplished through imaging. For each zebrafish, a brightfield and fluorescence microscopy image is acquired (Figure 5.1 and 5.2). Until recently, these images were analyzed manually. The analysis included localization of the zebrafish shape in a brightfield image and qualitative estimation of the granuloma cluster size and spread from a paired fluorescent image. The fluorescent signal is often of low intensity and noisy and therefore manual analysis is difficult and will not result in objective evaluation.

We have designed and implemented a framework for automated shape retrieval and

cluster analysis [43] (cf. Chapter 2). The recognition algorithm was based on a deformable template matching [23, 16] approach that distributes the zebrafish shape into vertical sub regions (slices). Besides our own work this framework has also been applied in large scale applications [54]. However, this algorithm used for zebrafish detection had certain drawbacks: the algorithm was slow due to high complexity of template matching [23]; the slices were fixed-width vertical regions that did not discriminate between the top and the bottom of the shape.

The results obtained with Chapter 2 have resulted in the development of another algorithm based on Anchor Regions. This is described in Chapter 3. By the use of this algorithm it is possible to distribute the zebrafish shapes into certain characteristic (Figure 5.10) but not size fixed regions. This algorithm is more true to nature, it can better deal with the large biological variation in shape and size of the zebrafish larvae (Figure 3.1).

Brightfield Imaging: Shape Localization and Annotation

In our previous analysis of infection spread we used *Approach 2* as described in Chapter 2 to localize the zebrafish shape and divide it into 11 regions counting from head to tail. This allowed for annotating the shape as well as doing spatial analysis. For this study we have abandoned this approach and developed a new template matching algorithm that localizes the zebrafish shapes without the computationally complex sub-region search approach.

This novel algorithm is based on mathematical morphology and probability theory. The algorithm estimates the location of the head, body and tail region of the zebrafish larvae and its orientation. This is done by distributing the template into possible sub templates, that are representing these regions. A probability tree is created to find an optimal shape resembling a zebrafish, cf. Chapter 3. Due to the rotation invariance of the algorithm the zebrafish larvae do not need to be aligned in a certain way. In Figure 3.14 the example of an output of the novel algorithm is shown.

Fluorescent Imaging: Analysis

The fluorescent signal is analyzed per fish mask, as it was found in the previous step. No granuloma formation is present at the not infected group *NI*, therefore this group was only used to obtain an intensity value that represents the noise level threshold n . All signal below n is considered noise, while signal above n represents granuloma presence.

The larva is modeled in 5 regions, listed in Table 5.6. The division into regions is an estimation and is proposed based on possible regions of interest. The detection of the regions is done in an automated fashion by *Approach 3* (by choosing a prototype template that consists of head, body and tail regions).

This result is presented in an infographic in Figure 5.10. The use of this kind of infographics is to illustrate and summarize the data gained from the statistical analysis in a very clear way. This way of data visualization will be used throughout this chapter.

Area Name	Notation
HEAD	H
BODY TOP	BT
BODY BOTTOM	BB
TAIL TOP	TT
TAIL BOTTOM	TB

Table 5.6: Automatic region assignment.

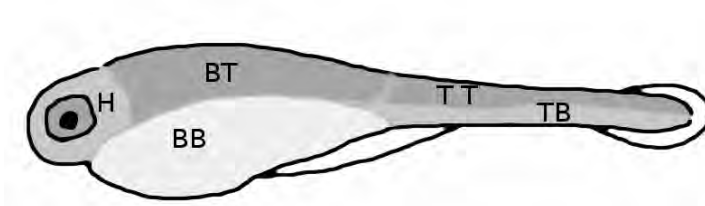


Figure 5.10: Graphical representation of the separation of a zebrafish into distinguished regions.

Output and Dataset Creation

Output is created in the form of an overlay image as in Figure 3.14, containing the retrieved zebrafish shape with the infection in the overlay and an entry in the *csv* file containing the granuloma spread in each of the regions; i.e., H , BT , BB , TT , TB , cf. Table 5.6. The granuloma spread for each region i is described by the amount of Clusters in a region, $CC[i]$, Definition 3, and the total amount of Infection in each region, $CS[i]$, Definition 5. Hence, we introduce the feature *Average Cluster Size*:

Definition 7. *Average Cluster Size [region]* is the average area covered by the infection in a single cluster at region i , denoted by $ACS[i]$. The area is expressed in units that represent the amount of pixels covering the region of interest within in the input images. The image pixels are classified as *infection* in an automated fashion as described in 4.3.2. Background pixels are automatically excluded from the analysis. Per region *Average Cluster Size [region]* is calculated by: $ACS[i] = CS[i]/CC[i]$.

5.3.1 Analysis and Results

In our study we set out to analyze the relationship between mutant and wild-type in the amount of clusters, spatial distribution and the cluster size. Note, that it is impossible to directly compare the amount of infection between different regions, since the regions are of different size. However it is possible to compare each region between the wild-type and the 714 mutant.

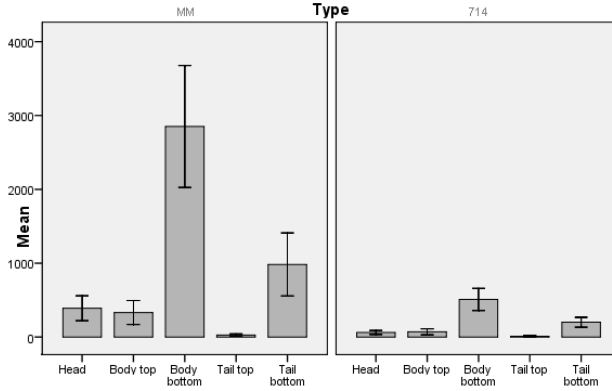


Figure 5.11: Spatial comparison of the mean amount of infection for MM and 714M in relation to the zebrafish regions.

Amount of Infection

First we compare the Cluster Size, $CS[i]$, Definition 5, measured as a numerical density in each region, compared between MM and $714M$ (cf. Figure 5.11).

At this point this distribution is not conclusive. This is due to the fact that MM make the fish more ill and thus overall produces more granulomas compared to $714M$. The mean amount of infection throughout the entire fish is 4558.24 (standard deviation $\sigma M = 569.51$) for MM and 825.69 ($\sigma M = 102.74$) for $714M$. The amount of infection in wild-type (MM) is roughly 5 times higher compared to $714M$.

In order to analyze the different batches in the same way we consider the normalized Cluster Size feature, $CSn[i]$, Definition 6. The normalization is done for each individual case and subsequently the mean is calculated. In Figure 5.12 the results are depicted.

From the graph we can observe that MM and $714M$ have approximately the same behavior. The mean and the 95% confidence interval suggest, that the two distributions in H , BT and TT can be considered similar. In BB (body bottom) and TB (tail bottom) the distribution behaves differently. Let us consider this in more detail. Our null hypothesis states that, under the assumption that the two groups are independent, their variances are equal. We therefore apply Levene's Test for Equality of Variances to the $CSn[i]$, $i \in \{H, BT, BB, TT, TB\}$, followed by the independent samples t-test. Based on the results from Levene's test we know which variances significantly differ. We observe that there is a significant difference for the total amount of infection (respectively $Sig.2t = 0.011$ and $Sig.2t = 0.018$) in the mean value for regions BB and TB (variance not equal).

Amount of Granuloma Clusters

We compare the average number of clusters, Definition 3, (variable $CC[]$) between MM and $714M$ (cf. Figure 5.13).

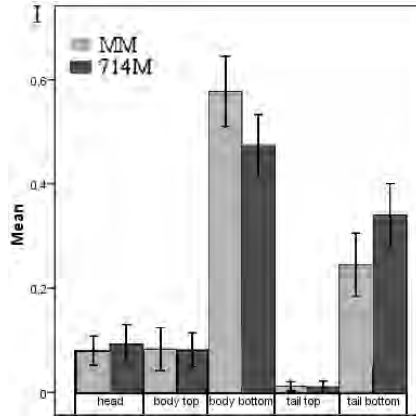


Figure 5.12: Comparison of spatial distribution of the normalized mean amount of infection ($CSn[i]$) for MM and $714M$ in relation to the zebrafish regions with a 95% confidence interval.

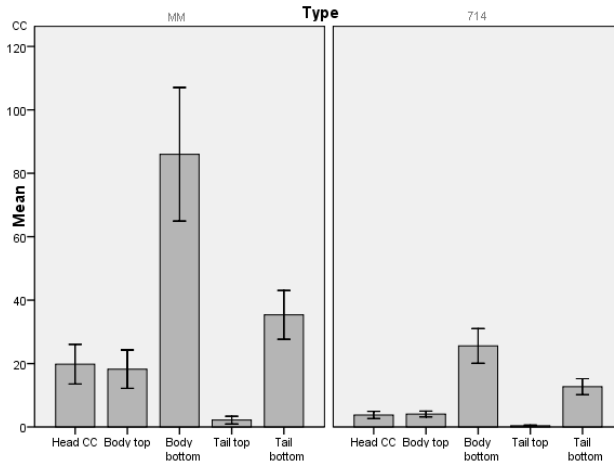


Figure 5.13: Comparison of spatial distribution of the mean amount of clusters for MM and $714M$ in relation to the zebrafish regions.

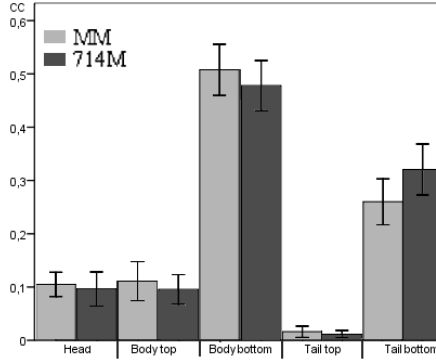


Figure 5.14: Comparison of spatial distribution of the normalized mean amount of clusters for *MM* and *714M* in relation to the zebrafish regions.

Throughout the entire fish for *MM* the $CC[i]$ is higher than for *714M*. *MM* has a mean amount of 161.94 ($\sigma M = 18.00$) throughout the entire fish and *714M* has a mean amount of 46.40 ($\sigma M = 4.13$). The amount of clusters in *MM* is about 3 times higher than in *714M*. This makes it difficult to compare the relative behavior of granuloma clusters. Therefore let us consider the normalized clusters amount $CCn[i]$, Definition 4, in Figure 5.14.

From the graph we can observe that *MM* and *714M* have approximately the same behavior. We analyze the graph in more detail by testing equality of variances and mean equality of each region, in the same way as described previously and find no significant difference in the mean values.

Average Size of Granuloma Clusters

We also look at the average size of the granuloma clusters, $ACS[i]$, Definition 7. Figure 5.15 shows the average cluster sizes for different regions of the zebrafish. We analyze the graph in more detail by testing equality of variances and mean equality of each region, in the same way as described previously. We observe that there is a significant difference (respectively $Sig.t = 0.000$ and $Sig.t = 0.009$) in the mean value for *BB* and *TB* regions (variance not equal).

Results

In all regions of the zebrafish for *MM* the amount of infection is higher than the *714M* with larger granulomas.

The same percentage of the granulomas in both *MM* and *714M* are located in the head region. A larger percentage of the granulomas of the *MM* than *714M* are located in the body bottom region.

A higher percentage of the granulomas of the *714M* than *MM* are located in the tail bottom region. In Figure 5.16 a graphical representation of the form of an infographic

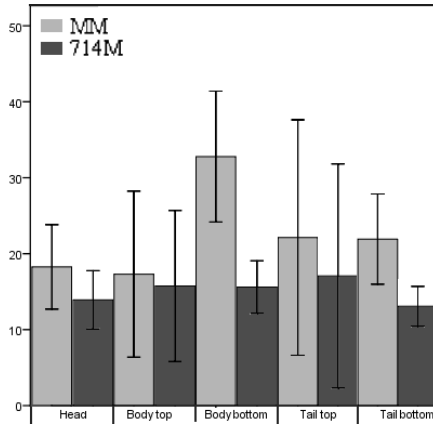


Figure 5.15: Comparison of spatial distribution of the average cluster sizes ($ACS[]$) for MM and $714M$ in relation to the zebrafish regions.

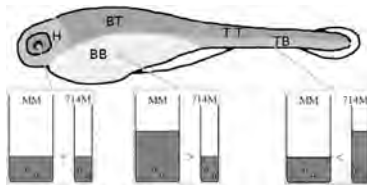


Figure 5.16: Graphical representation of the amount of infection. There is always more MM infection than $714M$; in the image this is illustrated by a wider bar. The percentages of their presence in certain regions are different; this trend is depicted through level height.

of these results is shown.

MM has a higher amount of clusters than $714M$ in all regions of the zebrafish (overall the amount of clusters is 3 times higher). There is no significant difference in cluster spread pattern for MM and $714M$.

In most regions the relative cluster sizes are not significantly different (though MM tends to have larger clusters). The exceptions are the body bottom and the tail bottom area, where MM gives significantly larger clusters than $714M$.

As a result of using *Approach 3* instead of *Approach 2* we could consider infection behavior in regions of interest. This resulted in more true to nature analysis that could be directly comprehended by the users.

5.4 In Depth Analysis of High Throughput Zebrafish Infections Screens

For a more in depth analysis of the data we consider a dataset that was created during 11 experiments performed on infection by different mutants of the *Mycobacterium marinum* strains visualized with fluorescence in *mCherry* red.

We analyze the data to find statistically relevant differences in behavior. We attempt to relate the result to biology. Analysis is performed on the same selection of mutants.

Mutants that were considered during the experiments were the following: 262, 308, 414, 415, 421, 423, 431, 730, 748, 801, 817, 885, 941, 943.

The following mutants were attenuated for early granuloma formation (infection levels were less than 50% of the E11 (wild-type) infection ([55, 64]): 262, 308, 421, 431, 730, 748 and 801 (cf. [55, 66]).

For mutants 748 and 801 it is known that both have a mutation in the genes from the ESX-1 region. Genes from the ESX-1 region are being knocked down in order to see which gene influence the virulence.

Mutant strains with knocked out genes that are involved in cell wall bio-synthesis (cf. [55], p64) are: 262, 431 and 730. Mutant strains 262 and 431 showed a substantial increased susceptibility to rifampicin (cf. [55, 65]). Other mutant strains are not specifically annotated and their specific functional profile is not known.

5.4.1 Experiments

The data that has been provided is separated into 11 experiments. Cases in each experiment are provided as images, i.e., a readout entails a brightfield image and a corresponding fluorescent image.

For retrieval, each experiment is identified by the user that took the images and a time stamp at the date the experiment was performed. The number of samples, i.e. zebrafish, that were exposed to the mutants differs per experiment.

5.4.2 Analysis and Results

We set out to analyze the difference in infectious behavior of the different mutants and attempt to discover distinguishable characteristics between the different mutants. We have established a more fine grained set of features that we measure; these are listed in Table 5.7.

5.4.3 Total amount of Infection

As a first comparison we consider the total area covered by the infection CS , Definition 5, measured for each mutant type.

Definition 8. *Cluster Size per mutant* is defined as average CS of all mutant m bacteria infected larva, denoted by $CSM[m]$.

Variable	Description
CS	Definition 5
CC	Definition 3
ClusterRegion	CS/CC (Average size of one cluster)
HClusterSize	Area covered by infection in the larva <i>Head</i> area
BClusterSize	Area covered by infection in the larva <i>Body</i> area
TClusterSize	Area covered by infection in the larva <i>Tail</i> area
HClusterSizeN	$HClusterSize/CS$
BClusterSizeN	$BClusterSize/CS$
TClusterSizeN	$TClusterSize/CS$
IPClusterSize	Area covered by infection in the larva <i>Injection Point</i> area
HRClusterSize	Area covered by infection in the larva <i>Heart</i> area
IPClusterSizeN	$IPClusterSize/CS$
HRClusterSizeN	$HRClusterSize/CS$

Table 5.7: Different variables that can be taken into account and their description.

It is preferred to compare each mutant infection pattern to the infection by the wild-type (*E11* or *MM*). We separate the data into groups, grouped by mutant id without considering the experiment number. The first step is to test each group for normality. Therefore we consider two well-known tests of normality, the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test. The sample size for each mutant is sometimes low (< 50), therefore the Shapiro-Wilk Test seems more appropriate, as it is typically suitable for handling both small and large sample sizes. We have tested the distributions per mutant for all experiments for being normally distributed by applying a Shapiro-Wilk Test. The results show that none of the mutant data was normally distributed. In Figure 5.17 this can be seen through inspection of the mean and the standard deviation. Explanation for the lack of not normally distributed values is the following: the amount of injected infection and imaging conditions were different for each experiment. A solution is to normalize the values per mutant per experiment.

The next step would be to normalize the groups one by one and then combine them back into one whole. The infection amount is normalized per experiment over the mean amount of infection of the wild-type in that experiment:

Definition 9. *Normalized Cluster Size per mutant* is the normalized $CSM[m]$ feature per mutant m bacteria infected larva, denoted by $CSMn[m]$. Normalization is done over the wild-type infected fish of a single experiment: $CSMn[m] = CSM[m]/CSM[MM]$.

The resulting dataset, again, contained no normally distributed variables. In order to be able to compare each mutant to the wild-type we checked the (not normal) distributions for similarity by the nonparametric Mann-Whitney test. Distributions similar to normalized E11 that were found were: 423, 885, 943. Suggested results are the following:

- Mutant 423 shows *no significant difference* in amount of infection as compared to the wild-type (according to a combination of all experiments).

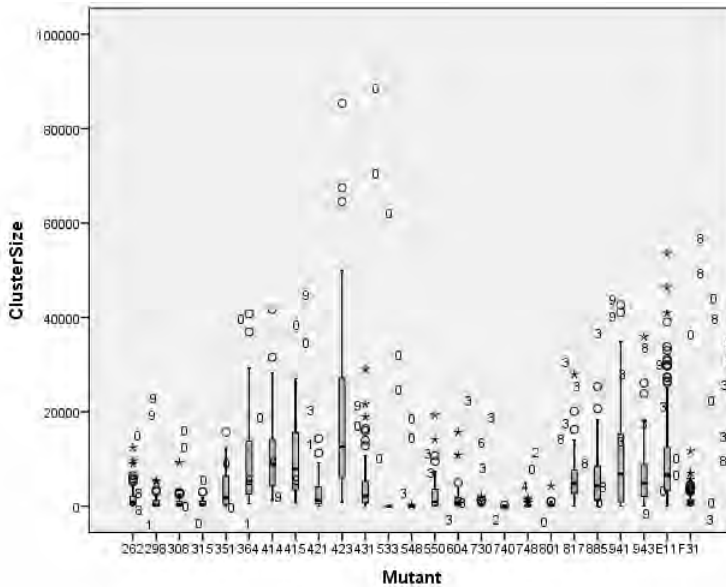


Figure 5.17: Graphical representation of the ClusterSize as it was measured for different mutants.

- Mutant 885 shows *no significant difference* in amount of infection as compared to the wild-type (according to a combination of all experiments).
- Mutant 943 shows *no significant difference* in amount of infection as compared to the wild-type (according to a combination of all experiments).

In order to be able to compare more mutants, we have attempted to separate the data into additional groups, namely clustered by mutant and experiment. To control if the data were normally distributed again, the Shapiro-Wilk Test was needed. Within the 11 experiments for the following mutants ClusterSize variables were normally distributed:

- Experiment 0: wild-type, 431, 748, 801, 941, 943
- Experiment 1: 423, 431
- Experiment 2: wild-type, 262, 941
- Experiment 3: wild-type, 421, 817, 943
- Experiment 5: wild-type, 421, 730, 817, 885, 941, 943
- Experiment 6: 262, 548, 730, 817, 885
- Experiment 7: wild-type, 943

- Experiment 8: wild-type, 423, 730, 817, 943
- Experiment 9: wild-type, 308, 414, 415
- Experiment 10: 298, 423, 941

In this manner the experiments where the infection pattern in wild-type is normally distributed can now be compared to the normally distributed mutant infection pattern. This was performed by Levene's test for variance followed by an Independent t-test for equality of means. With this test we could find the difference in infection and get an estimation of the difference.

For all the experiments where infection patterns were not normally distributed the Mann-Whitney U-test was used. This test could not be used to determine the actual difference of infection size, however it does indicate if the percentage of infection is significantly different. We use this test to support the results that were gained from experiments where the t-test could be used. The mutant strains which had no normally distributed data in any experiments are not considered. We also did not take into consideration those mutant infection patterns in which the amount of samples in an experiment was < 10 . The results of $CSMn[m]$ comparison are shown in Table 5.8.

It is difficult to extract direct conclusions from these results, since the conclusions presented here are only supported by limited experiments. However preliminary conclusions are shown in Table 5.9. For some mutant strains the results were not informative, as different experiments provided results that could not be integrated. Mutants that behaved in a similar way for different experiments were: 262, 308, 730, 748, 801 and 885.

Discussion

All considered, the mutant 885 behaves in the same way as the wild-type. No conclusive remarks can be stated for mutant 423 and 943, as in the current dataset the experiments provide contradictory results. A graphical representation of those results excluding the contradictions is shown in Table 5.10.

Mutant strains 748 and 801 are known to have a mutation in the genes from the ESX-1 region. For the ESX-1 it is known that the mutants are less virulent [55, 54].

In our results mutant strains 748 and 801 show a very large difference in infection percentage as compared to the wild-type (2% vs 100%). The low percentage of infection for both strains corroborates the known low virulence phenotype in zebrafish. We therefore assume that the results obtained with the other mutants can be used to determine their virulence phenotype.

Mutant strains 262 and 730 are gene knock-outs involved in cell wall bio-synthesis and show a similar behavior — that is a very large difference in infection percentage from the wild-type, respectively 25% and 6%. Since the knocked out genes are located at different locations in the pathway the difference in percentages suggests a different mechanism, yet similar effect.

5.4 In Depth Analysis of High Throughput Zebrafish Infections Screens

	0	1	2	3	4	5	6	7	8	9	10
262											
308											
414											
415											
421											
423											
431											
730											
748											
801											
817											
885											
941											
943											

Table 5.8: Variance and mean comparison between *MM* and each of the mutants. In cases where the t-test could be performed we also show the ratio of infection mean *mutant/MM*. For experiments which had a not normal distribution Mann-Whitney (*) is used to indicate if the percentage of infection is significantly the same or different.

5 Data and Pattern Analysis in Infection Studies in Zebrafish

Mutant	sig. diff	according to experiment	avg. percentage of MM infection
262	yes	2	25%
308	yes	9	11%
414	yes	10	
	no	0, 9	
415	yes	1	
	no	0, 9	
421	yes	3, 5, 7	
	no	2, 6	
423	yes	0, 1	
	no	8	
431	yes	1, 9, 10	
	no	0	
730	yes	5, 8	6%
748	yes	0	2%
801	yes	0	2%
817	yes	3, 7	
	no	5, 6, 8	
885	no		
941	yes	2, 9, 10	
	no	0, 3, 4, 5	
943	yes	3, 7	
	no	5, 6, 8	

Table 5.9: Experiment results: difference in amount of infection as compared to the wild-type throughout the entire fish.





	Mutant	Wild-type
262		
308		
730		
748		
801		
885		

Table 5.10: Difference in amount of infection as compared to the wild-type throughout the entire fish, depicted as bar height.

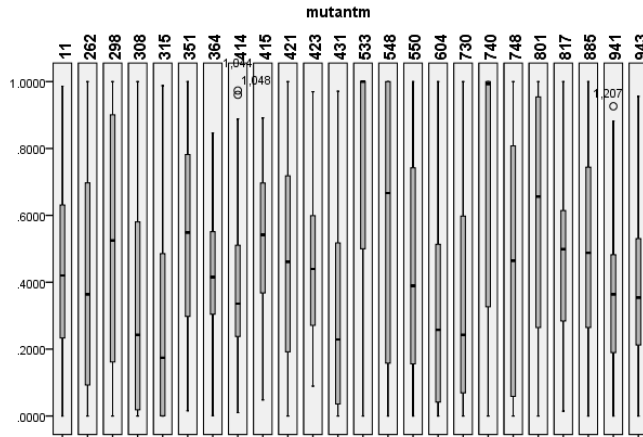


Figure 5.18: Box plot representation of the normalized cluster size CS_n for the *Body* region of the wild-type and the different mutants. A lot of difference can be seen in the behavior of the mutant strains.

5.4.4 Amount of Infection separated into Regions

We want to compare the amount of infection as it is present in the characteristic regions (*Head, Body, Tail, Heart Region, Injection Point*). Note that initially we consider the comparison of the amount of infection per region for different mutants (e.g., does the percentage of infection in the *Head* region differ for mutants A and B), instead of comparing the infection spread within each mutant. All cases can be compared in a correct way by normalizing the amount of infection in each region over the total infection amount in each fish. In this way the problem of a different total amount of infection is solved. For all objects the entire infection in each case is set to 1.0 and normalizes the amount of infection in each region. A normalized distribution of infection in the *body* area the fish is shown per mutant in Figure 5.18. We see a lot of difference in the behavior of the mutant strains, yet it needs to be proven to be statistically significant. We attempt to test the significance by an independent sample t test.

In order to perform this test the data under comparison needs to be normally distributed. Therefore we apply a Shapiro-Wilk test to test for a normal distribution. We have analyzed the following features:

- $CS_n(H)$ contained no normally distributed data for any of the mutants;
- $CS_n(B)$ contained normally distributed data for mutants: 351, 364, 414, 415, 423, 817, 943;
- $CS_n(T)$ contained normally distributed data for mutants: 351, 414;
- $CS_n(HR)$ contained normally distributed data for mutant 414;

5 Data and Pattern Analysis in Infection Studies in Zebrafish

- $CSn(IP)$ contained no normally distributed data for any of the mutants.
- Subsequently only the relationship between 351, 364, 414, 415, 423, 817, 943 in $CSn(B)$ and subsequently for 351, 414 in $CSn(T)$ could be considered for parametric statistical testing.

Our null hypothesis states that, under the assumption that the two groups are independent, their variances are equal. Therefore we need to test the variances first. This is done with Levene's test for equality of variances. If the variances are equal an independent samples t-test is performed in order to determine the equality of means. In Table 5.11 the results are shown.

The results of these tests are interpreted in the following way:

- Mutant 351 had statistically significant larger percentage of total infection ($53\% \pm 35\%$ of total infection) present at the *body area* as opposed to mutant 943 ($38\% \pm 27\%$ of infection)
- Mutant 415 had statistically significant larger percentage of total infection ($52\% \pm 20\%$ of total infection) present at the *body area* as opposed to mutant 414 ($39\% \pm 23\%$ of infection)
- Mutant 414 had statistically significant larger percentage of total infection ($39\% \pm 23\%$ of total infection) present at the *body area* as opposed to mutant 943 ($38\% \pm 27\%$ of infection)
- Mutant 817 had statistically significant larger percentage of total infection ($47\% \pm 27\%$ of total infection) present at the *body area* as opposed to mutant 943 ($38\% \pm 27\%$ of infection)

None of the variables of the wild-type were normally distributed. In order to be able to accomplish a comparison of the infection spread of each mutant infection and the wildtype we have chosen a different approach.

As an alternative for the parametric t-test, the Mann-Whitney U-test and the 2-sample Kolmogorov-Smirnov test are considered. For this test the variables do not need to be normally distributed. However the variables do need to follow an approximately same distribution and there is a loss of information, since values are presented as ranks. Mean equality can therefore not be predicted. This makes it a less powerful test than the t-test. We compare the control group MM to all the mutants and look for similar distributions. The null hypothesis is that the distributions of both groups are equal. The following distributions were found as the same. We used the Mann-Whitney U-test, since the amount of samples was fairly low and our data distribution is skewed (a lot of 0 values). We state H_0 : there is no difference in ratio of infection between the wildtype and mutant x . The results that reject H_0 ($p \leq 0.05$ and $|z| > 1.96$) are given in a graphical representation in Table 5.12, matching z and $p(2t)$ values for this table are given in Table 5.13.

Compare	Measure	Var	Means	N	Mean	σ	$\bar{\sigma}$
351/364	(B) <i>CSn</i>	!equal	no sig. diff				
351/414	(B) <i>CSn</i>	equal	no sig. diff				
351/415	(B) <i>CSn</i>	equal	no sig. diff				
351/423	(B) <i>CSn</i>	equal	no sig. diff				
351/817	(B) <i>CSn</i>	equal	no sig. diff				
351/943*	(B) <i>CSn</i>	equal	sig. diff	21/49	.533/.375	.292/.235	.063/.033
364/414	(B) <i>CSn</i>	equal	no sig. diff				
364/415	(B) <i>CSn</i>	equal	no sig. diff				
364/423	(B) <i>CSn</i>	equal	no sig. diff				
364/817	(B) <i>CSn</i>	equal	no sig. diff				
364/943	(B) <i>CSn</i>	equal	no sig. diff				
414/415*	(B) <i>CSn</i>	equal	sig. diff	33/30	.393/.526	.253/.205	.044/.037
414/423	(B) <i>CSn</i>	equal	no sig. diff				
414/817	(B) <i>CSn</i>	equal	no sig. diff				
414/943*	(B) <i>CSn</i>	equal	sig. diff	30/49	.526/.375	.205/.235	.0374/.033
415/423	(B) <i>CSn</i>	equal	no sig. diff				
415/817	(B) <i>CSn</i>	equal	no sig. diff				
415/943	(B) <i>CSn</i>	equal	no sig. diff				
423/817	(B) <i>CSn</i>	equal	no sig. diff				
423/943	(B) <i>CSn</i>	equal	no sig. diff				
817/943*	(B) <i>CSn</i>	equal	sig. diff	51/49	.473/.375	.237/.235	.033/.033
351/414	(T) <i>CSn</i>	equal	no sig. diff				

Table 5.11: Levene's test for equality of variances followed by the t-test for equality of means. Mutant combination with significantly different means are marked with a *.

5 Data and Pattern Analysis in Infection Studies in Zebrafish

Mut	Head	Body	Tail	Heart	Inj. Point
Location					
262					
298					
308					
315					
351					
415					
421					
431					
550					
604					
730					
740					
748					
801					
885					
941					

Table 5.12: Graphical representation of the percentage of infection measured in certain zebrafish regions. We compare each mutant (bars on the left of each comparison) to the wild-type (bars on the right of each comparison). Note, that not the amount of infection, but the percentages are compared.

5.4 In Depth Analysis of High Throughput Zebrafish Infections Screens

Mut	Head	Body	Tail	Heart	Inj. Point
262	$z = -2.105, p = .035$			$z = -2.342, p = .019$	
298			$z = -4.261, p = .000$		$z = -5.488, p = .000$
308		$z = -3.457, p = .001$		$z = -4.525, p = .000$	$z = -3.327, p = .001$
315	$z = -4.630, p = .010$	$z = -2.570, p = .000$	$z = -3.694, p = .000$	$z = -3.981, p = .000$	
351	$z = -2.145, p = .032$				
415		$z = -2.067, p = .039$			
421	$z = -2.660, p = .008$			$z = -3.462, p = .001$	$z = -2.484, p = .013$
431		$z = -2.791, p = .005$		$z = -4.100, p = .000$	
548	$z = -3.124, p = .002$		$z = -2.667, p = .000$	$z = -4.132, p = .008$	$z = -5.117, p = .000$
550			$z = -2.282, p = .022$		$z = -2.197, p = .028$
604	$z = -2.947, p = .003$	$z = -2.097, p = .036$	$z = -2.996, p = .003$	$z = -3.861, p = .000$	
730	$z = -2.175, p = .030$	$z = -2.102, p = .036$		$z = -3.405, p = .001$	
740	$z = -4.627, p = .000$	$z = -3.162, p = .002$	$z = -3.533, p = .000$	$z = -5.595, p = .000$	$z = -5.184, p = .000$
748	$z = -3.575, p = .000$	$z = -2.059, p = .039$		$z = -4.704, p = .000$	$z = -4.193, p = .000$
801	$z = -3.407, p = .001$	$z = -2.545, p = .011$	$z = -2.920, p = .003$	$z = -7.323, p = .000$	$z = -5.439, p = .000$
885	$z = -3.519, p = .000$			$z = -4.394, p = .000$	
941		$z = -2.186, p = .029$	$z = -2.826, p = .005$		

Table 5.13: Reported z and $p(2t)$ values of the Mann-Whitney U-test. We compare each mutant to the wild-type. Statistically significant differences are given here.

Discussion

As we have found earlier (Table 5.10) some mutants give large differences in infection amount throughout the entire fish, as compared to the wild-type. If we in addition consider the infection amounts within the regions, as presented in Table 5.12, we observe new patterns. For example, ESX-1 mutant strains 748 and 801, also on regional level, show similar behavior, namely for both lower percentage of infection is concentrated in the tail, head, injection point and heart region as compared to the wild-type. Mutant strains 262, 431 and 730 (genes knock outs which are involved in cell wall bio-synthesis) show a similar behavior at the injection point and heart region. All of them have a higher percentage of infection than the wild-type present in the heart and the same ratio as the wild-type at the injection point. Yet, the percentages of infection as measured over the head/body and tail regions do differ.

According to these results the following mutant strains behave the same as the wild-type when measured for the HR/IP sites: 351, 415, 941.

The measurements presented here can not be interpreted to directly compare the infection amount between regions in a single mutant strain. For example, although mutant strain 262 has a lower amount of infection as compared to the wild-type in *HR* and the same amount of infection as the wild-type in *IP*, which does not mean that for mutant 262 there is more infection in *IP* than in *HR*. The reason for it is that the results presented here are relative to the wild-type and the distribution of infection within the wild-type should be considered to make that comparison.

5.4.5 Infection Flow per Mutant Strain

For each mutant strain we want to investigate the difference in the response of the innate immune system in a comparison of different regions in a zebrafish embryo. For example, for a certain mutant strain, does the *heart* contain more infection then the *injection point*.

We use the same normalized dataset from the previous subsection. The data is not normally distributed and thus we use the Wilcoxon Signed-Rank Test instead of the *paired t*-test. We do this test only for the injection point *IP* and heart *HR* since the area they cover are approximately the same (while the head/body/tail region differs per individual case). Therefore no additional normalization over the region size needs to be performed.

Additionally these two regions are indeed dependent, since per fish the amount of total infection is always 1.0 and thus there is a connection between *HR* and *IP*, namely $HR + IP + rest = 1.0$. We run the Wilcoxon test on the data. Our H_0 states that there is no difference between the ratio of infection in *IP* and *HR*.

As a result we found that for most mutants the spread did not differ and therefore H_0 is not rejected. Significant difference was only found for mutants 298, 415, 550 and the wild-type. If we consider the median values of these mutants the following can be concluded:

- Mutants 298, 415, 550 and the wild-type had a significantly higher percentage of

Mut	Heart Region vs Injection Point
Wild-type	
298	
415	
550	
262	
308	
315	
351	
364	
414	
421	
423	
431	
604	
730	
748	
817	
885	
941	
943	

Table 5.14: Distribution of the percentage of infection within each individual, indicated by bar height. We consider the heart area and the injection point. The size of the areas is the same.

infection present in the *heart* as compared to the *injection point*.

- For other mutants there was no significant difference in the percentage of infection between the *heart* and the *injection point*.

A graphical representation of these results is shown in Table 5.14.

Discussion

Mutants 298, 415 and 550 seem to have the same flow of infection from the injection point to the heart. For all other mutant strains this mechanism works differently and the infection is evened out between the injection point and the heart.

We can see that the infection spread through the regions for the wild-type is: $HR > IP$ and therefore the results as they are presented earlier in Table 5.10 can indeed not be used for direct comparison of the distribution within a mutant strain.

In order to keep the statistics clear we have only compared regions of equal size to each other (HR and IP). Regions H , B , T have therefore not been analyzed in this experimental setup.

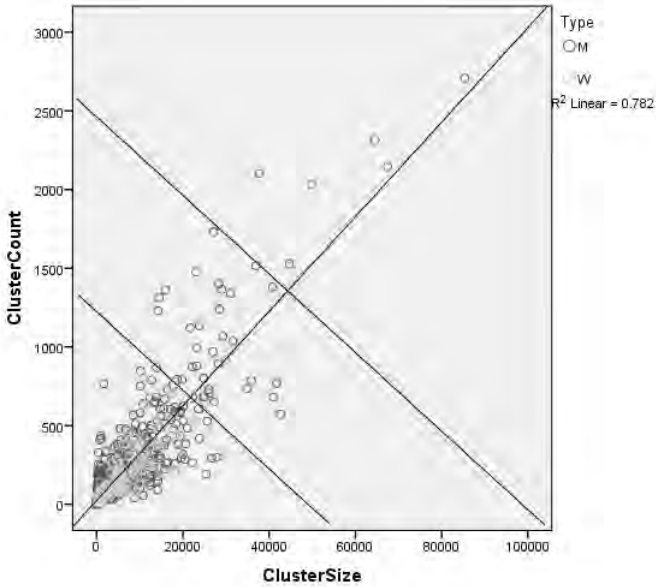


Figure 5.19: Scatter plot of the average cluster size. Spot location within each region does not provide locational information.

5.4.6 Average Size of Clusters

In our previous analysis we also took into consideration the normalized ClusterCount values. These values represent total infection area measured over all clusters at a certain location (cf. Section 5.2). Now we consider the average size of one cluster in a region i , $ACS[i]$ (Definition 7) separated over characteristic regions (*Head, Body, Tail, Heart Region, Injection Point*). Additionally we also consider the average size of one cluster within an entire fish.

Definition 10. *Average Cluster Size* is the average area covered by the infection in a single cluster, denoted by ACS . Area is expressed in units that represent the amount of pixels covering the entire object of interest within in the input images. The image pixels are classified as *infection* in an automated fashion as described in 4.3.2. Background pixels are automatically excluded from the analysis. *Average Cluster Size* is calculated by: $ACS = CS/CC$.

In Figure 5.19 we see that the correlation of the Cluster Size and Cluster Count shows a near linear increase. Pearson’s p -value is 0.000, indicating that this relationship is significant. Pearson’s correlation between the two groups is showing a significant positive correlation of 0.884, which illustrates that if the CS variable increases then CC also increases.

This also suggests that the interesting cases are the ones that do not follow this

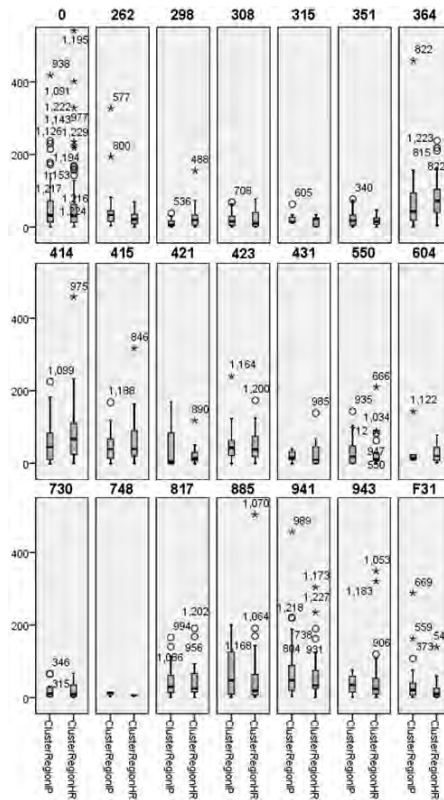


Figure 5.20: Box plot representation for the *ACS* is the injection point and the heart region for different mutants.

trend and are the outliers to the correlation regression (cf. Figure 5.19). A graphical rendition of the distribution for the injection point region and the heart region is shown in Figure 5.20.

We have performed normal distribution analysis on *ACS* for all the mutant strains and the regions *H*, *B*, *T*, *HR*, *IP*. From this approach we obtained the following outcome:

- *ACS* contained normally distributed data for mutants: 415, 885, 943;
- *ACS*[*H*] contained normally distributed data for mutant 351, 415, 548.
- *ACS*[*B*] contained normally distributed data for mutant 315, 604, 943;
- *ACS*[*T*] contained normally distributed data for mutant 423;
- *ACS*[*HR*] contained normally distributed data for mutants: 315, 351;
- *ACS*[*IP*] contained normally distributed data for mutants: 351, 415, 748, 943.

Compare	Measure	Var	Means	N	Mean	σ	$\bar{\sigma}$
415/885	<i>ACS</i>	\neg	\neg				
415/943*	<i>ACS</i>	=	=	30/49	36/27.79	18.85/14.98	3.4/2.14
885/943*	<i>ACS</i>	\neg	\neg	58/49	40/27.79	27.35/14.98	3.59/2.14
351/415*	<i>ACS[H]</i>	\neg	\neg	14/28	13.68/29.19	7.70/22.01	2.06/4.16
351/548	<i>ACS[H]</i>	=	=				
415/548*	<i>ACS[H]</i>	\neg	\neg	28/6	29.19/8.29	22.01/6.89	4.16/2.81
315/604	<i>ACS[H]</i>	=	=				
315/943	<i>ACS[B]</i>	=	=				
604/943	<i>ACS[B]</i>	=	=				
315/351	<i>ACS[HR]</i>	=	=				
351/415	<i>ACS[IP]</i>	=	=				
351/748*	<i>ACS[IP]</i>	\neg	\neg	14/14	28.07/7.94	25.68/5.72	6.80/1.53
351/943	<i>ACS[IP]</i>	=	=				
415/748*	<i>ACS[IP]</i>	\neg	\neg	24/14	50.33/7.94	41.76/5.72	8.52/1.53
415/943	<i>ACS[IP]</i>	\neg	=				
748/943*	<i>ACS[IP]</i>	\neg	\neg	14/38	7.94/37.58	5.72/23.85	1.53/3.87

Table 5.15: Levene’s test for equality of variances followed by the t-test for equality of means. Mutant combinations with significantly different means are marked with a *.

Our null hypothesis, again, states that, under the assumption that the two groups are independent, their variances are equal. The results are shown in Table 5.15. Note, that the wild-type is not present in this table, since it contained no normally distributed data for the *ACS*.

The results of these tests are interpreted in the following way (Arbitrary Units (AU) were measured in pixels):

- Mutant 415 resulted in statistically significant larger clusters (36 ± 22 AU) throughout the entire larva as opposed to mutant 943 (28 ± 17 AU).
- Mutant 885 resulted in statistically significant larger clusters (40 ± 31 AU) throughout the entire larva as opposed to mutant 943 (28 ± 17 AU).
- Mutant 415 resulted in statistically significant larger clusters (29 ± 26 AU) in the head region as opposed to mutant 351 (14 ± 10 AU).
- Mutant 415 resulted in statistically significant larger clusters (29 ± 26 AU) in the head region point as opposed to mutant 548 (8 ± 10 AU).
- Mutant 351 resulted in statistically significant larger clusters (28 ± 32 AU) near the injection point as opposed to mutant 748 (8 ± 7 AU).
- Mutant 415 resulted in statistically significant larger clusters (50 ± 50 AU) near the injection point as opposed to mutant 748 (8 ± 7 AU).








Location		Average size of a granuloma
entire larva	*	
entire larva	*	
head	**	
head	**	
injection point		
injection point	***	
injection point		

Table 5.16: Graphical representation of the average granuloma cluster size. Only comparisons of mutant infection strains that were found to be significant are shown. In each image the light gray represents the region where comparison is done. The dark grey circle depicts the standard deviation of the largest and the inscribed white circle (if present) the smallest granuloma size. The black line inside the circle represents the average size. It is used for reference.

- Mutant 943 resulted in statistically significant larger clusters (38 ± 28 AU) near the injection point as opposed to mutant 748 (8 ± 7 AU).

No other statistically significant relationships were found. A graphical representation of these results is given in Table 5.16.

Discussion

We would prefer to compare the mutant strains to the wild-type rather than to another mutant strain. However, since the wild-type was not present in the results we are considering mutants that mimic the behavior of the wild-type and that are present in the results of our analysis.

As we have found earlier (cf. Table 5.8) mutant strain 415 had approximately the same amount of infection according to experiments 0 and 9 as the wild-type as measured throughout the entire fish. This also applies for mutant 885 (cf. Table 5.8). We therefore use 415 and 885 to predict that the wild-type will have slightly larger clusters throughout the entire body as mutant 943 (cf. * in Table 5.16).

Strain 415 shares the same distribution of the infection between the *H* region as the wild-type (cf. Table 5.12). We therefore use 415 to predict that the wild-type will have larger clusters in its head as compared to the mutants 351 and 548 (cf. ** in Table 5.16).

Strain 415 shares the same distribution of the infection between the *IP* and *HR* region as the wild-type (cf. Table 5.14 and Table 5.12). We therefore use 415 to predict that the wild-type will have larger clusters at the injection point as compared to mutant 748 (cf. *** in Table 5.16).

In order to be able to do other comparisons with not normally distributed data we need to use an alternative for the parametric t-test. However, since the mean equality can not be predicted it is more difficult to say something about it. In the case of *ACS* we have chosen not to perform non-parametric testing as it is less powerful and makes reasoning about the results harder. However this test will be included in future work.

5.5 Conclusions

In this chapter we have described a recipe for the analysis of the data that is retrieved by our algorithm from Chapter 3 in order to retrieve interesting patterns. We are attacking the problem at hand in the following way.

- First, the measurements are collected and arranged into usable data.
- Second, the data is reduced by statistical analysis. During this step tradeoffs must be chosen between parametric and non parametric testing, the type of normalization and data clustering.
- Then the data can be reduced by the use of an infographic. Our infographics contain a lot of data, yet show it in a compact, understandable way. The meaning of the data can be illustrated much faster than from written statistical data. All the infographics used here are supported by statistical analysis. In this way the data can be presented to the user in a very clear way.