Proceedings of the 2015 International Conference on Machine Learning and Cybernetics, Guangzhou, 12-15 July, 2015

GENE REGULATORY EFFECTS INFERENCE FOR CELL FATE DETERMINATION BASED ON SINGLE-CELL RESOLUTION DATA

XIAO-TAI HUANG[1], LEANNE L. H. CHAN[1], ZHONG-YING ZHAO[2], HONG YAN[1]

[1]Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[2]Department of Biology, Faculty of Science, Hong Kong Baptist University, Kowloon, Hong Kong
E-MAIL: xthuang3@student.cityu.edu.hk, leanne.chan@cityu.edu.hk, zyzhao@hkbu.edu.hk, h.yan@cityu.edu.hk

Abstract:

Cell differentiation is a complicated biological process, involving lots of genes, which decide cell fate in early embryo development. These genes govern cell type formation (tissue formation). However, regulatory mechanism among these genes are not clear. In this paper, we infer gene regulatory effects for tissue formation to reveal gene regulatory mechanism based on state-of-the-art single-cell resolution data. Specifically, we disclose the gene regulatory mechanism about *pha-4* gene, a gene highly related to pharynx formation, in *Caenorhabditis elegans*. We quantify *pha-4* gene expression at cellular level to supervise the process of pharynx tissue development. Two types of data, wild-type data and mutant data, are exploited in our experiment. Wild-type data is *pha-4* gene expression data under normal development condition, while mutant data represents the gene expression data under abnormal development condition which means perturbation (gene knockdown) of some other genes. By comparing these two types of data based on two statistical hypothesis tests pipelines, gene regulatory effects have been investigated. In total, we have inferred gene knockdown effects of 579 genes. 74 of them possess either activating or inhibiting effect on *pha-4* gene. Finally, gene regulatory networks of pharynx tissue formation in *C. elegans* have been constructed based on these inferred gene effect results.

Keywords:

single-cell resolution data; cell lineage tracing; cell fate; gene expression; gene regulatory network

1. Introduction

In metazoan organism (multiple-cell organism), all cells are produced by cell divisions from zygote. These cells contain the same genome. However, the organism is generally constituted by several different types of cells for different tissues and organs. How are these different types of cells generated from the same genome? What is the mechanism that governs this automatic self-controlling system? The answer to these questions is cell differentiation which is an important biological process for tissues and organs formation. In early embryo, there is a type of cell named founder cell whose cell fate has been determined during cell differentiation. Descendant cells generated from different founder cells will eventually develop into different tissues or organs based on their founder cells' fate. Fate determination is closely related to the expression of specific sets of genes. In biological systems, expression of one gene is regulated by other genes. These genes regulate each other to form gene regulatory network which is fundamental for cell differentiation. Therefore, it is necessary to study gene regulatory networks for different tissues and organs formation.

Several cell differentiation gene regulatory networks have been reconstructed, such as human plasma cell differentiation [1], transcriptional networks in human blood stem [2], gene regulatory networks governing pancreas development [3], a regulatory network in *Caenorhabditis elegans* muscle and skin development in C lineage [4] and an inferred mechanistic network model for *C. elegans* cell development [5]. Most of those works in human research have not precisely detected gene's function at signal-cell. Brandilyn Stigler et al. only focused on C lineage in *C. elegans* and detected small part of cells in *C. elegans* embryo [4]. A recent research from Zhuo Du et al. used a single-cell resolution method to investigate changes of phenotype at 350-cell stage by RNA interference (RNAi) knockdown 20 genes [5]. They proposed an inferred mechanistic network which contains information of both gene interaction and cell division. However, the scale of their knockdown gene set is small compared to other developmental important genes in *C. elegans*. Here, we aim to reconstruct gene regulatory networks for *C. elegans* pharynx organ differentiation in different founder cells. We exploit state-of-the-art single-cell resolution data to detect pharynx formation related genes in *C. elegans*.

978-1-4673-7220-6/15/$31.00 ©2015 IEEE

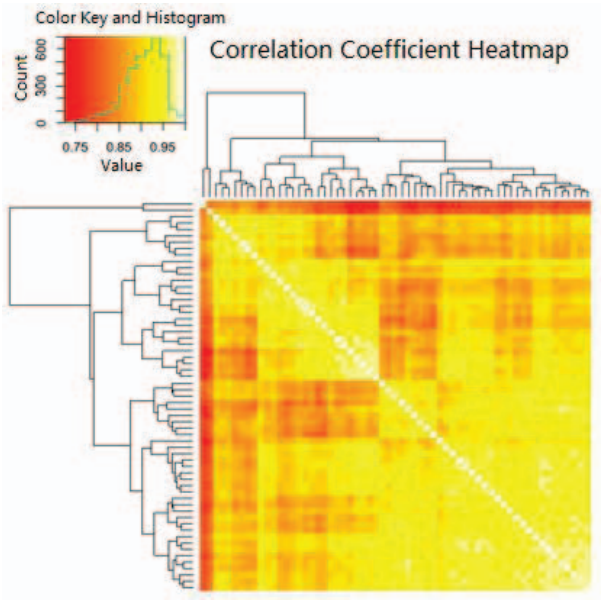

FIGURE 1. Correlation coefficient heatmap for 61 wild-type replicates.

2. Methods

2.1. Data acquisition

Single-cell resolution data is generated by using automated cell lineage tracing system [6]. In this system, it captures images from camera, to automatically trace *C. elegans* cells from one-cell stage (the beginning of embryo development) to 350-cell stage (the ending of embryo development) in *C. elegans* embryo. It identifies cells and tracks them over time from images of the embryo. The images are taken by multiple layers of the embryo for the purpose of capture all cells within the embryo. Besides tracking cells, it can also measure gene expression of specific interested gene, called marker gene, based on fluorescence intensity, at cellular level in the system. Because the data acquired by this system involves every single cells, it is called single-cell resolution data. The data, actually *pha-4* gene expression data, include two parts: wild-type data and mutant data. Wild-type data is from the experiment conducted under normal development condition. On the other hand, mutant data is acquired from RNAi single-gene knockdown experiment which means that perturbation of other single gene leads to abnormal development of the embryo. Details of experimental procedures can be found in [5, 7-8].

pha-4 gene is highly related to the development of pharynx. Cells express *pha-4* gene whose descendant cells will finally develop into pharynx cells. *pha-4* gene expression is measured for every cell per 1.5 minutes during embryo development. In this period, totally about 700 cells are measured both in wild-type and mutant experiments. Wild-type experiment is repeated 61 times which indicates the measurement of *pha-4* gene expression is conducted in 61 normal *C. elegans* embryos. In mutant experiments, totally, 579 genes are knocked down respectively in each single-gene RNAi experiment. For one specific gene knockdown experiment, it is repeated by 2-5 times which means the same specific gene is knocked down in 2-5 embryos. Therefore, for the total 700 cells in *C. elegans* embryo, each cell is with 61 sets of wild-type data and with 2-5 sets of mutant data per specific knockdown gene. The number of mutant data sets is much smaller than the one in wild-type, because RNAi single-gene knockdown experiment is time-consuming and expensive.

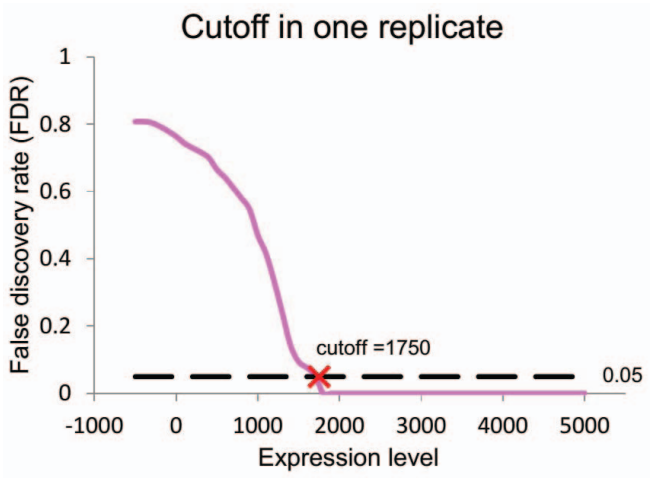

FIGURE 2. Determination of binary cutoff of marker expression in one replicate.

2.2. Data preprocessing

As we described in section 2.1, the data, both in wild-type and mutant, are gene expression value in every cell at every time point (per 1.5 minutes). For one cell, an average gene expression value is computed from every time point expression value to indicate an overall gene expression in the cell. Therefore, in one set of data, one cell is with one gene expression value for *pha-4* gene. Then, all gene expression values are normalized into a range from -500 to 5000 in one set of data. The larger this value, the higher gene expression is.

To examine the reproducibility in the data, correlation coefficients are computed in each pair of wild-type data sets, shown in Figure 1. In Figure 1, top two wild-type

Figure 4. *pha-4* marker gene expression in wild-type and mutant in every cell in *C. elegans*. Horizontal axis is all cells in cell lineage.

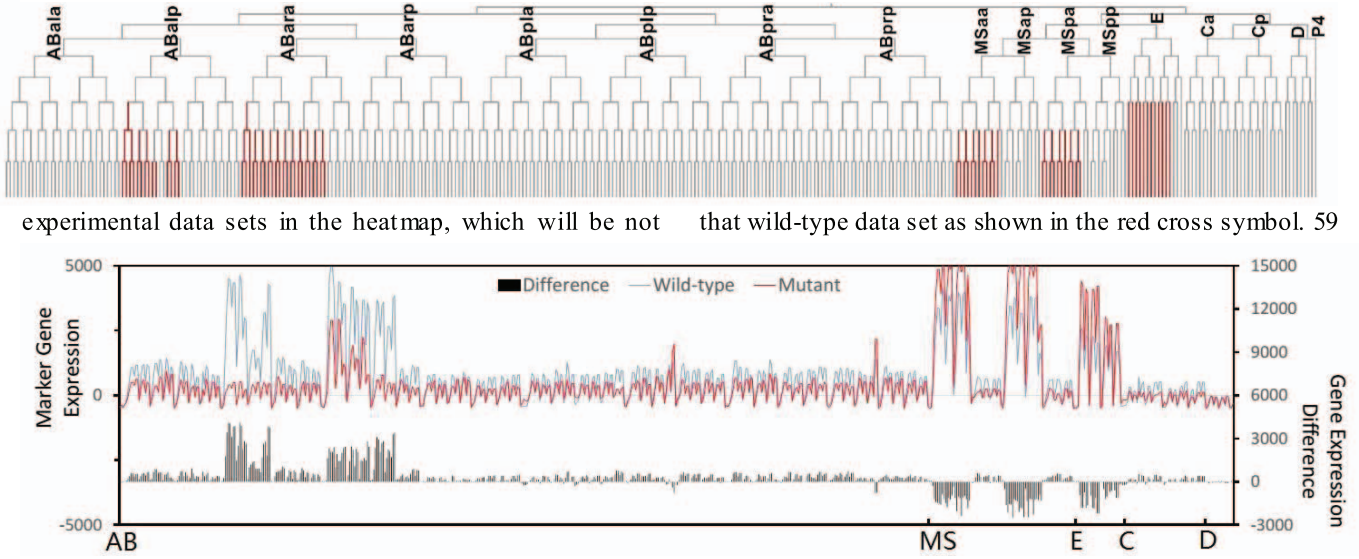

FIGURE 3. Binarized *pha-4* marker gene expression in cell lineage tree in wild-type *C. elegans*.

used in subsequent study, have a low correlation coefficient (< 0.8) due to variation of experiment conditions such as temperature. The correlation coefficients of remaining 59 experimental data sets are relatively high (> 0.9). Therefore, only those 59 remaining wild-type data sets are analyzed subsequently.

2.3. Identification of *pha-4* expressing cells

Due to variation of gene expression value, the value is binarized to control quantitative variability. A list of ground truth *pha-4* expressing cells (152 cells) is used as reference. The ground truth cells include terminal cells and their ancestors which will eventually develop into pharynx in wild-type based on previous literature [9]. A false discovery rate (FDR) of 0.05 is used to determine the *pha-4* gene expression level cutoff which can binarize gene expression [5]. The FDR control is a statistical method to control the expected rate of incorrectly rejecting null hypothesis. Here, in our study, the FDR is computed as formula (1).

$$\mathrm{FDR} = \frac{\text{the number of false pha4 expressing cells discovered by setting specific cutoff}}{\text{the number of pha4 expressing cells discovered by setting specific cutoff}} \quad (1)$$

The cutoff is initialized as -500. Then increase cutoff value and compute FDR every time. Iterate this step until the FDR equals to 0.05. The cutoff whose FDR equals to 0.05 is determined as the *pha-4* gene expression level cutoff. For example in Figure 2, the cutoff is determined as 1750 in that wild-type data set as shown in the red cross symbol. 59 cutoffs are determined in 59 wild-type data sets. Then, we average these 59 cutoffs as final *pha-4* gene expression level cutoff. By using this final cutoff, *pha-4* gene expression level is binarized into expressing or not expressing in one cell. Expression value greater than the cutoff is considered *pha-4* gene expressing in the cell. A wild-type cell lineage tree with *pha-4* gene expressing cells (in red color) is shown in Figure 3. The 'P0' is the original cell in one embryo of *C. elegans*. As development, from top to bottom, descendant cells are divided and differentiated into different types of tissues and organs. Red cells are expressing *pha-4* gene. Those *pha-4* gene expressing cells will eventually develop into pharynx.

2.4. Infer knockdown gene effect in single cell

In each mutant experiment, single gene is knocked down. Among 579 knockdown genes, some of them can not affect *pha-4* gene expression at cellular level, while others may have either activating or inhibiting effect. For example, the difference of *pha-4* gene expression between wild-type and mutant (*lin-12* gene knockdown) data is shown in Figure 4. Horizontal axis is all cells in cell lineage. They are arranged from left to right. Symbol 'AB' indicates a start in AB cell lineage. The cells on its right are belong to AB cell lineage until next symbol 'MS'. MS cell lineage is from symbol 'MS' to symbol 'E' in x axis. The rest E, C and D cell lineage is in the same manner. The top line chart, combining with left vertical axis, shows *pha-4* gene expression among different cells both in wild-type and

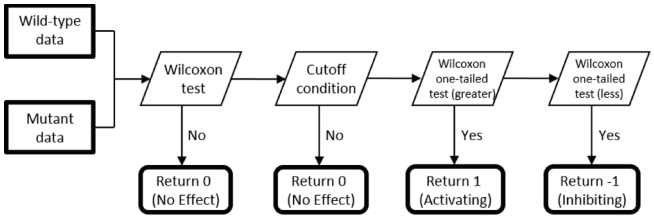

FIGURE 5. A pipeline for knockdown gene effect inference in single cell.

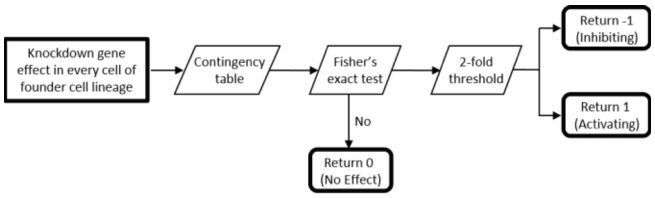

FIGURE 6. A pipeline for knockdown gene effect inference in a founder cell.

mutant (lin-12 gene knockdown) data set. The bottom bar graph, combining with right vertical axis, shows the difference of *pha-4* gene expression between wild-type and mutant (wild-type data minus mutant data) data set. When *lin-12* gene is knocked down in mutant, the *pha-4* gene expression is decreased in some cells of AB lineage. However, in E lineage and part of MS lineage, the *pha-4* gene expression is increased in some cells. This indicates knockdown gene has different effect on *pha-4* gene in different cells. To detect knockdown gene's effect on *pha-4* gene at cellular level, wild-type and mutant data in single cell are compared through following pipeline based on statistical hypothesis tests for their significant difference, shown in Figure 5. Since sample size in our mutant data sets is 2-5 (very small), traditional parametric tests are not appropriate with this extreme case [10]. Thus, we use a nonparametric method, Wilcoxon rank-sum test (Mann–Whitney U test). In one cell, a statistical difference between wild-type and mutant is tested with a p-value. If the p-value is not significant ($p > 0.05$), it indicates the knockdown gene cannot affect *pha-4* gene expression in the cell. On the other hand, we infer the knockdown gene having effect on *pha-4* gene when satisfying two conditions: 1) Significant difference between wild-type and mutant data; 2) Between average of wild-type data and average of mutant data, one of them should be greater than the cutoff (determined in section 2.3 previously), while the other should be less than the cutoff. Based on these two conditions, it can avoid two false positive cases: a. the *pha-4* gene is expressing both in wild-type and mutant; b. the *pha-4* gene is not expressing both in wild-type and mutant. That means we only consider those knockdown genes which can alter expressing state of *pha-4* gene in the mutant cell. Then, wild-type and mutant data are tested by Wilcoxon one-tailed test to detect which one is greater with a significant p-value. This can indicate the knockdown gene's effect (activating or inhibiting) on *pha-4* gene in the cell, shown in the pipeline in Figure 5. Finally, this statistical hypothesis tests pipeline is applied in every cell to infer knockdown genes' effect at a cellular level.

2.5. Infer knockdown gene effect in founder cell

During cell differentiation, the fate of all terminal cells (the cells generated at the end of embryonic period) have been determined in their ancestral founder cells previously [11]. Different sets of genes are responsible for specific cell fate determination. To decipher mechanism for different tissues and organs formation, it is necessary to detect those cell fate determination genes for different founder cells. However, most of those genes are expressed in descendant cells rather than in ancestral cells (founder cells) directly, such as *pha-4* gene in this study, see Figure 3 and [8]. Therefore, it is not sufficient only using founder cell data. Here, we exploit data both in the founder cell and its descendant cells. We select 17 founder cells based on previous study [5]. The names of founder cells are shown on the cell lineage tree in vertical direction in Figure 3, for example, 'ABala', 'MSaa', 'E' and so on.

The results of knockdown gene effect on every cell from previous section 2.4 are used here for inferring knockdown gene effect in one founder cell. A statistical hypothesis tests pipeline is proposed in Figure 6. Firstly, a 2×2 contingency table is constructed based on *pha-4* gene effect in the founder cell and its descendant cells. The two columns are the number of cells having significant knockdown gene effect and no effect respectively in the founder cell. The two rows are in wild-type and mutant respectively. Then Fisher's exact test [12] is performed in the contingency table to test whether this knockdown gene has a significant effect on the *pha-4* gene in the founder cell. If the p-value is significant (<0.05), a 2-fold threshold is applied to decide whether it is an activating or inhibiting effect. In other words, under the condition of significant p-value, if the number of activating effect cells is larger than 2-fold of inhibiting number, it indicates that this knockdown gene has an activating effect on *pha-4* gene in the founder cell, and vice versa. By applying this pipeline, effects of totally 579 knockdown genes are inferred in 17 founder cells.

3. Results and Discussions

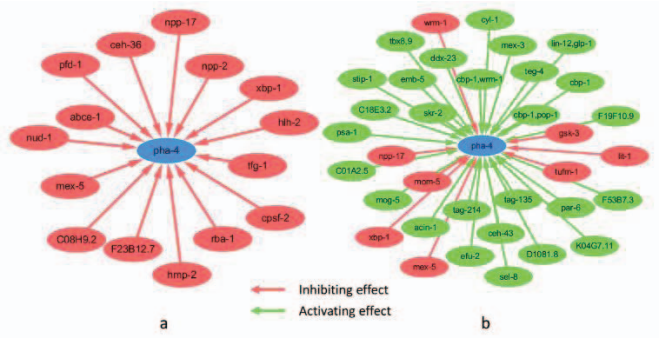

FIGURE 7. a. Inferred gene regulatory network for pha-4 gene in ABala founder cell. b. Inferred gene regulatory network for pha-4 gene in ABalp founder cell.

Based on data preprocessing, 59 wild-type data sets with a high reproducibility are used in gene effect inference. By controlling false discovery rate (FDR), a cutoff of 1732.37 is determined to binarize *pha-4* gene expression level. This cutoff is also used in subsequent knockdown gene effect inference. By using methods in the pipeline in Figure 5, 579 knockdown genes' effects are inferred in every cell in *C. elegans* cell lineage. The effects include activating effect, inhibiting effect or no effect. This information are combined with methods in another pipeline in Figure 6. Finally, effects of 579 knockdown genes on *pha-4* gene are inferred in 17 selected founder cells. Among them, 74 genes have either activating or inhibiting effect on *pha-4* gene in different founder cells. In other words, these 74 genes regulate expression of *pha-4* gene in founder cells. They are responsible for pharynx formation in specific founder cell. Moreover, they can also restrict some founder cells to develop into pharynx due to inhibiting regulation from part of 74 genes. Here, two example of constructed gene regulatory networks of *pha-4* gene in ABala and ABalp founder cells are shown in the Figure 7a and Figure 7b, respectively. 15 genes possess inhibiting effect on *pha-4* gene in ABala in Figure 7a, which are responsible to inhibit pharynx formation in this founder cell. On the other hand, in Figure 7b, 27 activating genes and 8 inhibiting genes cooperate to control *pha-4* gene expression in ABalp to facilitate to develop into pharynx. The number of activating and inhibiting genes among different founder cells are shown in Table 1. *pha-4* gene is expressed in the progenies of founder cells 'ABalp', 'ABara', 'MSaa', 'MSpa' and 'E', in shadow in Table 1. It is clear that more activating gene than inhibiting gene in these founder cells whose fate is to develop into pharynx. However, in other founder cells, there is no activating gene, while some inhibiting genes exists to repress *pha-4* expression to avoid these founder cells develop into pharynx. These two conclusions from our results are consistent with biological evidence. Moreover, some of results for knockdown genes are validated through compared with existing literature. For example, *pop-1* gene can activate expression of *pha-4* in ABara lineage [7]. *glp-1*, *tbx-8* and *tbx-9* are important for *pha-4* activation in AB lineage [13].

TABLE 1. Activating and inhibiting gene number in founder cells.

	ABala	ABalp	ABara	ABarp	ABpla	ABplp
Activating	0	27	40	0	0	0
Inhibiting	15	8	3	6	5	3
	ABpra	**ABprp**	**MSaa**	**MSap**	**MSpa**	**MSpp**
Activating	0	0	15	0	15	0
Inhibiting	8	4	0	5	0	2
	E	**Ca**	**Cp**	**D**	**P4**	
Activating	31	0	0	0	0	
Inhibiting	0	0	3	1	0	

4. Conclusions

In this paper, we use several methods to process state-of-the-art single-cell resolution data. It includes data preprocessing, determining cutoff for *pha-4* gene expressing cell identification, applying statistical hypothesis tests pipelines to infer gene effect both at cellular level and founder cell level. In our results, some genes have been identified for pharynx formation. Others can avoid cells differentiating in a wrong way. These genes possess either activating or inhibiting effect on *pha-4* gene in different founder cells. This reveals why tissues and organs formation could regularly and precisely be conducted in normal organism. Moreover, these genes may not regulate *pha-4* gene expression directly. Some hidden regulatory pathways could be possibly existed between the knockdown gene and *pha-4* gene. Study and identify these regulatory pathways will have insight into complicated cell differentiation process.

Acknowledgements

This paper is supported by Signal Processing and Biocomputing Lab of City University of Hong Kong, Neural Interface Research Laboratory of City University of Hong Kong, and Department of Biology of Hong Kong Baptist University.

References

[1] Alinikula, J. et al., "Gene interaction network regulates plasma cell differentiation" Scand J Immunol, 73, 512-9, 2011.

[2] Moignard, V. et al., "Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis" Nat Cell Biol, 15, 363-72, 2013.

[3] Arda, H. E. et al., "Gene regulatory networks governing pancreas development" Dev Cell, 25, 5-13, 2013.

[4] Stigler, B. et al., "A regulatory network modeled from wild-type gene expression data guides functional predictions in Caenorhabditis elegans development" BMC Syst Biol, 6, 77, 2012.

[5] Du, Z. et al., "De novo inference of systems-level mechanistic models of development from live-imaging-based phenotype analysis" Cell, 156, 359-72, 2014.

[6] Bao, Z. et al., "Automated cell lineage tracing in Caenorhabditis elegans" Proc Natl Acad Sci U S A, 103, 2707-12, 2006.

[7] Murray, J. I. et al., "Automated analysis of embryonic gene expression with cellular resolution in C. elegans" Nat Methods, 5, 703-9, 2008.

[8] Murray, J. I. et al., "Multidimensional regulation of gene expression in the C. elegans embryo" Genome Res, 22, 1282-94, 2012.

[9] Sulston, J. E. et al., "The embryonic cell lineage of the nematode Caenorhabditis elegans" Dev Biol, 100, 64-119, 1983.

[10] Siegel, S., "Nonparametric statistics" The American Statistician, 11, 13–19, 1957.

[11] Batra, R. et al., "Time-lapse imaging of neuroblastoma cells to determine cell fate upon gene knockdown" PLoS One, 7, e50988, 2012.

[12] Blevins, L. et al., "Fisher's Exact Test: an easy-to-use statistical test for comparing outcomes" MD Comput, 2, 15-9, 68, 1985.

[13] Mango, S. E., "The C. elegans pharynx: a model for organogenesis" WormBook, 1-26, 2007.