QUANTITATIVE ANALYSIS, IMAGE PROCESSING, AND HIGH-THROUGHPUT TECHNIQUES FOR NEURAL IMAGING IN *C. ELEGANS*

A Thesis
Presented to
The Academic Faculty

by

Charles Ling-zhi Zhao

In Partial Fulfillment
of the Requirements for the Degree
Biomedical Engineering in the
Walter H. Coulter Department of Biomedical Engineering

Emory University &
Georgia Institute of Technology
December 2016

QUANTITATIVE ANALYSIS, IMAGE PROCESSING, AND HIGH-THROUGHPUT TECHNIQUES FOR NEURAL IMAGING IN *C. ELEGANS*

Approved by:

Dr. Hang Lu, Advisor
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Dr. Robert Butera
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Patrick McGrath
School of Biology
Georgia Institute of Technology

Dr. Chris Rozell
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Kang Shen
Department of Biology
Stanford University

Date Approved: September 8, 2016

To my parents, the late Andrew Grove, the National Institutes of Health, the
Computational Neuroscience Training Grant, and the philanthropists, public and private,
who enable the pursuit of science

ACKNOWLEDGEMENTS

Getting a Ph.D. is a painstaking task, one that would be impossible without the support of an entire community. I would like to thank first my advisor, Dr. Hang Lu, whose guidance, support, and immense tolerance for my numerous eccentricities came always as a welcome surprise, as a luxury I will not have in the future, and helped to make the Ph.D. degree an enriching experience. I would like to thank Daniel Puleri, now a graduate student at Duke University, for being an extremely reliable second hand and sounding board for an astonishing four years—and I should not forget Yun-Hsuan "Stellina" Lee or the new graduate students Shivesh Chaudhary and Farhan Kamili, without whom Chapter 4 would have been substantively delayed. I would like to thank Dr. Adriana San-Miguel, now a professor at North Carolina State University, for mentoring me in the early years, and Dr. Kang Shen and Dr. Patrick McGrath, for their astonishing intellectual generosity as collaborators. I would also like to thank the rest of my committee for being there for advice and intellectual support when I needed them. I thank the Computational Neuroscience Training Grant and in particular Dr. Dieter Jaeger, for their funding and logistical support, and I thank Dr. Robert Liu and Dr. Lena Ting, for tolerating me in their labs for a rotation each. Finally, of course, there is the unspoken implicit debt I have to my past and present lab members, for all their work. Beyond the academic support, I must also acknowledge the less obvious social support, including the parents and sister who helped backstop the otherwise reckless financial decision of getting a Ph.D., and my roommate Lansing Wei, whose shocking willingness to cook and clean an unfair amount helped fuel the amount of time I spent on my Ph.D.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS AND TERMS

C. elegans	*Caenorhabditis elegans*
CGC	Caenorhabditis Genetics Center
Co-injection Marker	Placed in a strain to enable identification of desired genotype
D. melanogaster	*Drosophila Melanogaster*
E. coli	*Escherichia coli*
GCaMP	A GFP-Calmodulin fusion protein that fluoresces in the presence of Ca^{2+}
GFP/YFP/RFP	Green/Yellow/Red Fluorescent Protein
Mutagenesis	Deliberate introduction of mutations in a population
M9	Buffer suitable for washing off *C. elegans* individuals
M9-triton	M9 with 0.01% Triton X-100 surfactant
NIL	Near Isogenic Line, a type of introgression line
NGM	Nematode Growth Medium
OP50	Strain of *E. coli* used as feedstock for *C. elegans*
Outcrossing	Repeated mating with a strain to replace the genetic background
Overexpressed	Describes fluorescent marker so bright it obscures details
PDMS	Polydimethylsiloxane, used for microfluidic devices
Picking	Moving worms from plate to plate with a platinum wire instrument
QTL	Quantitative Trait Locus (or Loci)
RIL	Recombinant Inbred Line
Wild Isolate	A strain of *C. elegans* recently recovered from the wild

SUMMARY

The use of image processing and quantitative feature extraction in the biological sciences has become increasingly prominent in recent years, as advances in equipment and high-throughput imaging techniques allow the collection of increasing amount of high-quality images and video. With high-volume, quantitative phenotypic descriptors, it becomes possible to elucidate previously unseen aspects of the genotype-phenotype relationship and neural function, making the efficient parameterization and statistical analysis of large amounts of data. By developing and applying computational techniques to the rapid processing of *C. elegans* images and video, this thesis aims to explore both the relationship between synapse-affecting genes and synaptic morphology and the neural function of the *C. elegans* connectome as a whole. In this, I focus on the characterization of neural structure and development through the analysis of still images, and of neural function by the analysis of video data.

To address the first of these, in the first aim I examine a pre-existing imaging and processing pipeline in the Lu Lab, which had previously been used to characterize and sort bright, fluorescently-labeled synaptic markers. In order to expand the applicability of the pipeline to new strains, dimmer and more precise synaptic markers, and subtler phenotypes, I refined the image processing algorithms used to be more robust to different imaging conditions, in particular the presence of confounding objects. This is used to more broadly and accurately characterize the effects of already-established synaptic mutants by examining synaptic domains in a more statistical and quantitative manner. We

also illustrate the broad applicability of the segmentation approach used by touching upon applications outside *C. elegans*.

In the second aim, we perform a novel application of Quantitative Trait Loci (QTL) analysis to heterozygotes between synaptically-labeled strains and Recombinant Inbred Lines (RILs) between the N2 and CB4856 strains of *C. elegans*, demonstrating the ability to perform QTL analysis on a fluorescent marker phenotype for which this is otherwise infeasible. The advantages of the image processing pipeline established in the first aim allowed us to mitigate the technical downsides of using heterozygotes for QTL, allowing us to greatly expand the versatility of QTL analysis for a variety of markers. This enables us to use our quantitative, high-volume data to statistically predict the locations of synapse-influencing natural variants in the *C. elegans* genome, particularly those that drive differences between the laboratory strain N2 and the true wildtype CB4856, identifying a potential QTL between the strains on chromosome IV.

In the final aim, I turn my attention from structure to function, examining the problem of monitoring many neurons in the head ganglion of *C. elegans* simultaneously using the GCaMP family of fluorescent calcium markers. By developing a segmentation, tracking, and data processing pipeline, I demonstrate that the tracking of the individual activity of many neurons simultaneously can be performed algorithmically without manual correction, decreasing the time invested per video by two orders of magnitude. I first illustrate and validate the algorithm using videos and manually-curated neural traces provided by Manuel Zimmer, for which an analysis was published in October 2015. I first show that our algorithm generates results of comparable accuracy and can reproduce much of the published results. I then demonstrate the ability to process a large number of videos quickly, using our own videos of a lightly anesthetized worm to illustrate our ability to collect neural recordings on the whole head of the worm.

Taken together, this thesis illustrates the scientific power of high-throughput, computer vision based methods to explore new aspects of subtle phenotypes, including synaptic phenotypes that have previously proven resistant to these kinds of study.

CHAPTER 1

INTRODUCTION

This thesis explores the use of image processing and computation, as applied to the high throughput imaging of *C. elegans* neural phenotypes, both structural and functional. A major challenge in the field is the development and exploitation of computational and high-throughput methods in order to study the nervous system of *C. elegans* in new, deeper ways. Here, we seek to advance the state of the art in this field by directly addressing three open problems. For candidate gene approaches, phenotypes are often evaluated qualitatively by eye, or quantitatively for large scale phenotypes, neglecting subtle phenotypes visible only high-power microscopy[18]. Similarly, genetic mapping with Quantitative Trait Loci (QTL) analysis is only done with phenotypes that can be quantified rapidly *en masse*[19]. Finally, whole ganglion calcium imaging is a burgeoning new topic in the field, but a stringent need for manual supervision of neuron tracking prevents the use of this technique for large sample size experiments[3]. This thesis intends to directly address these challenges; to motivate the development and application of the aims herein, this chapter provides an introduction to the relevant techniques, technologies, and questions, discussing the current state of the field as well as challenges in methodology and experimentation.

1.1 *C. elegans* as a Model Organism

Caenorhabditis elegans (C. elegans) is a free-living, transparent nematode, about 1mm in mature length. Organizationally, the body of the animal consists of a long, tough cuticle surrounding a fluid-filled pseudocoelom, alongside which reside the core tissues of the animal. It possesses musculature to enable movement by crawling, a pharynx and grinder at the mouth to enable feeding on bacteria, an intestinal system, a nervous system consisting of exactly 302 neurons[15], and a variety of other tissues. Notably, populations of *C. elegans* are in large part hermaphroditic and self-fertilizing, possessing a pair of gonads that generate both the sperm and ova necessarily to produce progeny, which are laid as eggs. A small percentage of the population is male, producing only sperm[10]. For the purposes of this document only hermaphroditic individuals are examined or discussed, although the male will be relevant for the purposes of genetics. *C. elegans* has four larval stages, referred to as L1 through L4, followed by an adult stage when they are reproductively mature. Under conditions of environmental stress and low food, L1 and L2 may also transition to an alternate dauer stage of development, a resilient and very long-lived state that can persist for months until conditions improve and the nematode may continue to L4[10].

As a model organism, *C. elegans* possesses a number of advantages that recommend it to the experimenter, tabulated in Table 1.1.

> For neural studies specifically, the combination of optical transparency and stereotyped neuron development allows for advantageous imaging frameworks that cannot be achieved for other multicellular organisms. The consistent recurrence of the same neurons allows for the examination of neurons in different individuals that are *known* to have the same function, guaranteeing cells that are directly comparable. The optical transparency of the worm allows for the direct observation of fluorescent markers within neurons, both for structural and functional markers, without disturbing or damaging the animal[10].

The combination of these two properties allows for repeated, high-throughput experimentation on the same neuron in a large number of different individuals, but the practical implementation of this kind of high-throughput experimental presents and experimental, and often computational, challenge. Much of this thesis is devoted to the automatic and accurate analysis of data gathered in this way.

Table 1.1: Experimental advantages to working with C. elegans

Advantage	Details
Genetic Homology to Humans	60-80% of human genes have a *C. elegans* ortholog[6] All three germ layers, most major tissues Some common neurotransmitters (dopamine, serotonin)[9]
Ease of Culture	Fast-growing; resilient to contamination; eats *E. coli*; can be cultivated on agar plates
Ease of Genetic Manipulation	Hermaphroditic and self-fertilization allows for consistent, isogenic lines Rare males allow for breeding between strains Rapid reproductive cycle: 3 days to reproductive maturity, 300 eggs laid per cycle Easy uptake of RNAi, extrachromosomal DNA, etc. via feeding or microinjection[10]
Optical Transparency	Easy imaging of internal landmarks, such as neurons or fluorescent markers
Stereotyped Development	All 959 somatic cells have an unvarying, mapped lineage all the way back to the zygote, including the 302 neurons[11-13]. The connectome between these 302 neurons is well-mapped[15]. Synapse formation is very consistent between isogenic individuals; synapses are over 75% reproducible[16].

In the wild *C. elegans* dwells in the soil and in rotting fruit. The most common laboratory strain, the Bristol Strain N2, was collected by Sydney Brenner from the soil of Bristol, England, but a number of other variants—Wild Isolates—exist. Most of these were collected relatively recently, once it was clear that, due to laboratory adaptation, N2 contained a number of significant genotypic and phenotypic differences from wild strains[20].

1.1.1 A Brief History of *C. elegans*

The use of *C. elegans* as a model organism was first proposed and popularized by Dr. Brenner in the late 1960s. Dr. Brenner, a molecular biologist and geneticist who would later go to win the 2002 Nobel Prize in Physiology or Medicine[21-23], was interested in examining the relationship between genes and behavior. Considering the nervous systems of larger animals to be too complex to readily study this connection, Dr. Brenner sought an organism that was small, easily manipulated genetically, and had a simple but discernable nervous system. *C. elegans* came to be the answer he chose[10].

Over the next few decades, researchers set out to rigorously characterize the development and function of *C. elegans* using electron microscopy and fluorescent labeling to trace the cell lineage of each of the nematode's somatic cells, finding that in the case of *C. elegans*, the fate of each individual cell is fully-encoded in the genome[11-13]. This examination let to the discovery that a number of cells in the L1 stage never made it to adulthood, dying without producing any daughter cells. The characterization of this process and the examination of *C. elegans* mutants that did not lose these cells led to key discoveries about cell division and apoptosis[24, 25], and became the basis of Brenner's Nobel Prize, shared with Howard Robert Horovitz and John Edward Sulston, who worked with him[21-23].

Since then, the number of labs working in *C. elegans* has proliferated. It would be beyond the scope of this thesis to discuss in detail all of the developments that have originated or involved work in *C. elegans*. Here, we confine ourselves the brief observation that *C. elegans* played a key role in the discovery of RNA interference (RNAi)[26-28] and has become one of the key platforms for the use of green fluorescent protein (GFP)[29, 30] due to *C. elegans*'s optical transparency, both Nobel Prizes. *C. elegans* has also become one of the key research organisms in studying axonal growth[31], synaptic trafficking, and synapse formation[32], due to the presence of a large number of stereotyped axons and synaptic domains in the adult nematode.

Despite substantial progress, the programme laid out originally by Dr. Brenner remains only partially fulfilled—the full spectrum of *C. elegans* genetics and neural behavior is only partially mapped. The drive for high-throughput imaging and subtle trait quantification is thus driven by a need to expand the field of inquiry even farther, exploring new aspects of the genotype-phenotype connection. It is this need that helps drive much of the motivation for Chapter 1 and 2.

1.1.2 Genetic Manipulation in *C. elegans*

The genome of *C. elegans* contains 6 diploid chromosomes, including 5 pairs of autosomes and 1 pair of sex chromosomes, referred to as X. The sex-determination system is X0, with hermaphrodites having 2 copies of the sex chromosome and males having only one. Because of this, hermaphrodites will have solely hermaphroditic progeny, except for a few males resulting from X chromosome nondisjunction during meiosis, while the progeny of male-hermaphrodite matings will be 50% hermaphrodite, 50% male. *C. elegans* populations thus have very few males in the long-term and are primarily self-fertilizing, though the exact male ratio differs depending on conditions and is generally higher in non-N2 strains[10].

The self-fertilizing hermaphroditic nature of *C. elegans* plays a big role, with populations tending to become homozygous at all loci in the long-term. It is thus possible to get inbred strains just by moving one individual to a new plate and allowing it to found a population. This makes *C. elegans* genetics often much more straightforward than in other species. Appendix A describes the genetic nomenclature conventions in the *C. elegans* community and which are adhered to in this thesis[33]. These are unfortunately unique to the field.

This section is devoted to describing the generation of novel strains in *C.* elegans, in order to provide background for and help motivate Chapter 3, where the use of heterozygotes is used to circumvent the creation of hundreds of novel strains. Discussion of the topic here is relatively brief and tailored to this motivation, as relatively few new strains are generated in this thesis; Appendix B may be referred to for a much more complete explanation.

Generating Novel Mutants

The traditional, and most common method, of generating new strains of *C. elegans* is via forward genetics[18]. This consists of random mutagenesis, followed by identification of novel phenotypes and isolation of the mutation responsible. Finally, these genotypes are repeatedly mated with the original parent strain (usually N2), selecting for progeny with the desired phenotype, a process called outcrossing[18].

While this approach to generating mutants is fruitful and provides mutants of use to *C. elegans* community as a whole, it is usually unhelpful for generating mutants in a specific gene of interest. If the exact mutant desired is not already available, a more targeted approach may be used, involving a zinc-finger nuclease or CRISPR-CAS9[34-37].

Fluorescent Marker Insertion

One of the most useful aspects of *C. elegans* for the experimenter is its optical transparency. This enables the visualization of fluorescently-labeled landmarks within the animal without needing to cut open or otherwise physically manipulate the animal. As such, the successful inclusion of genetically-encoded fluorescent markers is an important aspect of *C. elegans* genetic manipulation. As mentioned in Appendix A, *C. elegans* strains which have been genetically transformed can be labeled with an abbreviation such as “Ex” or “Is”—although other abbreviations, e.g. “IR” for introgression lines, exist.

The use of “Ex” refers to the presence of an extrachromosomal array that has been introduced by the injection of foreign DNA into the gonads of a healthy hermaphrodite. The primary advantage of this approach to genetic transformation is its speed and efficacy, but the level of expression of the injected genes and co-injection markers is extremely variable, and even sibling worms from the same parent show substantially variable expression[38-40]. The use of the “Is” labeled indicates that the genetic transformation has been “integrated” into the genome. A number of techniques exist to do this, usually reliant on generating random breaks in the genome and relying on DNA repair to integrate foreign genes. Compared to the extrachromosomal strains, these integrated strains carry a number of advantages, the most principle of which is stable expression, but care has to be taken about the possibility of mutations from the integration process. This can allow for much more reliable quantitative comparisons between individuals, and is the reason why integrated strains are used for much of the work in this thesis[38, 41, 42].

Combining Existing Strains when the Background Strain is the Same

A very common scenario facing the researcher is the need to hybridize specific existing loci into one new strain. In many cases, this can be done without resorting to gene-editing tools by exploiting the favorable interbreeding properties of *C. elegans*. In the simplest scenario, when the two genotypes are each confined to specific genetic loci against the same genetic background, the procedure is relatively straightforward and will be outlined below; the fundamental experimental techniques are the same as the more complex case. This assumes the two loci are on different chromosomes; two loci on the same chromosome will require chromosomal recombination rather than Mendelian genetics for mixing, requiring repeated matings and other complications[43, 44].

First, the two strains are interbred, and F2 progeny that are homozygous for both of the parent genotypes isolated. The most general, worst-case protocol involves moving F2 individuals onto new agar plates, one individual per plate, to found new populations. Each new population may then be evaluated for the presence of one of the two genotypes, either by sequencing or, if possible, direct inspection of phenotype. With a probability 25%, the population will show the desired genotype, meaning it must have had a homozygous parent and by homozygous itself. This population then has a 75% chance of containing at least some of the other genotype, and the other genotype may then be refined by repeating the same procedure. A variety of common tricks can shorten this procedure if the genotype has certain properties, or if a co-injection marker is used[43, 44].

Combining Existing Strains with Difference Backgrounds

Another common scenario occurs when it is necessary to integrate a genotype at a specific locus into a different target genetic background. For the purposes of this thesis, this is particularly relevant for Aim 2, when considering the problem of performing a QTL analysis using a phenotype that requires a fluorescent marker to measure, which would require the integration of a fluorescent marker into a variety of different backgrounds. It is noteworthy that the downsides of this procedure, noted below, make it so that in many cases it is superior to repeat on the target background the original procedure that generated the genotype in the first place—for example, by just inserting the fluorescent marker as described previously. In the case of genetic insertion of fluorescent markers for QTL purposes, however, this is inadmissible, as no such technique is reliable enough to ensure quantitative comparability between strains, given the potential for off-target insertions, uncertainty about copy number, and randomized insertion into the genome[38]. Thus, standard QTL analysis requires either using the procedure below for integrating the marker in each of many Recombinant Inbred Lines (RILs), or repeating the entire procedure for generating RILs each time a study requires a new fluorescent marker, a prohibitive downside. We save extensive discussion of RIL generation for Chapter 3, where we propose a procedure that avoids the need to generate all of these RILs.

Cursory thought reveals that a single mating is insufficient to perform the integration of a given gene loci into a new background, because one of the paternal chromosomes will always contain the original background of the gene loci being integrated. Once the mating has been performed, it becomes necessary to outcross the strain into the target background, while still maintaining the gene being integrated, a nontrivial task if the phenotype of the gene cannot be easily seen.

In the simplest case, where it is possible to observe the phenotype in the heterozygote, then outcrossing may be performed by repeatedly mating males of the target background into the strain, selecting for heterozygous progeny that contain gene.

This may be done until the background has probably been fully integrated (>7 matings), and then individuals may be picked onto individual plates and evaluated for homozygosity. In the other cases, when the heterozygous phenotype cannot be observed but the homozygote can, it is necessary to perform a longer protocol. The homozygotes can be found in the F2 generation after mating, and males of the target background can be used to mate with these. Because recombination can only potentially occur in the heterozygote, however, the number of necessary matings is unchanged. In the worst case, where even the homozygote cannot be easily phenotyped, it becomes further necessary to pick individuals onto their own plates and sequence some of progeny, as it is not possible to non-destructively sequence *C. elegans* individuals[43, 44].

The reliance of this procedure on recombination introduces a number of downsides which should be discussed. A co-injection marker, for example, can no longer be used as a fully reliable proxy for the gene of interest, as the probability that it has become separated during recombination can no longer be neglected, and care must be taken to either sequence the strain regularly or not allow the population to bottleneck one individual. Perhaps more importantly, recombination occurs properly only among homologous regions of the chromosome. If the gene of interest is an insertion, then it cannot itself undergo recombination and is prone to causing errors in recombination in its immediately vicinity. Finally, of course, it can never be fully guaranteed, only probabilistically guaranteed, that the entire target background has truly been transferred, and any potential defects in the overall process lead to a requirement for more crossings to ensure success[43, 44].

1.1.3 Laboratory Mutations in *C. elegans*

The history of some of the most commonly used *C. elegans* strains, in particular N2, the Bristol Wildtype used as the laboratory standard, provides a fascinating foray into the potential effects of laboratory adaptation on the genotype and phenotype of a commonly used model organism. The modern technique for long-term storage and preservation of *C. elegans*, freezing of starved freshly hatched L1s in liquid nitrogen, was only introduced in the late 1960s, giving plenty of time for lineages to diverge, both from each other and the wildtype[20]. While genetic drift may have played a role in introducing changes, the most interesting changes arise from laboratory adaptation, the quasi-natural selection of *C. elegans* strains for optimal propagation in the laboratory[45]. The question of exactly what changes have developed in *C. elegans* over its long stay in laboratory incubators, and whether subtle variations in neural phenotypes exist and can be detected by QTL analysis, serve as a major motivating factor for Aim II of this thesis.

A number of significant laboratory adaptations are now known to exist in N2 relative to all known strains recently gathered from the wild, termed wild isolates[20]. The most significant of these are mutations in *npr-1* and *glb-5* that significantly diminish their activity. In wild isolates, the protein NPR-1 regulates the social feeding behavior of *C. elegans*, with decreased activity leading to increased aggregation of individuals into gregarious social clusters and adventurous foraging behavior, whereas increased activity leads to solitary feeding and careful exploitation of local resources. In N2, significantly increased NPR-1 activity leads to solitary individuals that are very reluctant to leave a local source of food[45]. This is almost certainly an adaptation to culturing on agar plates,

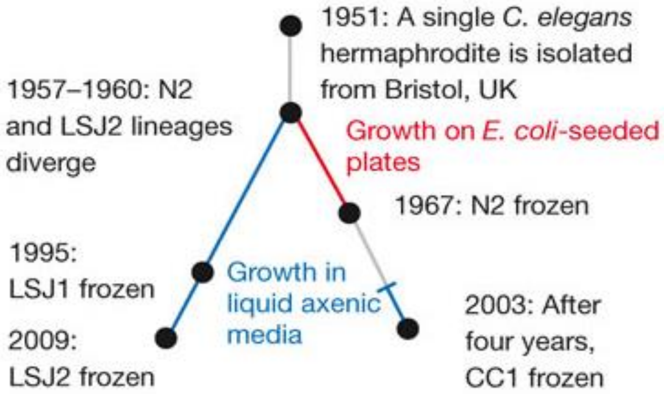

1951: A single *C. elegans* hermaphrodite is isolated from Bristol, UK
1957–1960: N2 and LSJ2 lineages diverge
Growth on *E. coli*-seeded plates
1967: N2 frozen
1995: LSJ1 frozen
Growth in liquid axenic media
2003: After four years, CC1 frozen
2009: LSJ2 frozen

Figure 1.1 A history of the divergence between the strains N2, LSJ1, LSJ2, and CC1. Figure adapted from McGrath, et al[1].

where there is only ever one *E. coli* lawn as a source of food and worms are frequently picked for passaging—it is easier to pick solitary individuals rather than those aggregated in clumps, which has resulted in a form of accidental artificial selection. This effect is further enhanced by a decrease in the activity of GLB-5, a globin involved in oxygen sensation that in the wild isolates leads to a marked preference for the low oxygen environments (5-12%) of its natural habitats and social aggregation in the presence of high oxygen[46]. The behavioral and physiological impact of these two mutations is not limited to just what is stated here, and are profound enough that these two genes are repeatedly detected in QTL mappings of differences between N2 and other wild isolates.

It is vital to understand what exact differences exist between N2 and wild isolates, given the role N2 plays as the background strain for nearly every *C. elegans* study. At the very least, any loss of function mutants found in N2 relative to wild populations would lead to that particular loci being very difficult to detect using forward genetics. It was for this reason that a range of wild isolates were examined for differences in synaptic morphology in Chapter 3, leading to one of the motivations for performing a QTL analysis.

1.1.4 Synaptogenesis in *C. elegans*

In *C. elegans*, synapse formation happens *en passant*, with synapses budding off the axon as it passes by a neighboring neuron or muscle. Like all synapses, this is characterized by a presynaptic density, consisting of a distinct region of the membrane heavily populated by neurotransmitter-bearing vesicles. Unlike vertebrate synapses, there is no obvious postsynaptic density filled with ion channels and signal transducers[15, 32].

The classical synapse-labeling fluorescent marker is a fusion of the protein synaptobrevin (SNB-1), an integral membrane protein of synaptic vesicles, and GFP[47], and is used, for instance, in the genotype *wyIs92*[48], which we use in chapter 1 and 2 and will describe there. Much of what is known about synaptic assembly derives from mutant screens of *C. elegans* conducted using this marker, which identified a number of sets of genes whose mutations are associated with unusual synaptic vesicle patterns[32, 49-51]. Notably, these studies were conducted by visual inspection of synaptic domains, the manual analog of what is advanced in Chapter 2, and the best understood aspects of synaptic formation and assembly derive from the gene families identified in these studies (*sad, sam, syd, and syg*) and relevant follow-up studies[32].

The least well-understood aspects of synapse formation deal with its regulation and coordination. It is believed that many of the components of the pre-synaptic density have at this point been identified, but it remains relatively mysterious why synapses form where they do, or what coordinate the many synaptic proteins together into a presynaptic density[32]. The genes involved here, particularly genes involved in regulation as part of a gene network, are likely much more difficult to find in a mutant screen, as the effects caused by their loss may cause only subtle effects on the ultimate synaptic phenotype. For example, the JNK/JKK kinase pathway, whose molecular role is only beginning to be understood, has only a subtle effect on synaptic phenotype[52] (Section 2.4).

The ability to detect gene-gene relationships in putative synaptic regulatory genes would thus be invaluable, even as a bare epistatic relationship such as established chapter 2, which would enable further follow-up studies. Further, by searching wild populations for synaptic-affecting genes, it is likely possible to turn up subtle genes that would not be noticed in a mutant screen due to low penetration, or too subtle an effect to be detected in a single mutant animal. Thus, a synapse-focused QTL, such as conducted in Chapter 3, might prove invaluable in discovering new, potentially crucial synapse-influencing genes.

1.2 Microfluidics for the Manipulation of *C. elegans*

The development of microfluidics, the class of techniques for manipulating fluid flow on a micron-level scale, has been spurred in the past two decades both by technical developments and the realization of its value for the imaging and manipulation of small biological organisms, whether these be mammalian cells or *C. elegans*[53-57]. This is done by engineering the length scale of microfluidic devices to correspond to the organism being studied, aided by the predictable, laminar flow induced by these low Reynold's number channels. In addition to providing for manipulation of organisms, these devices allow for the rapid and precise insertion and removal of chemical agents and other forms of experimental manipulation[55, 58-60]. In biology, the silicon MEMS once used have now been almost entirely supplanted by devices made of polydimethylsiloxane (PDMS), a soft silicone polymer. While not as robust and reusable, PDMS devices are easy to fabricate using a silicon master, are much more biocompatible, are permeable to oxygen and carbon dioxide, and are soft enough to enable to actuation of internal valves merely by the application of pressure to the right locations[54, 61]. These properties are extremely valuable for biological applications. We discuss here briefly the fabrication and use of these microfluidic devices, without going into extensive detail on the fairly standard protocols involved.

1.2.1 Device Fabrication and Preparation

The first step in the construction of a PDMS is the design of the device itself, typically in a computer-assisted design (CAD) software such as AutoCAD. Microfluidic device design is an entire field, one that will not be substantively explored here, though many of the fundamental ideas are implicitly explored in Figure 1.3. Once designed, a silicon negative mold (or master) of the device is produced, most commonly with photolithography, and is then coated with dimethylchlorosilane or a similar compound ("silanization"), which prevents too much adhesion of PDMS to the master. The height of the features on this master varies by device design, but is typically in the range of tens of microns for *C. elegans* devices[53].

While highly important, the design and fabrication of the silicon master is a one-time affair, barring trial and error for the refinement of the design. Far more common is the fabrication of the devices themselves, which can be done over and over on the same silicon master, which is often large enough to mold as many as two dozen devices at once. The PDMS pre-polymer is mixed with cross-linker, stirring vigorously, and the gas bubbles thus formed are eliminated by placing the resulting mixture into a vacuum chamber until it is observed that the bubbles are gone. This mixture is then poured directly onto the silicon master within a large petri dish and allowed to spread out evenly, and this is then baked in a 70 °C for roughly four hours[53, 61]. It is common for devices used in the Lu Lab, particularly the devices discussed in the thesis, to use two different ratios of pre-polymer to cross-linker: 20:1 for a small layer poured directly onto the master, suitable for valve actuation, and 10:1 for a large layer poured above that, for structural support and manipulation. A short baking of 20 minutes is carried out between these pours. Depending on device design and the needs of the user, there are many more elaborate and involved protocols, but they will not be discussed here.

Once the baking is complete, the PDMS layer, usually about 0.5 cm thick, is carefully peeled off the silicon master. This layer is then sliced into individual devices and syringe needles of the proper size are used to punch small holes into the device in pre-planned locations, providing access to the microfluidic channels for the later insertion of needles and tubing for the insertion and extraction of liquid, pressurization of valves, and so forth. Along with a clean, thin glass slide, 0.16-0.19 mm thick, this hole-punched device is cleaned, then bathed in an oxygen plasma for a short period of time, ~20 s. This creates oxygen radicals on the both the surface of the glass and PDMS, and the side of the device with the microfluidic channels is then adhered to the glass, forming permanent covalent bonds The grooves in the PDMS formed by the silicon master now become closed channels bounded on one side with glass and accessible via the previously punched holes[53, 61].

1.2.2 Operation and Design of a Single-Layer *C. elegans* Imaging Device

Figure 1.3 illustrates the microfluidic device used throughout Aims I and II of this thesis, designed by Adriana San-Miguel[62]. Pressure control is provided by an off chip valve box that allows for the toggling of individual pressure sources via a computer

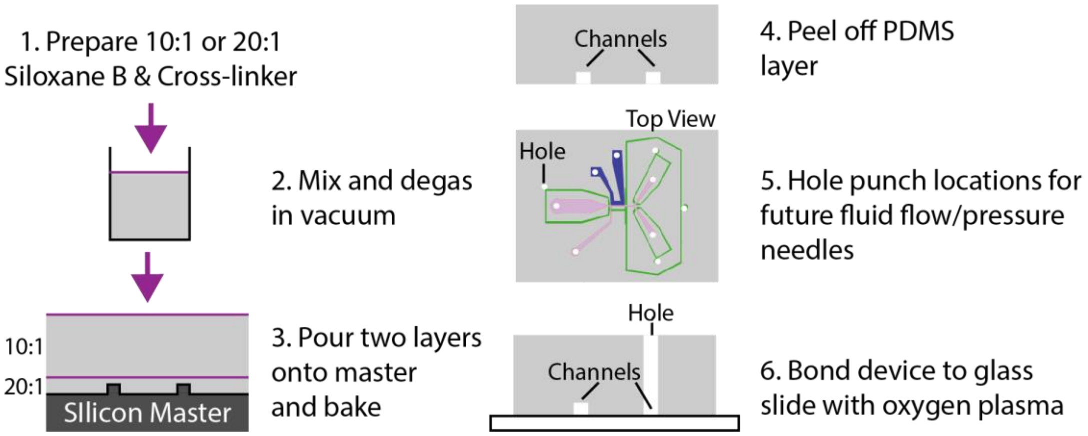


Figure 1.2 PDMS device fabrication in summary, omitting some details. Drawings are not to scale.

interface. Fluid flow is driven by a relatively low pressure (3-10 PSI) pumped into vials of 0.01% Triton-X in M9 buffer, one of which is used for flush control and the other for the in-flow of *C. elegans* individuals; and flow control is provided by separate solenoid pinch valves. Flow within the device is controlled by pneumatically drive valve chambers adjacent to the flow channel, such that pressurization of the chamber (at ~35 PSI) restricts flow. By placing the entire device setup above an inverted microscope, *C. elegans* individuals may be imaged by flowing the animals in, restraining them in the imaging region by closing the valves in front and behind, then releasing them again by opening the valve in front. A separate channel on the side allows for the flow of chilled 4 °C 50/50 glycerol/H_2O through the device, which temporarily immobilizes individuals in the imaging channel for imaging without recourse to paralytic drugs. This device may be operated manually, by toggling various arrangements of the valves through a custom GUI, or even, with a sufficiently well-synchronized worm population, on full automatic, as discussed in previous work [63, 64]. Unfortunately, because of the thickness of the channel and of the worm body itself, features within the worm body can only be reliably clearly imaged on an epifluorescent microscope when the proper side of the worm is pressed against the glass slide through which imaging occurs. To achieve this, the end of the imaging channel, near the imaging valve, narrows substantially to force the animal against the glass. There is unfortunately no currently known way to control whether the dorsal or ventral side of the worm faces the glass (though it will be either dorsal or ventral), or even to ensure that the proper end of the worm (head or tail) enters the exact imaging region. Since whether or not this occurs is roughly uniformly random, only 25%

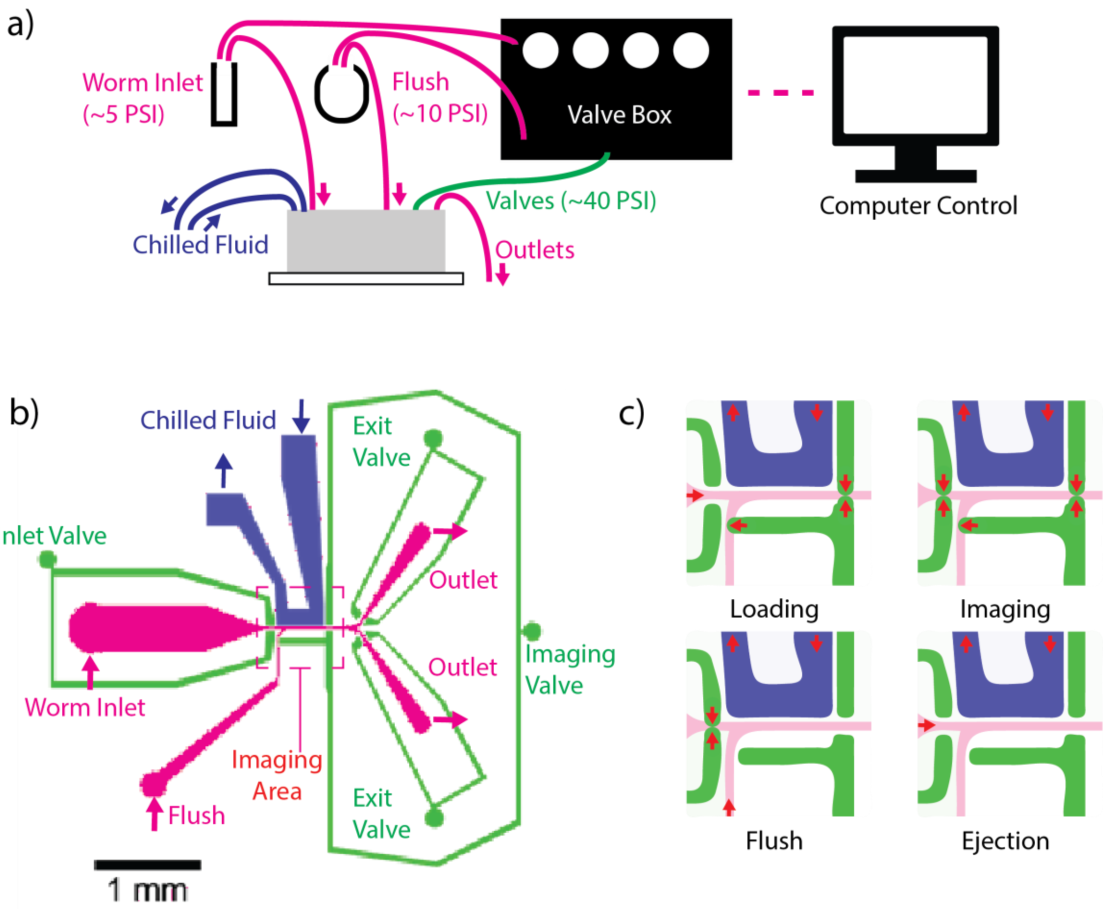


Figure 1.3 PDMS device operation and design. Part A shows the general setup, where a computer controlled valve box is used to regulate the pressure inputs and valves for the device, controlling flow to and from the device. Part B shows the device of the channels on the microfluidic chip itself. Part C shows the primary modes of operation of the device. First, an individual worm is loaded into the imaging region in the center. Flow is briefly stopped to enable imaging, and then the worm is flushed out and ejected. Note that this operation mode neglects the two exit valves, which may optionally be used to sort which outlet a given worm leaves by, in case sorting is needed. Shifting between these modes may be done manually on the computer or fully-automatically, by detecting and imaging worms without manual input.

of the worms that enter the device are suitable for automatic imaging—this is, however, no issue, as the number of worms that may be fed into the device can be enormous, and worms may be discarded automatically. This illustrates the power of the microfluidic approach, although this loss of images becomes relevant if the population being imaged is limited in number, as occurs in Chapter 3. It should also be noted that even with manual imaging, 50% of the worms would have the wrong side of the worm facing the objective.

1.3 Microfluidics for Neural Imaging in *C. elegans*

Much of the text in this section was adapted from *"Trends in High-throughput and Function Neuroimaging in C. elegans"* in WIREs Systems Biology and Medicine, a review paper I co-authored and which is currently in review.

The combination of *C. elegans's* natural advantages as a model organism and the advantages of microfluidics as an experimental platform has spurred the development of a number of microfluidic platforms intended to examine neural structure and activity under a variety of different conditions[53, 63, 65-68]. This takes advantage of the most valuable experimental aspects of *C. elegans*, exploiting its optical transparency to gather information on neural structure and function on a large scale. However, to do this, it is necessary to overcome a number of challenges. For instance, in order to obtain detailed quantitative information, it is necessary to immobilize worms effectively, collect images efficiently and rapidly, and robustly process the images obtained. With images or video in hand it becomes necessary to accurately track and characterize what may be a large number of neurons and neural features.

For high resolution neural imaging, one important technical challenge is the immobilization of individual animals during imaging. Even for very short exposures and bright markers, slight movements in the animal can drastically decrease image quality. While paralytic drugs are traditionally used to limit this, these drugs often have unknown effects on the phenotypes observed, and may damage the animals, limiting further experimentation and experimental throughput. On microfluidic devices, however, the use of these drugs can be avoided through techniques such as cooling, physical restriction, carbon dioxide, and gelation[64, 69-71].

Here, I discuss both structural and functional imaging of neurons in *C. elegans*, particularly as it pertains to the aims of this thesis. The first involves the detailed, high-throughput imaging of neural structures, usually taking only a static view of the worm, enabling new types of genetic screens and gene association studies. The latter requires dynamic imaging of the worm over time, using a marker for neural activity such as the fluorescent calcium marker GCaMP (which represents a fusion of GFP and calmodulin)[72, 73]. It is worth noting that both of these take advantage of the unique characteristics of *C. elegans*.

1.3.1 High-resolution, High-throughput Imaging of Neural Structure in *C. elegans*

Static imaging, particularly of fluorescent markers, is the workhorse of many developmental studies. However, many of the most powerful techniques for mapping the genome and performing mutant screens require the accurate, large-scale quantitative characterization of the phenotype under study, something that was previously only done on phenotypes that could be rapidly graded by eye. Recently developed techniques in high-throughput and automated imaging have allowed the extension of these kinds of studies to subtle and dim fluorescent features, including neural structures, such as synapses, that can only be evaluated under high magnification[63, 64].

Traditional genetic approaches require the examination of a large number of individual animals, either searching a population of mutagenized individuals for a change in phenotype, or screening a diverse array of strains for the source of a difference phenotype, such as is done in QTL[18]. While a number of techniques have been developed for the rapid screening of fluorescent neural markers, thus far only microfluidics has proven capable of doing so while also possessing the resolution to examine fine structural features such as synapses[63, 64].

A key innovation here is the use of automation; by using a microfluidic device such as the one introduced earlier, these microfluidic devices expedite the examination and retrieval of worms by drastically simplifying the imaging process. To achieve total automation, image processing algorithms can automatically examine the phenotype of interest and decide whether a worm is mutant, or simply whether or not a worm should be imaged. Examples of success with this approach in performing mutant screens, both from our lab and others, include the identification of synaptic and metabolic mutants by variants of a single-channel confinement device[74] and the identification of chemotaxis mutants with a device capable of generating controlled gradients[75]. Success with mapping the genome for the source of a particular phenotype, then, is a natural continuation of this work, and the subject of Aim II of this thesis.

1.3.2 Functional Imaging of *C. elegans* Neural Activity with GCaMP

The optical transparency of *C. elegans* enables the observation of neural activity without damaging the worm, via the use of the calcium marker GCaMP. While calcium levels within the neuron are only an indirect marker of neural activity, and questions remain about the effect of using a fluorescent marker that itself sequesters calcium, the use of GCaMP has been an invaluable tool in the understanding of simple circuits and stereotypical neural relationships in *C. elegans*, gradually displacing the use of FRET-based markers like cameleon for applications that require high dynamic range and do not require millisecond temporal resolution[5, 8, 14, 17, 76]. This enables the optical measurement of neural activity *in vivo* for extremely long periods of time, in a context where electrophysiology is both challenging and very damaging to the animal[77]. Here, too, microfluidics has a role to play, enabling the high-precision measurement of calcium activity with greatly lowered use of muscle paralytic drugs such as tetramisole, simply by the use of the microfluidic confinement methods already discussed—though the use of cooling or carbon dioxide is inappropriate in this case[69, 70]. While there are, of course, downsides to examining the animal under confined conditions rather than e.g. freely-roaming, there are substantial upsides, including better controlled imaging conditions, the ability to use higher magnifications, and the ability to deliver reliable, controlled stimuli.

Traditionally, limitations on the spatiotemporal resolution of microscopy techniques have prevented observation of more than a few neurons at a time at a sufficiently high time resolution (~0.1 s or less)[78], prompting unavoidable dissatisfaction with the limitations of examining *C. elegans* neural processing a few neurons at a time. In recent years, with the advent of a new generation of microscopy techniques, including spinning disc confocal[79] and light-field microscopy[4, 8], a number of research groups have turned to the idea of “Whole Brain” (or “Pan-neuronal”) imaging, where as many neurons as possible are imaged at once, either under rest or under deliberate stimulus[4, 5, 8, 14, 17, 80, 81]. This provides simultaneous information about a large number of neurons at once, for instance enabling the direct correspondence of neuronal activity with stimulus. A spate of new research in this direction promises new advances in the field, but recent papers have proven to be primarily proof-of-concept, providing data on only 4 or 5 worms at a time (Table 4.1). One of the main bottlenecks on throughput here, the need to reliably track and analyze hundreds of neurons at a time, can be effectively resolved by the judicious use of segmentation, tracking, and post-processing techniques, as demonstrated in Chapter 4 of this thesis.

1.4 Thesis Outline

Many of the challenges in the field of *C. elegans* neuroscience research can be addressed by the development and exploitation of computational and high-throughput methods. This thesis seeks to advance the state of the art in this field by directly addressing these open problems, demonstrate new experimental methodologies enabled by effective image processing, both directly through new forms of accurate image analysis, and indirectly through new experimental methodologies that could not previously be attempted. For the analysis of subtle differences in phenotype, it is necessary to develop accurate and robust methods for quantifying large amounts of high-throughput imaging data. For the genetic mapping of features that can only be visualized under a high-power microscope, it is necessary to develop both experimental and computational techniques for the rapid assessments of large numbers of different strains. Finally, for the analysis of functional calcium traces in large numbers of neurons simultaneously, automated methods must be developed to replace manual or semi-manual data curation.

The chapters of this thesis are organized around the idea of examining both structural and functional imaging in *C. elegans*, with Aims I and II pertaining to structural, single-image analysis and Aim III pertaining to the analysis of dynamic functional information from global brain videos taken of calcium signals in the *C. elegans* head ganglion. In Aim I **(Chapter 2**), previously developed algorithms for the automated segmentation and imaging of the synapses of a *C. elegans* motor were completely redesigned for application to a new, more difficult problem, in particular the imaging of very dim markers, and the robust automated imaging of animals even when the age distribution of the population may be substantially broader than usual. This has direct relevance to Aim II **(Chapter 3**), where these new algorithms are used to enable QTL genetic mapping on the synaptic morphology of the DA9 motor neuron, which was observed to differ between the laboratory wildtype N2 and the wild isolate CB4856. The motivation here is two-fold, demonstrating both the first application of high-throughput microfluidic imaging to genetic mapping, and the ability of focused algorithm development to take on an otherwise technically infeasible project. In Aim III (**Chapter 4**), we turn our attention from structure to function, demonstrating the ability to accurately track and analyze calcium traces from hundreds of neurons at a time over a long period of time. We show that with this automated technique we can substantively replicate the conclusions of a previous study carried out primarily by manual correction and annotation in a fraction of analysis time, and extend it to dozens of additional animals, rather than the 4-5 that have been reported per study so far. Finally, we summarize the contributions of this thesis, drawing conclusions and providing suggestions for future work following up the work herein (**Chapter 5**).

All experiments were carried out and devices fabricated either by the author or under his supervision, unless explicitly indicated otherwise in the text. However, for clarity, the image data in Chapter 2 of *D. melanogaster* embryos and T cells was taken by Dr. Thomas Levario and Dr. Ariel Kniss-James of the Lu Lab. Except for the last set of data explicitly taken under the author's supervision, the video data used in Chapter 4 for analysis was generously provided by the laboratory of Manuel Zimmer at the Research Institute of Molecular Pathology (IMP) in Vienna, Austria from Kato *et al*[5]. The majority of strains used were generated by the laboratories of Kang Shen at Stanford University, Patrick McGrath at the Georgia Institute of Technology, Eric Andersen at Northwestern University, and Cori Bargmann at Rockefeller University. All members of the Lu Lab, however, owe an implicit debt to previous members of the lab and the research community as a whole for microfluidic chip designs, pre-fabricated silicon masters, established microscopy setups, legacy computer code, and so forth. The image processing and data analysis techniques will be frequently borrowed from the field of computer vision and machine learning, but this thesis does not intend to break new ground in the field of fundamental algorithm development. It instead intends to break new ground in the accurate and novel application of previously unused techniques to neural phenotypes in *C. elegans*.

CHAPTER 2

ROBUST IMAGE SEGMENTATION AND ANALYSIS FOR NOISY AND DIM FLUORESCENT IMAGING

Much of the work in this chapter is in preparation for publication as Zhao *et al.*, *"Rapid, Simple, and Versatile Quantitative Phenotyping of Fluorescent Reporters Enabled by Relative Difference Filtering and Clustering."*.

2.1 Motivation, Background and Overview

One of the key goals of studying the model organism *C. elegans* has been the elucidation of the complex and multi-faceted relationship between genotype and phenotype. Phenotype is a broad term, used to describe everything from nuances of behavior to levels of protein expression; thus, understanding the relationship between the outward qualities of an organism and its encoding genotype is one of the most complicated tasks one can undertake. Fortunately, the geneticist's toolbox is filled with methodologies for investigating the relationship between the two, and the application of these to *C. elegans* has yielded multiple Nobel prizes[21-23, 27-29], as well as the first understanding of a number of key genes and processing, including apoptosis during development[12, 22, 24, 25], the Notch signaling pathway[82] and numerous participants in Ras/Map-kinase signaling[83].

However, using traditional methods, many of the most powerful such techniques cannot be practicably applied to subtle features or features that require high magnification to observe, in particular those that require fluorescent markers for labeling. The reasons here are two-fold: on the one hand, many techniques, such as forward genetics via mutagenesis[18, 43] and QTL mapping[19], require examination of hundreds or thousands of individual animals to saturate, an endeavor that cannot traditionally achieved for phenotypes that require inspection under high magnification, as this would require the inspection of hundreds of individuals on agar pads within a very short time window. On the other hand, accurate phenotyping with quantitative precision is also a necessity for mapping techniques like QTL[19], and can enhance the usefulness of candidate genetics, examining mutant phenotypes with more precision than is possible with standard observation.

To address these deficiencies, the Lu Lab developed a robust and efficient microfluidic device intended precisely to allow the rapid imaging of high-magnification phenotypes in *C. elegans*[63, 64, 74]. By designing a valve box that allows for computer control of each of the individual valves on the device, and incorporating useful experimental features such as a separate cooling channel for the immobilization of animals, it enabled rapid, automated imaging of animals[64]. By automatically segmenting and measuring the properties of fluorescently-labeled synapses, this further sorting of mutagenized animals based on synaptic phenotype[63, 64], and was successfully used to conduct a mutant screen on synaptic phenotypes.

Support vector machines (SVMs)[84] on a large number of synaptic features were used to segment synapses before quantification. To very briefly summarize the procedure, a wide variety of different filters were applied to every image, and the values given by the filter for every pixel, were used to train a SVM to identify synapses based on filter values. Positive (synaptic pixels) were selected by a human operator selecting the region of each synapse, with a hand-tuned filtered thresholding procedure used to segment pixels out of this region. Negative pixels were selected randomly from the regions not selected[64].

Further improvements made later drastically simplified the fabrication process of the microfluidic device, introduced additional SVM classifiers for identifying the region of the worm being imaged, and reduced the number of images taken of inappropriate sections of the worm[62].

Despite these successes, it became clear that the system in its current form had a number of deficiencies, impacting both its robustness and applicability to forms of traditional genetics where careful examination of potentially different strains is necessary. The SVM classifiers the existing methodology relied on were highly sensitive to the exact conditions of the training set used to train the algorithm, and performed poorly under new conditions and for new fluorescent markers. Further, the unapproachability of the technique for external labs made it a liability in expanding its use as a tool beyond just the Lu Lab. Finally, even with a properly constructed training set, the pixel-based SVM had difficulty distinguishing between fluorescently-labeled synapses and auto-fluorescent fat droplets whenever the two objects were about equally intense, as occurs frequently when the fluorescent marker is dim, rather than heavily over-expressed as the markers used in the original experiments were.

With all of this in mind, this portion of the thesis aimed to address these deficiencies. First, we developed a new robust and untrained image segmentation technique for identifying relevant fluorescent dots in an image in the presence of confounding, similar-looking objects. Not only does this technique successfully separate fat droplets from synapses in this use-case, it is more appropriate for situations with little data available for training, including situations where the experimental conditions or condition of the animals may change frequently, and can even be generalized beyond this particular application in *C. elegans*. As such, it is more proper for accurate quantification of large numbers of different strains, such as in the QTL analysis in Chapter 3, which does not have the luxury of sorting out everything that looks substantially different from the training set as “mutant”. To fully characterize the process in a variety of situations, and as fruitful use of data available to our lab, we demonstrate this segmentation technique on more than just *C. elegans,* extending it to both *Drosophila melanogaster (D. melanogaster)* embryos and human T cells in a microfluidic device, using images acquired by Dr. Thomas J. Levario and Dr. Ariel Kniss-James, formerly of this lab.

Some of the *D. melanogaster* embryo work is published as Levario *et al.*, *"An integrated platform for large-scale data collection and precise perturbation of live Drosophila embryos"*[7].

Secondly, in order to demonstrate the usefulness of high-throughput techniques for subtle phenotypes in candidate approach genetics, and for quantitative genetics, we expand the range of quantitative features gathered, study necessary corrections for accuracy, and use new post-processing methods from the field of convex optimization to mitigate sparse noise in the dataset. We demonstrate the validity and accuracy of the processing pipeline by examining known synaptic mutants, demonstrating results consistent with manual observation as well as quantitatively evaluating a novel epistasis between two genetic loci.

2.2 Robust Segmentation of Fluorescent Phenotypes with Relative Difference Filter and Clustering

The first goal was to develop a segmentation method for fluorescently labeled synapses that was independent of the pixel-based SVM segmentation method previously-used. As discussed, this was motivated by the observation that this method rarely generalizes beyond a particular experimental setup and marker, necessitating manual curation of at least some images before it can be used, feature selection relevant to the problem at hand, as well as a parameter search to find optimal values for parameters. This raises questions about the ability of the segmentation to accurately label synapses in mutant animals, and the implementation of the full segmentation workflow can be daunting. Finally, and crucially, it was found that this segmentation method dealt poorly with the presence of confounding objects like fat droplets when the synapses were labeled with a dim marker, unless heuristic features were developed to detect the relatively straight and compact synaptic domain for a given arrangement of synapses.

To address this, Z-stacks were taken of the synaptic domain of the neuron DA9 in the genotype *wyIs92 [Pmig-13::snb-1::yfp; Podr-1::rfp]*, obtained from the Kang Shen lab[48]. Here, SNB-1::YFP is a fusion between synaptobrevin-1, a protein consistently found in the pre-synaptic density of *C. elegans* neurons, and the yellow fluorescent marker YFP; *Podr-1::RFP* is an extremely bright co-injection marker localized to neurons in the head. The promoter for *mig-13* ensures that the synaptic fluorescence is expressed only in the VA and DA subset of motor neurons. DA9 was chosen because it has a consistent synaptic domain always found along the dorsal side of the tail, and also because it is a very common neuron used in this kind of study, including in previous work from this lab[48, 52, 62-64, 74].

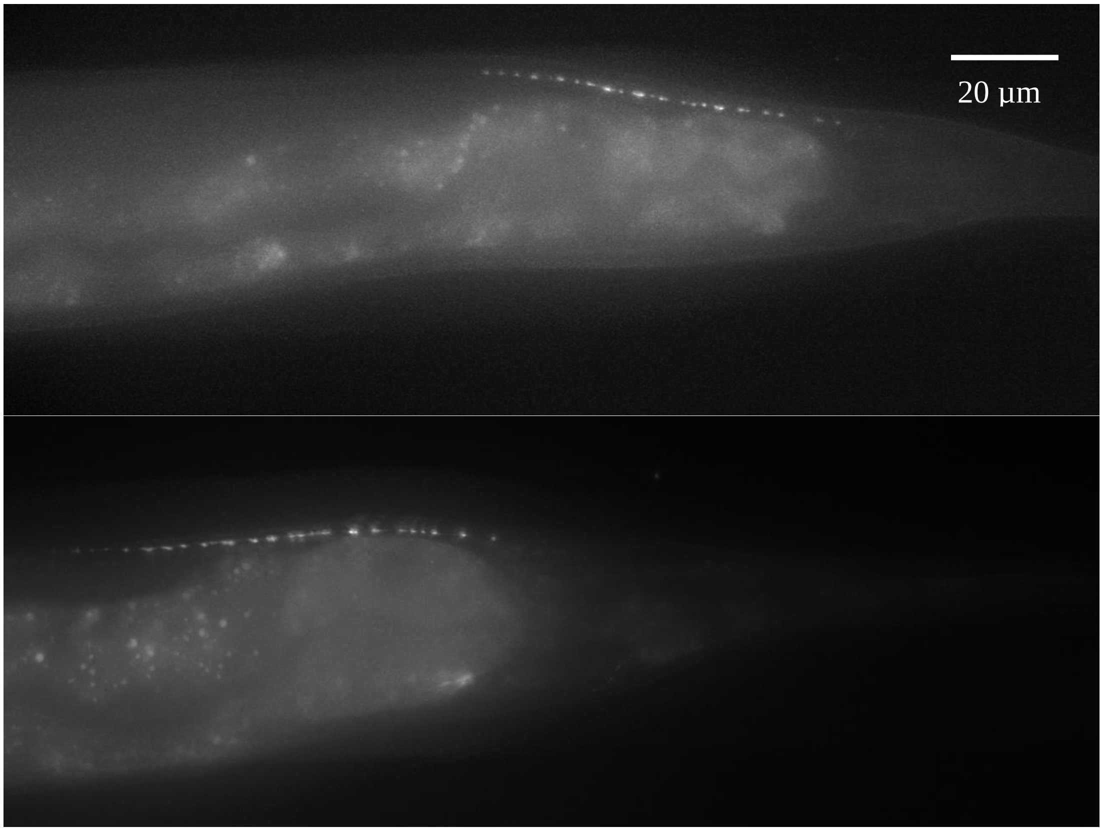


Figure 2.1 Representative images of the synaptic domain of the motor neuron DA9, in the dorsal nerve cord. The cell body lies in the ventral nerve cord, sending an axon in the direction of the tail that immediately curve backs around and extends towards the head on the dorsal, as picture here. These are images of the genotype *wyIs92* on an inverted fluorescent microscope at 40x magnification; synapses were fluorescently labeled with *Pmig-13:snb-1::yfp*. Horizontal length of the images is about 206 µm, and these images have been contrast adjusted for visibility. Note the visible fat droplets in both images.

In addition, images were acquired from others in the Lu Fluidics Lab in order to validate the success of the procedure on very different imaging conditions, and also to solve vexing image processing questions relevant to the lab. In the first case, time-sequence images were taken Dr. Levario of *D. melanogaster* blastulas with nuclei labeled by histone-GFP and imaged at 40x in a microfluidic device by a confocal microscope. From an image-processing standpoint, the objective was to successfully segment only the outer ring of nuclei along the dorsal-ventral axis, ignoring the nuclei in the middle that are frequently visible. Embryos were imaged with either single or double-photon imaging; here, one example of each is presented. The results of extensive further experimentation by Dr. Levario using this algorithm are published[7]. These results are briefly presented in Section 2.2.2 as proof of the algorithm's accuracy, but are not presented in the results section of this chapter, as it was not part of the original goal of the thesis.

In the second case, time-series images were taken by Dr. Kniss-James of Jurkat T cells labeled with a cystolic calcium indicator in a microfluidic array of cell traps on an inverted fluorescent microscope[85]. These were segmented to find single-loaded T cells, ignoring unusual loading and T cells suspended out of the cell traps, where quantification is unreliable. Detailed methodology may be found in Appendix C.1.

Four goals were set for the new segmentation procedure:

1. **Accuracy**: Above all, of course, it is necessary that any segmentation developed be accurate, successfully identifying synapses while ignoring confounds and producing a plausible segmentation of the pixels within a given synapse
2. **Robustness**: The procedure should generalize well, beyond a particular set of imaging conditions or a given marker
3. **Few Parameters; No Training:** The procedure should be relatively straightforward to use on new sets of images, with no explicit training. It is

unavoidable that a few parameters require manual selection or heuristics, but this should ideally be as straightforward as possible.

4. **Resilience to Confounding Objects:** One of the main motivations of developing this was, of course, the problems that confounding fat droplets posed to the previous segmentation method. While this is technically an aspect of accuracy, this was an important consideration in design.

2.2.1 Algorithm Design

With the previous in mind, we set out to design an algorithm suitable for not just this set of images, but for the general class of problem. As such, many of the examples in this section come not just from *C. elegans*, but also from the embryos of *D. melanogaster* and from arrays of T cells, imaged by Thomas Levario and Ariel Kniss-James, also from the Lu Lab. As will be illustrated, the algorithm was able to provide useful results for this situations as well as the DA9 synaptic domain.

In many fluorescently-labeled biological images, objects of interest can be characterized as regions of intensity brighter than the local surroundings, organized into clear spatial patterns. Generally speaking, images often contain sparse and Gaussian noise, uneven illumination, as well as extraneous fluorescent objects that are not of interest. These kinds of noise are pervasive in biological contexts, resulting from optical blur, light-distorting aspects of live sample, stochasticity of photon arrival in low-lighting conditions, and so on—biological samples are rarely pristine. In low-light applications, such as fluorescent imaging, sparse noise often presents a particular issue, with the magnitude of the sparse noise sometimes comparable to the magnitude of the signal. In addition, even when considerable care is taken to only fluorescently label objects of interest, background autofluorescence can easily conceal or obscure objects of interest, or even present spurious objects that fool image classification algorithms—for instance, when small pieces of debris contaminate the image, or when unrelated structures

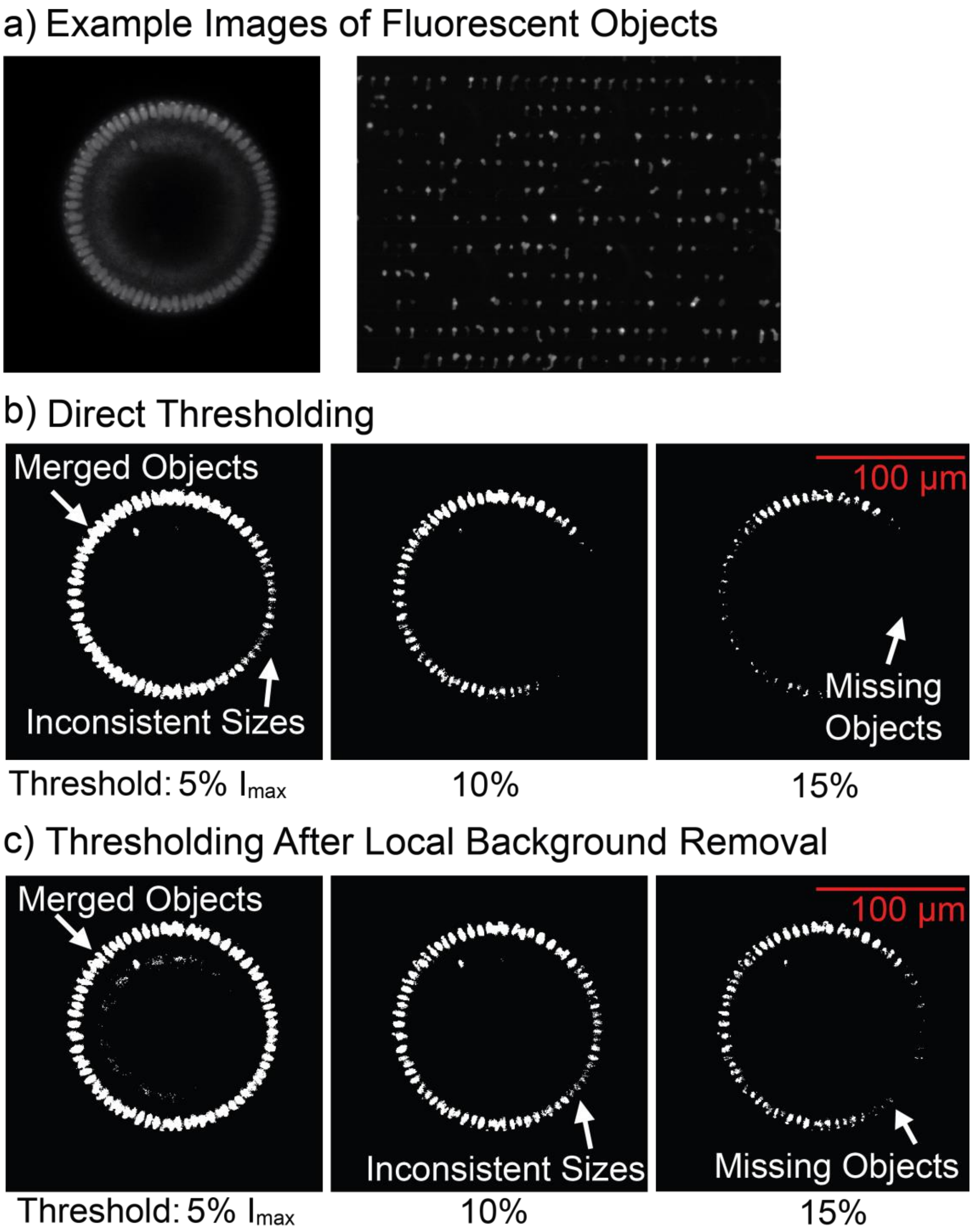


Figure 2.2 Direct thresholding often fails to segment biological images, even after local background removal. Part A shows two example images of fluorescent objects against a dark background. These are nuclei in a *D. melanogaster* embryo and T cells in a microfluidic array, imaged by Dr. Thomas Levario and Dr. Ariel Kniss as discussed in the main text. Part B shows the result of thresholding the embryo directly with 3 different thresholds. Part 3 shows the result of thresholding the same embryo after subtracting local background intensity. Neither thresholding is satisfactory.

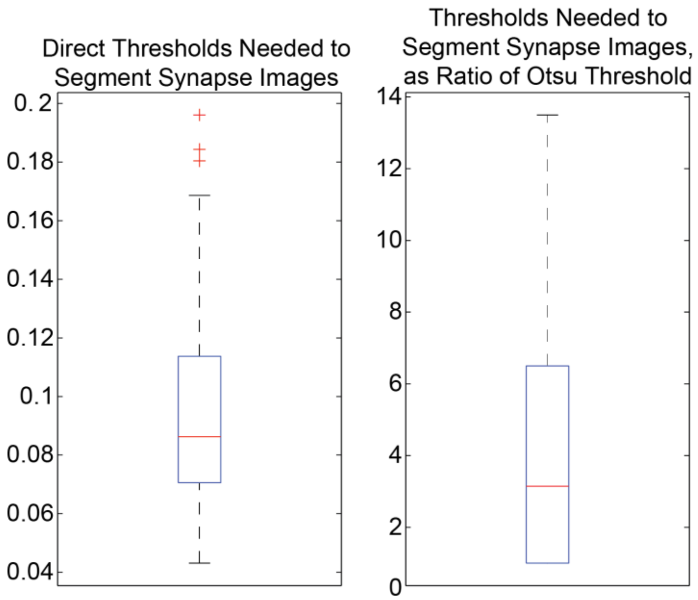


Figure 2.3 No consistent optimal direct thresholds or ratio of the Otsu threshold can be chosen that effectively segments the synapses of DA9. The thresholds shown in these boxplots were chosen manually for a random subset of the synapse images used in this chapter.

particularly uneven illumination, would confound thresholding-based techniques, as can be easily demonstrated in a representative example. Even subtracting the local average is often insufficient; under uneven illumination, both the objects and background are dimmer, and compensating for only the dimmer background still leaves dimmer objects that do not threshold well in combination with the brighter regions (Fig. 2.2 and 2.3).

We thus targeted these particular sources of noise. The general approach is outlined in Fig. 1C. Briefly, we first pass the image through a standard 3x3 median filter to lower the amount of sparse noise in the image, removing the sharp single-pixel oscillations in intensity that are common in low-light imaging. We also preemptively zero out regions of the image with intensity lower than a certain percentage of the maximum; this both reduces sparse noise and alleviates numerical issues in the next step. Uneven illumination is then dealt with by passing the image through a relative difference filter, a pixel-level filter defined as:

$$I_n = \frac{I - \mu_{50}}{\mu_{50}} \ (0 \ if \ \mu_{50} = 0) \qquad (1)$$

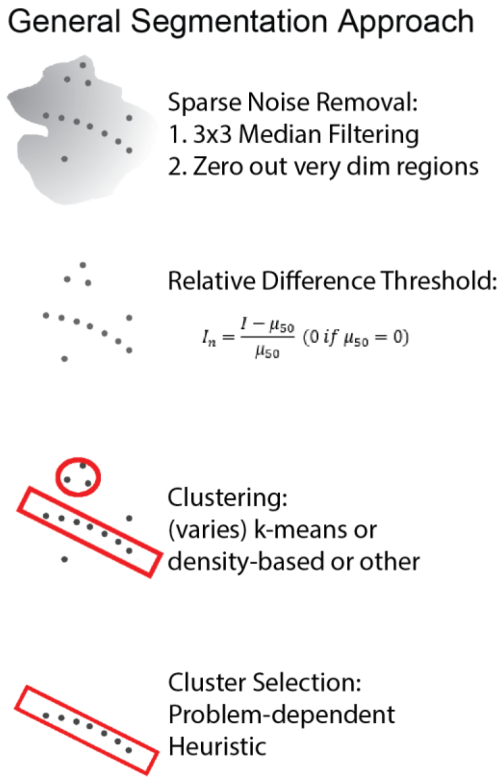

General Segmentation Approach
Sparse Noise Removal:
1. 3x3 Median Filtering
2. Zero out very dim regions
Relative Difference Threshold:
$I_n = \frac{I - \mu_{50}}{\mu_{50}}$ $(0\ if\ \mu_{50} = 0)$
Clustering:
(varies) k-means or
density-based or other
Cluster Selection:
Problem-dependent
Heuristic

Figure 2.4 Summary of the general approach used here for robust segmentation

where I_n is the new pixel value, I is the original pixel value, and μ_{50} is the average value of all pixels within a 50x50 region. This effectively replaces each pixel with its relative difference from the local average. Applying a threshold to the filtered image then effectively selects for pixels that are unusually bright compared to the local background. This is of course mathematically identical to $\frac{I}{\mu_{50}} - 1$, normalizing by the local background, but we prefer this version as it is clearer in meaning.

While this given procedure reveals objects more clearly than simple methods, it is also prone to generating anomalous objects in an image, often in dim regions of the image. Most of these objects can be removed by detecting and removing collections of pixels that are either too small or have too small a solidity (i.e. are too irregularly shaped). Fig. 2.5b-c illustrates this; initial filtering produces many incorrect objects in the center of these images, which are mostly eliminated by removing small and irregular objects. In many cases, however, these extraneous objects are not just processing artifacts, but represent actual objects in the images—special yolk nuclei (Fig. 2.5a-c), or fat droplets in the gut of *C. elegans* that resemble synapses (Fig. 2.1)—that are nonetheless not objects of interest. To alleviate this problem, we recognize that objects of interest that are typically found inside biological structures are by nature well-structured to allow for specific functions to occur; for example, cells in developing embryos are organized into highly stereotyped patterns. These are often the objects that confound pre-packaged algorithms or generic thresholding methods. However, taking advantage of the spatial patterns present in most biological images, we can exploit the structure present in biological images to efficiently segregate the objects into groups by clustering based on their locations in the image. Once this is done, we may select for only the relevant clusters by applying sorting criteria based on expected properties of the objects of interest. While performing this selection does require some custom algorithm development, the choice can often be quite easy; for instance, in the case of the *D. Melanogaster* embryos presented in next section, the nuclei of interest are usually arranged in a circle at the edge of the embryo.

The given procedure is flexible and easily adjusted; for example, as shown later, k-means clustering, which clusters objects to minimize the spread of individual clusters,[86] is chosen for the cell traps. In the other examples, where the objects of interest exhibit a uniform density and extraneous objects represent outliers, density-based clustering, using the algorithm DBSCAN[87], is used instead. It is of course also possible to use other clustering algorithms to suit the application, such as G-Means[88], though they were not used here. While it is desirable to avoid excessive amounts of calibration, these kinds of changes are often straightforward. This particular choice can be made in a principled manner—use DBSCAN when outlier removal is desired or the objects of interest are clearly more structured, and k-means when the exact location of the clusters is the more important factor.

2.2.2 Algorithm Accuracy over Three Experimental Conditions

In this section I test the general implementation of this procedure on three different experimental conditions. In the first two cases the examples come from outside *C. elegans*, and were imaged by others in the Lu Fluidics Lab. In the third case, images were taken by the author. Some specific details of algorithm implementation are provided in each case; full details are provided in Appendix C.2.

<u>Segmenting the Histone-GFP-Labeled Nuclei in Developing Drosophila Embryos</u>

Imaging *D. melanogaster* embryos is a challenge that often introduces uneven illumination and other types of imaging noise, but by using genetically-encoded fluorescently-tagged histones, we can obtain images relatively free of extraneous objects, making this system a good test of the algorithm under relatively clean conditions (Fig. 2.5a). Indeed, in the case of single-photon imaged embryos, no clustering was needed, as the number of extraneous nuclei was minimal. Images were filtered as described and thresholded with a single value chosen by manual inspection of a small subset of the images in each video, making sure to include both the beginning and end, due to photobleaching. The values chosen were 0.8 in the case of the single-photon images and 0.4 in the multiphoton images. In the multiphoton case, the center of embryo was identified by evaluating the centroid of the largest area after a simple threshold of the image set at 10% of the maximum intensity.

Fig. 2.5a-d shows two representative embryo images, taken from one single-photon and one double-photon imaged embryo. The filtering process is effective enough that the single-photon images are segmented accurately as is, but the additional artifacts (i.e. yolk nuclei) seen in the multiphoton images make it necessary to perform clustering. In this case, this is most efficiently done by estimating the centroid of the embryo and clustering based on the distance of objects from this centroid. To cluster, we use DBSCAN, since the objects of interest clearly differ in density from the anomalous objects, and because DBSCAN is also capable of excluding outliers, unlike k-means. The DBSCAN parameter used was 5 objects in the neighborhood of each point.

Once clusters are obtained, it is still likely that some clusters will be found containing anomalous objects. These can usually be excluded by excluding clusters with a small number of objects—in this case of nuclei in stages 5 and beyond of developing fly embryos, less than 10. After clustering, the segmentation of nuclei is accurate enough for analysis. We evaluated the average length of the detected nuclei at every time point, as well as their average distance from the centroid, in both embryos imaged with single-photon excitation and those imaged with multiphoton excitation. The accuracy of segmentation in the single-photon excitation images was sufficient that clustering was not used for these images. As is apparent if overlaid with embryo staging done by an expert, the information obtained from the nuclear segmentation clearly shows transitions in the embryo between stages, and can be used for the automated staging of embryos (Fig. 2.5e-f).

Quantifying the Effects of Perturbing *D. melanogaster* Embryos with Anoxia

Figures and data here are from Levario *et al.*, *"An integrated platform for large-scale data collection and precise perturbation of live Drosophila embryos"*[7] or Levario *et al. "Statistical comparison of dynamic phenotypes enabled by microfluidics and computer vision"* (In review). I developed the image segmentation and some of the analysis used herein, helped edit the manuscripts, and was credit as second author, but was uninvolved in data collection. Only relevant results are presented here; the interested reader is referred to the published manuscripts for further details.

The progression of *D. melanogaster* embryonic development is a topic of key interest in development biology. Using the segmentation procedure described, the nuclear areas of embryos imaged in a novel on-chip platform for developmental imaging were quantified, enabling the mitotic progression of the embryos to be tracked over stage 4 through 8 of development, using the information to accurately time the entry of the embryos into each mitotic cycle (Fig. 2.5e-f). In addition, by pulsing the embryos with 10 minutes of humidified nitrogen gas during nuclear cycle 13, it was possible to precisely quantify the effects of anoxia on embryonic development, showing that these anoxic pulses substantially delay later development, but without obvious permanent damage to the embryos (Fig. 2.6 and 2.7)[7].

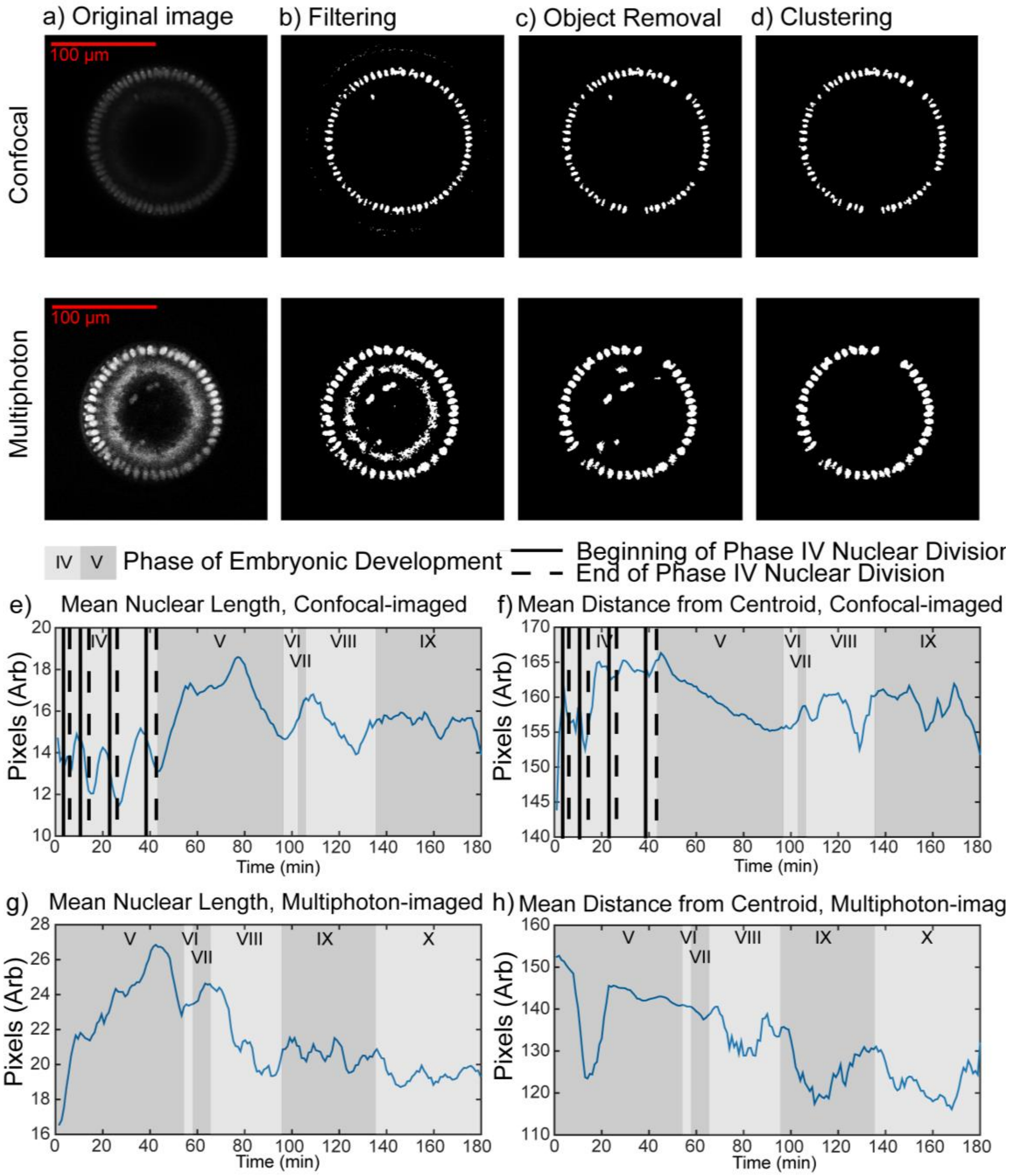


Figure 2.5 Segmentation and analysis of dividing nuclei in two *D. melanogaster* embryos, imaged with two different methodologies. Parts A-D illustrate the stages of the segmentation algorithm as applied to these two fluorescently-labeled embryos. While some nuclei are lost. The segmentation obtained is more than good enough to measure average properties of the nuclei. In Part E and F, dips in measurements of mean nuclear length and distance from the centroid correspond exactly to mitotic cycles in stage IV of embryonic development, and the combination of the two help mark the occurrence of later stages of development. In the Parts G and H, stages V-X of embryonic development can be effectively marked instead.

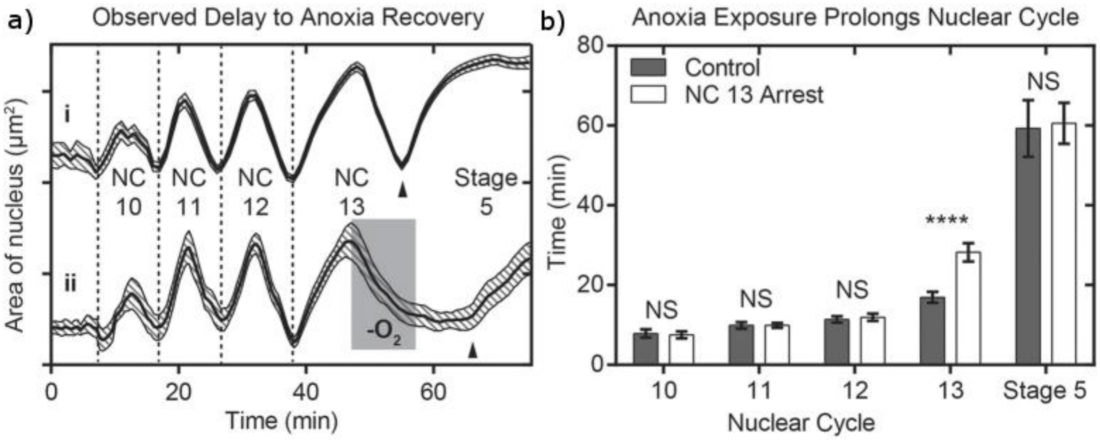


Figure 2.6 Anoxia induced delay in Stage IV *D. melanogaster* embryonic development. a) Average ± S.E.M. nuclear area trajectory for (i) 35 embryos grown in normoxia (ii) 14 embryos experiencing 10 minutes of anoxia during nuclear cycle 13. Black triangle indicates telophase to interphase 14 transition. b) Average ± S.D. durations for nuclear cycles 10-13 (stage 4), and stage 5 for these same embryos. Nuclear cycles 10–12 and stage 5 durations are not significant (NS) while nuclear cycle 13 is statistically different from control (****p < 0.0001. T-test). Figure and caption adapted from Levario et al., with permission[7].

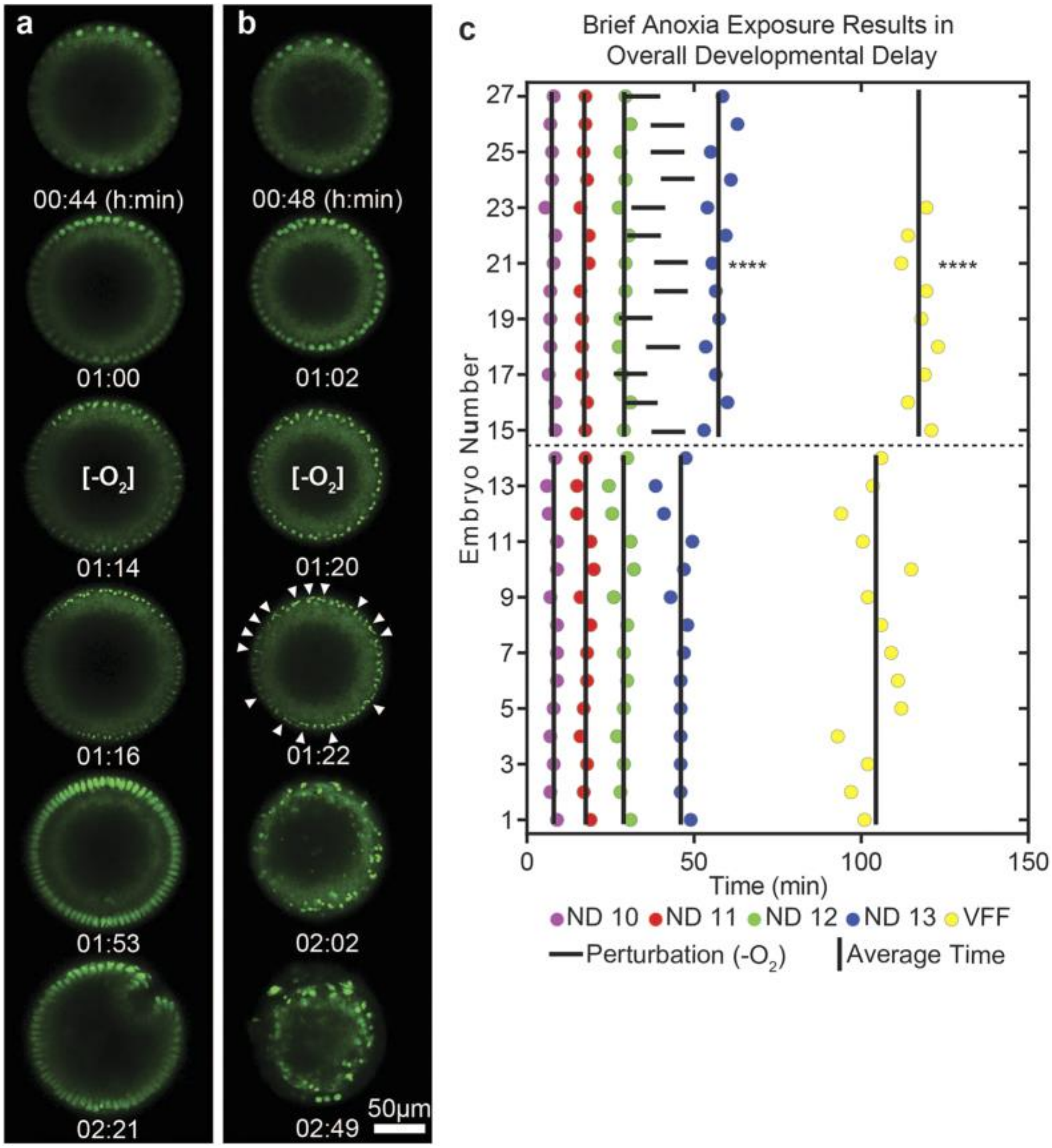


Figure 2.7 Recovery from anoxia-induced developmental arrest. a**)** Frames from a Histone-GFP expressing embryo that recovers from anoxia-induced arrest. Top to bottom: nuclear cycle 12, cycle 13, cycle 13 arrest in metaphase, cycle 13 anaphase-telophase transition, stage 5, and ventral furrow formation. b**)** Frames from a Histone-GFP expressing embryo that does not recover from anoxia-induced arrest. Top to bottom: nuclear cycle 12, nuclear cycle 13, nuclear cycle 13 arrest in metaphase, nuclear cycle 13 anaphase-telophase transition (white triangles indicate fused daughter nuclei), nuclear delamination (final two frames). c**)** The timing of milestones. Milestones include nuclear division (ND) 10, 11, 12, and 13, and ventral furrow formation (VFF). Embryos 1–14 are grown entirely in normoxia, and embryos 15–27 are exposed to brief anoxia in nuclear cycle. The timing of nuclear division 13, and ventral furrow formation are statically different from control (****$p < 0.0001$. T-test). Figure and caption adapted from Levario et al., with permission[7].

Segmenting and Tracking Fluorescent Cells in Microfluidic Arrays

To evaluate the effectiveness of the algorithm in a second, very different experimental system, we imaged Jurkat T cells labeled with the cytosolic calcium indicator, Fluo-3 AM, and loaded in a previously characterized microfluidic cell-trap array(6, 34). Images were acquired every 6 s for a total of 60 minutes while cells were stimulated with an oscillatory treatment of H_2O_2. Here, the organization of the system imposed by the microfluidic device provides a natural structure for clustering to segment the cells, but the primary challenge lies in identifying T cells that have been properly

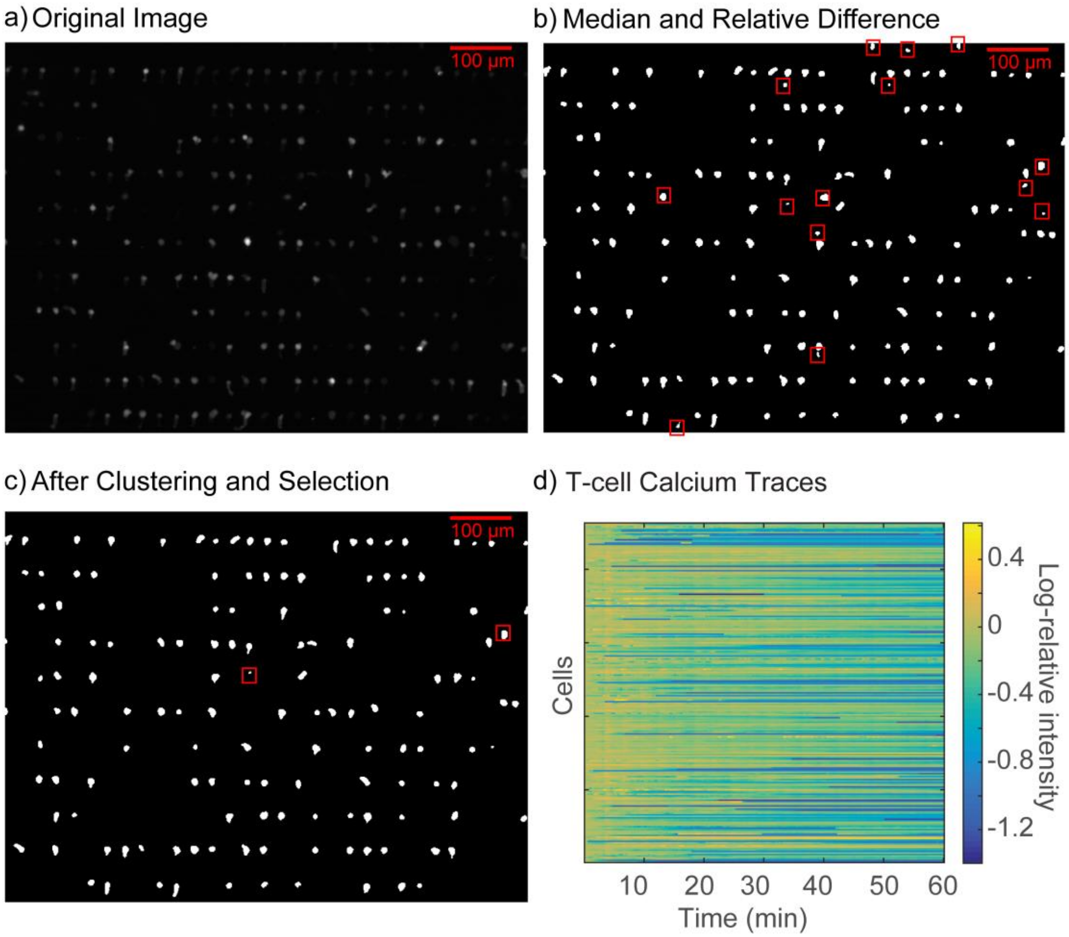


Figure 2.8 Segmentation and clustering for the T cell Microfluidic Array. Part A shows the original image. Part B shows the resultant binary image after filtering. A large number of cells outside of the traps still show up in the image, boxed in red. Part C illustrates how the majority of these improperly loaded cells can be removed with a clustering technique. Part D shows the resultant T cell calcium traces, unsorted and relative to the average intensity of all cells.

loaded, rather than T cells which are merely suspended in the channel or trapped in other parts of the device. Indeed, these out-of-focus T cells are sometimes trapped in patterned rows parallel to the T cells of interest, presenting a particular challenge for automatic analysis (Fig. 2.3). To avoid inaccurate segmentation, manual identification was previously used to identify only cells of interest, a procedure that took a substantial increase in time and was prone to inconsistencies in cell identification from person to person. This system provides a more stringent test of the use of clustering to remove extraneous objects, and to resist changes in intensity due to photobleaching, which was prominent in the data.

As before, we use the filtering to rapidly detect cells against the dark background (Fig. 2.4). In this case, we are not interesting in removing outliers with DBSCAN, because the anomalous objects often have a density similar to the objects of interest. By using k-means clustering on the vertical coordinate of the object centroid, we can effectively sort the objects into rows, choosing a clustering parameter of 18. Since there are 11 rows in the device, this accommodates both the expected properly loaded cells and the expected clusters of extraneous objects.

Removal of anomalous objects can then be done effectively by estimating the average spacing of the rows and removing the clusters that are too close to the neighboring rows. Specifically, this was done by merging clusters within 5 pixels of each other vertically, identifying the row spacing using a 2D-fourier transform, and discarding clusters that failed to be near a multiple of this spacing from the top row. This achieves a precision and recall of 90% and 95%, respectively (Fig. 2.11a). Note that the clustering parameter used for the data contained herein was chosen without reference to this charts, which show that an even better parameter can be chosen with a parameter scan.

From the segmented image for each frame, we can then calculate the calcium intensity for each cell throughout the dynamic experiment. In Fig. 2.8, we show the individual fluorescent calcium traces.

Characterizing the Synaptic Domain of the *C. elegans* Neuron DA9

Finally, we illustrate the use of the algorithm for the analysis of large sample-size, subtle features in *C. elegans*, where accurate quantification is especially important, since this is the application the algorithm was originally developed for. *C. elegans* individuals with a marker labeling the synapses of the tail neuron DA9 were imaged in a microfluidic channel and their synaptic morphology evaluated with the filtering and clustering algorithm. Here, the images are obtained under low-light (SNB-1::YFP is a dim reporter) high magnification conditions, highly prone to sparse noise; in addition, the images contain a large number of autofluorescent fat droplets that misleadingly resemble synapses. Despite this, while various types of noise in the image make direct thresholding inadequate for detecting synapses, the median and relative difference filtering process can readily identify synapses within the image (Fig. 2.9).

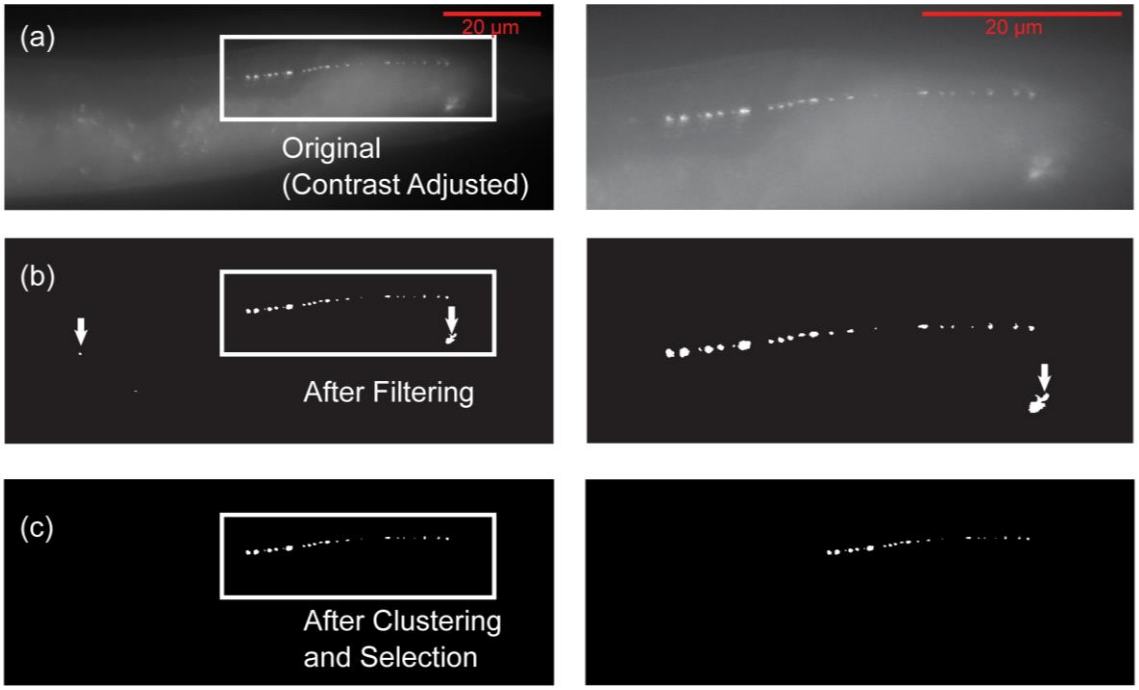


Figure 2.9 Segmentation of a DA9 synaptic domain with the filtering and clustering algorithm. For each part of the figure, the right side is a zoomed-in inset of the figure on the left side. After initial filtering, we see in part B that some extraneous objects still remain, but these are eliminated by clustering, cluster selection, and merger.

Here, there is little need to filter out small or misshapen objects, and synapses themselves are often very small. Thus, no size or shape-based object removal is applied. Filtered synaptic images contain a number of extraneous objects, however, and it is still necessary to distinguish the actual synaptic domain from the often very similar-appearing autofluorescence (e.g. from fat droplets within the intestine) in the samples. As before with *Drosophila* embryos, this is done with DBSCAN (neighborhood parameter 4) and cluster selection; clusters are selected by their linearity and lack of vertical self-overlap. DBSCAN is chosen here rather than K-means because of the clear difference in density between the synapses and other objects, and because of DBSCAN's ability to discard outliers. It is also necessary to detect clusters that are arrayed in nearly a line, for those occasions when missing synapses in the middle cause a synaptic domain to be separated into more than one group.

We find that clustering and selection is sufficient to identify the synaptic domain for analysis; the clustering step has a precision of 89% and a recall of 98% (not including images discarded for poor image quality—wrong neuron or synaptic domain out of focus).

In order to characterize the synaptic morphologies detected, a set of 29 features were measured (Table 2.1). Features were chosen so as to summarize the properties of the synapses or of the domain as a whole, without specifically targeting known differences between the strains. Of note, these features are different from those chosen previously for this kind of study.

2.2.3 Algorithm Speed and Parameter Robustness

Table 2.1 Features measured from the synaptic domain of ***C. elegans***

Synaptic Feature (Over Synapses in Image)	**Category**	**Mean**	**SD/Mean**
Area (Pixels in Synapse)	Size	Feature #: 1	14
Area/(Area Calculated from F6 and F7)	Shape	2	15
Perimeter/(Perimeter from F6 and F7)		3	16
Eccentricity of Approximate Ellipse		4	17
Diameter of Circle with Same Area	Size	5	18
Major Axis Length of Approximate Ellipse		6	19
Minor Axis Length of Approximate Ellipse		7	20
Mean Intensity of Pixels in Synapse (arb)	Intensity	8	21
SD/Mean Intensity of Pixels in Synapse (arb)		9	22
Perimeters (Pixels along edge of synapse)		10	23
Max Intensity in Synapse		11	24
Min Intensity in Synapse		12	25
Total Synaptic Intensity (F1*F8)		13	26
Additional Synaptic Domain Features			
Synapse Number	Misc.	Feature #: 27	
Mean Distance between Synapses (pixels)		28	
Synaptic Domain Length (pixels)		29	

For many applications, including the processing of large numbers of images, it is desirable that image characterization be done as a fast as possible, ideally on the order of a few seconds or less, and that it be robust to a wide variety of possible parameter choice. It is unavoidable that some parameters (i.e. the threshold used after filtering) must be changed for each experimental condition. Most of these may be rapidly estimated with visual testing of one or two representative images—the filtering parameters can be rapidly determined in this manner. For other parameters, such as the clustering parameter, that cannot be determined so readily, it is desirable for ease of use that the accuracy of the post-filtering procedure be relatively insensitive to changes in these parameters; a good algorithm should work adequately regardless of parameter choice, so that a reasonable choice will likely work for all applications. Beyond ease of application, having insensitive parameters also provides reassurance that the results obtained are not an artifact of the exact parameters chosen.

The filtering and clustering procedure fulfills these criteria. Even on a relatively low-end processor (a 1.6 GHz Intel Core i7 Q720M), both the filtering procedure and clustering take at most two seconds, and the overall algorithm for all our experimental conditions take at most a few seconds on the same platform, even taking into account feature measurement, *etc*. (Fig. 2.10).

Table 2.2 shows the set of calibration parameters that are used for each of the experimental conditions discussed in this chapter, including those which they have in common and those they do not (primarily in the cluster selection stage). Varying the non-

Table 2.2: Manually calibrated parameters in the three algorithm implementations

Parameters Shared by All	**Description**
Blackout Threshold	% of maximal intensity below which pixels are set to 0
Relative Difference Threshold	Areas with relative difference greater than this are considered objects
Solidity Criterion	Objects with a solidity less than this are rejected
Clustering Criterion	For K-means, the expected number of clusters, for DBSCAN, the number of objects expected in a neighborhood
T cell Array Parameters	**Description**
Merge Criterion	Clusters closer than this in vertical difference are merged
Spacing Criterion	Used to identify clusters in-between two rows of cells
Synapses Parameters	**Description**
Merge Criterion	Clusters closer than this in horizontal distance and that meet the criteria are merged
	Total # Calibrated Parameters
Nuclei (Confocal)	3
Nuclei (Multiphoton)	4
T cell	6
Synapses	5

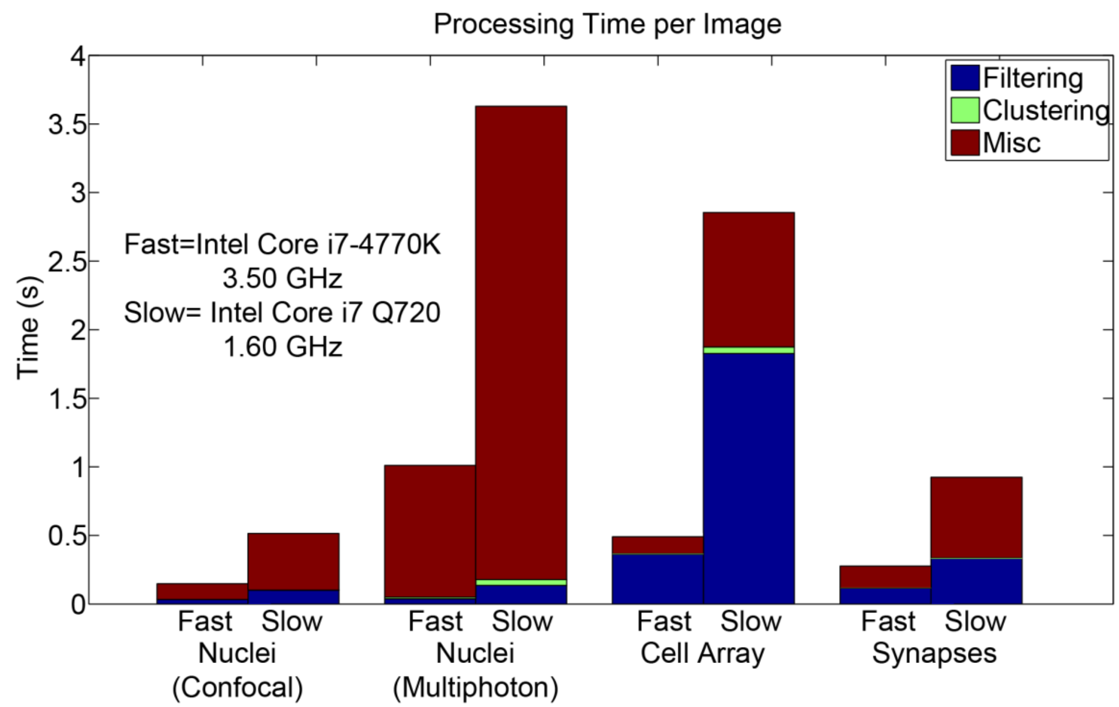


Figure 2.10 Processing time per image using the filtering and clustering algorithm, divided into steps. No variation of the algorithm takes more than a few seconds, or more than a second on a desktop processor. Often, the long step is the heuristic selection of clusters.

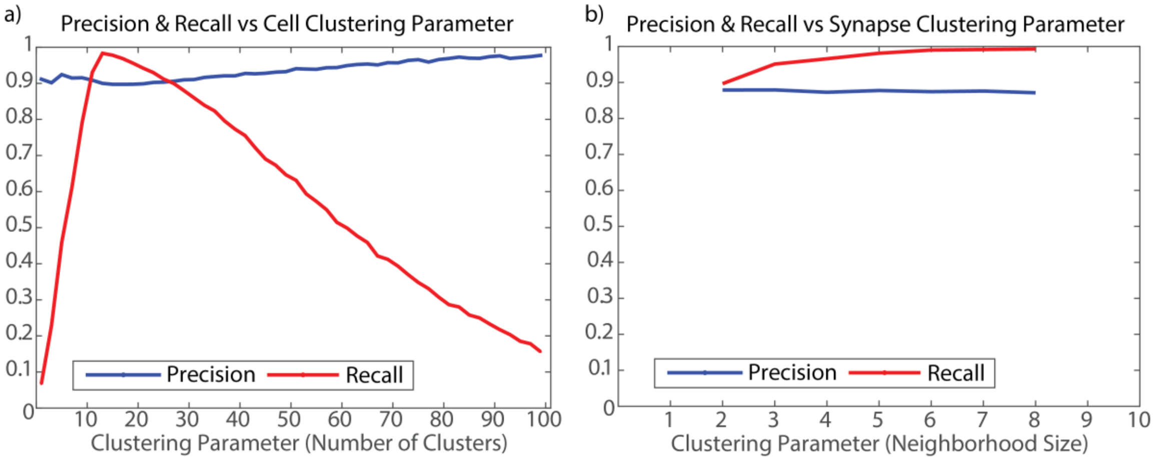


Figure 2.11 Precision and Recall versus choice of clustering parameter for (a) the T cell microfluidic array segmentation and (b) the synaptic domain segmentation. Both segmentations are reasonably stable within the reasonable set of parameter choices (11-25 for the cell trap segmentation and 2-8 for the synapse segmentation). Details of accuracy evaluation can be found in Appendix C.3.1.

filtering parameters over a wide range has relatively little influence on the accuracy of post-filtering clustering (Fig. 2.11), with little change in the precision and recall for both examples over a reasonable parameter range. The recall of the cell trap algorithm is more sensitive, but we note that the number of clusters input into K-Means should be somewhat the number of rows (about 11), to accommodate both the expected clusters and clusters of extraneous objects, so the value is still stable within the reasonable range of parameters.

2.3 Detection of Epistasis in Previously Studied Synaptic Mutants

In order to demonstrate the ability of this pipeline to answer interesting questions about specific phenotypes, and to illustrate its applicability to candidate gene approaches, we took a focused look at two synaptic mutants in the *wyIs92* genotype already described (with the SNB-1::YFP marker in the neuron DA9)[48]. We examined Day 1 adult individuals from four different strains, the base strain with just the marker, two single mutant strains with either the gain-of-function *unc-104 (wy673)* or the loss-of-function *jkk-1 (km2)*, and a double mutant strain with both mutations). These strains were generated by and received from the Kang Shen Lab at Stanford University[52].

2.3.1 Experimental Methodology

Worms were cultured on nematode growth medium (NGM) plates seeded with OP50 *Escherichia coli* bacteria according to standard methods at 20°C

As described in the introduction, *C. elegans* imaging and sorting is done within the single-layer sorting device. Fluid-suspended worms are pressure driven through the inlet into the imaging area, where pneumatic valves restrict worms for imaging. Then, worms are sorted into one of two exits. A cooling channel is used to flow a solution of 50% glycerol cooled to ~ 4°C, preventing worms from moving during fine imaging.

Valve control is done with custom software and an automated system of valve control, again as previously described. Fluid cooling is done with a custom-built peltier and peristaltic pump assembly[63]. Imaging was done at 40x on a Leica DMI3000B and Leica DM4599 inverted scopes, with a Hamamatsu C9100-13 EM CCD Camera.

By gently centrifuging the animals after removal from the plate and removing the supernatant, and also raising the inlet pressure on the device from 3 to 5 PSI, it was possible to cause adult animals to appear much more consistently in the imaging region on loading. To compensate for the greater density of animals and greater inlet pressure, valve pressure was raised to from 35±5 PSI to ~40±5 PSI to prevent animals escaping (the range on PSI values is due to variance in the stiffness of the device resulting from additional crosslinking during long-term storage), and the timing on the automated control steps was modified to enable continued automated imaging.

2.3.2 Additional Changes to the Imaging and Quantification Pipeline

In order to improve the efficiency and efficacy of the quantification pipeline a number of methodological changes were introduced. The most important changes were made to choices of features and the post-processing of the feature data.

The most important innovations were in the feature data and its post-processing. A small set of basic features relating to the size, shape, and intensity of the synapse were devised, chosen to reveal as much as possible about potential changes in synapse formation. Except for a few full-synaptic-domain features, these were calculated per worm as a mean over all detected synapses. Crucially, however, after substantial trial and

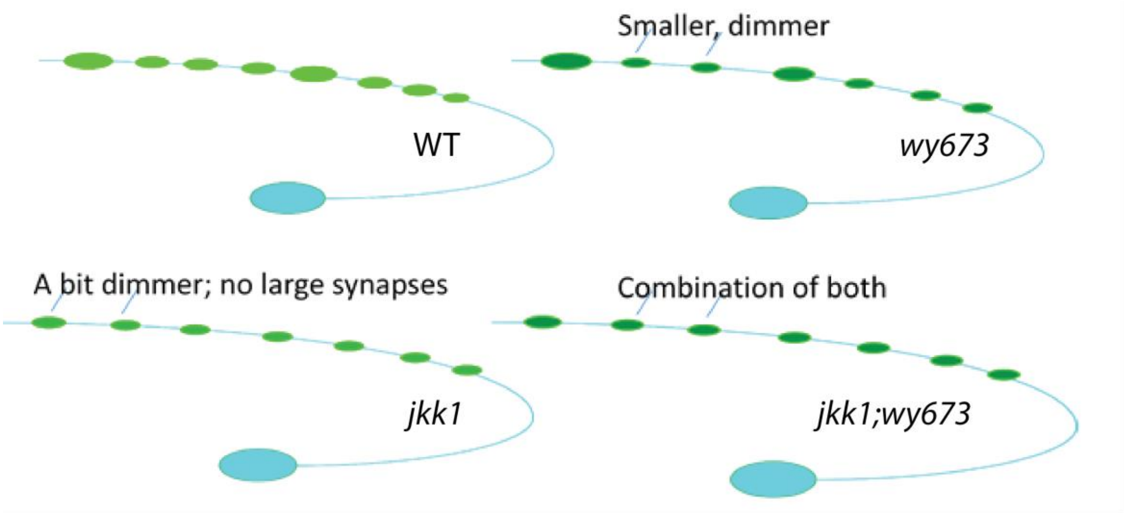


Figure 2.12 Illustration of the differences in the synaptic domain between the wildtype and mutant strains involved in this study.

error, the variability of each feature (the standard deviation divided by the mean) was also included as a feature, in order to capture differences between the synapses as well as global properties. The standard deviation itself was not used, as it was found to have a very high correlation with the mean.

Subsequent to this feature extraction, a number of outliers were observed, despite the strains in this case not being mutagenized populations. In the mutant screens previously conducted by the lab, these had probably been sorted as mutants and then failed to show penetrance of the trait to the next generations. In this case, however, these individuals added substantial additional variance to the data, and were sorted out by a common outlier removal criterion; that is, individuals were sorted out of the dataset if in any individual feature they fell outside the range $(25\, percentile - 1.5 * IQR, 75\, percentile + 1.5 * IQR)$, where IQR is the interquartile range.

2.3.3 Results

The purpose of the study is to compare the wildtype strain with the mutants, demonstrating that the results of this algorithmic analysis are consistent with previous manual characterization. By conducting a detailed, higher-sample size analysis, we find that we can in addition draw biologically important conclusions that are only feasible due to the quantitative nature of the analysis. Previous qualitative observations by the Shen Lab have noted that both the *unc-104* gain of function and *jkk-1* mutants substantively reduce the intensity and size of the synapses in the neuron DA9, while the double mutant appears to show a combination of the two phenotypes (Fig. 2.13 and 2.14)[52].

When analyzed with the image processing pipeline, all three mutant strains show clear differences from the wildtype (Fig. 2.15a-c). By examining that features that show the most prominent differences, it is clear that *unc-104* greatly decreases synaptic size and intensity, while *jkk-1* reduces the variance of synaptic size, while exerting a much milder effect on synaptic intensity (Fig. 2.15a a-b). This is consistent with previous manual characterization of the strains, which indicates that both mutant strains possess

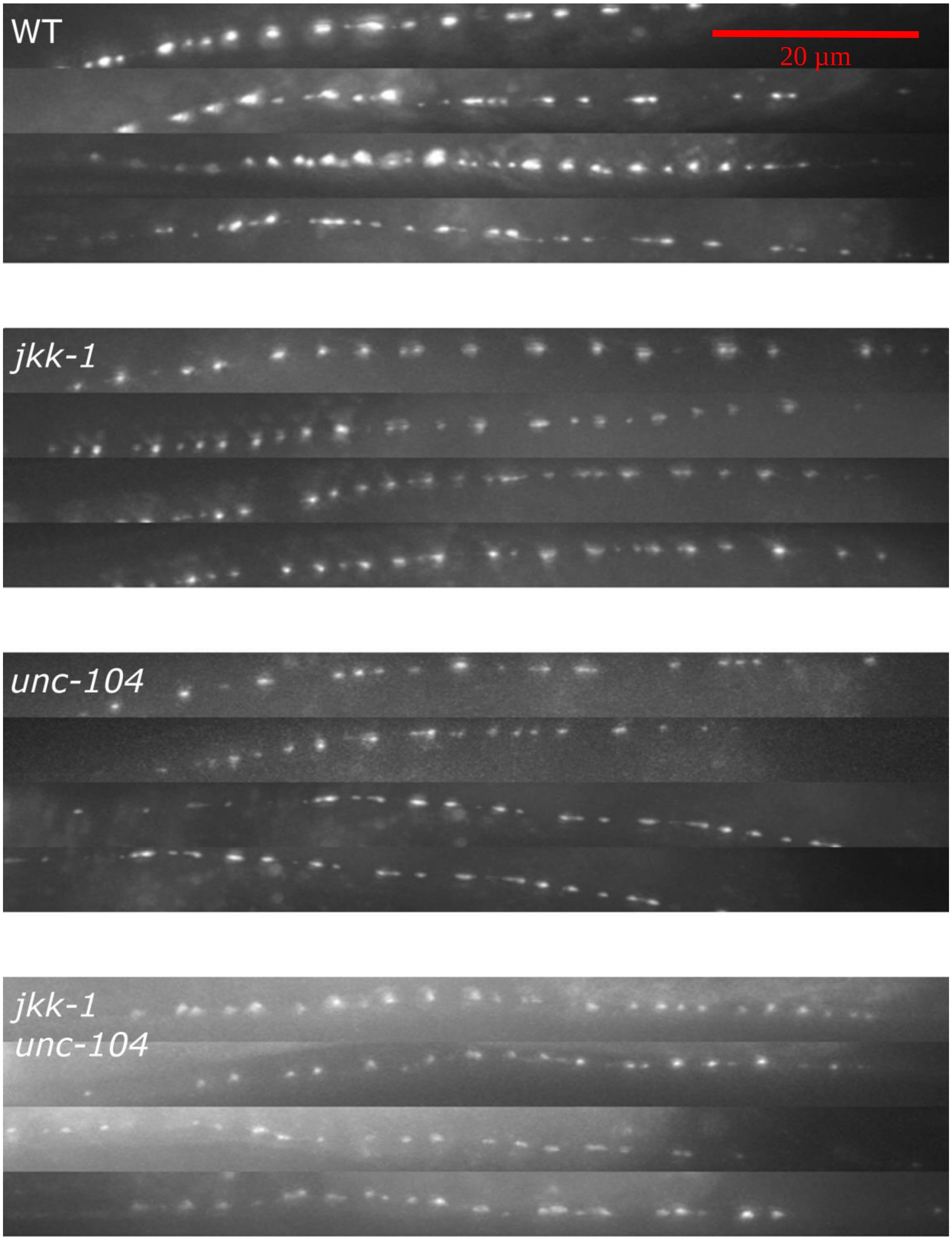


Figure 2.13 Representative images of DA9 synaptic domains from the wildtype, *jkk-1*, *unc-104*, and double mutant genotypes. Images have been contrast adjusted for visibility.

dimmer and smaller synapses. Moreover, *jkk-1* shows a loss of the brightest synapses in the domain, which is apparent from the different size and variance and can be shown explicitly with additional post-hoc analysis (Fig. 2.16). While this phenotype had been suspected qualitatively, this is the first clear demonstration of the fact, and would not have been initially noticed without the variability features I introduced in Section 2.3.

Previously, the double mutant *unc-104;jkk-1* has been observed to be dimmer and smaller than either single mutant. While Fig. 2.15c is suggestive of this, this difference is only evidently significant in *jkk-1* (Fig. 2.15d-e). Indeed, a comparison between the strains reveals that *unc-104;jkk-1* is very similar to *unc-104*, far more similar than *jkk-1* is to the wildtype. This shows that *unc-104* is epistatic to *jkk-1*. As a further, more detailed study of these epistatic effects, the summed effect of the two single mutants on the

$$(unc104; jkk1_n - \mathrm{WT}\ _n) \qquad (3)$$

The results are shown in Fig. 2.15f. It is apparent that the intensity of fluorescence in the synapses in the double mutant is far lower than would be expected without epistasis, whereas the physical size of the synapses in the double mutant shows little epistasis. Details of the estimation of statistical significance can be found in Appendix C.3.2.

The clear epistatic effect of *unc-104 (wy673)* on *jkk-1 (km2)* has not been previously characterized, due primarily to a lack of detailed quantitative measurements in hand-curated data, but is clearly indicated in a quantitative analysis such as done here. Further, within this quantitative analysis, it is clear that this epistatic effect is most pronounced on the concentration of the marker-labeled SNB-1::YFP within the synapses, while having little effect on the size of the synapses. Previous work indicates that both UNC-104 and JKK-1 play a role in the trafficking of synaptic material into and out of synapses[52], but this epistatic effect may illustrate to the poorly understood functions of *jkk-1*, perhaps indicating a separate mechanism for regulating the amount of synaptic

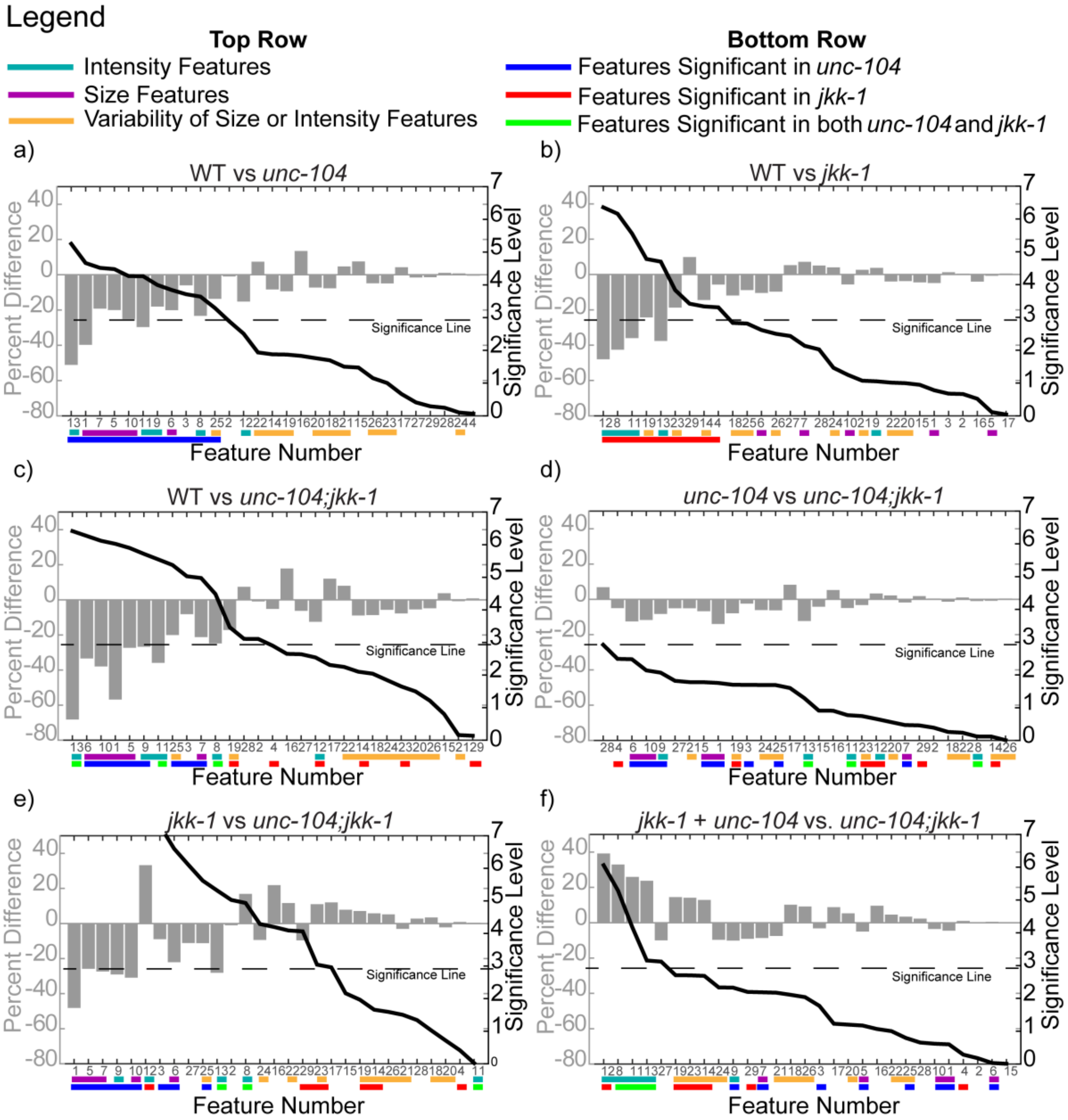


Figure 2.14 Comparisons of synaptic features between the mutant strains show expected effects and epistasis. Plots here show the relative difference between strains. The grey bars show the percent difference (latter minus former) for a given feature, while the dark black curve shows the corresponding significance level (two-tailed Welch's t-test). The horizontal dotted line indicates the significance level after the Bonferroni correction. Features are sorted from highest significance level to lowest. Colored bars illustrate features of interest, as indicated in the Legend. (a-c) show that the mutant strains are distinguishable from wildtype, with the phenotypes expected. (d-e) shows that the double mutant is highly similar to *unc-104*, far more than would be expected given the effect of *jkk-1* in (b). *unc-104* is thus epistatic to *jkk-1*. (f) illustrates this by comparing the double mutant with the mathematical sum of the differences in *jkk-1* and *unc-104*. The actual double mutant has much higher values in intensity measures than would be expected from a linear combination of differences. The sample sizes are 34 (WT), 35 (*unc-104*), 61 (*jkk-1*),70 (*unc-104;jkk-1*).

additional gain-of-function mutations in *unc-104* (*wy798*, *wy865,* and *wy873*). While some subtle differences were noted, the strains appeared similar to *unc-104 (wy673)* (Fig. 2.16).

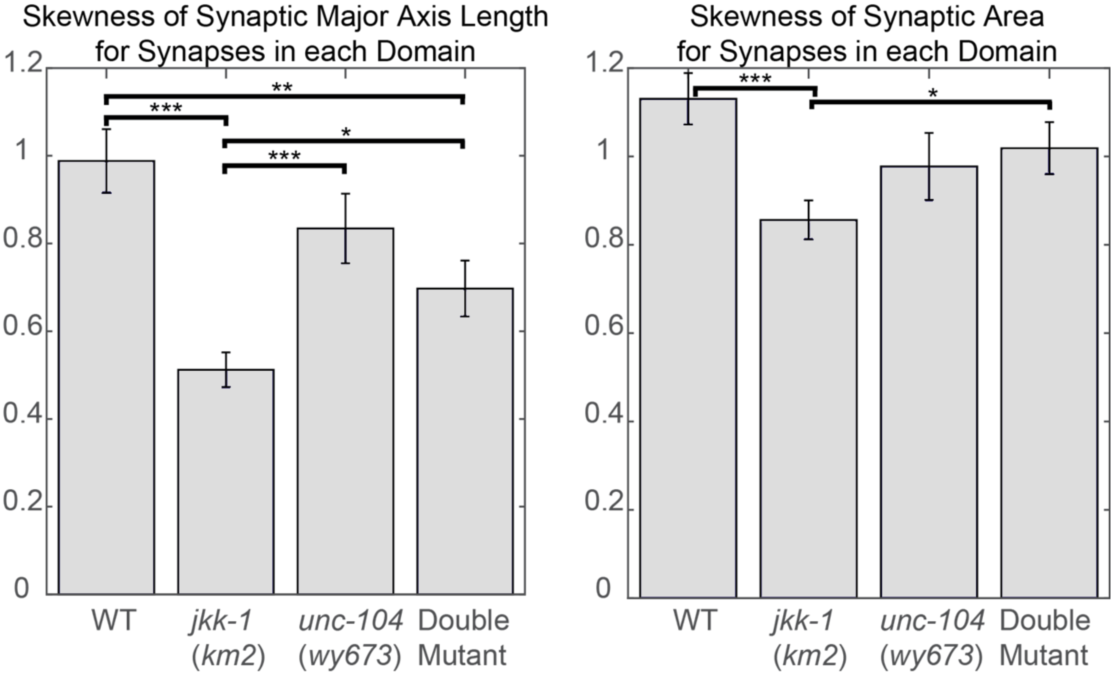


Figure 2.15 *jkk-1* loses large synapses relative to other strains. Skewness is a measure of the left-right balance of the histogram of values. *jkk-1* is significantly less biased towards larger synapses than the other strains.

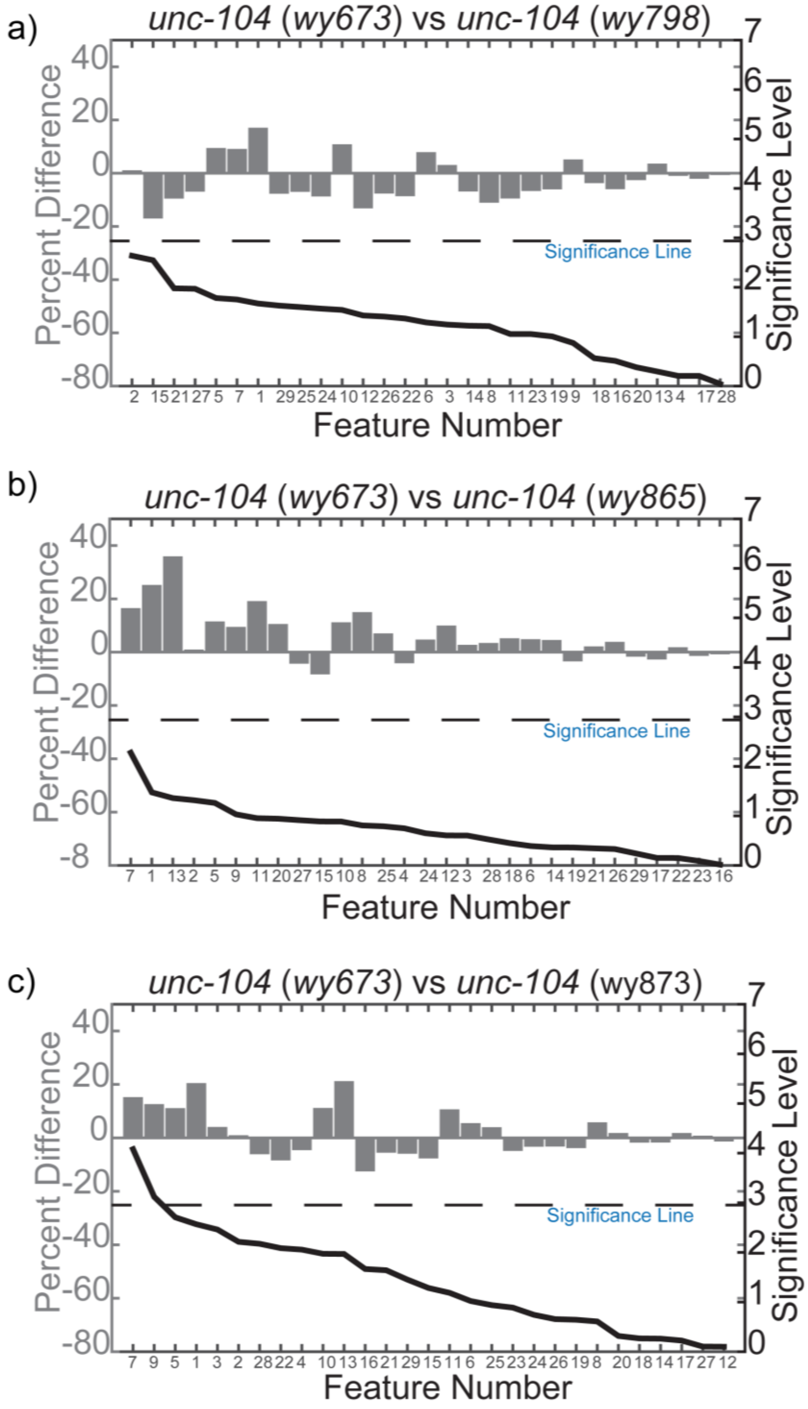


Figure 2.16 Comparison of the *wy798*, *wy865*, and *wy873* gain of function alleles of *unc-104* with the *wy673* used thus far in this chapter. Except for two features in *wy873*, these strains are statistically indistinguishable from *wy673*. Since these are still different alleles, it's not too surprising that *wy873* is slightly different.

2.4 Discussion and Conclusions

In this chapter of the thesis I developed a new robust and untrained image segmentation technique, specialized for identifying relevant fluorescent dots in an image in the presence of confounding, similar-looking objects, and use it to demonstrate the first application of our high-throughput imaging technique to this kind of candidate genetic approach.

The algorithm was targeted towards the deficiencies of the previous SVM-based image segmentation pipeline, including difficulty separating fat droplets from synapses under high exposure conditions, overfitting for individual experimental conditions, and complexity of implementation. Thus, it is appropriate for situations where the experimental conditions or condition of the animals change frequently, and was successfully generalized to applications beyond *C. elegans*. It is thus a much more suitable approach for the quantification of large numbers of different strains, which will be directly useful for the QTL analysis in Chapter 3. We demonstrate this segmentation technique on more than just *C. elegans*, extending it to both *Drosophila melanogaster (D. melanogaster)* embryos and human T cells in a microfluidic device, and extracted biologically valuable data in all three cases.

2.4.1 Limitations and Considerations

While demonstrably fruitful, it is not likely that the image processing pipeline as outlined is *optimal*. While the accuracy of the segmentation and clustering is clearly more than sufficient for the given application, it is somewhat below that of the trained SVM method, sacrificing some accuracy for easier implementation and better generalization between strains and conditions—in particular the segmentation is vulnerable to inaccurately lumping multiple synapses together as one synapse. While some different segmentation techniques were examined during initial testing, no systematic evaluation of all possibilities was carried out; the emphasis was on obtaining one that was accurate enough suited the desired criteria, not necessarily the best one. In the future, if exacting accuracy is required, such a systematic study may be carried, but, as I have shown, the current approach is more than sufficient to demonstrate results in several biological systems.

With regards to broad generalization of the segmentation and clustering approach to different model systems, the biggest drawback is the reliance on heuristic expert knowledge to select the final, correct clusters. There is unfortunately no way to produce a simple cluster selection technique to all possible arrangements of fluorescent objects and confounds, and the user is forced to produce their own technique for a given situation—e.g. clustering by distance from the center for *D. melanogaster* embryos or by row distance for the cell trap arrays. One possibility for future investigation is to train a machine learning algorithm to detect correct *clusters* rather than the original images, a problem that should be considerably easier to solve. This, however, removes some of the key upsides of performing segmentation and quantification in this manner.

2.4.2 Implications for Candidate and Quantitative Genetics

The emphasis on robustness to experimental conditions, particular high exposure and dim marker conditions arose out of more than just a desire to improve the algorithm. The application of this high-throughput imaging approach to candidate genetics and quantitative genomics, requires the accurate segmentation of relevant phenotypes in as many conditions as possible, including mutant conditions. This is the reason it was considered necessary to re-examine the features used to characterize the synaptic domains, and to include features such as variability—although the choices made are of course specific to this particular phenotype—and to perform a layer of outlier detection.

The culmination of this effort produces feature data that is precise and accurate enough to verify qualitative observations of synaptic domains, and to produce a biologically plausible account of epistasis between two different synaptic proteins. This serves as a crucial demonstration that high-throughput phenotyping can be just as valid in for a candidate approach as in mutant screening, and can be relied upon to measure subtle features in a phenotype as precise as synaptic domains.

This reassurance is necessary to carry the technique forward beyond characterization of mutants into the realm of genetic mapping and QTL analysis, where exact quantitative accuracy is necessary. Further, the applicability of the segmentation approach to dim markers confounded by objects such as fat droplets is of critical importance to a study that relies on the use of heterozygotes, as Chapter 3 of this thesis does.

CHAPTER 3

HIGH-THROUGHPUT IMAGING AND QTL MAPPING FOR THE IDENTIFICATION OF SUBTLE SYNAPTIC MORPHOLOGY-AFFECTING LOCI

3.1 Motivation, Overview and Background

In this chapter, we give a new dimension to the high-throughput phenotyping methods developed in Chapter 2, seeking to take it beyond the mutant screens and phenotypic studies where it has already been applied, into the realm of quantitative genetics—powerful, sample-size dependent techniques such as linkage mapping[89], QTL analysis[19], and genome wide association studies[90]. *C. elegans* was the first multicellular organism to have its genome completely sequenced, enabling techniques first developed for single-celled organisms to in principle be applied to identify the genetic loci that influence key phenotypes[91]. However, the application of techniques like QTL analysis to *C. elegans* has been constrained by the reliance of the method on obtaining large volumes of quantitative phenotype data, in particular on a large number of different RILs. Because of the difficulties inherent in adapting existing RILs to the use of new genetic markers, and the burdensome task of generating a new set of RILs for a given problem, the use of QTL has thus far been restricted to phenotypes that can be observed readily and quantitatively by eye, without reliance on additional genetic manipulation or genetically-encoded markers—for instance number of progeny, pharyngeal pumping, lifespan[92-95] and so forth[96]. This restriction limits the reach of what is otherwise a fruitful technique for discovering genome-wide and unsuspected genetic loci for a given phenotype. It is this restriction that we address in this chapter of the thesis, applying the techniques of QTL loci to fluorescently-labeled synapses, a subtle and genetically-encoded phenotype whose genetic origins remain only partially explored, but where understanding is of critical importance to some of the most common neurological disorders, such as autism spectrum disorder or schizophrenia[97-101].

We open this chapter with a brief background on RILs and QTL analysis.

3.1.1 Recombinant Inbred Line (RIL) Generation

Speaking most generally, a Recombination Inbred Line is a strain that represents what is essentially a hybridization of two different pure-breed lines, strains that have been bred to themselves so extensively that they are isogenic and homozygous at all loci[102]. At its most basic level, by taking two such strains, mating them to each other, and purifying the resulting genetic recombination events, it is possible to make a new inbred line that is the combination of the two parent strains. The concept of inbred lines and RILs was first developed in mice[103], but has only recently been extended to *C. elegans*, despite the genetic and mating problems that make *C. elegans* ideal for genetic manipulation, e.g. easy breeding and naturally "inbred" lines[104, 105].

A number of different techniques exist for generating RILs. The fastest and most straightforward involves mating the two parent strains and producing as many F2 progeny as possible, followed by stabilizing homozygous loci by picking individual progeny and allowing extended self-propagation. While relatively convenient—but still labor and time-intensive—this method of generating RILs produces a relatively low level of mixing between the two parent genotypes, since a single mating event provides only one chance at recombination. This produces RILs that are individually of relatively little statistical value in genetic mapping—each individual strain is on average only capable of localizing a given phenotype trait to a very large region of the chromosome, and mapping any individual phenotype in detail requires either luck or examining a prodigious number of strains[102].

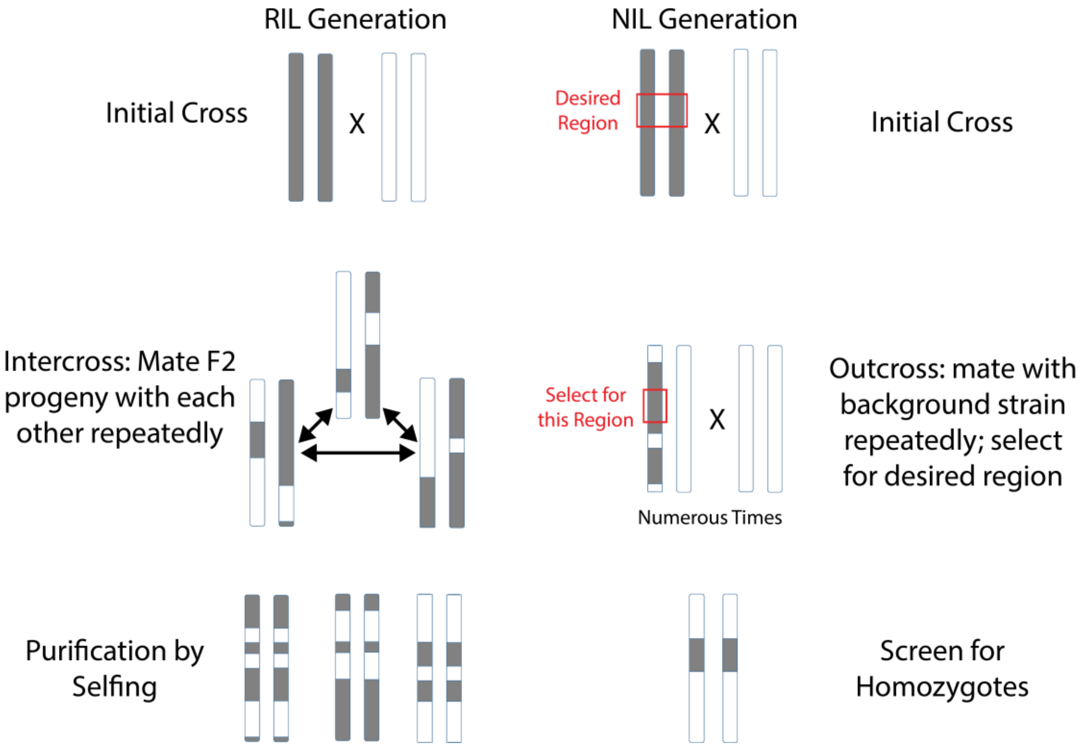


Figure 3.1 Recombinant Inbred Line (RIL) generation (left) and Near Isogenic Line (NIL) generation (right). For RIL generation, after intercrossing, lines may be purified by picking individual worms and allowing self-mating for a large number of generations (7+). For NIL generation, selecting for a given genetic region may be done by picking individual progeny and screening a sample their progeny for genetic markers flanking the desired region. The final NIL homozygote may be selected for by individual worms for which *all* the progeny sampled have the desired genetic markers. Much of the relevant techniques involved are discussed in Appendix B.

Consequently, the most useful modern RIL strains are generated by a so-called advanced intercross, where the F2 progeny and subsequent generations are mated to each other an extensive number of times, in order to generate new recombination events and merge together varying combinations of already existing recombination events[102, 106]. This enables the final RILs generated to contain highly mixed, nearly random mixtures of the two parent genotypes, greatly increasing the mapping power of each strain imaged. In order to efficiently sequence the RILs generated without requiring whole-genome sequence for every strain, single nucleotide polymorphisms (SNPs) between the two parents are identified and used as “markers”—not to be confused with fluorescent markers—to indicate which parent a given locus in a strain originated from. These SNPs can be readily identified in each strain without resorting to whole-genome sequencing[102].

A number of different schemes exist for maximizing the effectiveness of the subsequent intercrosses, including circular mating, inbreeding avoidance, or random assortment, but in most cases maintaining a large population of different intermediate strains is much more important than adhering to any particular scheme[106].

One of the major methodological hurdle to QTL analysis that this chapter addresses is the time, resources, and manpower that goes into constructing a new set of RILs, a necessity if the phenotype being examined requires a marker not already present in an existing set of community RILs. Constructing a set of RILs is an endeavor that requires extensive planning, careful repeated mating, and the simultaneous maintenance of dozens or hundreds of strains[102]. It is for this reason that the use of genetically-encoded tools—such as fluorescent markets—to study particular traits in RILs is rare, as this requires either the successful uniform integration of the genetically-encoded modification into a large number of different strains—something beyond the reach of even many of the most advanced genetic manipulation tools[38], or the laborious recreation of a new set of RILs from scratch, in which the consistency of the exogenous fluorescent marker cannot be assured due to the number of recombination events. Because of this, the use of fluorescent markers to label useful phenotypic features in RIL strains is nearly nonexistent, despite the tremendous advantages the use of RILs provides—something this chapter intends to change.

3.1.2 Quantitative Trait Loci (QTL) Mapping

Of course, the ultimate purpose of having a large set of RILs is to enable quantitative genetic mapping, the use of statistical and associative techniques to map phenotypic traits to specific regions of the genome. One of the most powerful such methods is referred to as Quantitative Trait Loci (QTL) Analysis, or QTL mapping, where quantitative phenotypic measurements on a large number of RILs are combined with genotypic data to statistically infer the genetic loci driving variation in the given phenotype[19, 104]. With methods like composite interval mapping and sufficient sample size, it is sometime even possible to infer the influence of more than one loci[107, 108].

General Idea

Broadly speaking, a quantitative trait locus is a region of the genome that is associated with a particular phenotype; that is, differences in that region of the genome can cause changes in the value of particular phenotypic trait either directly or combined with changes in some other region of the genome. Given measurements taken on populations with variation on some parts of their genome, QTLs can be found by examining the statistical association of each individual loci with changes in the phenotype. QTL analysis techniques, generally speaking, take as inputs phenotypic measurements on a wide variety of different populations with known genotypes, and produce as an output an estimate of the how likely it is that each locus in the genotype affects the given trait[19].

Very roughly speaking, the key idea is that if variation A of a given loci is more frequently found in individuals with a higher value in a given trait, while variation B is more frequently found in individuals with a lower value, then this given locus has a chance of being a QTL that affects this particular trait.

Limitations and Requirements

While this is a very crude formulation, it immediately illuminates some key properties of QTL analysis: QTL analysis can only search for QTLs within the restricted genotypic space represented by the variation that exists in the populations that have been studied. It does not matter if variation C at a given locus changes the phenotype if variation C is not present in the study, or if there is no other variation to compare it to. At the same time, not all the variation that could be studied might be interesting for a given study—if the study is intended to find QTLs that affect the size of synapses in a particular neuron, variations in loci that cause the neuron to fail to develop entirely are non-germane. To perform a QTL analysis on the entire genome, then, it is thus critical to have as much relevant variation as possible at as many loci as possible, while excluding variation that is likely to be irrelevant.

It is also clear that cross-correlation between the individual loci must be minimized—for example, if two loci are very frequently found together in any given strain, it is difficult to distinguish between the effect the two loci have on a phenotype. If there are many loci all with significant cross-correlation, then the problem of inferring a relationship between each genetic loci and the given phenotype becomes intractable.

It is also clear why QTL analysis focuses on quantitative phenotypic traits, rather than categorical or qualitative phenotypic traits, as the ability to use a continuous or at least ordinal variable for the phenotype provides vastly more statistical power for association testing. It is absolutely necessary to accurately examine as many different genotypes as possible, again for reasons of statistical power. Because of this, QTL analysis is confined to phenotypic traits that can be practically quantified with high efficiency, on as many strains as possible. In *C. elegans*, this has typically meant whole-worm traits observable under a low magnification microscope[93, 109].

Need for RILs

It is precisely to address these needs that RIL strains are generated, sequenced, and used for QTL analysis, particularly those that are generated by advance intercrossing. By construction, provided that the two parent strains contain the important variation being sought, RIL strains are designed to contain variation at as many loci as possible, since each recombination event introduced during RIL generation creates new genetic regions which travel separately from their original surroundings, representing new loci for which variation can be examined. Via extensive interbreeding, the effect of genetic linkage, the primary source of correlation between different loci, is mitigated. Consequently, the difficulty in making RIL strains that use genetically-encoded fluorescent markers is an enduring roadblock to investigating the genetic origins of phenotypic features which require these markers to be examined.

Analytical Methodology

While it is possible to use a technique like Analysis of Variance (ANOVA) to perform the actual statistical analysis necessary to extract QTLs from a given phenotypic and genotypic dataset, ANOVA carries the significant downside that it is unable to examine locations in the genome other than the exact locations of the SNP markers, leaving potential gaps in the coverage of the whole-genome QTL analysis. The most common method is instead interval mapping, introduced in 1989 by Lander and Bolstein[19]. In this method, regularly spaced intervals in the genome are considered as potential QTLs. For each such interval, a model is constructed assuming that the interval is the single true QTL, and the probability that this model leads to the observed results is calculated as a likelihood of odds (LOD) score. Loci with LOD scores that achieve statistical significance are then considered as the location of putative QTLs. The desired level of statistical significance is usually calculated by permutation testing, by randomly shuffling the relationship between the genotypes and phenotypes in the data and finding the LOD score that excludes all but a certain percentage of the random datasets, corresponding to the desired p-value.

One weakness of standard interval mapping is that the underlying models assume only the existence of only one QTL, which can cause substantial distortion to the location of the putative QTL. Newer methods, in particular composite interval mapping, are capable of handling the possibility of multiple QTLs. This is of great value in phenotypes with heavy multigenic inheritance, into which synaptic phenotypes almost certainly fall—though this is dependent on the degree of variation found in the parent strains[107, 108].

3.1.3 Motivation and Goals of this Study

While a powerful methodology for unraveling the genetic origins of phenotypic features, data collection for QTL analysis in *C. elegans* has thus far been limited by the technical requirements required to enable the statistical methods involved. The situation is, however, far worse for the study of subtle phenotypes such as synaptic morphology, which has been inhibited by the need to generate new RIL strains with the requisite fluorescent markers for study, and by the difficulty of quantifying such a subtle phenotypic features on a large scale, since this requires the accurate high-throughput imaging of numerous RIL strains. This is particularly disappointing, since it is precisely these kinds of subtle synaptic phenotypes that are hypothesized to drive some human diseases[97-101].

The difficulty of quantifying these subtle features on a large scale has been addressed by Chapter 2 of this thesis, however; it remains only to consider methodologies for using the existing RILs in conjunction with a fluorescent marker. As we shall see, it is possible to examine fluorescent features in the F1 progeny of a marker strain and RIL strain, enabling QTL mapping on synaptic morphology in *C. elegans*. Key to this again are the contributions of chapter 2: heterozygote markers are very dim, about the same intensity as autofluorescent fat droplets in *C. elegans*, and the previous generation of automated methods would have balked at segmenting them accurately. Traditional manual imaging and segmentation, on the other hand, would be prohibitively labor and time-intensive.

By using a collection of advanced intercross lines generated by the laboratory of Erik Andersen[105] and near isogenic lines[110] (NILs, see Fig. 3.1), we identify a region of chromosome IV that is a QTL for differences in synaptic morphology between the Bristol Wildtype N2 and the Hawaiian isolate CB4856, demonstrating the feasibility and value of this approach to QTL mapping in a previously infeasible context.

3.2 The Parent Strains N2 and CB4856 have Subtle, but Measurable Differences in Synaptic Morphology

Before setting out to use a modified methodology to perform fluorescent marker-based QTL, it was necessary to first discover two potential RIL parent strains with a verifiable difference in synaptic morphology. As discussed in the previous section, it is most likely futile to perform a QTL mapping on a set of strains that may not even have variation that affects the phenotype of interest. To ensure that a given a set of RILs has variation affecting a given phenotype, it is necessary to first verify that the parent strains differ in some way in the given phenotype—*particularly* in heterozygote crossing with the marker strain.

To do this, I conducted an initial examination of two non-N2 strains of *C. elegans*, the liquid culture strain LSJ2 and the Hawaiian Wild Isolate CB4856, using the same marker strain and neuron as study as Chapter 2 (the genotype *wyIs92* [*Pmig-13::snb-1::yfp*][48] and the motor neuron DA9), so as to enable me to re-use the same overall methodology with as little adaptation as possible. Rather than integrate the marker from *wyIs92* into these strains via repeated outcrossing, as would be typical, I instead examined the F1 progeny of crosses between the two strains and *wyIs92*, as this would most closely match the scenario envisioned for the future imaging of RILs. In order to make the germane comparison with the Bristol wildtype N2, it was also necessary to image the F1 cross between *wyIs92* and N2. In addition, I also imaged the F1 progeny for crosses with the genotype *wyIs92;jkk-1;unc-104*, the double mutant from Chapter 2[52], as our experience with epistasis in the double-mutant suggested that it might be useful to examine both the base strain and strains with already existing synaptic mutations.

Figure 3.2 summarizes the results of this initial study. We had intended to examine other wild isolates of *C. elegans* to find one with a difference in synaptic morphology, but this proved to be unnecessary, as CB4856 already showed a substantial different in synaptic morphology from N2. Specifically, *wyIs92*xCB4856 has synapses that are larger and bright in intensity pretty much across the board, but with little change in variability or evident change in the distribution of these features.

It is worth noting that in this study we expanded the number of features to 48; our previous experience with the mutant *jkk-1* in Chapter 2 had indicated to us the importance of studying the distribution of synaptic sizes, rather than just the area, so we added 19 features to our feature set designed to detect such deviations. The new feature set is summarized in Table 3.1.

Table 3.1 Features measured from synaptic domain of *C. elegans* for QTL mapping. Features 1-29 are identical to Table 2.1.

Synaptic Feature (Over Synapses in Image)	Category	Mean	SD/Mean
Area (Pixels in Synapse)	Size	Feature #: 1	14
Area/(Area Calculated from F6 and F7)	Shape	2	15
Perimeter/(Perimeter from F6 and F7)		3	16
Eccentricity of Approximate Ellipse		4	17
Diameter of Circle with Same Area	Size	5	18
Major Axis Length of Approximate Ellipse		6	19
Minor Axis Length of Approximate Ellipse		7	20
Mean Intensity of Pixels in Synapse (arb)	Intensity	8	21
SD/Mean Intensity of Pixels in Synapse (arb)		9	22
Perimeters (Pixels along edge of synapse)		10	23
Max Intensity in Synapse		11	24
Min Intensity in Synapse		12	25
Total Synaptic Intensity (F1*F8)		13	26
Skewness of Intensity		42	N/A

								Distributional Features
Base Feature	Category	10th %tile	25th %tile	50th %tile	75th %tile	90th %tile	Skewness	Max
Area	Size	30	31	32	33	34	40	41
Major Axis Length		35	36	37	38	39	43	44

Base Features	Category	Mean 0-25 %tile	Mean 25-50 %tile	Mean 50-75 %tile	Mean 75+ %tile
Mean Intensity	Intensity	45	46	47	48

Misc. Features			
Synapse Number	Misc.	Feature #: 27	
Mean Distance between Synapses (pixels)		28	
Synaptic Domain Length (pixels)		29	

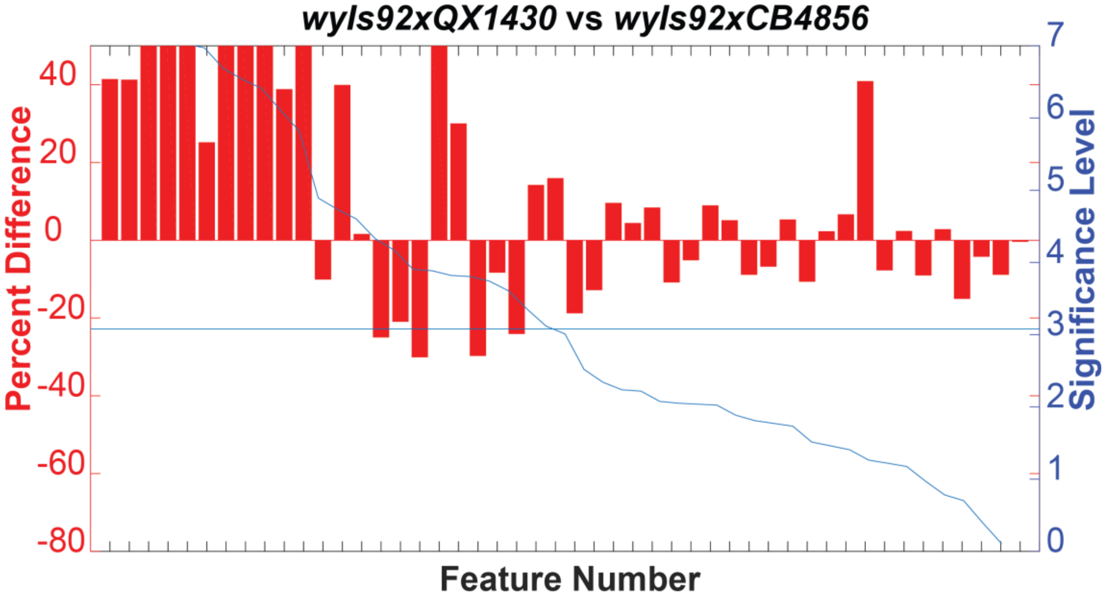


Figure 3.2: Comparison between *wyIs92*xQX1430 and *wyIs92*xCB4856. In both figures, the red bars and the left side bar show the percent difference in various features (positive values mean CB4856 is higher) while the blue line and right side bar show the significance level according to Welch's T-test. The horizontal blue line shows the 95% significance level after the Bonferroni correction for multiple comparisons. Features are sorted by significance and labels at the bottom are not shown for clarity; the significant features on the left in part A are, in order, features 7, 5, 1, 43, 10, 3, 48, 13, 47, 6, 46, 12, 44, 2, 15, 27, 16, 45, 9, 30, 8, 25, 18, 28 (See Table 3.1).

With this data in hand, it was apparent that N2 and CB4856 would serve as fertile ground for a search for synapse morphology-affecting QTLs.

3.3 Experimental Protocol and Methodology

In this study, we perform a QTL analysis on RILs between the parent strains N2 and CB4856, targeting QTLs that influence the morphology of the synaptic domain of the motor neuron DA9. We requested from the laboratory of Erik Andersen an initial sample of 80 published RIL strains, generated by an advanced intercross between the strains CB4856 and QX1430[105]. QX1430 is a variant of N2 that carries the wildtype version of *npr-1*, to suppress the significant effects of the laboratory *npr-1* on QTL analyses, and a transposon knockout of the *peel-1/zeel-1* genetic element, which drives hybrid incompatibility between N2 and CB4856[111]. This genetic element is critical to remove, as it otherwise substantially suppresses recombination frequency in its vicinity, drastically reducing the number of separate QTL intervals in that section of the genome. The marker we used again *wyIs92*(*[Pmig-13::snb-1::yfp])*[48]; it should be noted that while this marker still contains *peel-1/zeel-1*, there is no embryonic lethal effect unless the pair of genes is separated by recombination during meiosis, something that cannot happen before the F2 generation in a cross.

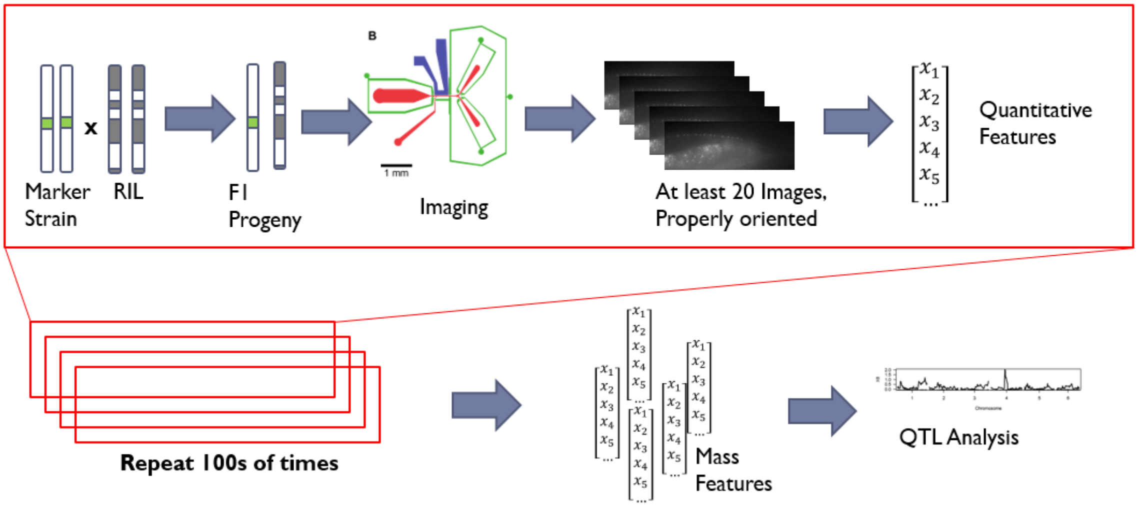


Figure 3.3 Overview of the approach to QTL analysis taken here.

In order to conduct a QTL analysis without generating a full set of RILs with the desired marker, we designed a protocol for imaging the Day 1 adult F1 progeny of the marker genotype *wyIs92* and each of the RILs. In order to image a reasonable sample size of worms derived from each of RILs, we applied the high-throughput imaging pipeline developed in Chapter 2 of this thesis. By doing this, we bypass the two largest hurdles to conducting QTL analysis of fluorescently-labeled synaptic domains. By quantifying the phenotypic properties of the synaptic domains of at least several dozen progeny from each of the RIL strains, we gathered the phenotypic data necessary to perform a QTL mapping of the entire *C. elegans* genome.

3.3.1 Preparation of F1 Progeny for Imaging

In the methodology we lay out below, one consideration we must keep in mind is sample size. By using F1 progeny, we limit the number of animals we have that are of the right age for imaging at any given time, with the result that we must consider how many animals are lost in every step of the process. Given the number of animals available for homozygote studies, many of the experimental procedures traditionally used have no consideration for the number of animals lost—but this something we must keep in mind. It is not, however, a truly critical consideration, as in the worst case we could make more mating plates per cross—this just introduces more labor and logistical overhead.

To prepare F1 progeny for imaging, I crossed males of the genotype *wyIs92* with hermaphrodites of the RIL strain being studied by picking adult males and L4-stage hermaphrodites onto a 35 mm diameter nematode growth medium (NGM) plate freshly seeded with 50 µL of OP50 *E. coli*. This is a standard mating protocol[112]; the small, freshly seeded source of food causes the *C. elegans* individuals to crowd a very small area, drastically increasing the chance of mating, and the use of L4, pre-reproductive hermaphrodites ensures that they receive packets of sperm as early as possible, which suppresses self-fertilization. Substantially more males are picked than hermaphrodites to further assure mating.

By using males of the fluorescent marker strain, we can guarantee that any progeny that are observed to have the fluorescent marker are legitimate F1 progeny, since any progeny produced by hermaphroditic self-fertilization will not have the marker and can be simply ignored during imaging. This eliminates one of the common concerns about crossing *C. elegans* populations, and enables us to adjust the usual mating proportion—20 males to 3 hermaphrodites—to a less cautious 20 males to 5 hermaphrodites. This increases the number of progeny obtained at the risk of producing some non-mated progeny, which is no longer a concern. After the initial experiments, I further increased the number of progeny obtained by crossing 2 sets of 20:5 simultaneously for any given RIL, rather than 1. As we will see, this is roughly what is necessary to produce reasonable number of processed images in the final data set.

It was also observed that with the given protocol, the plates often ran low on *E. coli* 3 days after initial mating, with a measurable impact on the final results compared to plates that did now. After discarding the data where this issue occurred, I adopted a protocol where on day 3 the entire agar plate is sliced in half and transferred physically to two new 60 mm diameter pre-seeded with *E. coli*, flipping it over so the worms land on the lawn. Another option would have been to add more liquid *E. coli*, but this method guarantees an ample supply of food, preventing potential influence on the final experimental results.

Given the design of the microfluidic device, which was optimized to allow Day 1 adults to fit neatly into the imaging channel[62], we were able to image worms of the desired age (roughly Day 1 Adult) simply by refusing to image worms that were substantially narrower or thicker than the size of the channel (thicker worms can be prevented from clogging the device by the use of the flush channel). Both traditional and high-throughput imaging of *C. elegans* is typically done on synchronized populations of individuals hatched from their original eggs at roughly the same time. However, both of the common methods for generating synchronized populations, hypochlorination ("bleaching") and the "lay-off"[113], drastically reduces the number of progeny obtained relative to the full egg-laying productivity of the adults, with the number of progeny dependent on the number of adults transferred. For our purposes, the number of progeny lost can easily become unacceptable. Thus, we instead used the microfluidic device's dimensions as a novel method of resolving this methodological hurdle, providing synchronization that we estimate to be within ±6 hours, comparable to bleaching or a less stringent lay-off. This does assume that the RIL hybridization and the F1 crossing process do not cause substantial changes in the growth rate or size of the adult worms, so it was necessary to monitor the populations for evidence of such an effect, but this was not observed.

We found that the optimal time to image the population of F1 progeny was on the fifth day after initial mating, which permits 1 day for the L4 P0 parents to reach maturity, an additional day to reach peak egg laying, and 3 days for the progeny from that day to reach Day 1 adulthood. These progeny were washed off the plate with M9-triton and used for imaging; the P0 parents could be excluded on device due to their evident enormous size.

3.3.2 Imaging and Quantification

Once the F1 progeny were obtained, imaging proceeded much as described in Chapter 2, using chilled fluid for immobilization, with a few key modifications. The key difference is that with the limited number of F1 progeny available for imaging, we maximized the numerical efficiency of imaging by using a semi-manual procedure, with the device states controlled manually while the details of the valve arrangements were handled automatically. This was motivated by the follow considerations:

1) As discussed previously, the random orientation of worms entering the imaging region imposes a 75% attrition on the number of images. In worms that arrive head-first, we can mitigate this by moving the device stage to focus the objective on the tail. While only some of the immobilized worms will have the DA9 neuron

synaptic domain up against the glass, this is sufficient to recover a substantial number of the lost worms.

2) It is now necessary to perform mild size selection on the individuals as they enter the imaging region of the device, excluding worms that are noticeably too small or too large.
3) The fully automated methodology sometimes takes poor images, particularly since worms immobilized by cooling still twitch occasionally
4) It is necessary to calibrate automated imaging initially, losing a small number of animals setting automation parameters for a given device and experiment.

Further Changes to Features and Quantification

To perform this QTL mapping, a number of additional changes were made to the quantification pipeline, both planned early on and due to experimental observations about the dataset.

1) As mentioned in Section 3.2 and illustrated in Table 3.1, An additional 19 features were added to the feature, designed to examine the distribution of synaptic intensity and size, after it was observed that these were important to measure the phenotype of the known synaptic mutant *jkk-1*.
2) For the intensity features, background normalization was added, replacing background subtraction. Over the course of experimentation, it was observed, for instance, that the intensity of the synapses as originally measured showed a strong correlation with the passage of time—more specifically, intensity tended to vary with whether an imaging run was taken earlier or later in the overall set of experiments. It was found that this phenomenon closely tracked a similar trend in the *background* intensity of the images, and could be eliminated by dividing this out. It is not clear what the source of these long-term correlations were, but one likely explanation is, for instance, a gradual change in the intensity of the fluorescent light source over long-term use.

3) It was also observed that the number of synapses reported by the algorithm was somewhat less than the 25 average found in previous manual observations. On detailed review of the segmentation process, it was found that a common segmentation error was the inappropriate merger of neighboring synapses into one larger blob during segmentation. After some experimentation with other methods, it was decided to adopt a consensus method—by using the SVM methodology we already had on hand for the detection of DA9 neurons to select pixels, it was possible to substantially refine the segmentation of individual synapses, without reintroducing the anomalous fat droplets and other problems of the SVM approach. Unfortunately, this is a trained method which we already know to be very specific to given imaging conditions, as discussed in Chapter 2, and the combination of the two different methodologies greatly increases the complexity; we decided this was acceptable given the sensitivity of QTL mapping to inaccurate data and noise.

With these changes in the methodology established, I set out to gather as enough data from the RIL crosses as possible, in order to perform QTL analysis. After false starts where data was lost due to problems with starvation (as mentioned in 3.3.1) and due to a shift in experimental locales, data from 47 strains was gathered over the course of 4.5 months, interrupted by a major equipment failure. The results of this data collection, still ongoing, will be presented in section 3.4.

With the full experimental procedure in hand, it is now possible to discuss sample size—specifically, the reasons why sample size for imaging sessions is a significant concern for this methodology and why pains were taken to mitigate worm losses whenever possible. Unfortunately, starting from the very beginning of the experimental procedure, there is a steady attrition in experimental animals, with the ultimate empirical result that the average number of usable images at the end of the procedure was optimally about 60-80, with many imaging sessions suffering further loss due to device failures or simple human error.

The sample size breaks down as follows:

1) The average N2 *C. elegans* individual lays about 291 eggs in its life span over nearly 3 days, subject to environmental conditions. With the experimental timing we have chosen and ±6 hour precision on our judgment of worm age on the device, about 75 of these progeny will be the right age at the imaging time[114]. Using a total of 10 hermaphrodites, we then have 750 progeny that may potentially be imaged.
2) Significant losses occur when transferring worms from the plate into the microfluidic vial. This is done with washing with M9-Triton, but even rigorous washing leaves a large number of adults still on the plate. This is complemented by unknown losses during the microfluidic loading process, usually due to worms that stick to parts of the tubing or debris. It is unclear how much loss this represents, but a 20% loss rate would leave 300 individuals remaining.
3) Of these 300 individuals, roughly 50% will be male progeny, which we do not image. Some small number—around 5%—will be unmated progeny as well. This leaves 135.
4) Of these 135 remaining individuals, about half will enter the device head instead of tail-first, and another half of the remaining will fail to have the dorsal side pressed against the glass as required for imaging. An additional subset will enter with curled tails or some other inappropriate orientation. Despite attempts to recover some of the head-first imaging by manual imaging, we estimate only about 60% of the worms that enter the device produce valid images. This leaves about 81 individuals.

5) Of the 81 such images taken, an additional 20% or so are lost to issues with segmentation, leaving about 64 images remaining.

In practice, many imaging sessions produce fewer than this, due to experimental contingencies and occasional clogs in the device that must be dealt with.

Principal Component Pursuit for Sparse Noise Reduction

To reduce sparse noise outliers in the dataset, particularly in a few particularly unstable features that vary substantially in the event of poor segmentation (e.g. the average distance between synapses), an implementation of principal component pursuit (PCP) was written, based on the algorithm provided in Candès *et al.*, "*Robust Principal Component Analysis?*"[115]. Given a data matrix M whose expected rank is substantially lower the dimensionality of the data matrix, but which is known to be substantially contaminated by high magnitude sparse noise, Principal Component Pursuit divides the matrix into two components L and S, such that $M = L + S$, L is as low-rank as possible and S is as sparse as possible. This effectively removes the sparse noise, provided that the data is *a priori* known to be low rank, something which is almost certainly true for synaptic phenotyping data. The convex optimization can be performed exactly by an alternating iterative algorithm which we will not discuss here, referring the interested reader to Candès *et al*[115]. While the algorithm was not ours, the implementation was my own. I used the suggested default:

$$\lambda = \frac{1}{\max(N_1, N_2)}, where\ M\ is\ N_1 \times N_2 \qquad (1)$$

as the cost parameter controlling the relative importance of the sparsity of S and the low-rank of L, following the advice of Candès *et al.*, and did not find a need to change it.

PCP was then performed on the entire dataset at once, combining data from individuals for every strain into one large data matrix. While for many features this post-processing of the dataset had little effect, the most volatile features, particularly integrated intensity and synaptic domain length, which contain significant outliers due to merged synapses and missing regions of the synaptic domain during segmentation, were substantially normalized by the application of PCP, reducing intrastrain variance (Fig. 2.12). Since these features were known *a priori* to be most problematic and sensitive to segmentation and imaging issues, and were frequently observed to be incorrect (e.g. in many cases individual synapses are concealed by bending of the worm tail or an intervening fat droplet), PCP is behaving as desired, and salvaging a number of images which would otherwise have to be manually discarded. It should be noted, however, that this form of data conditions carries the implicit assumption that the data gathered from each strain shares the same underlying eigenvectors. While we believe this is likely true, since the description of the synaptic domain is likely low rank, it does introduce the possibility that PCP would conceal some differences between strains.

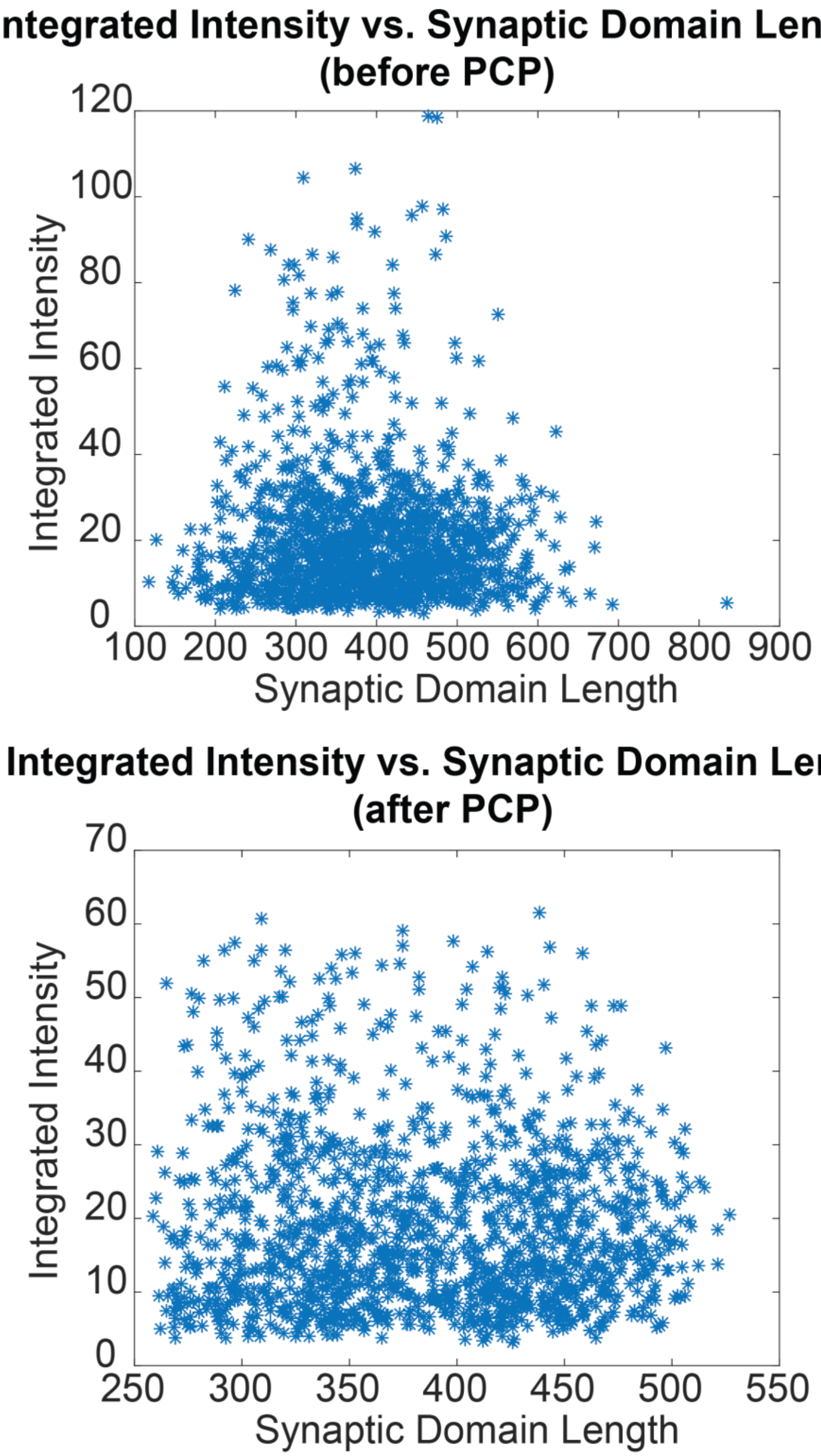


Figure 3.4 Scatterplots of Integrated Intensity vs. Synaptic Domain Length, before and after PCP application. Before PCP, a substantial number of data points have implausibly large integrated intensities, with an implicit synaptic area as high as 40 µm^2 and an implausibly large range in synaptic domain length (roughly 16 to 100 µm). After PCP, the corresponding values are about 20 µm^2 and 40 to 80 µm, which is much more biologically plausible.

3.3.3 Additional Design Considerations

The proposed experimental methodology, particularly the use of F1 crosses, introduces a number of new potential concerns to the QTL analysis. Fortunately, these are not serious or were already addressed in the design.

Firstly, the use of F1 progeny rather than a full integration of the synaptic marker into the RILs adds a subtlety to what can be detected—because the marker genotype *wyIs92* itself has N2 as a background, recessive synapse-affecting variants found in CB4856 cannot be detected with the currently given protocol. In our case, since the synaptic variation we are seeking was already detected in the parent strains (comparing *wyIs92*xN2 with *wyIs92*xCB4856), a QTL dominant in CB4856 seems to already be present, so this subtlety doesn't damage us.

It would be possible to detect QTLs recessive in CB4856 by integrating the synaptic marker into homozygous CB4856—a one-time procedure—and then comparing the parent crosses *wyIs92(CB4856 background)*xN2 and *wyIs92(CB4856 background)*xCB4856. If a recessive QTL needed to be found, this CB4856 background marker strain could then be crossed into the RILs in the exact mirror of the protocol we used in our study. In the hypothetical scenario where no dominant QTL had been found, recessive QTLs are only a short methodological distance away.

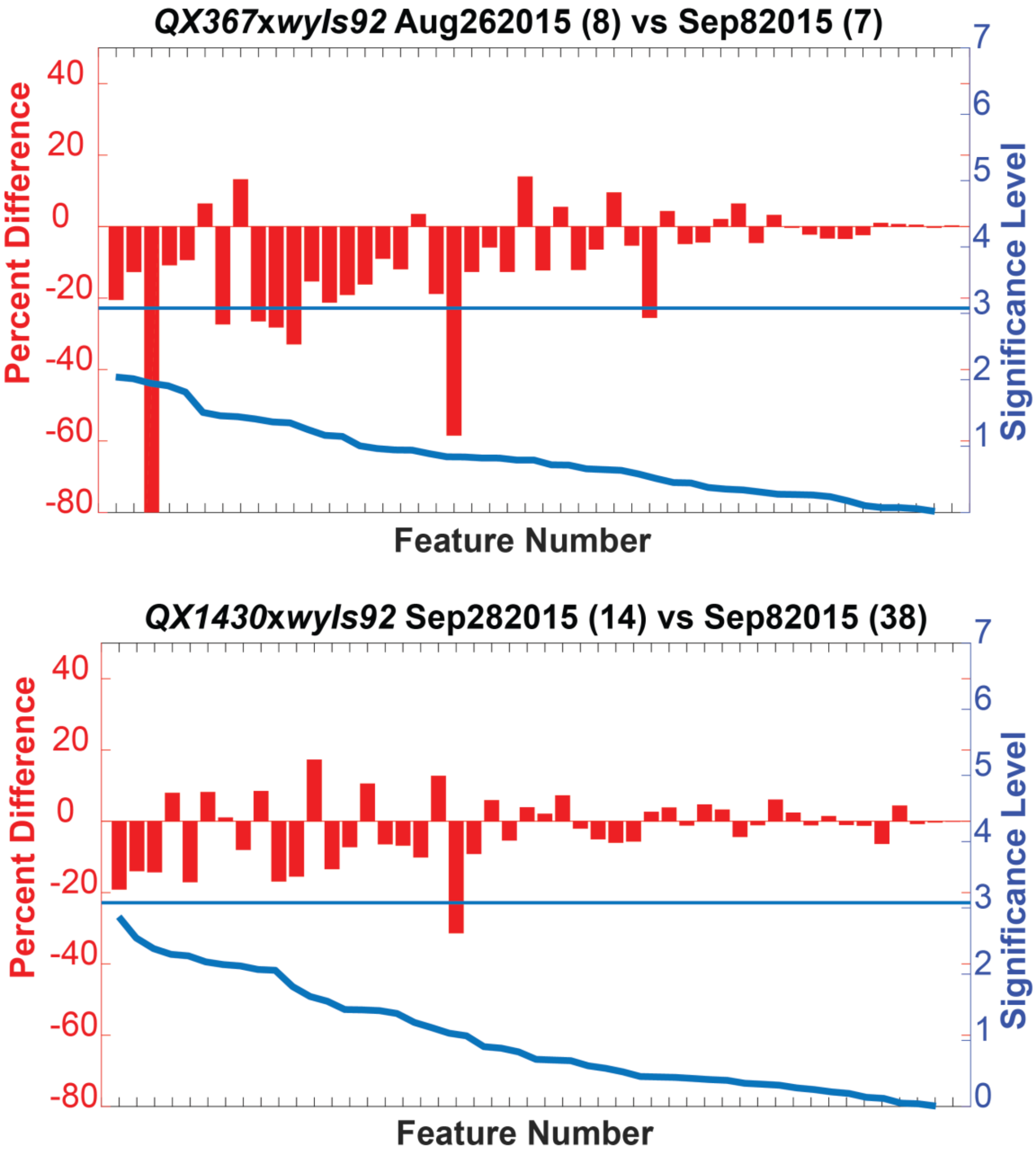
Secondly, the given approach does not completely rule out distorting interactions between the heterozygous N2 background of the marker and the RIL background being studied, which would cause a heterozygote-only effect that doesn't show up in the homozygote. A relatively short follow-up study, examining the full homozygous introgression of the relevant QTL into *wyIs92*, would be necessary to show the effect is not heterozygote-only, and we fully intend to carry one out. It is worth noting, however, that this kind of heterozygote-only effect is not only an unlikely scenario, but would itself be of scientific interest. In addition, since a significant effect was already found in the F1 crosses of the markers with the parent strain, it is at least not possible for the phenotypic change being sought to disappear entirely.

Finally, we also face the question of what phenotypic feature to use for QTL analysis. Traditional interval mapping uses only one quantitative measure of the phenotype, but we have 48 such measures, and could devise as many additional measures as desired. As also discussed in 3.4.1, to avoid doing too many comparisons, we choose three of the most representative features from the set that was found to differ in the parent strains, synaptic area (#1), mean synaptic intensity (#8) and the Top 25% of integrated synaptic intensity (#48), as well the first two principal components of the features, which contain the substantial majority of the variation.

3.3.4 Inter-trial Controls and Auxiliary Studies

In order to properly validate the ongoing study, and to ensure consistency across multiple trials and across multiple imaging sessions, I discuss here a subset of the additional data analysis that is necessary to guard against potential fluctuations in the consistency of the imaging.

One potential source of concern is variability between different imaging sessions. An obvious way to test for potential differences between imaging sessions is to compare different imaging sessions for the same strain, but taken on different days. In this, we have a natural source of experimental data to make this comparison: due to experimental issues—device rupture, human error, cooling system breakdown—or concern about insufficient sample size, a subset of the data for certain strains already draws upon more than one separate experiment. I tested different imaging sessions for the same strains whenever more than one imaging session was available, and whenever the sample size on each individual session between compared was greater than 10 individuals. There were 2 such sessions, one involving the parent strain *QX1430*. Figure 3.5 shows comparisons between the different trials, showing that inter-session variation was not statistically significant and lower than the already measured difference between the parent strains.

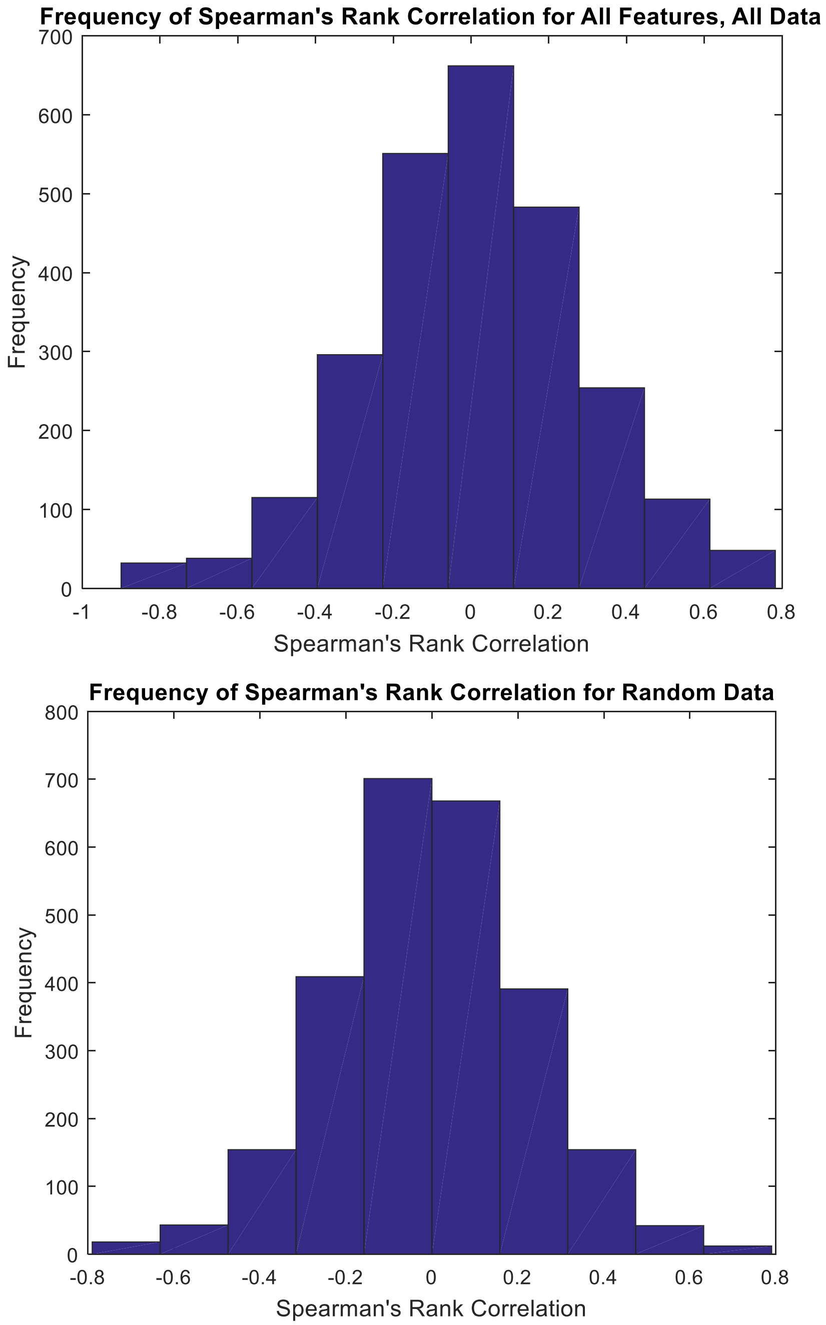


*QX367*x*wyIs92* Aug262015 (8) vs Sep82015 (7)

*QX1430*x*wyIs92* Sep282015 (14) vs Sep82015 (38)

Figure 3.5 Inter-trial comparisons for two different imaging sets: *QX367*x*wyIs92* and *QX1430*x*wyIs92*. These comparisons show no significant difference. Charts are the same style as in Figure 3.2. Labels in the titles indicate the dates on which images were taken; the numbers in the parenthesis are the sample size of images after full processing. The red bars and the left side bar show the percent difference in various features (negative values mean the second set of data is lower) while the blue line and right side bar show the significance level according to Welch's T-test. The horizontal blue line shows the 95% significance level after the Bonferroni correction for multiple comparisons. Features are sorted by significance and labels at the bottom are not shown for clarity.

In order to address the possibility that leaving worms suspended in fluid in tubing awaiting imaging in the microfluidic device—sometimes for a few hours—would affect the results of imaging, I calculated the Spearman's Rank Correlation Coefficient between every feature of every imaging run and the order the worms were imaged in, whenever the number of worms was greater than 10. The histogram of correlation coefficients is shown in Figure 3.6, along with random data of the same size for each imaging run; the two are indistinguishable. Examining each of the features separately produces similar histograms (data not shown).



Frequency of Spearman's Rank Correlation for All Features, All Data

Frequency of Spearman's Rank Correlation for Random Data

Figure 3.6 Histograms of Spearman’s Rank Correlation Coefficient for (top) all of our imaging data in every feature and (bottom) size-matched random data. The two histograms are nearly indistinguishable. The total number of vectors for which the correlation coefficient was calculated was 63264.

3.4 QTL Analysis and Results

With the fully-processed data in hand for 47 strains, we turned our attention to the actual QTL mapping. Before doing so, it was necessary to decide what quantitative features to do the QTL mapping *on*. QTL mapping typically focuses on only one quantitative, phenotypic measurement, but we had 48 features in hand. It would not have been wise to perform QTL mapping on all 48—this would have resulted in either overly favorable statistical testing, due to the multiple comparisons, or an *under*-powered test, if a correction such as the Bonferroni correction were used—since many of the features are substantially correlated with each other, a full multiple comparisons correction would understate the level of statistical significance.

Thus, we chose to narrow the scope of our focus to two classes of features that we felt were well-justified: 5 features that had shown a statistically significant difference between the two parent strains, chosen to be as distinct as possible, and the magnitude of the first two PCA components, with PCA being performed on a matrix containing all of the individual animals measured. This formulation provides more statistical power for detecting the true QTLs responsible for differences between the parent strains, without performing an undue number of comparisons.

The first two principal components were chosen because they are by far the most explanatory principal components (35 % and 33% of the explained variance, respectively, compared to 9% for the third principal component). Besides the first two principal components, the features chosen for this were the mean of the synaptic area, the mean synaptic intensity, and Top 25% of Integrated synaptic intensity (Features 1, 8, 13 in Table 3.1, respectively). There is obviously a substantial relationship between some of these features, but each of these examines synaptic intensity in a slightly different light, so we feel it is appropriate as long as we proceed with caution.

3.4.1 QTL Analysis

To perform the QTL analysis, we used a standard interval mapping, using the R/qtl library[116] and modifying code provided to us by Patrick McGrath[19]. Significance level was tested for with permutation testing, using 100 permutations each. We did not elect to use composite interval mapping, as manual inspection of the intensity values suggested the strong possibility of there being one strong QTL affecting the phenotype (Fig. 3.7).

As can be observed in Figure 3.8, while an intriguing peak is consistently observed in chromosome IV, one which has persisted and grown slightly as more RIL data has been accrued, this peak has not yet achieved statistical significance with the given amount of data.

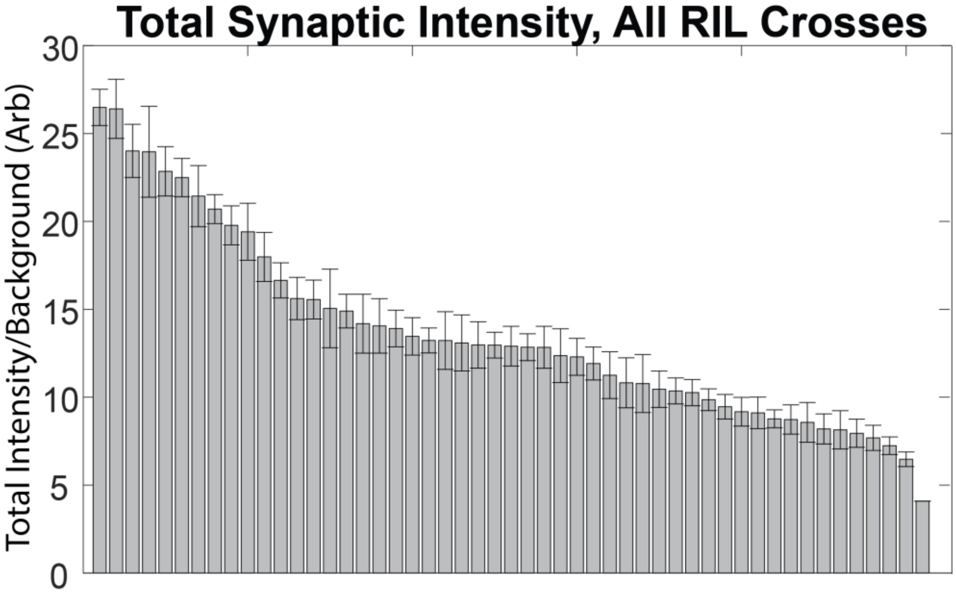


Figure 3.7 Integrated Synaptic Intensity for all of the RIL crosses so far. While there is some degree of non-bimodal variation, the data is suggestive of one, strong-effect QTL driving the main difference in phenotype. RIL strain labels are omitted for clarity; strains are sorted from highest feature value to lowest.

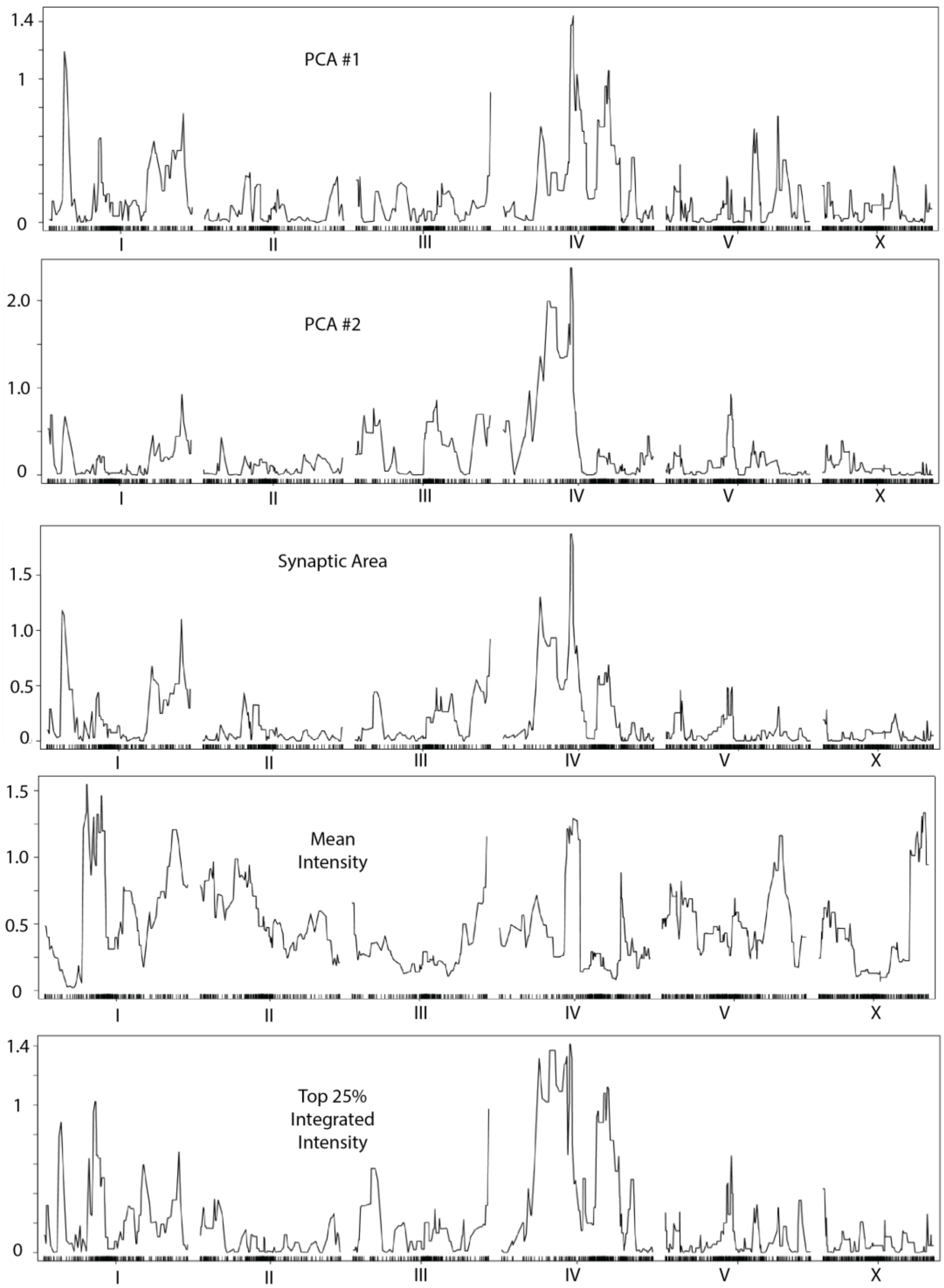


Figure 3.8 QTL mappings for PCA Components #1 and 2, as well as Features 1, 8 and 48. Y-axis is LOD score, x-axis is locations on chromosomes I through V and X. Tick marks at the bottom show the locations of QTL markers. There is a consistent peak near the center of chromosome IV in all measures, with the highest peak occurring between nucleotide 2,914,279 and 3,737,430 of chromosome 4 (version WS244 of Wormbase[2]). 95% significance is at LOD score ~3, but this is not consistent between measures and should not be relied upon as more than a guideline.

3.4.2 Near-isogenic Lines (NILs) for Detailed Examination of a Potential QTL

One option to proceed given the results of the QTL mapping so far would be to continue imaging more RILs, and that is one avenue which we intend to pursue. However, in order to perform a focused examination of the most interesting LOD score peak, as seen in chromosome IV, 4 near-isogenic lines (NILs) to N2 were requested from the laboratory of Cori Bargmann[110], introgression lines composed primarily of the N2 genotype with parts of the CB8456 genotype covering the specific region or nearby regions of chromosome IV found in the current QTL study. Specifically, CX11901 covers the region from 151,889 to 3,920,366 base pairs (b.p.)., CX11879 from 2,761,525 to 3,347,952 b.p., CX12777 from 1,799,032 to 3,920,366 b.p. and CX11932 from 3,347,952 to 13,049,020 b.p.

Unfortunately, these strains are based on N2, not QX1430 as the RILs in the QTL mapping were, and thus carry the N2 copy of *npr-1*, a potential concern if the RILs we find turn out to be on chromosome X where *npr-1* is. The F1 progeny of these strains with *wyIs92* were imaged using the same protocols as the RILs.

As can be seen in Figure 3.8, in mean synaptic intensity, the average brightness of the pixels in each synapses, the 3 NILs containing the LOD peak on chromosome IV have the same phenotype as *wyIs92* x CB4856, while the one that doesn’t, CX11901, does not, a result strongly suggestive of a synaptic morphology-affecting QTL residing in the region covered by these 2 NILs, in this case nucleotide positions 2,761,525 to 3,347,952 of Chromosome IV, as labeled in the WS244 version of Wormbase[2]. Incorporating details from the location of the peak in the previous QTL further refines the selected position to the region from 2,914,279 to 3,347,952.

This effect, however, does not carry over exactly into synaptic area or the top 25% integrated intensity (which reflects the average intensity of the 25% of synapses with the largest integrated intensities, which is mean intensity multiplied by the synaptic area). Here, the two NILs which contain small regions containing the QTL show the effect (i.e. they match *wyIs92* x CB4856), but CX11901, which contains nearly the first quarter of chromosome IV containing the QTL, does not. This is suggestive of a potential multigenic effect, and not entirely unusual. For example, the same effect was seen in a study of chemosensation for bacterial peptides using these same NILs[110]. It is worth noting that the region of chromosome IV covered by CX11901 contains large LOD score regions of chromosome IV (Fig. 3.8), raising the change of another QTL in the area.

Some of the data from one of the NIL crosses here was imaged by Farhan Kamili, a graduate student in the Lu Lab.

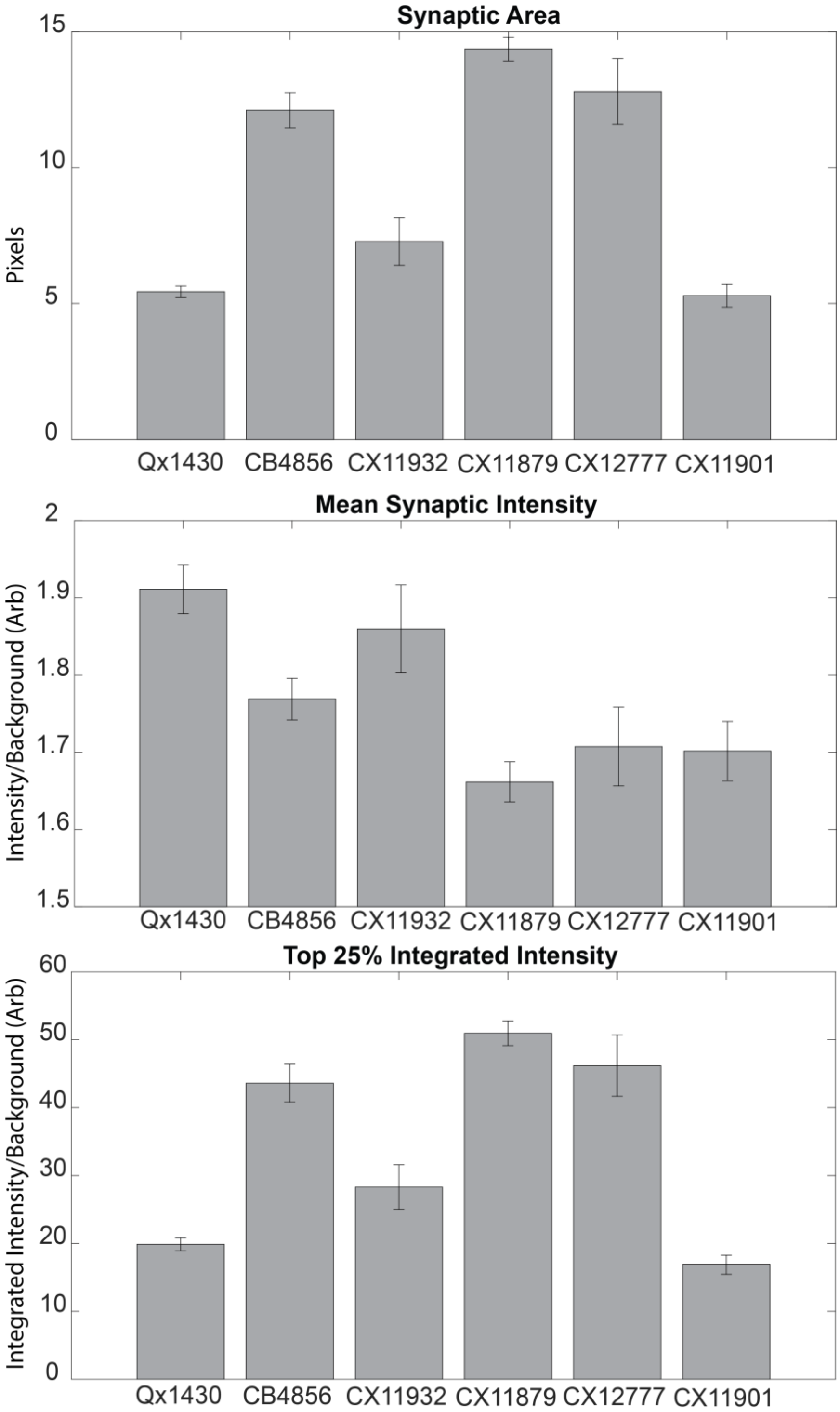


Figure 3.9 The parent strains and NILs after crossing with *wyIs92*, measured in features 1, 8, and 48, respectively. The mean synaptic intensity shows an effect suggestive of a QTL in the suggested region, but the other two features show a more complicated, potentially multigenic story.

To fully verify this effect, it is profitable to integrate the *wyIs92* marker directly into the NILs and image the homozygotes, a process which is considerably simpler than integration with an RIL or with CB4856, because of the limited number of markers that must remain consistent. Preliminarily, this has been done for three of the NILs, CX11879, CX12777, and CX11901, which can be seen in Figure 3.10. The result here is consistent with the results from the heterozygote, showing a substantial difference between these three strains and the N2 base *wyIs92*, but is not yet complete, lacking the full integrant with CB4856 and CX11932, but is suggestive. The partial effect seen in CX11901 synaptic area and Top 25% integrated intensity stands in intriguing contrast to the effect seen in the heterozygote.

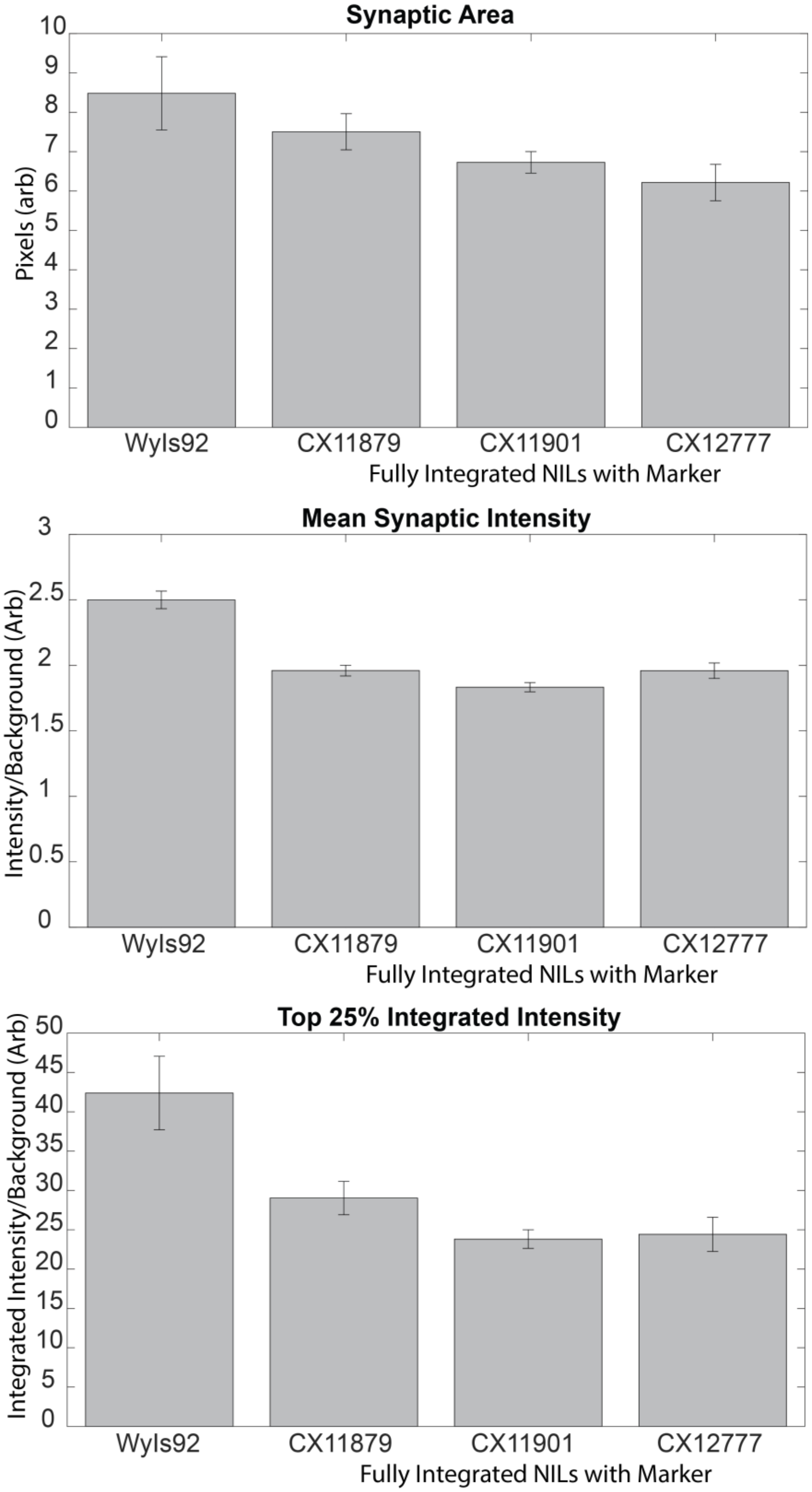


Figure 3.10 The homozygote *wyIs92* and the NILs CX11879, CX11901, and CX1277, integrated with the *wyIs92* marker. All three show significant differences with the N2-based *wyIs92*, hinting at a synapse-affecting QTL in the suggested region.

3.5 Discussion

In this chapter of the thesis I demonstrate the application of our high-throughput microfluidic imaging pipeline to QTL analysis, mapping a difference in the size and intensity of the largest synapses of worms with the CB4856 genotype down to a narrow putative QTL on chromosome IV. In doing this, I help illustrate the value of these techniques to quantitative genetics, a domain previously unaddressed by this kind of microfluidic imaging. By using microfluidics to overcome the otherwise pernicious methodological limitations of imaging heterozygote crosses between established RILs, we were able to establish a technique for effectively imaging and mapping fluorescently-labeled phenotypes in RILs, something which had previously been hampered by severe methodological limitations. This is a substantial expansion of a lucrative genetic mapping technique to a new domain, one of key importance to understanding multigenic, subtle phenotypes.

The potential targets of such a new approach to QTL mapping are far-ranging, going far beyond synaptic phenotypes or even *C. elegans*. Firstly, the demonstration that heterozygotes may be used to extract novel genotypes raises the possibility of performing a similar QTL study whenever the process of generating a large population of RILs with a single marker or mutation is difficult, such as in slower-breeding organisms like mice. It also raises the possibility of genetic studies that specifically focus on variation that exists, rather variation generated by mutagenesis—for instance, one might search an existing genome for QTLs that specifically affect the phenotype generated by a particular mutation, by crossing this particular mutation into a large set of RILs and examining the progeny.

Secondly, the specific demonstration that fluorescently-labeled phenotypes may be used for this kind of quantitative genetics, using high-throughput microfluidics, has obvious applications for any *C. elegans* phenotype that requires a fluorescent marker to properly identify. Rather than focusing on macroscopic or easy to quantify phenotypes, QTL could be done on specific, difficult to detect structures, or on the localized expression of certain proteins or mRNA, something which is already done on a whole-worm scale but difficult to achieve on individual cells or structures[117, 118].

Finally, the techniques introduced in this chapter have application beyond just QTL mapping specifically, but can generalized to other types of quantitative genetics, or indeed any study where the detailed phenotyping of a large number of different strains is desired. It could be used for broad studies of wild isolates, for example. Much could be done to illuminate the still nebulous portions of the *C. elegans* genome, getting us closer to a full understanding. In particular, a deeper understanding of multigenic, subtle phenotypes like synaptic morphology will help drive in turn a deeper understanding of related multigenic human diseases such as autism spectrum disorder and schizophrenia. it is our hope that this demonstration helps to spur similar work on a variety of topics.

3.5.1 Limitations and Considerations

The methodology outlined and performed in this chapter nonetheless has a number of limitations. Some of these limitations are shared with QTL itself and inherent to the statistical methodology—for instance, the need for the parent strains to have meaningful differences in the phenotype being studied, and for these parent strains to have an existing population of RILs between them.

Despite the methodological innovations of this chapter, performing a QTL mapping in this fashion is still time-consuming and laborious relative to many types of experimentation. While this is in some sense inherent to the statistical requirements of QTL mapping, and certainly an easier task than many types of long-term experimentation, there is room for improvement. Improvements in the device design, such as direct orientation control of the animal, either in a head vs. tail or dorsal vs. ventral sense, would greatly increase the numerical efficiency of imaging, relieving the methodological restrictions imposed by the need to manage sample size. Even a device with the ability to reject worms in the worm orientation back into the inlet channel could be sufficient.

From the point of view phenotype quantification, there are some additional improvements that might be made. The reliance on a SVM method we know overfits the imaging setup is disappointing but may be necessary—that case, perhaps a method with better regularization might be appropriate. The reliance on manual, carefully chosen features is also unfortunate, and may leave out potential phenotypic characteristics. While it is unlikely that expertly-chosen features can be avoided, it is possible that more objective feature selection could be performed.

CHAPTER 4

FULLY-AUTOMATED 3D TRACKING AND ANALYSIS OF WHOLE-BRAIN NEURAL ACTIVITY IN *C. ELEGANS*

Much of the material in this chapter is adapted from a manuscript in preparation, Zhao *et al. "Fully-automated 3D Tracking and Quantification of Neurons in C. elegans Global Brain Imaging"*. A debt is owed to the lab of Manuel Zimmer at the Research Institute of Molecular Pathology (IMP) in Vienna, Austria, who provided the hand-curated calcium imaging videos used in much of this chapter, from the publication Kato et al, *"Global Brain Dynamics Embed the Motor Command Sequence of* Caenorhabditis elegans*"* in Cell[119]. This chapter discusses the development of an accurate and fast algorithm for the automated segmentation and tracking of neurons in whole brain calcium imaging, enabling the evaluation of neural activity in many neurons over a large number of worms, without the lengthy manual curation that currently bottlenecks this approach to functional neural imaging.

4.1 Motivation, Background and Overview

One of Sydney Brenner's original reasons for selecting *C. elegans* as a model organism was the expectation that its small, stereotyped nervous system would provide crucial insight into the origins of behavior, insights that were obscured by the relatively enormous neural systems of even *D. melanogaster*. In *C. elegans*, however, electrophysiology was discovered to be technically extremely challenging, due to both the small size of the animal and its pressurized pseudocoelom, which ruptures explosively upon puncture. While heroic efforts have been made to make electrophysiology possible, the necessary investment in training and resources has limited it to only a few specialized labs[77, 120-123]. Instead, much of the focus for functional imaging of neural activity in *C. elegans* has been on Genetically-encoded calcium indicators (GECIs) such as cameleon[73, 124] or GCaMP[125], which fluoresce in tune with the local concentration of Ca^{2+} ions, thus serving as a proxy for neural activity that exploits the optical transparency of *C. elegans* to avoid damaging the animal.

4.1.1 Calcium Imaging, Whole Ganglion Imaging, and Motivation

Calcium imaging has traditionally been done on at most a few neurons a time, with research groups focusing on individual neurons or circuits in an effort to deduce or analyze their function. While the interrelationships of small subsets of the *C. elegans* nervous have been decoded in this fashion, particularly in the sensory processing of individual sensory stimuli or in the direction of core motor behaviors[67, 121, 126-130], more holistic understanding has proven elusive, especially when it comes to the difficult to correlate interneurons. It would be useful to observe the activity of many neurons at once, perhaps even the entire nervous system at once (302 neurons), but until recently the tradeoffs between temporal resolution, spatial resolution, and field of view have prevented successful imaging. This requires imaging at a rate of at least 10 Hz, at a magnification low enough to capture a large part of the worm in the field of view (usually 40x or less), but retaining enough spatial resolution fine enough to segment individual neurons, not achievable with most fluorescent scopes at the given microscope.

With recent innovations in microscopy, for instance the advent of spinning-disc confocal and light-field microscopy, it has finally become possible to perform this kind of “whole brain”, “global brain”, or “pan-neuronal imaging”, recording calcium traces from a large number of neurons at once; the number of neurons imaged depends on the imaging method, how much the movement of the worm is restricted, and the quality of neuron identification within the collected videos. Both freely-moving and confined imaging has been reported, with the number of cells recorded varying from around 60 to over 120 (Table 4.1)[4, 8, 14, 17, 119].

The processing of these videos remains a substantial challenge, however, limiting published studies to, thus far, at most 5 individuals. The requirement that imaging be carried out on a wide field of view limits the ability of microscopy setups to resolve individual cells, making it difficult to separate the tightly packed neurons of the head ganglion. Additional issues include the segmentation of objects in 3D and the tracking of objects that disappear and reappear in the segmentation due to variance in GCaMP intensity. Such errors in segmentation then compound with movement of the neurons over time to cause errors in tracking each individual neuron over the time frames of a very long video, leading to confounding numerical difficulties in properly tracking the fluorescence of the neurons over time.

For these reasons, fully-automated analytical methods have thus far not been used for data analysis, and the proper annotation and analysis of these volumetric videos continues to rely on extensive and exhaustive hand correction. The logistical burden of this manual correction, particularly for long videos, leads the in-depth analysis of large numbers of individuals to be impracticable. Further, it is possible that manual correction introduces unwanted subjectivity into the data gathered, leading to potentially inaccurate calcium traces if, for example, a manual observer were to correct a segmentation to favor the brightest sections of a neuron with dimming activity.

For these reasons, we aim to design an algorithm that eliminates as much as possible this need for manual curation. It must of course also be high in accuracy, run in

Table 4.1: Publications on Whole Ganglion Imaging, illustrating the types of microscopy, the diversity in imaging conditions, and the number of neurons and worms imaged. Table reproduced from Cho et al[3].

Lead Investigator	Freely-behaving?	Microscopy Setup	Worms Anesthetized?	# neurons observed	# worms reported
Edward S Boyden & Alipasha Viziri[4]	No	Light-field Microscopy	Partial	74	1
Manuel Zimmer & Alipasha Viziri[8]	No	Two-photon with sculpted light	Partial	~99	5
Manuel Zimmer[5]	No	Spinning-disc confocal	Partial	107-131	5
Andrew M Leifer[14]	Yes	Spinning-disc confocal	No	56-77	4
Aravinthan DT Samuel[17]	Yes	Spinning-disc Confocal	No	26-84	5 (1 control)

reasonable time, and, for a given set of imaging conditions, require no adjustment of parameters.

4.1.2 Goals of this Work

To address the technical limitations impeding large-scale functional neuronal studies, I set out to develop a segmentation, tracking, and post-processing pipeline capable of automatically processing arbitrarily-long, volumetric videos of *C. elegans* individuals labeled in all neurons by GCaMP. This pipeline was designed to require little to no manual correction, take at most a few hours to run, and have an accuracy comparable to hand-curation, with no need to re-adjust parameters between individual experiments. As an initial step, for comparison, we requested published, hand-curated data from the laboratory of Manuel Zimmer, using it as the baseline for our pipeline[119]. This data was automatically segmented and tracked, but required extensive hand-curation over the course days to achieve sufficient accuracy

Once the pipeline was established on the videos provided, we sought to use the hand-curated data provided to us to show that the pipeline extracted the same cell traces and reached the same conclusions, but with a procedure that at most a few hours to run no hand-correction. We were able to do this, generating a pipeline that was able to process the videos in 150 minutes each, and for which replicating the analysis seem previously in Kato *et al.* produces comparable results.

With that in hand, we sought to demonstrate that it was possible to use this pipeline to analyze a much larger number of worms than in previous studies. By imaging a large number of worms of the same strain ourselves, I then apply the pipeline to videos produced in our own lab, analyzing a 16 additional worms, demonstrating the extraction of whole worm scale neural activity for a large number of worms. We show that many of these worms undergo a cyclical rhythm in neural activity similar to Kato *et al.*[119], with a period of roughly 36 s. We anticipate the pipeline herein to enable a wide variety of previously infeasible large-scale calcium studies.

4.2 An Algorithm for Rapid and Accurate Tracking of GCaMP-labeled Neurons in *C. elegans*

To develop the tracking algorithm, data was obtained from the Zimmer Lab as previously discussed. Worms were immobilized with 1mM tetramisole and loaded into a microfluidic device[69], then imaged for 18 minutes using a spinning disc microscope equipped with an EMCCD camera at a frame rate of 2.6-3 Hz at 40x. Each frame consisted of 11 or 12 z-slices taken at a spacing of 2 µm. Kato *et al.* used a region-of-interest segmentation and greedy tracking, along with extensive hand-correction, to generate the hand-curated data—only the final data was used in this work. The strain used for imaging with ZIM504 [*memEx199; lite-1 (xu-7)*]. Here, the *lite-1* mutation eliminates the usual *C. elegans* sensory response to high-frequency light[131]—necessary for fluorescent functional imaging—and *memEx199* is *Ex[Punc-31::NLSGCaMP5k; Punc-122::gfp]* where the former is the pan-neuronal nuclear-localized GCaMP5 and the latter is a co-injection marker. Further details on this procedure may be found in the relevant paper, but I do not focus on it here[119].

After initial algorithm design, the algorithm was first verified by both by our own hand-annotation and by careful comparison with the previous hand-curated date. The algorithm was then used on 16 of our own videos, taken on the same strain but using agar pad, demonstrating our ability to rapidly and effectively process a large number of calcium imaging videos.

4.2.1 Algorithm Design

The overall algorithm was divided into 3 stages: 3D Segmentation to identify neurons, tracking to identify neurons over the course of a long video, and post-processing to mitigate any errors that might remain.

Figure 4.1 shows an example frame of the video.

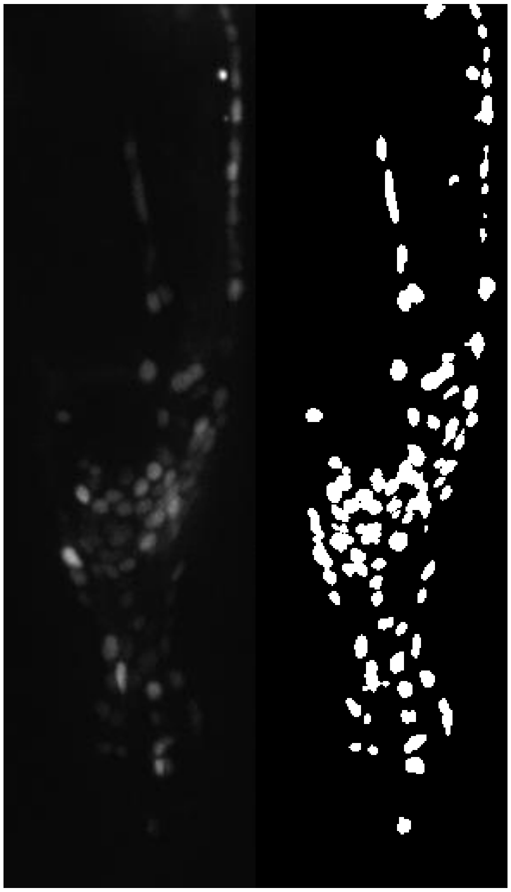


Figure 4.1 Representative maximal Z-projection of a single frame from one of the calcium imaging videos. Left: The original Z-projection. Right: The same frame after thresholding (for visualization).

3D Segmentation

Each 2D slice of the videos was filtered by a Laplacian of Gaussian filter with a filter size of 20 and a standard deviation of 2 and thresholded to select all pixels with a value lower than -0.005. The 2D z-slices of each timestep were then combined into a single 3D volumetric image and separated into blobs. The blobs were inspected for those with a size greater than a preset number of voxels (125). Those greater than this size were locally re-thresholded via a binary threshold search until all objects were below the given size. If no such separation was possible, the best separation was kept, including the original object if no separation was possible—as was the case occasionally when a cell did appear as greater than 125 voxels in size. Objects less than 20 voxels in size were discarded.

This segmentation method was chosen after consideration experimentation and examination of previously developed cell segmentation techniques. While the neuronal nuclei labeled by the GCaMP were roughly elliptical, neither a circular nor elliptical Hough transform[132, 133] did a satisfactory job of detecting the objects, obviously missing a large number of neurons in the image while generating many spurious false positives. While manual inspection of the image suggested that standard filters and techniques would be sufficient, none of the standard approaches—local thresholding such as in Chapter 1, edge detection with morphological closure, or second-derivative filters like the Laplacian of Gaussian—were satisfactory, since all were unable to distinguish two merged objects that were in about the same location.

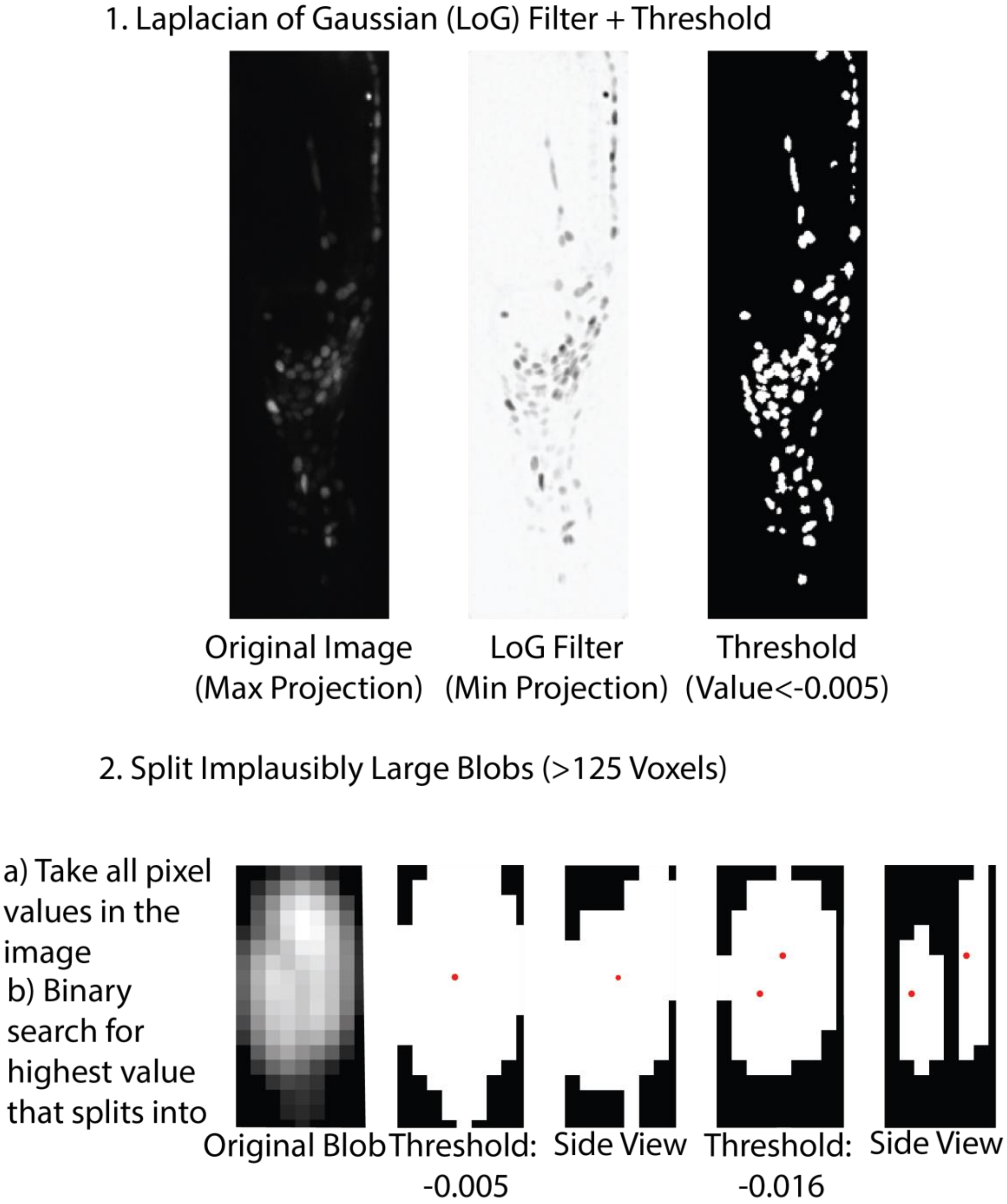

1. Laplacian of Gaussian (LoG) Filter + Threshold
Original Image
(Max Projection)
LoG Filter
(Min Projection)
Threshold
(Value<-0.005)
2. Split Implausibly Large Blobs (>125 Voxels)
a) Take all pixel values in the image
b) Binary search for highest value that splits into
Original Blob
Threshold: -0.005
Side View
Threshold: -0.016
Side View

Figure 4.2 Illustration of the Segmentation Procedure. In Step 2, the second and third image shows what was originally a single blob after initial thresholding. The fourth and fifth image shows what are now two. The fifth image presents a side-view of the final separation, showing that these are two distinct objects in the Z-plane. The centroids of the detected blobs are shown as red dots.

The usual solution to this, a localized watershed transform based on the distance transform, was tested but was not empirically as successful as repeatedly re-thresholding the object based on the size criterion, dealing poorly with neurons that were not roughly circular. While a relatively blunt and slow method, repeated rethresholding performs very well, likely because it takes intensity into account—a watershed-based separation would have difficulty with a nearly spherical combination of objects such as in part 2 of Figure 4.2. Separating objects based on other parameters like solidity were considered but discarded when they did not significantly improve results and substantially increased runtime. A completely separate convolutional neural net approach had also been used to segment the image, but was discarded when it was discovered that this separation approach worked better.

With a method for separating merged objects in place, it was judged not terribly important empirically which choice of segmentation filter was used—Laplacian of Gaussian was chosen as a natural choice for this kind of problem, and also for producing the smoothest nuclei. Parameters were chosen primarily by manual selection, though the maximal size of the objects was chosen by examining a histogram of object sizes for objects that were deemed correct. It is worth noting that under constant imaging conditions, the level of GCaMP would be consistent between animals and these parameters would not change much.

Figure 4.3 shows the result of some example segmentations.

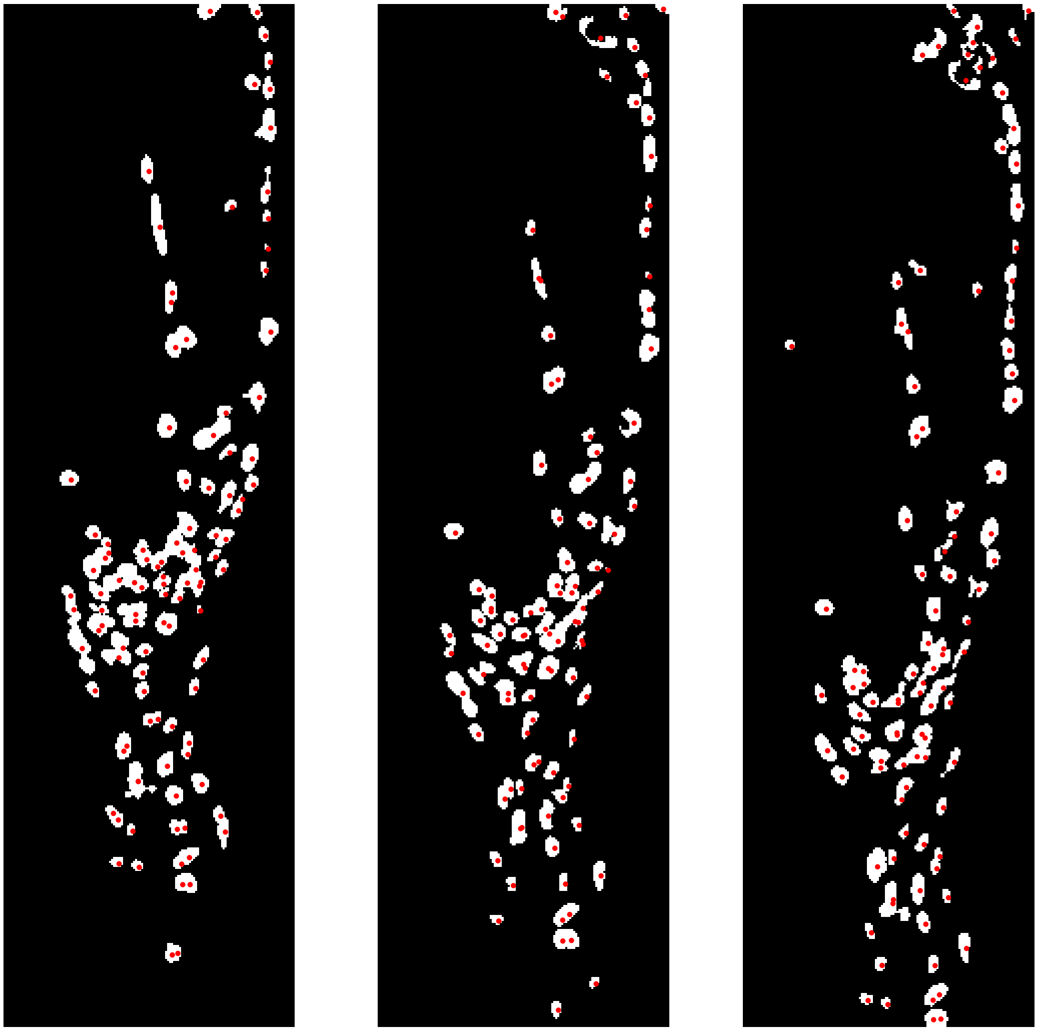


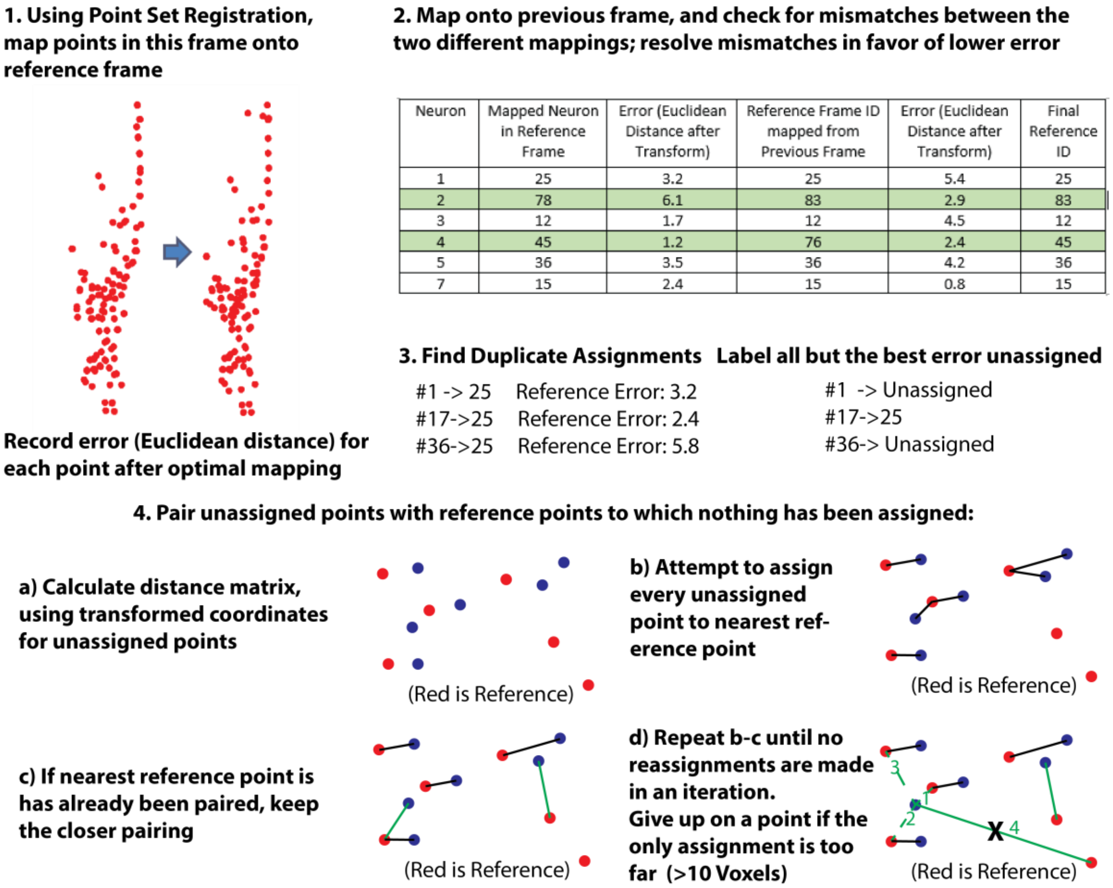
Figure 4.3 Three example segmentations of different frames from the same video, displayed as maximum Z-projections. The red dots indicate the X-Y location of individual cells identified by the segmentation. Note the need to separate merged objects, and the relative efficacy of the re-thresholding procedure at doing so.

Tracking with Point Set Registration

With the segmentation in hand, we considered it likely that a standard greedy tracking algorithm, using the Hungarian algorithm[134] or auction algorithm[135] to match points from frame to frame, might work tolerably well under these circumstances, since the videos from Zimmer involved heavily confined worms. We knew, however, that greedy tracking algorithm would perform poorly over long videos such as those from Kato *et al.*, since any single error in tracking would derail the rest of the data from a given neuron—and it was already observed that as the GCaMP signal faded in some neurons they'd briefly be lost in the segmentation. Further, with an eye to the future, it was considered important to be able to handle substantial motion and distortions of the worm, so that it might be possible automatically track worms that were not heavily confined, such as already existed in some studies (Table 4.1).

With this in mind, we considered a very different approach to the tracking problem. A greedy tracking algorithm would consider each neuron in the animal as an individual object to be followed, but each neuron is actually a single location embedded in a larger deformable object. Given that, the problem of identifying corresponding 3D points between different frames of a video strongly resembles another class of deeply-studied problems: point set registration[136-141].

Taken most generally, point set registration consists of finding a transformation that best converts one set of points into another set of points according to some criterion, usually one that corresponds well to *a priori* knowledge about how the transformation should behave. Most applications relate to points gathered from specific locations in a real-world object, and a number of algorithms have been introduced over the years to



1. Using Point Set Registration, map points in this frame onto reference frame

Record error (Euclidean distance) for each point after optimal mapping

2. Map onto previous frame, and check for mismatches between the two different mappings; resolve mismatches in favor of lower error

Neuron	Mapped Neuron in Reference Frame	Error (Euclidean Distance after Transform)	Reference Frame ID mapped from Previous Frame	Error (Euclidean Distance after Transform)	Final Reference ID
1	25	3.2	25	5.4	25
2	78	6.1	83	2.9	83
3	12	1.7	12	4.5	12
4	45	1.2	76	2.4	45
5	36	3.5	36	4.2	36
7	15	2.4	15	0.8	15

3. Find Duplicate Assignments Label all but the best error unassigned

#1 -> 25	Reference Error: 3.2	#1 -> Unassigned
#17->25	Reference Error: 2.4	#17->25
#36->25	Reference Error: 5.8	#36-> Unassigned

4. Pair unassigned points with reference points to which nothing has been assigned:

a) Calculate distance matrix, using transformed coordinates for unassigned points

(Red is Reference)

b) Attempt to assign every unassigned point to nearest reference point

(Red is Reference)

c) If nearest reference point is has already been paired, keep the closer pairing

(Red is Reference)

d) Repeat b-c until no reassignments are made in an iteration. Give up on a point if the only assignment is too far (>10 Voxels)

(Red is Reference)

Figure 4.4 Tracking neurons using point set registration, including both the consensus approach used (Step 2) and the fallback Gale-Shapley correction for duplicate assignments (Steps 3 and 4). In 4D, the numerals 1-4 indicate the first 4 attempted assignments for the blue point on the center-left. The last assignment, #4, was greater than 10 voxels and discarded, leaving the point unassigned. All values and drawings in this figure are fictitious for the purposes of illustration.

perform either rigid point registration, where the spatial transformation is restricted to be isometric[137-140], or non-rigid, where some other restriction is applied, and which is more suited for deformable objects[136, 141]. By treating the centroids of the identified neurons as individual points, it was found that the non-rigid coherent point drift algorithm could be used to reliably match neurons from frame to frame. Coherent point drift was chosen for supporting non-rigid registration, an important consideration in the deformable body of the worm, as well as being robust against noise and missing points. We used a freely available Matlab implementation from the authors, and do not explore the details of the algorithm here, but the interested reader may refer to the original paper on coherent point drift[136]. One downside to the use of coherent point drift is that, unlike methods such as robust point matching[139], it does not guarantee a one-to-one correspondence, which is highly desirable for cell tracking. However, it provides non-rigid matching, robustness to missing points, and a transformation model for the 3D space in which the points are embedded. We discuss mitigation of the downside later in this section.

Applying point set registration to tracking objects through a video requires some adaptation, however. As usually provided, point set registration maps one single set of points to another single set, not a large number of point sets to themselves. There are three natural solutions to this:

1. Register each frame to every other frame. Use a consensus approach to identifying points in each frame, taking the mapping onto every other frame into account. For instance, one could attempt to find the set of point assignments that would minimize the number of “mismappings” according to the point set registration.
2. Register each frame to the previous frame. Use the mappings to keep track of point assignment over time.
3. Define a reference frame, and compare all other frames to this frame, use the points in the reference frame as absolute identities.

We may immediately discard the first method as being impracticable—for a 3000 frame video, it would require 8.997×10^6 applications of point set registration—at roughly 0.1 s each, this would take 10.5 days. Methods 2 and 3 each have significant downsides. Matching each frame to the previous frame reintroduces one of the major downsides of greedy frame-by-frame tracking: the tendency towards accumulating errors over time as each transient misassignment permanently damages the accuracy of the tracking. However, matching each frame to a single reference frame discards all temporal information present in the video. While this may be desirable in cases where the positional correlation between adjacent frames is weak, that is not the case in the videos provided, where each frame of the video strongly resembles the previous one.

I found that a mixed approach, using a consensus of approaches 2 and 3, worked best. For any given frame (except the reference frame and first frame), the frame was compared to both the reference frame and the frame immediately previous. Whenever the two point set registrations disagreed about the identity of a given neuron, precedence was given to whichever one produced the least error—that is, the lowest distance between the transformed point and the actual point to which the mapping was attempted. Empirically, this gave a good combination of the advantages of both approaches, serving to limit the number of neuron assignments that don’t make sense given the previous frame, while also preventing long-term drift in neuron identity.

It is possible to imagine other, more complex approaches, for example using a consensus of regularly spaced reference frames, or of a number of frames immediately around the given frame, weighting by temporal proximity. However, neither of those approaches generated appeared to generate accuracy improvements relative to the given approach, and were discarded as needlessly complex.

One question that this approach engenders regards the choice of reference frame. A number of potential approaches suggest themselves, but I settled on the most conservative approach: choosing the frame with the lowest number of putative neurons.

Since this anchors the number of neurons the algorithm expects to find, this effectively places the least burden on the segmentation to identify all neurons in the image and on the tracking algorithm to recognize missing points, at the obvious cost of decreasing the number of neural traces identified. It is also possible that a reference frame with more neurons would be more informative about the geometric arrangement of neurons in a given frame.

As mentioned, coherent point drift does not produce one-to-one correspondences between frames. This is problematic, as tracking neurons and extracting calcium traces is dependent on the expectation that any given neuron is only found once per frame. This is resolved by using a greedy distance-based tracking as a fallback. Here we use a custom-modified Gale-Shapley algorithm[142] rather than the Hungarian algorithm that would be more typical, because we do not want the algorithm to be too concerned with decreasing the distance of a particular far-lying point (that may be an outlier or not in the other set) at the expense of other matches.

1. Examine the correspondence generated by the registration procedure for duplicate assignment—points that have been assigned to the same original point in the reference frame. For each set, preserve only the correspondence for the point with the lowest distance to the transformed point predicted by coherent point drift; discard the rest.
2. Compile a list of orphan points, points in the reference frame and current frame with no partner. For each possible pairing between the first and second group, calculate the distance between the point in the current frame and the transformed point from the reference frame predicted by coherent point drift.

   > *Note: this is an expensive calculation, but for all the videos examined, only takes place on a few points per frame. Calculation time is thus negligible.*

3. Perform modified Gale-Shapley matching between the two groups:

a. Repeatedly iterate through the list of unmatched points in the reference frame, pairing each to the lowest distance point in the current frame, recording the distance. If this distance is greater than 10 voxels, ignore this pairing.

    i. If this lowest distance point is already taken by another reference point, then transfer it to this point if the distance to this point is lower than the distance to the other reference point; otherwise, examine the next lowest distance point. Repeat until a pairing is made or the lowest distance is greater than 10 voxels.

    ii. Terminate iteration when the list of pairings is unchanged after an iteration. Gale-Shapley provides the guarantee that this will occur.

    iii. *Note: the 10-voxel limitation prevents matching of points over implausible distances, since neurons never move more than this distance in one frame. The termination condition is necessary as this matching no longer guarantees that all points will find partners, the usual ending condition for Gale-Shapley. It is likely that another registration algorithm, particular one that exploits the use of symmetric Euclidean distances here, would be faster, but the time cost of this matching is negligible; this also has the advantage of being reusable in 4.2.2, where the distances are not Euclidean.*

Post-processing and Principal Component Pursuit for Removal of Sparse Noise

After tracking and segmentation, it is necessary to perform additional post-processing, both to correct errors in the previous steps, and to adjust for noise intrinsic to the data collection. Some adjustments are made:

1. Neurons that had a variance in position of more than 800 voxels^2 were removed from the dataset. This was on the observation that in the videos given, no single neuron ever moved more than about 10 pixels per frame, but

that a few erroneously tracked neurons moved far more than this. This is specific to this set of videos, of course, and would not be admissible for a much more set of images.

At this stage, the calcium traces corresponding to the activity of each neuron are calculated. The raw calcium signal, F, from any given neuron in a frame is calculated by taking the average voxel intensity over all voxels in a segmented neuron. This is the standard approach in the literature[119], though it is possible to imagine that a different approach—e.g. taking the average of the top 20 voxels—would be less sensitive to the details of segmentation; empirically, it does not seem to matter much. The relative calcium signal, $\Delta F/\overline{F}$, is then calculated, where $\overline{F}$ is the average signal over all time frames and ΔF is the difference from this average in a given frame. As these are unstimulated neurons, $\overline{F}$ is the typical measure used to estimate the baseline intensity level, though it is possible to imagine more accurate measures. Additional corrections are then made:

2. Missing point interpolation and median filtering. Neither coherent point drift nor the fallback tracking guarantees that each putative neuron from the reference frame will be found in every frame, resulting in a small number of missing neurons in some frames. The $\Delta F/\overline{F}$ signal of these missing points is estimated by pchip (shape-preserving piecewise cubic) interpolation from the rest of the time trace, after applying a 3-point median filter.
3. Principal Component Pursuit is applied to the entire dataset to reduce sparse noise[115]. Final data are shown in Figure 4.5.
    a. This carries the same assumptions as in Chapter 3, i.e. that the data is low rank and the noise is sparse. While the latter is almost certainly true, the first is debatable, especially given that each "observation" here is a different neuron, rather than different measurements of the

same neuron. Implicitly, this assumes strong correlations between neuron behavior. Empirically, this is true, and PCP performs quite well at conditioning the data (Fig. 4.6b-c), but this has the potential to obscure sharp single neuron behavior.

b. With a reimplementation of the algorithm, PCP can be made robust to missing points as well as sparse noise[143]. This would obviate #2 above, though #2 would still be useful if PCP were not used.

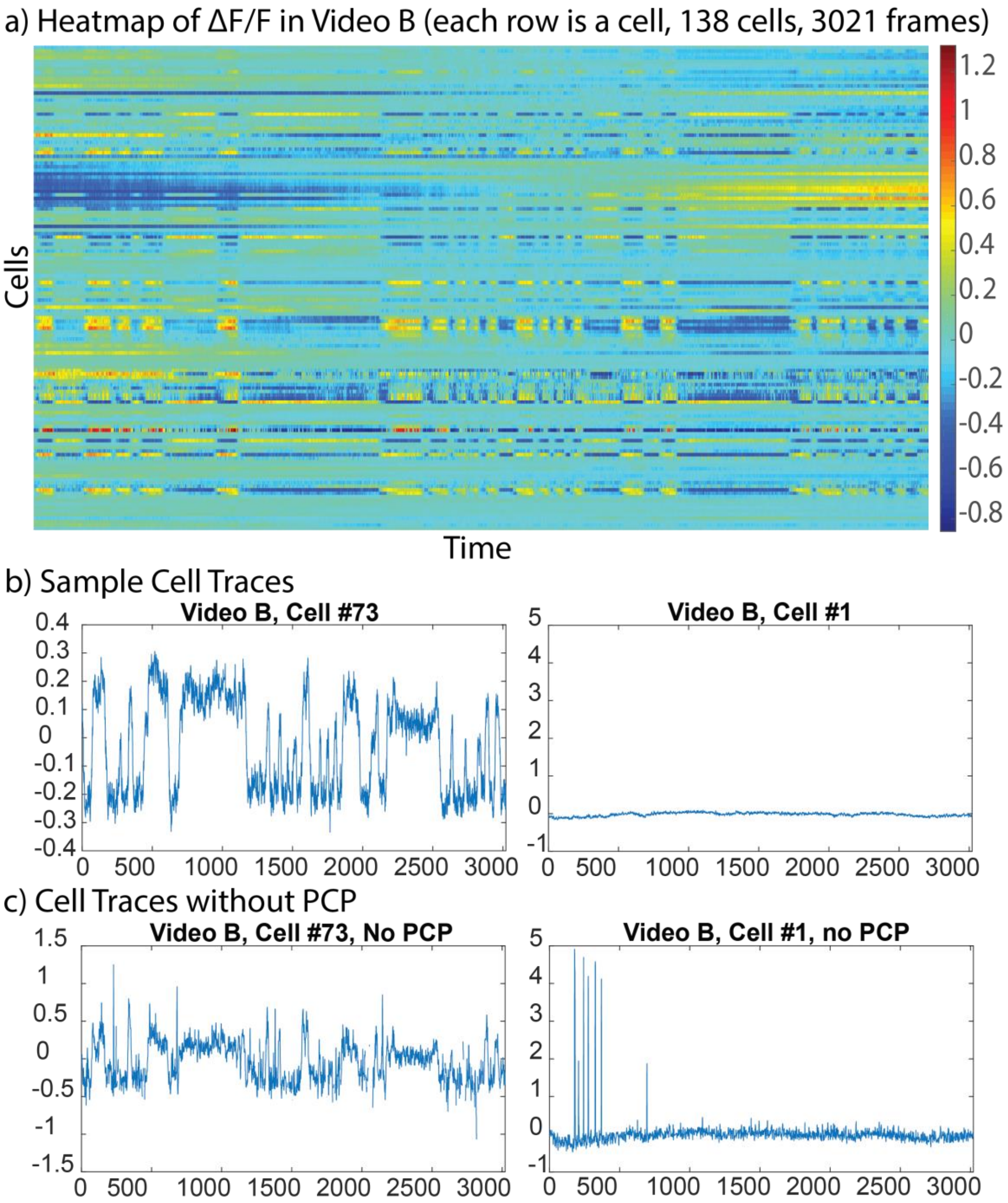


Figure 4.5 Sample calcium imaging data from a single video (labeled Video B based on the chronologically of when the recording was taken). Part A shows a heatmap of all 138 cells detected by the algorithm, in arbitrary order. The high degree of correlation between the activities of some neurons is evident. Part B shows two representative cell traces, the first of an active neuron and the second of an inactive neuron. Part C shows these same data traces without conditioning by PCP—the data contains considerably more sparse noise, particularly in the case of the inactive neuron. The magnitude of the peaks in the active neuron is reduced by PCP, however. Note that the y-axis between part b and c differs for cell #73.

Accuracy by Hand-Annotation

In order to evaluate formally the accuracy of the segmentation, randomly-selected frames of a video, segmented by the algorithm, were evaluated by eye. It was immediately observed that it was difficult to reliably count how many neurons the algorithm was “failing” to segment, as the question was too subjective and dependent on the observer queried. It was, however, somewhat possible to observe errors the segmentation made in separating merged objects, which remained a serious concern. Based on a detailed count of 10 of the sampled frames, an average of 10.5±2.67 (SD) cells, or 8.6±2.2% of the cells, were judged to be in fact more than one cell, clearly illustrating that the problem of separating merged objects is not yet entirely solved. It is important to note, however, that this is a problem that also plagues even hand-annotated data sets, which still rely on some form of automated segmentation.

In order to estimate the accuracy of the tracking algorithm by direct manual curation, *all* of the tracked cells from a single video were examined for tracking errors, using a custom-coded GUI, by Stellina Lee in our lab. Attention was not paid to the quality of segmentation, as that had already been addressed; the only question was whether or not each frame of the cells extracted from the tracking algorithm was accurate or not. As a first step, 7 cells that were missing from large segments of the video were removed, as these were clearly due to issues with the segmentation. The tracking algorithm performed well on the remaining cells, with only 3.3% of time frames judged to be incorrect. Of this 3.3%, 1.0% occurred when the tracker failed to properly follow a cell that was, in fact, present. The remaining 2.3% consisted of actual misassignments. Of

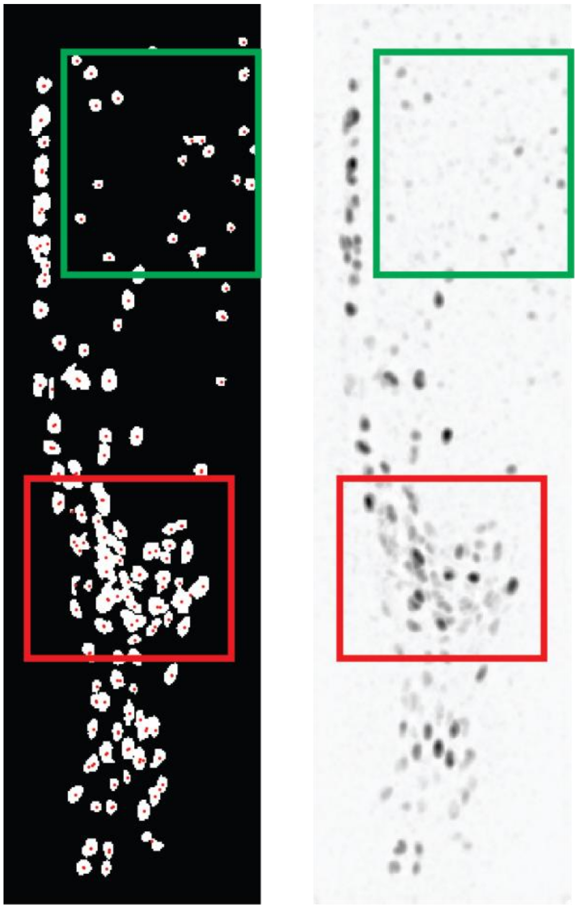


Figure 4.6 Example segmented and filtered image, illustrating where most of the segmentation errors occur (red and green boxes). Based on my manual curation, nearly all of the segmentation errors occur because of difficult separating merged cells in the central nerve ring (red box) and in accurately resolving neurons in the anterior mid-body (green box). It is worth noting that these are difficult problems to solve even by eye, and are not unique to this algorithm. Even manual observers would have difficulty segmenting these areas exactly.

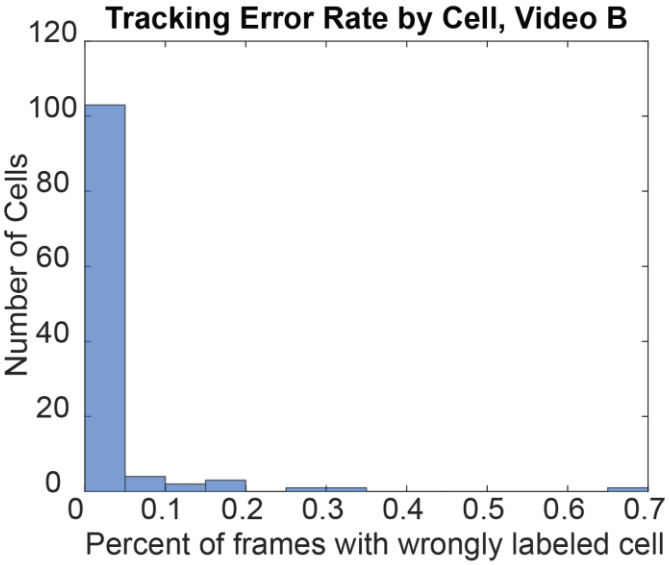


Figure 4.7 Frequency histogram of the misassignment error rate per cell, illustrating that the majority of errors occur in only a few cells.

these misassignments, 0.31% involved a cell that was not segmented on this particular frame, and could perhaps be blamed on the segmentation.

As illustrated in Figure 4.7, detailed examination reveals that most of the misassignments still occur in only a small subset of cells; removing 8 of these drops the misassignment percentage to 0.57%. Few of these were removed by the initial stages of post-processing, indicating that a greater effort can still be made to remove problematic cells.

Algorithm Speed and Cost

The combined process takes about 2 hours on the videos provided by Zimmer (150-170 x 512 pixels, 11-12 Z-slices, 2800-3250 frames) with a i7-4770k processor (3.50 GHz) and 16 GB RAM, and no optimization has thus far been attempted. Figure 4.8 shows a detailed breakdown of the time cost of this algorithm on one of the videos from Kato *et al.* and for one of the videos we collected ourselves (see Section 4.3). This breakdown is representative of all of the videos collected.

As can be seen, no one step is particularly dominating in terms of computation time. If *F* is the number of frames in a given video and *V* is the number of voxels in a frame (Width*Height*Depth), then the theoretical time complexities for each step are:

1) *O(VF)* for file loading and segmentation. (Dominant cost: Convolution and reading parts of the image file from disk)
2) *O(VF)* for Blob Extraction and Splitting. (Dominant cost: Blob Extraction)
3) *O(MF)* for tracking, where *M* is the number of putative neurons in the reference frame. (Dominant Cost: Coherent Point Drift)
4) ~*O(FN)* for PCP, where N is the number of iterations needed to complete PCP, which is directly related to the prevalent level of sparse noise in the dataset. This is an extremely rough estimate, and N itself varies by as much as a factor of 3 in our testing for each class of video (ranging from roughly 2500 to 7500 and 200 to 500 for Kato *et al.* video and our videos, respectively) but cannot be predicted *a priori*. (Dominant cost: Singular Value Thresholding within PCP).

In steps 2 and 3, the dominant cost is related to a Matlab built-in or imported algorithm which is already heavily optimized and difficult to improve on. The efficiency of step 1 is partially controlled by the efficiency of reading frames of the image from disk, which the built-in Matlab function is not particular efficient at, and seems to deteriorate as the file size increases into the multi-gigabyte range. As Step 1 is overall the shortest step, the cost-benefit of working to improve this file read-in is likely not worth it, especially as the built in convolution is heavily optimized.

Step 4, Principal Component Pursuit, has the most potential for improvements in speed. Substantially speedier versions of PCP already exist[143-145], relying on fundamental algorithm improvements, improvements to optimization, and aggressive use of optimization. It is likely that a choice of a substantially more efficient algorithm could render the time cost of step 4 unimportant compared to the other steps.

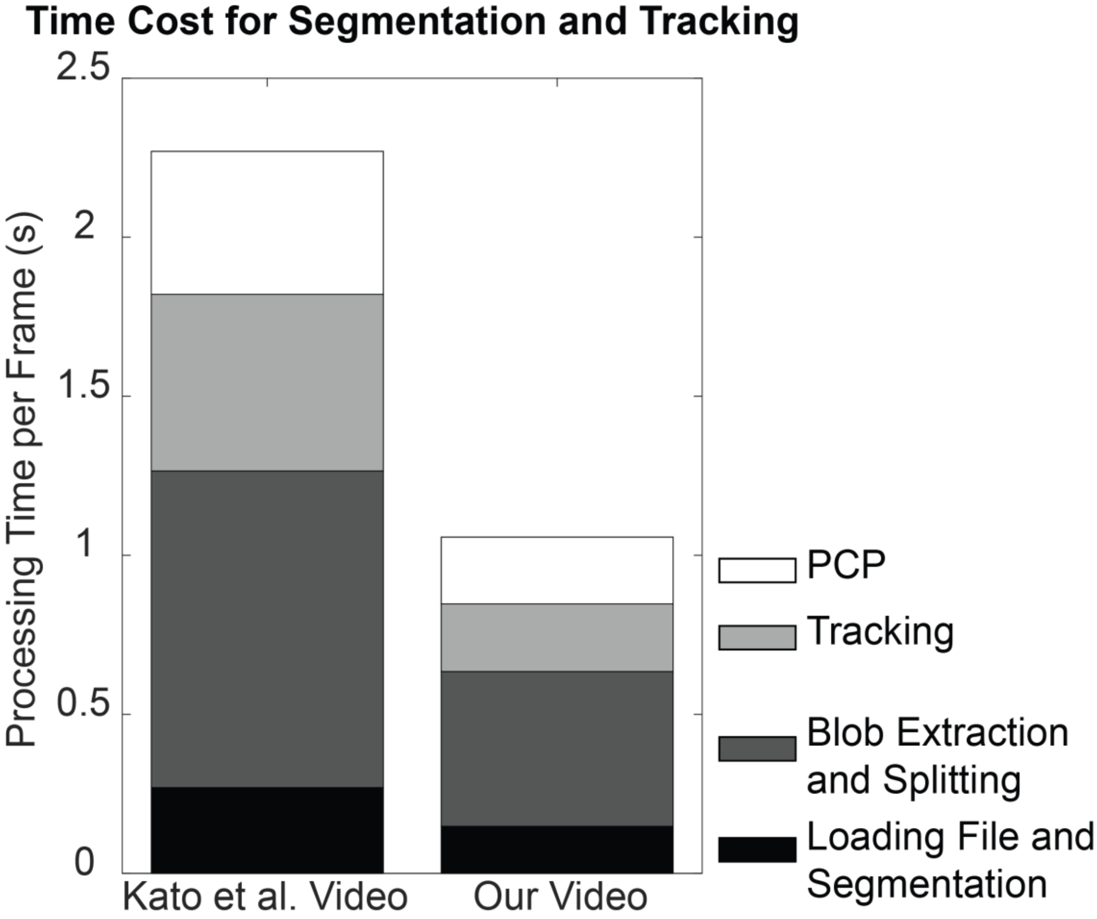


Figure 4.8 Time cost for each stage of the algorithm, for a representative video from Kato et al. and taken by our own lab (see Section 4.3). These represent the total time for each step divided by the number of frames in the video being analyzed (3021 for the Kato et al. video and 357 for ours), representing 114 minutes and 6.3 minutes total, respectively. Frame dimensionality was 148 x 512 x 12 for the former and 422 x 336 x 10 for the latter, with 147 and 36 putative cells identified, respectively.

4.2.2 Validation using Hand-Curated Data

Making a direct comparison between the data generated by my analysis and the previous analysis by Kato *et al.* cannot be a simple matter of examining individual traces for equivalency between the two sets of data—with different segmentation approaches used, the chance of two traces being identical is effectively zero, even when both traces measure the same cell. It must further be noted that while Zimmer's data is used here as a baseline for comparison, it cannot be assumed to be perfectly accurate, even with manual curation, as it still reliant on the quality of the segmentation used and the manual curation itself.

<u>An Automated Method for Finding Similar Cells in Two Separate Analyses</u>

From a technical standpoint, what we are most interested in is that the same cells are being found, and that the same general calcium activity traces are being measured. While the apparent accuracy of the algorithm with regards to cell selection was been verified by manual inspection in the previous section, we are still interested in the broad similarity of our data to the published data.

In order to identify the same cells in both data sets, we realize that we are once again facing a one-to-one matching problem, suitable for solution with a registration algorithm such as Gale-Shapley[142], almost exactly analogous to the fallback greedy tracking used in the tracking methodology above. It only remains to choose an appropriate distance measurement between cells in Zimmer's analysis and ours. Since we know the coordinates of the centroids of both our cells and Zimmer's cells, it is tempting to use the simple Euclidean distance—but this would make it difficult to disambiguate cells that are set very close to each other, especially given the focus of Zimmer's segmentation om relatively large regions of interest. Instead we choose a mixed distance:

$$(1 - R) + d/\lambda$$

Where R is the product-mean correlation coefficient (PMCC) between the data traces associated with a given cell, d is the Euclidean distance, and λ is an arbitrary parameter controlling their relative importance. This is of course no different in practice from the more usual $\lambda R + d$, but this formulation implies a useful way to think about lambda. Specifically, since $1 - R$ is at most 2 in the worst case where $R = -1$, then the d term is absolutely dominating if $d \geq 2\lambda$. In practice, even $d \geq \lambda/2$ is more than sufficient to make d the dominant term in this distance. As a rule of thumb, then, we can set $\lambda = 2d$, where d is the roughly the distance where we stop believing two cells are the same, no matter how similar their calcium traces. This is an important consideration, since distant cells in both datasets, particularly those that correspond to symmetric neurons of the same type, often show very strong correlations in their activity, as can be observed in both our and Zimmer's analysis (personal communication, Manuel Zimmer). In our case, we set $\lambda = 100$, though this is likely an overestimate, since the average diameter of cell in voxels is only at most 10. As we shall see, this does not seem to matter much.

It is important to note that the correlation distance $1 - R$ does not satisfy the triangle inequality, and hence neither does the mixed distance; however, this does not matter for the modified Gale-Shapley approach taken here, which matches that used in the tracking above, except that it does not have a maximal distance beyond which matches are not made. The choice of the PMCC as a measure of similarity between calcium traces is carefully done; it serves as an intuitive measure of similarity while remaining insensitive to differences in relative magnitude between the two traces, something which would be very sensitive to the exact, but unimportant details of segmentation.

Figure 4.9 illustrates the results of applying this approach to one of the five videos. The median PMCC between matched cells for all five videos ranged from 0.68 to 0.76, with a median distance between 2.4 and 7.8 voxels. Empirically, PMCCs over about 0.6 were reliable matches, provided that the cells in question were within 30 voxels of each other. A majority of the cells, 59.7%, fit this criterion, with 75.2 % of cells with a distance criterion < 0.7.

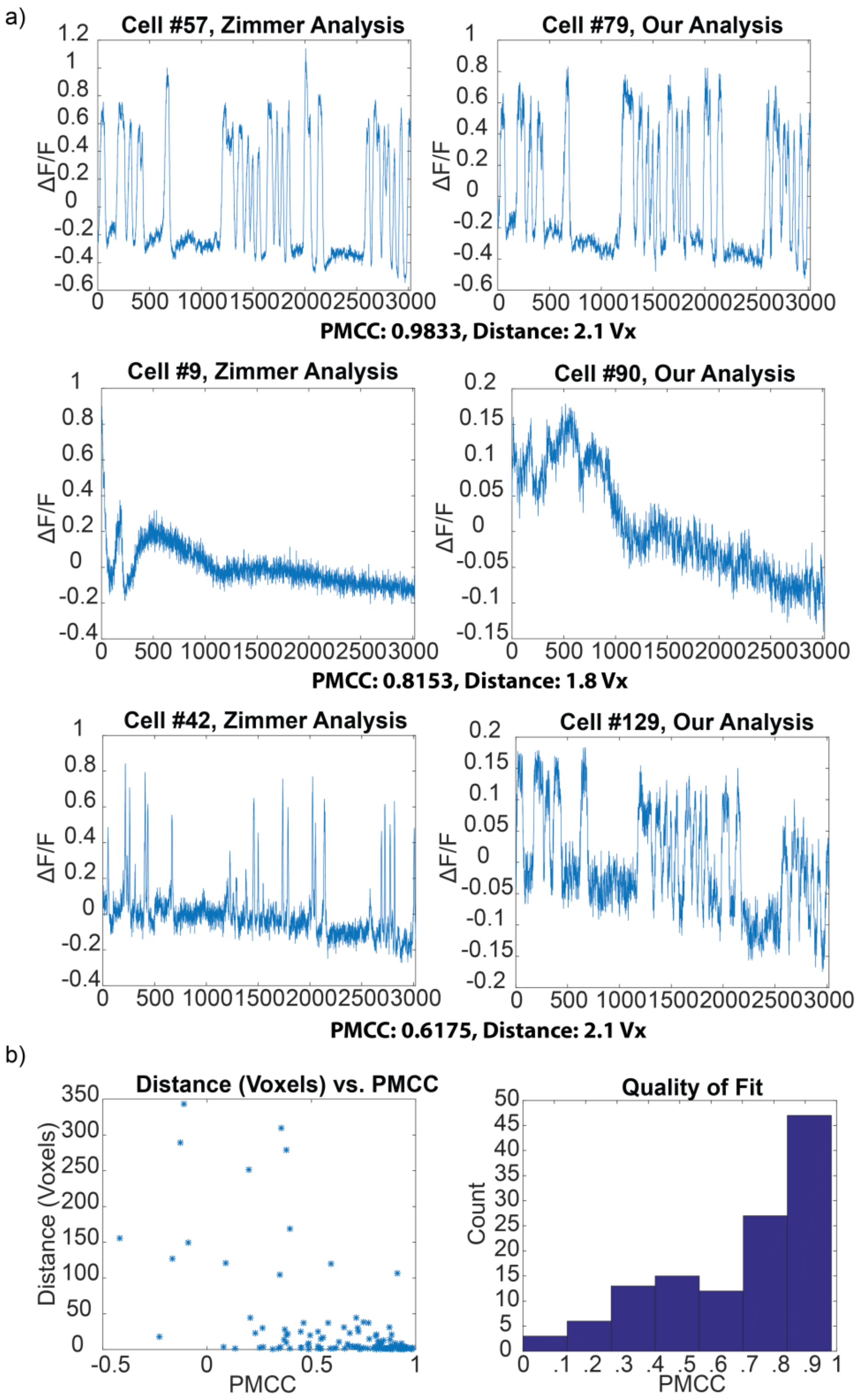


Figure 4.9 Example cell pairings between the two analyses, with Zimmer's analysis on the left and ours on the right. Part A shows an example of a PMCC that is nearly 1, one that is around 0.8, and one that is around 0.6. All three pairs of traces look broadly similar, though with some distortions likely due to differences in segmentation. Part B shows summary pairing data for a representative video.

PCA and Replication of Previous Results

In order to fully validate the developed algorithm from a scientific perspective, it would be ideal to replicate the conclusions of Kato *et al.* as much as possible. The centerpiece of their analysis was a temporal principal component analysis (PCA) which redimensionalized the activity of the worm ganglion into a three-dimensional manifold. By using the activity of key motor neurons as indicators for forward and backward worm movement, different parts of this manifold could be shown to correspond to different activity states, supporting a hypothesis that the entire worm brain is involved in different states of motor planning.

In order to sufficiently validate our results relative to the original analysis in Kato *et al.*, we consider it necessary to replicate the salient points of the temporal PCA analysis, illustrating that the 3D manifold generated has broadly the same features and shape, and that different parts of the manifold still correspond to different activity states. In this, we may use the same activity states identified in Kato *et al.* and even the same cell IDs, using the matching algorithm from the previous section to identify what are putatively the same neurons. To summarize the procedure, which was matched as carefully as possible to that used by Kato *et al.*:

1) Remove key high-responding and erratic sensor neurons from the data set ('BAGL' 'BAGR' 'AQR' 'URXL' 'URXR' 'AVFL' 'AVFR' 'ASKL' 'ASKR' 'ALA' 'IL2VR' 'IL2L').
2) Normalize all traces to their highest absolute magnitude.
3) Take the time derivative by using Total Variation Denoising[146], using manually chosen alphas matching those chosen by Kato *et al.*
4) Perform PCA, treating each time point as an observation and each neuron as a dimension.
5) After redimensionalizing the data into PCA dimensions, reintegrate each PC dimension time course.

The movement state of the worm and cell IDs were determined from the Zimmer data, with the cell IDs based on finding the match to each given cell ID in the Zimmer using the method in the previous section. Figure 4.10 summarizes the results for a single video and gives an illustrative comparison with equivalent figures calculated based on the data from Kato *et al*.

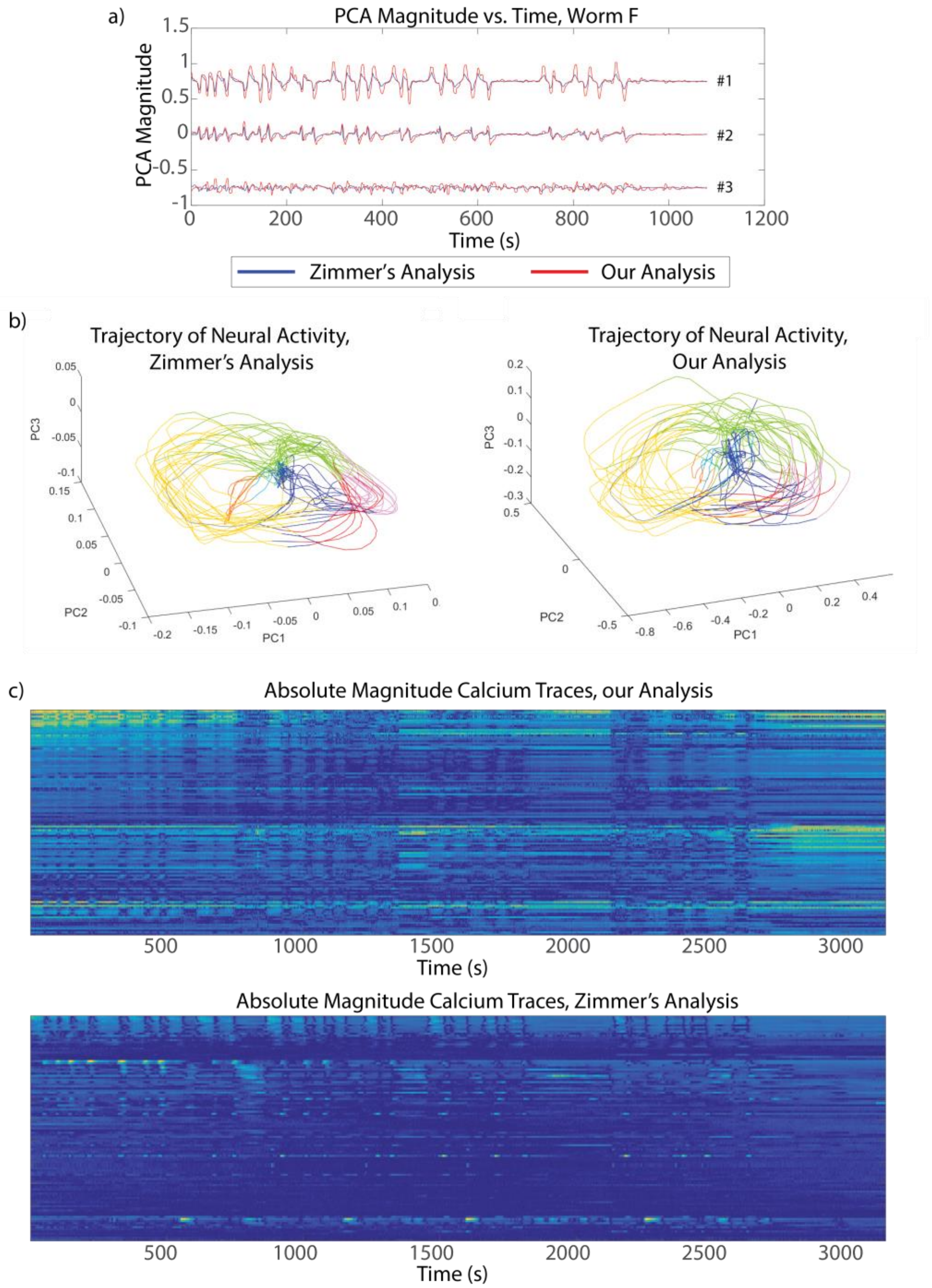


Figure 4.10 A replication of some of the results from Kato et al.[5], illustrating the similarity that results when using our analysis and theirs. Part A shows an overlay of PCA component magnitudes over time for both versions. Part B shows the time trace of neural activity in principal component space. Colors show the RISE/FALL states identified by Kato et al. for each time point and represent the same time points in both traces. Part C shows heatmap plots of the absolute value of the calcium traces in both analyses. Each horizontal line is one cell, and the cells are grouped into three groups based on whether they have the highest weight in principal component 1, 2, or 3, and sorted from highest to lowest weight. The same patterns and cells are evident, though this is unfortunately obscured by the differences in relative intensity between the two segmentations.

Most of the PCA analysis is courtesy of Shivesh Chaudhary, from our lab, though the author was involved in planning and interpretation.

4.3 Imaging and Analysis of a Large Sample-size Set of *C. elegans*

In order to demonstrate the applicability of the algorithm to a large sample size collection of worms, and to test its use under slightly different imaging conditions. We immobilized ZIM504—the same strain as used by Kato *et al.*—day 1 adults on agar pad in 5 mM tetramisole. Using a spinning disc confocal microscope, we then imaged each worm for 10 minutes each, at a frame rate of roughly 0.6 Hz, not imaging more than 6 worms a session to prevent the worms from being on the pad for too long before imaging. This protocol was deliberately chosen to be both simple and traditional, to highlight the lack of reliance on any particular protocol or technique. This has two downsides: the difficulty of getting individuals into an ideal posture for imaging (left or right side pressed into the glass), which substantially reduces the number of detectable neurons in some animals, and the reliance on a relatively high concentration of paralytic tetramisole to ensure smooth imaging. Once videos were obtained, they were run through the previous algorithm to extract calcium traces, a process that took only a few hours.

4.3.1 Results and Preliminary Analysis

Figure 4.11 shows summary data obtained for 2 of the 16 videos obtained. In this case, unlike in 4.2, there is no cell ID or activity state information. Evaluating the underlying similarity between the activity of many individual animals is be a topic of considerable interest, as it would establish a firm baseline for the global neural activity of *C. elegans*.

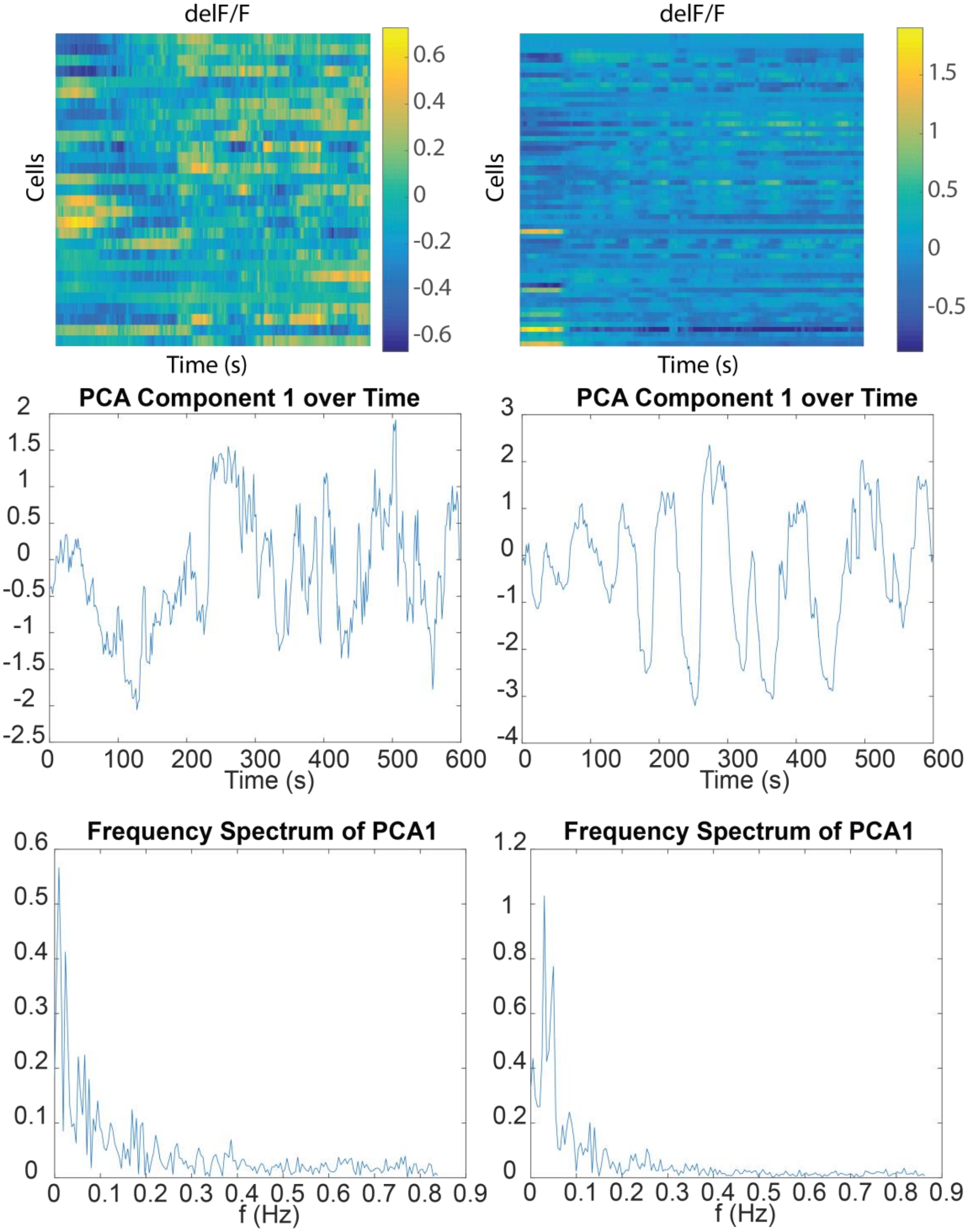


Figure 4.11 Best-result calcium traces from two of the individuals imaged on agar pad. The top row is shows raw heatmaps of the calcium traces as ΔF/F. The second row shows the value of the first principal component over time; the first principal component here is calculated in the same way as the last section of 4.2.2, but with no application Total Variation Denoising. The last row shows the frequency spectrum of this first principal component trace. The presence of two or more low frequency peaks is fairly typical, and was seen in 11 of the 16 videos examined. The number of neurons detected here is 36 for the left set of figures and 64 for the right set.

By performing much the same PCA analysis as in 4.2.2, except for the use of Total Variation Denoising and conducting a Fourier analysis of the time trace of first principal component, we see a moderately consistent frequency peak at roughly 0.03±0.01 (SD) Hz (36±10 s period), present in 11 of 16 of the videos analyzed. Unfortunately, we are limited by the low temporal resolution of our current imaging approach, which after all only has a maximal resolution of ~1.7 s, as well as the limitations of imaging on agar pad rather than microfluidic device, with the result that many of the neurons in the videos were obscured by the orientation of the worm, and that the intensity traces were occasionally contaminated by Z-shifts in the worm head, which increased or decreased the intensity of all neurons simultaneously. In the future, this could be resolved straightforwardly by imaging animals on a microfluidic device as in Kato *et al.*, or being more selective in the videos chosen for analysis.

4.4 Discussion

In this chapter of the thesis, I develop an efficient, fast, and accurate algorithm for automatic segmentation and tracking of "whole ganglion" videos in *C. elegans.* By using previously published, hand-curated videos, I demonstrate its accuracy by direct annotation, by cell to cell comparison with the previous hand curation, and by substantive replication of previous results. This algorithm takes only about 2.5 s to run per 150-170 x 512 x 11-12 frame, so it is not only easier than the manual annotation, it is also faster, and likely more objective in its assessments, making the analysis of whole ganglion imaging videos substantially easier.

While great pains are taken to demonstrate the algorithm's accuracy and its comparability to hand curation, the main value of algorithm lies not in superior accuracy or the mere proof of concept for automated tracking, but in its combined labor and time-saving value in eliminating what is otherwise a multi-day, intensively manually-supervised process. We illustrate this by applying it to analyze 16 videos, an immediate leap in sample size compared to any previous study, with an overall protocol that takes only a few hours of data gathering and a day of processing, giving a glimpse of the relatively huge amounts of data waiting just over the horizon. Even with a poorly-refined experimental protocol and low temporal resolution, it is already possible to detect a very slow ~36 s oscillation in neural activity present in *C. elegans*.

Thus, in its current state, despite the limitations described below, the given algorithm is sufficient to substantially advance the state of the field, and would serve as a valuable tool for future whole brain imaging. With further improvement to resolve the issues with segmentation of closely clumped cells, the simultaneous analysis of all of the neurons in the head ganglion during sensory stimulus could finally help to untangle the difficult to find relationships between interneuron behavior and external stimulus, or understand the activity of the *C. elegans* connectome during poorly understood whole brain activity, such as lethargus or learning. It will help to extend our understanding of *C. elegans* neural processing beyond relatively simple and well-defined circuits to the nebulous processes that direct the animal as a whole, and also to help elucidate the global effect of genetic changes to behavior, or of neuromodulators such as serotonin. It is our hope and expectation that this segmentation and tracking algorithm, or a close variant, will drastically accelerate the rate of progress in studying large-scale neural activity in *C. elegans*.

4.4.1 Limitations and Considerations

The process outlined in this chapter of the thesis has a number of clear limitations which are important to address. There are two major caveats, pertaining not to the algorithm itself, but to the limitations of its scope. Firstly, while the algorithm performs quite well on immobilized worms such as used in Kato *et al.* or in the imaging performed here, there is considerable interest in imaging freely-behaving worms, embodied in a number of already published studies. We have not yet attempted to apply our algorithm to a freely-behaving situation, but there is reason to believe that adjustment would be necessary, given built-in aspects of the algorithm that rely on a relatively immobile worm. That being said, our experience with some of our own calcium imaging videos with more mobile worms offers reason to believe that the tracking algorithm is robust to significant movement (Fig. 4.12 and 4.13).

Secondly, segmentation and tracking only addresses one of the two most difficult steps of calcium video analysis. While the analysis algorithm provided here provides a fast way of generating calcium traces for a large number of videos, providing considerable fodder for large-scale, neuron identity-independent analysis, many

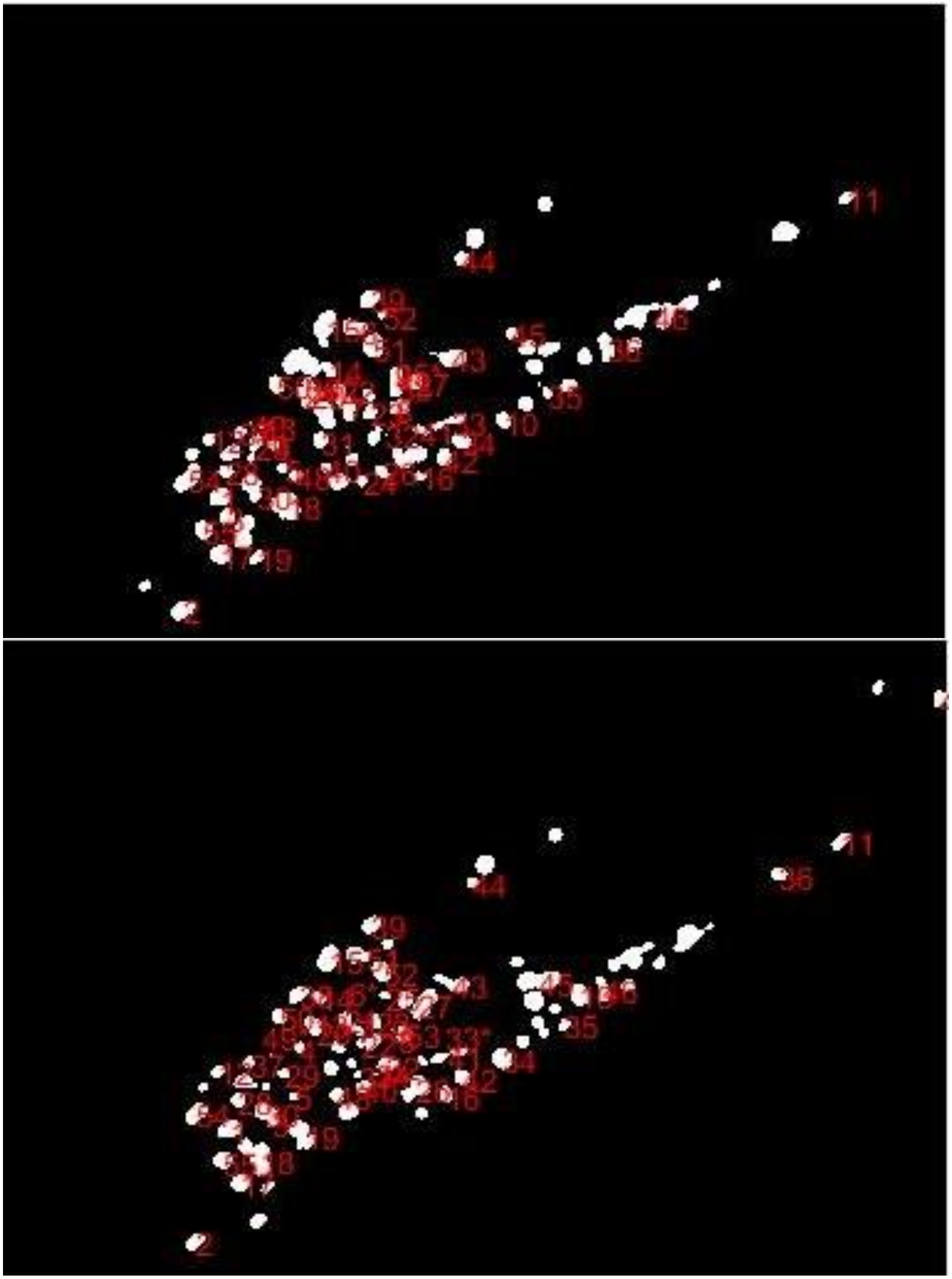


Figure 4.12 Tracked neuron identities for segmented frames of a single calcium imaging video in which the individual moved significantly. Neuron identities were tracked reasonably well, though clear errors show where improvement is still possible. This is frame 42 and 85.

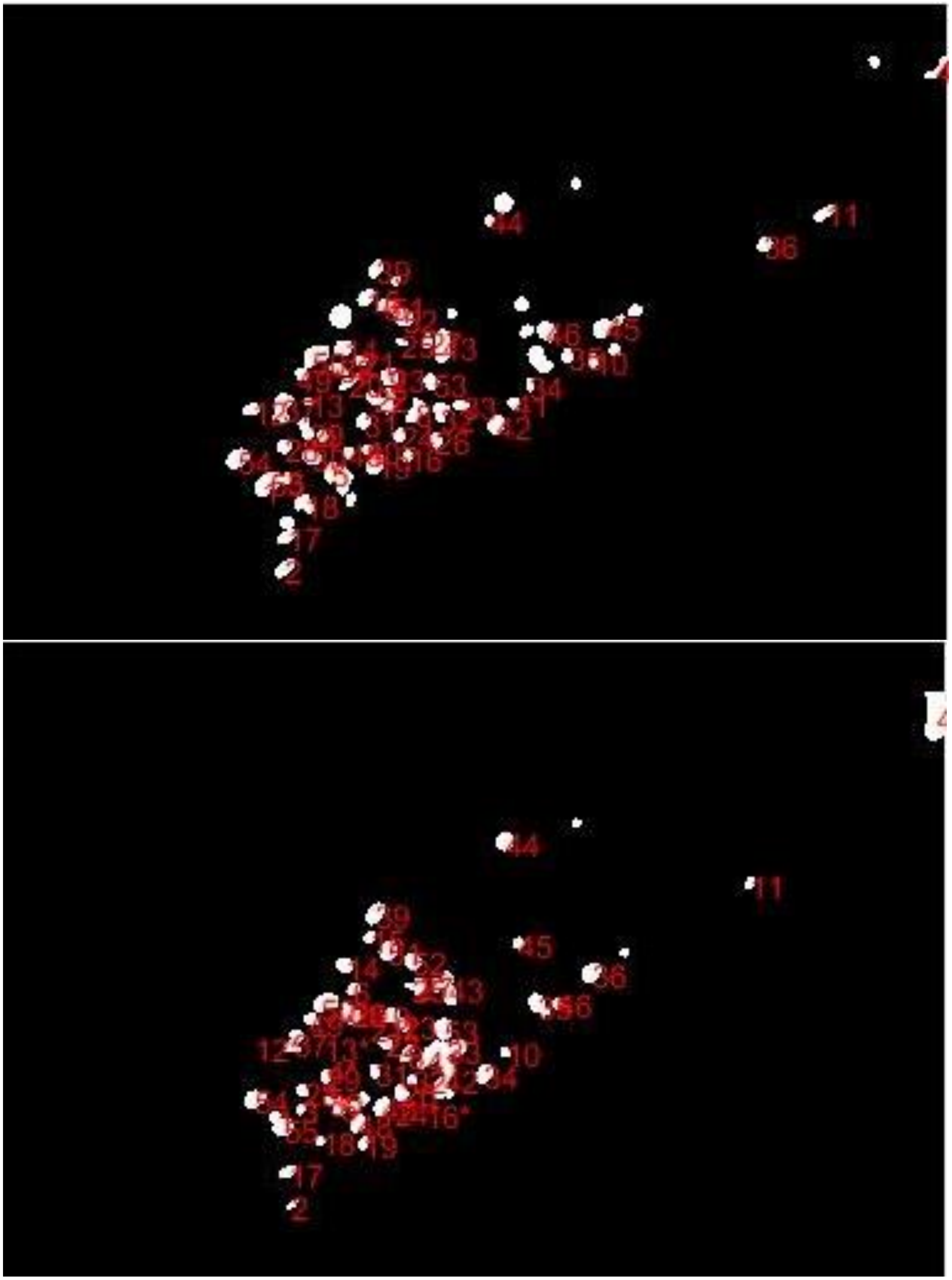


Figure 4.13 (Continued from 4.12) Tracked neuron identities for segmented frames of a single calcium imaging video in which the individual moved significantly. Neuron identities were tracked reasonably well, though clear errors show where improvement is still possible. This is frame 186 and 285.

biological questions are reliant on knowing the precise identities of the neurons found. Considerable attempts have been made at automating the identification of individual neurons, including by this author, but substantial variation in neural morphology between worms even of the same genotype have greatly impeded efforts in this regard. However, two separate potential solutions to this problem present themselves. On the one hand, long-ongoing efforts to generate worm strains with a combinatorial pattern of genetically-encoded fluorescence in each neuron, suitable for identification use, may finally reach the stage of widespread usability, experimentally solving a problem that has not been solved algorithmically. Alternatively, by imaging a large number of different strains, each labeled pan-neuronally by one color while labeled in specific landmark neurons by a different color, it might be possible to build a database of many different worm neural landscapes, each with key neurons identified. This would be suitable for training a machine learning algorithm to learn to identify neurons of interest.

Outside of these limitations, there are also a number of considerations in the current algorithmic pipeline that should likely be addressed in the future. For instance, the separation of merged objects used here is no longer state of the art in the field; the laboratory of Yuichi Iino has demonstrated that a more refined separation of merged objects with isointensity surfaces performs well, almost certainly better than the blunt re-thresholding used here[147]. It is also possible that a watershed transform using an intensity-aware distance transform would also perform better. In videos where the frame-to-frame variation is low, it is also possible that using adjacent frames to help inform a segmentation would have value.

A number of technical improvements can be also made in the tracker; some of the parameters were set empirically, but it is possible that a proper parameter search might find a more optimal set of parameters. The choice of the frame with the minimum number of neurons for a reference frame places the least pressure on the segmentation, but is probably not the optimal choice of reference frame—a better automated way to choose one might be found. Finally, rather than interpolating over missing points and using PCP, we could implement a version of PCP that directly handles missing points, simplifying and likely improving the approach[143].

` The field as a whole is fast-moving, and it is likely that aspects of the work here have already been duplicated or are being improved upon elsewhere. Moving forward, it will be necessary to sample the literature and compile the best-performing, best-practice versions of the techniques embodied here, rather than being bound to the specifics of algorithmic implementation. Nonetheless, we believe the successes demonstrated here will serve as an impetus to jolt the field forward.

CHAPTER 5

THESIS CONTRIBUTIONS AND FUTURE WORK

5.1 Thesis Contributions

This thesis set out to advance the state of the field in the use of image processing and computation in *C. elegans* research, particularly as it applies to high-throughput imaging and neuroimaging. As the questions asked by the field grow in subtlety and ambition, technical advancement in both imaging and computation have kept pace, awaiting only adaptation to the biological domain. Building on the techniques, equipment, and designs of previous researchers in this lab, this thesis sought to advance both the breadth and depth of its applicability to the questions of interest in the field, both by broadening its usability and by advancing the methodological state of the art.

We began in Chapter 2 by examining the existing high-throughput microfluidic imaging pipeline and addressing outstanding issues in its performance and methodology. To address the problem of imaging particularly dim markers in the presence of confounding droplets, and to address the problems the original segmentation approach had with overfitting a particular set of imaging conditions, an alternate, simpler segmentation approach was developed, and a new set of quantitative features was chosen. To demonstrate the value of this new approach, and to illustrate the applicability of high-throughput imaging to candidate gene studies in addition to forward genetics, we characterized a pre-selected set of dim synaptic mutants, demonstrating the ability to repeat manual characterization and go beyond, by demonstrating numerically the existence of epistasis between two synaptic mutants where this could not be determined manually. As a further demonstration of the scope of the segmentation approach, we applied the general approach to two novel situations, developing *D. melanogaster* embryos and arrays of Jurkat T cells, illustrating the ability of a filter and clustering approach to generate biologically useful results in a number of different situations.

In Chapter 3, we use the techniques and approach developed in Chapter 2 to further extend the reach of our high-throughput imaging technique to QTL mapping of synaptic morphology, with the two-fold goal of demonstrating the applicability of the technique to quantitative genetics, and to demonstrate the ability of the technique, coupled with focused algorithm and methodological development, to take on what is otherwise a technically infeasible task, the QTL mapping of fluorescently-labeled synapses within *C. elegans*. Here the technical difficulty was also two-fold, lying first in the difficulty of generating the large number of RILs containing the requisite fluorescent marker, and in the imaging of the micron-level synapses for a large number of individuals for a large number of strains. We overcome the first difficulty by adopting a novel process of imaging F1 crosses instead of the fully-integrated strains. This introduces a number of technical difficulties which we overcome with the use of the high-throughput imaging pipeline, which also takes care of the second difficulty, after additional modifications to aid in quantitative accuracy. By examining 47 RILs between the strains N2 and CB4856, crossed with the marker genotype *wyIs92*, we are able to identify a putative QTL on chromosome IV (though not at the significant level) and begin to verify it by studying introgression lines containing the putative QTL region.

Finally, in Chapter 4, we turn our attention from structural imaging to functional imaging, spurred by the development of new microscopy techniques that enable the so-called whole brain imaging of the *C. elegans* head ganglion, enabling the ~3 Hz monitoring of many neurons simultaneously in a bid to discern large scale functional patterns in *C. elegans* neural behavior. While a number of labs have performed this kind of whole brain imaging[4, 5, 8, 14, 17], the analysis and scientific payoff of this data collection has been bottlenecked by the processing and quantification of the 3D videos generated—segmentation and tracking of neurons typically requires days of manual hand correction and curation before the data is usable. We set out to design an improved segmentation and tracking algorithm capable of generating results comparable to manual curation without the heavy investment of time and effort, enabling the processing of whole brain videos on a large scale. We achieve this with a combination of a simple filter, a rethresholding approach to separate closely crowded cells, coherent point drift for tracking[136], and post-processing that includes principal component pursuit[115]. With this, we are able to replicate manually curated results on published calcium imaging videos, as well as perform analyses of videos taken by our lab.

In summary, this thesis takes the microfluidic and automation techniques in this lab and extends them into new domains by developing new algorithmic and methodological approaches. It is our hope that the techniques developed in this thesis will allow the detailed future interrogation of the *C. elegans* nervous system, both structurally and functionally, and of the detailed relationship between the *C. elegans* genotype and phenotype. More generally, we hope that the algorithmic approaches embodied here spur both continued innovation and a new generation of experimental studies.

5.2 Future Directions

In this section I outline a limited selection of potential future work elaborating or continuing the work in this thesis. I focus here on the most obvious experimental continuations of the work done, having discussed much of the potential technical improvements in the respective chapters.

5.2.1 Reverse Genetics on Synaptic Mutations

The successful and clear demonstration of epistasis between synaptic mutations demonstrated in Chapter 2 served as an illustration of the value of quantitative phenotyping in extracting difficult to characterize properties of subtle phenotypes, but as a scientific conclusion it is relatively limited. It is of some interest that the genes *unc-104* and *jkk-1* are epistatically linked, but this observation would only form a very minor part of the much greater patchwork that controls synaptic regulation and development.

One immediate possibility that emerges from the work in Chapter 2 is simply the direct evaluation of a large number of known synaptic mutants. In each case there are potential subtleties in the phenotypic that may lie undiscovered, or may have previously only been suspected by manual observers. The scope of manual description is usually limited—synapses are brighter, or dimmer, or larger, or smaller—but a full study with a wide range of carefully chosen features would likely reveal additional subtleties, subtleties that could help illuminate the exact role of specific genes if they phenocopy another, better understood gene.

The work in Chapter 2 also points the way forward to a potentially much deeper study of gene-gene interactions as they pertain to synaptic morphology and development. Many synaptic development genes are known already, but their exact role and interaction with each other is often poorly understood, if it is understood at all. A detailed study of the phenotypic outcomes of combining numerous pairs of different synaptic mutants could reveal much in the way of unexpected or unknown interactions between different genes, or could help rule out suspected interactions. The inferential value connecting epistasis to detailed genetic pathways is limited, of course, but the such a study could be conducted on a very large scale using the existing techniques, perhaps even using genes whose influence on the synaptic domain is only suspected.

One last possibility would be to expand the scope of the synapses studied beyond a single neuron to multiple or numerous neurons. Analogous to the situation in whole brain imaging, pan-synaptic strains are being developed that distinctly label individuals synapses throughout the whole worm, with markers where the expression level is controlled by a precise gene editing technique like CRISPR/CAS-9, preventing the overexpression and washout of individual synapses often seen in densely packed regions like the head ganglion. The study of synaptic mutants on a larger number of neurons, instead of the one neuron where they were often originally characterized, has the potential to illustrate the very poorly differences in synaptic regulation and gene expression between the different neurons of *C. elegans*.

5.2.2 Biological Conclusions from the QTL Analysis

The chapter for which the logical continuation is the clearest is Chapter 3. While this thesis has demonstrated that QTL mapping on fluorescent synaptic markers is possible with the given methodology, much remains to be said about the biological consequences of what has been found. A putative QTL is not the same as a verified synaptic morphology-affecting gene, and much work remains to be done to achieve a full verification, including the further identification of the likely genetic region involved and a study of genetic knockouts or mutants of probable genes in the area to characterize whether or not they have the expected phenotypes. Additional work would have to follow, characterizing the exact phenotypic effect, its potential interactions with other known genes and, if possible, its exact role and function. Likely the techniques from Chapter 2 would have a role to play, but the mainstay here is traditional biological work.

The implications of what has been found to what can be learned about synaptic morphology from studies on N2 would heavily depend on what exactly has been found. It is intriguing that the laboratory strains would carry a morphological difference in synapses when compared with wild isolates, but it is currently an open question why this occurred, or what effect it would have on studies in N2. Most excitingly, this is exactly the kind of synaptic variation that may play a role in certain kinds of human disease[97-101].

Further out in the future is the potential of carrying out this kind of QTL analysis on other phenotypes or other wild isolates. That would require another study on at least the same scale as this, with a methodology hopefully improved by the experiences of this one, as laid out in the conclusions of Chapter 3. This thesis is likely not the final word on the application of QTL to fluorescent markers.

5.2.3 Calcium Imaging of the *C. elegans* Head Ganglion under Stimulus or in Mutants

With the advent of efficient large-sample size whole ganglion functional imaging comes a whole host of potential questions for the field to address. The vast majority of questions about neural function that have been asked about individual neurons or small groups of neurons has a valid, unanswered analogue when applied to the neural connectome of *C. elegans* as a whole.

Focusing on the studies that are within ready reach of the methodology as already developed, it would be very straightforward with the current setup to study the effect of different chemical stimulants on *C. elegans* neural behavior and activity—indeed, the study by Kato *et al.* has already investigated to some degree the effect of oxygen concentration on the animal[5]. One could imagine providing any one of a number of chemoattractants, repellants, or pheromones and examining the effects, looking through the data with reverse correlation for novel neurons responding to the stimulus, or to examine the activity of the ganglion as a whole—it would be surprising if the resting pattern found by Kato *et al.* holds steady under heavy stimulus[5]. With the use of a microfluidic device to deliver mechanical stimulus (Cho et al. in preparation), it would be possible to even study the effect of touch, provided the algorithm can mitigate the effect of the body deformation. Another possibility would be to provide the worm with a heavy dose of a neuromodulator such as a serotonin and observe its effect on neurotransmission.

Another straightforward study would be to examine the effect of known synaptic mutants or neural activity. Transmission or neuromodulator mutants, or even morphological mutants, could be examined for their effect on the behavior of the neural population with the simple step of using an existing mutant strain for imaging rather than the default N2. Many mutants have subtle or unnoticeable behavioral effects, and this would be an intriguing way to more rigorously characterize their effects on the worm, perhaps unveiling neurons not previously suspected to have been involved.

This is by no means an exhaustive listing of the potential studies enabled by Chapter 4, but only an illustration of the clear scientific value of even the simplest of modifications to the study protocol—the lowest-hanging fruit, in other words. The field of whole brain imaging will have a lot to do, and we hope that this thesis has provided the tools.

APPENDIX A

NOMENCLATURE IN THIS THESIS

Gene names are italicized and are formed by letters followed by a hyphen followed by a number—for example, the gene *unc-104*. Genes are very frequently named for the phenotype caused by a mutation in the gene, since this is often how the gene was discovered. This rule is broken for transgenes, which are named based on their original name, e.g. *gfp*. Fusions between two different genes are referred to by connecting the parent genes with a double colon, e.g. *gfp::syd-2* is a fusion of the proteins *gfp* and *syd-2*.

Mutant genotypes are, confusingly, often referred to by the same name as the gene that is mutated. When clarity is desired, the specific name of the allele will be placed in parenthesis afterword. Alleles are usually 1-2 letters followed by a number. For example, in *unc-104 (wy673)*, the allele name is *wy673*. The letters are a special designation identifying the lab which isolated the allele; the numbering is assigned by the lab. Optionally, the chromosome of the gene will be identified immediately after designation, e.g. *unc-104 (wy673) II* indicates that this gene is on chromosome II.

In this thesis, the allele of a mutant will always be identified when confusion is possible; otherwise, it will be identified the first time the mutation is mentioned and the abbreviated version without the allele name will be used afterward. The chromosome number will not be used unless it is relevant to the discussion at hand.

Transformed genotypes, which include extra genes compared to the base strain—e.g. strains with a fluorescent marker—are named similarly to mutant allele, except than an “Ex”, “Is”, or other abbreviation is inserted between the lab code and number, indicating whether the extra genes are in an extrachromosomal array (Ex) or have been integrated into the genome (Is). In the special case where a targeted gene editing approach has been used to insert a gene “in-line” with another gene—that is, directly afterward, under the control of the same promoter—the genotype is referred to instead as a specialized allele of the gene where the insertion, e.g. *syd-2 (wy1073[gfp(no introns)::syd-2])* indicates that this is the *wy1073* allele of *syd-2*, noting that this contains in addition a fused *gfp::syd-2*. Note the use of parentheticals to enclose additional information.

Strains are named based on 2-letter lab code followed by a number much as mutant alleles are—for instance CB4856 or MY14—with a few exceptions dating back to the start of field (such as the very common N2 and LSJ2). These are written capitalized and non-italicized.

Protein names are based on the gene name, and are just capitalized non-italicized versions of the gene name. e.g. the *unc-104* gene encodes the protein UNC-104.

APPENDIX B

GENETIC MANIPULATION IN *C. ELEGANS:* AN OVERVIEW

This appendix discusses in detail common methods for the genetic manipulation of *C. elegans* and is an extended version of the discussion in Chapter 1, focusing on the generation of novel mutants, the insertion of fluorescent, and the combination of different strains into one. While this section is intended to provide additional information for the interested reader, it is not intended to serve as a complete overview, which would be beyond the scope of this thesis. Wormbook may be used for an even more detailed review of the topic[18, 38, 43, 44, 112].

Appendix B.1 Generating Novel Mutants

The traditional, and most common method, of generating new strains of *C. elegans* is via forward genetics[18]. This consists of random mutagenesis, followed by identification of novel phenotypes and isolation of the mutation responsible. The first step is most commonly carried out by temporary immersion of L4 individuals in a solution of ethyl methylsulfonate (EMS), a powerful carcinogen that frequently induces replacement of G:C nucleotides[148], but a variety of other methods exist that achieve a more uniformly random set of mutations, including frameshifts[149-151]. F2 individuals with notable differences in the phenotype of interest are identified, and their progeny screened for persistence of the phenotype. Finally, these genotypes are repeatedly mated with the original parent strain (usually N2), selecting for progeny with the desired phenotype, a process called outcrossing. This is done a number of times, usually at least 7, to isolate the mutation responsible and eliminate background mutations that are non-germane. Finally, the mutant is sequenced, the mutated gene determined, and the new strain may eventually be sent to the Caenorhabditis Genetics Center *(CGC)* for provision to the rest of the community.

While this approach to generating mutants is fruitful and provides mutants of use to *C. elegans* community as a whole, it is usually unhelpful for generating mutants in a specific gene of interest. If the exact mutant desired is not already available, a more targeted approach may be used, involving a zinc-finger nuclease or CRISPR-CAS9[34-37]. Because even the most targeted approach generates off-target mutations, outcrossing and sequencing is still required.

Appendix B.2 Fluorescent Marker Insertion

One of the most useful aspects of *C. elegans* for the experimenter is its optical transparency. This enables the visualization of fluorescently-labeled landmarks within the animal without needing to cut open or otherwise physically manipulate the animal. With the use of genetically-encoded markers that can be directed by the right promoter to specific cells or features, this becomes even more valuable. As such, the successful inclusion of genetically-encoded fluorescent markers is an important aspect of *C. elegans* genetic manipulation. The design of appropriate markers, e.g. fusion proteins that combine a marker like GFP with a native protein in order to monitor the expression of the native protein, is somewhat beyond the scope of this thesis, but I devote some space here to the inclusion of these markers into *C. elegans* strains, a topic of relevance to understanding the origins of the numerous strains used in this thesis.

As mentioned in Appendix A, *C. elegans* strains which have been genetically transformed can be labeled with an abbreviation such as “Ex” or “Is”—although other abbreviations, e.g. “IR” for introgression lines, exist. The first refers to the presence of an extrachromosomal array that has been introduced by the injection of foreign DNA into the gonads of a healthy hermaphrodite. The use of highly repetitive sequences containing the gene of interest and a co-injection marker leads to nearly guaranteed recombination in the gametes, allowing for the generation of progeny that contain both. Alternatively, the inclusion of fragmented genomic DNA can create so-called “complex arrays” instead, which mitigate the strong suppression of tandem repeat expression in the germline[38].

The primary advantage of this approach to genetic transformation is its speed and efficacy—with the proper training, a laboratory technician can transform many strains in a matter of hours, given properly aged worms and the right DNA fragments on hand. It carries, however, a number of downsides. The level of expression of the injected genes and co-injection markers is extremely variable, dependent on the number of copies of the genes in an array and the number of arrays in a given animal. The transmission of these arrays via mitosis and meiosis is extremely variable, and even sibling worms from the same parent show substantially variable expression. Further, any such genes expressed are usually overexpressed—expressed well in excess of their constitutive expression, with frequently unknown phenotypic effects. It is necessary to routinely pick for individuals with high levels of the co-injection marker (justifying its inclusion) to maintain population expression, and quantitative comparisons of expression intensity cannot be made between individuals[38].

The use of the “Is” labeled indicates that the genetic transformation has been “integrated” into the genome, forming an integrated strain. A number of techniques exist to do this. Methods that begin with a pre-existing extrachromosomal strain rely on gamma ray or UV irradiation to generate random DNA strand breaks, after which DNA repair enzymes will occasionally incorporate the extrachromosomal strain into the genome. Selection for individual homozygous in the co-injection marker and isolation of the desired genotype via outcrossing follows. Other, potentially superior methods exist, including coinjection of the desired genes with oligonucleotides or bombardment of the gonads with DNA-coated gold nanoparticles[38, 41].

Compared to the extrachromosomal strains, these integrated strains carry a number of advantages, the most principle of which is stable expression. While precise control of the expression level of the transforming gene is not achieved, expression is substantially more stable in descendants of the same ancestor, and also stable for an indefinite number of generations without any special maintenance. This can allow for much more reliable quantitative comparisons between individuals, and is the reason why integrated strains are used for much of the work in this thesis. Transforming genes are still overexpressed, however, and the random nature of the gene insertion into the genome raises the possibility of undesired phenotypical effects due to effects on native genes, in the worst case by directly interrupting and mutating a native gene, which cannot be solved by outcrossing.

A final and relatively new addition to the arsenal of *C. elegans* transformation arises from the advent of efficient targeted gene-editing techniques, particularly the more convenient CRISPR/CAS9 methodologies[34-37]. Used judiciously, these can be used to very carefully insert transforming genes into carefully controlled locations, with control over the number of copies inserted and even the possibility of, for instance, placing fusion proteins directly in-line with the native protein, achieving control by the same promoter. While the potential for off-target insertions still exists, follow-up sequencing can be used to verify the location of the insertion in a given single-parent population. These techniques possess all the advantages of the usual integrated strains while substantially reducing the problem of off-target insertions and providing precise control of expression level and copy number, which was not previously possible. Strains generated in this manner represent an exciting future direction for accurate quantitative imaging, but these techniques have only matured in less than a year before the presentation of this thesis.

Appendix B.3 Combining Existing Strains when the Background Strain is the Same

A very common scenario facing the researcher is the need to hybridize specific existing loci into one new strain. In many cases, this can be done without resorting to gene-editing tools by exploiting the favorable interbreeding properties of *C. elegans*. In the simplest scenario, when the two genotypes are each confined to specific genetic loci against the same genetic background, the procedure is relatively straightforward and will be outlined below; the fundamental experimental techniques are the same as the more complex case. This assumes the two loci are on different chromosomes; two loci on the same chromosome will require chromosomal recombination rather than Mendelian genetics for mixing, requiring repeated matings and other complications[43, 44].

First, the two strains are interbred: a large number of males of one strain are placed with a small number of hermaphrodites from the other onto an agar plate seeded with a small amount of *E. coli*—a common ratio is 20 males to 3 hermaphrodites. The unbalanced gender ratio and confined conditions created by the animals seeking the small spot of food serve to ensure that as many progeny as possible are product of matings.

If, as is common, neither of the two strains has a significant number of males, the ratio of males may be increased by stressing an L4 population with a brief heat shock, typically 30 °C for 6 hours. *C. elegans* is typically cultured at 20 °C and cannot thrive at 30 °C. The developing gonads in the L4 worm undergo active meiosis to produce sperm, before later switching to ova, and the heat stress significantly increases the rate of chromosomal nondisjunction, leading to larger number of sperm with no X chromosome. These lead to male progeny in the next generation. If the number of males is still small, the ratio of males may then be further boosted by performing a self-mating using the progeny. Since the progeny of a male/hermaphrodite mating is 50% male, even a relatively balanced mating plate consisting of, for instance, 3 males and 3 hermaphrodites will eventually generate a large number of males suitable for interbreeding[112].

One common complication is that one or both strains are slow-moving, weak, unusually heat-sensitive, or otherwise unsuited to generating males for this process, due to the mutation they care. In such cases, the more robust of the two strains must provide the males.

After initial mating, the challenge then becomes to isolate only those progeny that are homozygous for both of the parent genotypes. This is of course only possible starting in the F2 generation. The most general, worst-case protocol involves moving F2 individuals onto new agar plates, one individual per plate, to found new populations. Each new population may be sequenced and evaluated for the presence of one of the two genotypes, and for the presence of the wildtype genotype. With a probability 25%, this population will show the desired genotype and no wildtype, meaning it must have had a homozygous parent and by homozygous itself. This population then has a 75% chance of containing at least some of the other genotype, and the other genotype may then be refined by repeating the same procedure: picking individuals to new plates and examining the next generation for homozygosity in the other genotype. Since this is evidently a time-consuming, probabilistic procedure bottlenecked by the number of plates established after the original F2 generation, it clearly behooves the experimentalist to establish as many of these as practicable. At least 12 is recommended, with more suggested if it is suspected that the mating might have gone poorly, e.g. if the mutations involved severely impact the efficacy of mating.

In many, or even most, cases, this long procedure may be significantly abridged by the properties of one or the other genotype involved. If either genotype has a visible phenotype, even if only visible under a microscope, this may be used to skip the sequencing steps, although sequencing the final product is still recommended. It is for this reason that certain genotypes, particularly those that were formed by gene transformation as in fluorescent strains, often include a co-injection marker, either a very bright and obvious fluorescent marker that be inspected under a benchtop fluorescent dissecting scope or a dominant gene that causes an obvious phenotype; the possibility of recombination separating the co-injection marker from the actual gene of interest is low in only one mating, and may be ruled out by final sequencing. Even better, if the genotype is only partially dominant or recessive, as most mutations are, it may be used to evaluate homozygosity in an individual without examining its progeny.

In another case, one of the genotypes is located on the X-chromosome and the other is not: by mating males of the other strain with the hermaphrodites of the strain with the X-chromosome genotype, it is guaranteed that male progeny will carry the X-chromosome genotype. These may be mated with hermaphrodites from the X-chromosome genotype to guarantee homozygous progeny, though the non- X-chromosome genotype will be relatively dilute and cannot be homozygous until the generation afterward.

Appendix B.4 Combining Existing Strains with Difference Backgrounds

Another common scenario occurs when it is necessary to integrate a genotype at a specific locus into a different target genetic background. For the purposes of this thesis, this is particularly relevant for Aim 2, when considering the problem of performing a QTL analysis using a phenotype that requires a fluorescent marker to measure, which would require the integration of a fluorescent marker into a variety of different backgrounds. It is noteworthy that this kind of integration procedure carries with it a number of downsides; in many cases, it may be superior to repeat on the target background the original procedure that generated the genotype in the first place. In the case of genetic insertion for QTL purposes, however, this is inadmissible, as no such technique is reliable enough to ensure quantitative comparability between strains, given the potential for off-target insertions, uncertainty about copy number, and other considerations[44].

Cursory thought reveals that a single mating is insufficient to perform this kind of integration, because one of the paternal chromosomes will contain the original background of the gene being integrated. Once the mating has been performed, it becomes necessary to outcross the strain into the target background, while still maintaining the gene being integrated, a nontrivial task if the phenotype of the gene cannot be easily seen.

In the simplest case, where it is possible to observe the phenotype in the heterozygote, then outcrossing may be performed by repeatedly mating males of the target background into the strain, selecting for heterozygous progeny that contain gene. This may be done until the background has probably been fully integrated (>7 matings), and then individuals may be picked onto individual plates and evaluated for homozygosity. In the other cases, when the heterozygous phenotype cannot be observed but the homozygote can, it is necessary to perform a longer protocol. The homozygotes can be found in the F2 generation after mating, and males of the target background can be used to mate with these. Because recombination can only potentially occur in the heterozygote, however, the number of necessary matings is unchanged. In the worst case, where even the homozygote cannot be easily phenotyped, it becomes further necessary to pick individuals onto their own plates and sequence some of progeny, as it is not possible to non-destructively sequence *C. elegans* individuals.

The reliance of this procedure on recombination introduces a number of downsides which should be discussed. A co-injection marker, for example, can no longer be used as a fully reliable proxy for the gene of interest, as the probability that it has become separated during recombination can no longer be neglected, and care must be taken to either sequence the strain regularly or not allow the population to bottleneck one individual. Perhaps more importantly, recombination occurs properly only among homologous regions of the chromosome. If the gene of interest is an insertion, then it cannot itself undergo recombination and is prone to causing errors in recombination in its immediately vicinity. Finally, of course, it can never be fully guaranteed, only probabilistically guaranteed, that the entire target background has truly been transferred, and any potential defects in the overall process lead to a requirement for more crossings to ensure success.

APPENDIX C

DETAILED EXPERIMENTAL PROCEDURE FOR IMAGING OF *D. MELANOGASTER* BLASTULAS, T CELL ARRAYS, AND *C. ELEGANS* SYNAPSES

The purpose of this appendix is to give brief but more detailed experimental procedures for the imaging, image processing, and testing of the *D. Melanogaster* blastula and T cell array data used in Chapter 2 for algorithm experimentation. For exact experimental procedures, the interested reader may consult Levario, *et al. "An integrated platform for large-scale data collection and precise perturbation of live Drosophila embryos"*[7] or He, Kniss, *et al., "An automated platform enabling dynamic stimuli delivery and cellular response readout for high-throughput single-cell signaling studies"*[152]. I am indebted to Dr. Thomas J. Levario and Dr. Ariel Kniss-James for the text of this Appendix, which has been adapted from Zhao *et al. "Rapid, Simple, and Versatile Quantitative Phenotyping of Fluorescent Reporters Enabled by Relative Difference Filtering and Clustering" (in Submission).*

Appendix C.1 Imaging Protocols

Appendix C.1.1 Imaging of Histone-GFP *D. melanogaster* Embryonic Nuclei

Adult flies expressing histone-GFP were allowed to mate and lay eggs on a fresh agar plate for 2.5 hours at 25°C. Embryos were collected from the agar plate, dechorionated with 2.5% sodium hypochlorite for 1 min, rinsed with deionized water, and suspended in 0.3% Triton X-100 containing phosphate buffered solution (PBST). Dechorionated *Drosophila* embryos were loaded into a previously described microfluidic array that automatically orients the embryo for directly imaging the dorsal-ventral plane from either anterior or posterior[57, 153]. Once embryos were loaded, the fluidic connections were removed and the device was mounted onto a Zeiss LSM 710 confocal microscope with a Zeiss 40x oil immersion objective. Z-slices were obtained ~80 µm from either anterior or posterior pole at one image per minute for 3 hours. Embryos were maintained at 25^{o}C throughout imaging via an environmental chamber. Single-photon imaging was achieved via 488 nm excitation source while multiphoton imaging was achieved via 920 nm excitation source. Both embryos were staged such that imaging would encompass the early events of embryogenesis that includes gastrulation and ventral furrow formation in the developing embryos.

Appendix C.1.2 Imaging of Array-loaded T cells with Calcium Dye

Jurkat T cells, from the Jurkat E6-1 human acute T cell lymphoma cell line (American Type Culture Collection), were labeled with the cytosolic calcium indicator, Fluo-3 AM, cell permeant (Life Technologies). Cells were incubated at 37°C for 40 minutes with 5 µM Fluo-3 and 0.05% w/v Pluronic F127, washed 3 times with PBS, and subsequently loaded into a previously characterized microfluidic device in RPMI without Phenol Red(34). Once cells were loaded into the device, images were acquired using a FITC filter cube (Omega XF22) with a Nikon Eclipse Ti inverted fluorescent microscope. Elements Software (Nikon) was used for time-lapse microscopy with images taken every 6 s for a total of 60 minutes while cells were stimulated with an oscillatory treatment of 100 µM H_2O_2 at a frequency of 2.78 mHz, corresponding to a period of 6 minutes.

Appendix C.2 Details of Algorithm Implementation

Image processing is done using Matlab™ R2011a software with custom code. Filtering and object removal algorithms are simple, and based on functions included with the Matlab Image Processing Toolbox (particularly the function regionprops); these are provided as their own functions. Object area is defined straightforwardly as the number of pixels within an object; object solidity is defined as area of the object divided by the area of the object's convex hull. Clustering is done using either the Matlab built-in k-means algorithm or DBSCAN, using the Matlab implementation provided by Daszykowski *et al*[87, 154].

While most of the steps for the algorithm are identical for all our experimental conditions aside from configurable parameters, the last step, cluster selection differs for the T cell Array and *C. elegans* Synapse conditions. In addition, for the *D.* Melanogaster nuclei it was also necessary to identify the center of the embryo for clustering. The details are described here:

Appendix C.2.1 *D. melanogaster* Embryonic Nuclei

For the confocal images, no clustering was used, as it was deemed unnecessary. For the multiphoton images, to identify the center of the embryo, the following was performed:

1) **Identify the center of the embryo**: the median-filtered image was thresholded to find all regions with intensity less than 10% of the maximum intensity in the image. We then evaluated the centroid of the region with the largest area (typically the middle of the embryo).

DBSCAN was then performed on the distance of the segmented objects from the filtering step from this centroid, using a neighborhood parameter of 5.

Appendix C.2.2 Array-loaded T cells

The K-means clustering parameter used was 18. In order to discard unwanted clusters, the following steps were performed:

1) **Identify the row spacing**: The 2D-Fourier transform was done on the binary image obtained from the filtering step. This was then thresholded according to:

$$log10|I| + 1 > 4.5$$

Because of the nature of the array, the Fourier transform has peaks at [0,0] and regularly at the frequency embodied by the row and column spacing of the device, with each subsequent peak being dimmer. The chosen threshold eliminates all but the first peaks (and the peak at 0). The row spacing is obtained from the vertical value of this peak.

2) **Merge Clusters that are Close**: k-means clustering with the given parameters occasionally generates multiple clusters in the same row. Sets of clusters where the cluster centers are closer than 5 pixels in the vertical direction are merged. This aids in the next step

3) **Discard Anomalous Rows of Cells between the Actual Rows**: Because of the design of the device, rows of out of focus cells often form between in-focus cells. These end up in a single cluster after the previous steps. These anomalous rows are discarded by first sorting the clusters by the vertical location of their centers, then discarding clusters that have the following properties:

$$d_{i,i+1} + d_{i,i-1} - s < 5 \ (d_{i,i+1} \geq 5, d_{i,i-1} \geq 5)$$

Where s is the spacing determined from step 1 and $d_{i,i+1}$ is the y-distance between the current (ith cluster) and the next.

Appendix C.2.3 *C. elegans* Synapses

The DBSCAN clustering parameter chosen was 4. In order to discard unwanted clusters, the following steps were performed:

1) **Discard Outlier Objects in each Cluster**: For each cluster, the interquartile range (IQR) was calculated. Objects that were more than 1.5 times the IQR below the 25th percentile or above the 75th percentile in either the x or y-direction were deleted from the cluster.

2) **Discard Clusters that Don't Look like Synaptic Domains**: Clusters that failed the following criteria were removed:

   a. More than 3 objects

   b. Horizontally oriented ($|a| > 0.3$ where a is the slope of the regression line obtained by performing a linear regression on all object centers)

   c. Fewer than 5% of objects that overlap if only the horizontal coordinate is considered

    d. Linearity (residue after linear regression <50 and $r^2 < 0.1$)

3) **Select Most Linear Cluster**
    a. Cluster with smallest residue after linear regression, if this is less than 50
    b. Cluster with smallest r^2, if this is less than 0.1
    c. Otherwise, select no cluster (segmentation failure)
4) **Merge Clusters that look like they Connect with the Chosen Cluster**: The body of the worm often obscures part of the synaptic domain, disconnecting a synaptic domain and causing it to end up in separate clusters. These are reconnected by the following procedure:
    a. The leftmost and rightmost object in each cluster is obtained
    b. A biased distance is calculated between the rightmost object of the chosen cluster and the leftmost objects of the remaining clusters, as well as vice versa. The biased distance is:
        i. $d_{bias} = \sqrt{\Delta x^2 + 4\Delta y^2}$
    c. The first cluster that is found to have a biased distance less than 150 is merged, and this procedure is repeated until no such cluster is found.
5) **Test whether Objects in Cluster Overlap Vertically**: Repeat step 2c for the main cluster. Discard if cluster fails.
6) **Discard Cluster if there are less than 10 or more than 30 synapses**
    a. This synaptic domain is known to usually have 20-25 synapses. Going far outside this range usually indicates a bad image/failed segmentation

Appendix C.3 Miscellaneous Analytic Methods

Appendix C.3.1 Evaluating Clustering Accuracy

Pre-clustering binary images were characterized manually, with each individual object within the image labeled as either an object of interest or not. With this manual characterization stored, the clustering procedures for each time of image were run for a broad range of parameters, and the output images compared with the manual characterization.

Appendix C.3.2 Extension of Welch’s t-test for Evaluation of Epistasis

In order to apply Welch’s t-test for the test of epistasis in Fig. 5F, it was necessary to evaluate both effective standard deviation and degrees of freedom the combined variable:

$$\left(wy673_{\;n} - WT_{\;n}\right) + (km2_n - WT_{\;n}) - (wy673;km2_n - WT_{\;n})$$

$$= wy673_{\;n} + km2_n - wy673;km2_n - WT_{\;n}$$

where as before $\mathrm{A}_n$is taken to be feature n of strain A. The standard deviation is readily calculated using the linearity of the variance, such that:

$$\sigma_{eff}^2 = \sigma_{wy673_{\;n}}^2 + \sigma_{km2_n}^2 + \sigma_{wy673;km2_n}^2 + \sigma_{WT_{\;n}}^2$$

The corresponding degrees of freedom can be estimated using the Welch-Satterthwaite equation:

$$\frac{\left(\frac{\sigma_{wy673\;n}^2}{N_{wy673}} + \frac{\sigma_{km2\;n}^2}{N_{km2}} + \frac{\sigma_{wy673;km2\;n}^2}{N_{wy673;km2}} + \frac{\sigma_{WT\;n}^2}{N_{WT}}\right)^2}{\frac{\sigma_{wy673\;n}^4}{N_{wy673}^2(N_{wy673}-1)} + \frac{\sigma_{km2\;n}^4}{N_{km2}^2(N_{km2}-1)} + \frac{\sigma_{wy673;km2\;n}^4}{N_{wy673;km2}^2(N_{wy673;km2}-1)} + \frac{\sigma_{WT\;n}^4}{N_{WT}^2(N_{WT}-1)}}$$

where N_A is the sample size of the data gathered for strain A.

Welch’s t-test may then be performed as usual.

REFERENCES

1. McGrath PT, Xu Y, Ailion M, Garrison JL, Butcher RA, Bargmann CI. Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes. *Nature* 2011, 477:321-325.
2. WormBase web site, http://www.wormbase.org, release WS244, date Sep 1 2014.
3. Cho Y, Zhao C, Lu H. Trends in High-throughput and Function Neuroimaging in C. elegans. *WIREs Systems Biology and Medicine (In submission)* 2016.
4. Prevedel R, Yoon YG, Hoffmann M, Pak N, Wetzstein G, Kato S, Schrodel T, Raskar R, Zimmer M, Boyden ES, et al. Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat Methods* 2014, 11:727-730.
5. Kato S, Kaplan HS, Schrödel T, Skora S, Lindsay TH, Yemini E, Lockery S, Zimmer M. Global Brain Dynamics Embed the Motor Command Sequence of Caenorhabditis elegans. *Cell* 2015:656-669.
6. Kaletta T, Hengartner MO. Finding function in novel targets: C. elegans as a model organism. *Nat Rev Drug Discov* 2006, 5:387-398.
7. Levario TJ, Zhao C, Rouse T, Shvartsman SY, Lu H. An integrated platform for large-scale data collection and precise perturbation of live Drosophila embryos. *Sci Rep* 2016, 6:21366.
8. Schrödel T, Prevedel R, Aumayr K, Zimmer M, Vaziri A. Brain-wide 3D imaging of neuronal activity in Caenorhabditis elegans with sculpted light. *Nature methods* 2013, 10:1013-1020.
9. Chase DL, Koelle MR. Biogenic amine neurotransmitters in C. elegans. *WormBook* 2007:1-15.
10. Corsi AK, Wightman B, Chalfie M. A Transparent Window into Biology: A Primer on Caenorhabditis elegans. *Genetics* 2015, 200:387-407.
11. Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol* 1977, 56:110-156.
12. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* 1983, 100:64-119.
13. Kimble J, Hirsh D. The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. *Dev Biol* 1979, 70:396-417.
14. Nguyen JP, Shipley FB, Linder AN, Plummer GS, Liu M, Setru SU, Shaevitz JW, Leifer AM. Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* 2015:33.
15. White JG, Southgate E, Thomson JN, Brenner S. The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci* 1986, 314:1-340.
16. Durbin RM. Studies on the development and organisation of the nervous system of C. elegans.”. *Ph.D thesis. University of Cambridge.* 1997.
17. Venkatachalam V, Ji N, Wang X, Clark C, Mitchell JK, Klein M, Tabone CJ, Florman J, Ji H, Greenwood J, et al. Pan-neuronal imaging in roaming Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* 2015:201507109.

18. Kutscher LM, Shaham S. Forward and reverse mutagenesis in C. elegans. *WormBook* 2014:1-26.
19. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 1989, 121:185-199.
20. Barriere A, Felix MA. Natural variation and population genetics of Caenorhabditis elegans. *WormBook* 2005:1-19.
21. Brenner S. Nature's gift to science (Nobel lecture). *Chembiochem* 2003, 4:683-687.
22. Horvitz HR. Worms, life, and death (Nobel lecture). *Chembiochem* 2003, 4:697-711.
23. Sulston JE. Caenorhabditis elegans: the cell lineage and beyond (Nobel lecture). *Chembiochem* 2003, 4:688-696.
24. Avery L, Horvitz HR. A cell that dies during wild-type C. elegans development can function as a neuron in a ced-3 mutant. *Cell* 1987, 51:1071-1078.
25. Hedgecock EM, Sulston JE, Thomson JN. Mutations affecting programmed cell deaths in the nematode Caenorhabditis elegans. *Science* 1983, 220:1277-1279.
26. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 1998, 391:806-811.
27. Fire AZ. Gene silencing by double-stranded RNA (Nobel Lecture). *Angew Chem Int Ed Engl* 2007, 46:6966-6984.
28. Mello CC. Return to the RNAi world: rethinking gene expression and evolution (Nobel Lecture). *Angew Chem Int Ed Engl* 2007, 46:6985-6994.
29. Chalfie M. GFP: lighting up life (Nobel Lecture). *Angew Chem Int Ed Engl* 2009, 48:5603-5611.
30. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. Green fluorescent protein as a marker for gene expression. *Science* 1994, 263:802-805.
31. Hobert O. Neurogenesis in the nematode Caenorhabditis elegans. *WormBook* 2010:1-24.
32. Jin Y. Synaptogenesis. *WormBook* 2005:1-11.
33. Hodgkin J. Genetic nomenclature guide. Caenorhabditis elegans. *Trends Genet* 1995:24-25.
34. Kim H, Kim JS. A guide to genome engineering with programmable nucleases. *Nat Rev Genet* 2014, 15:321-334.
35. Dickinson DJ, Ward JD, Reiner DJ, Goldstein B. Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. *Nat Meth* 2013, 10:1028-1034.
36. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013, 339:819-823.
37. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012, 337:816-821.
38. Evans TC. Transformation and microinjection. *WormBook, ed. The C. elegans Research Community* 2006.
39. Mello C, Fire A. DNA transformation. *Methods Cell Biol* 1995, 48:451-482.

40. Kelly WG, Xu S, Montgomery MK, Fire A. Distinct requirements for somatic and germline expression of a generally expressed Caernorhabditis elegans gene. *Genetics* 1997, 146:227-238.
41. Mello CC, Kramer JM, Stinchcomb D, Ambros V. Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. *EMBO J* 1991, 10:3959-3970.
42. Praitis V, Casey E, Collar D, Austin J. Creation of low-copy integrated transgenic lines in Caenorhabditis elegans. *Genetics* 2001, 157:1217-1226.
43. Fay DS. Classical genetic methods. *WormBook* 2013:1-58.
44. Fay D. Genetic mapping and manipulation: chapter 7--Making compound mutants. *WormBook* 2006:1-4.
45. de Bono M, Bargmann CI. Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans. *Cell* 1998, 94:679-689.
46. Gray JM, Karow DS, Lu H, Chang AJ, Chang JS, Ellis RE, Marletta MA, Bargmann CI. Oxygen sensation and social feeding mediated by a C. elegans guanylate cyclase homologue. *Nature* 2004, 430:317-322.
47. Nonet ML. Visualization of synaptic specializations in live C. elegans with synaptic vesicle protein-GFP fusions. *J Neurosci Methods* 1999, 89:33-40.
48. Klassen MP, Shen K. Wnt signaling positions neuromuscular connectivity by inhibiting synapse formation in C. elegans. *Cell* 2007, 130:704-716.
49. Schaefer AM, Hadwiger GD, Nonet ML. rpm-1, a conserved neuronal gene that regulates targeting and synaptogenesis in C. elegans. *Neuron* 2000, 26:345-356.
50. Zhen M, Jin Y. The liprin protein SYD-2 regulates the differentiation of presynaptic termini in C. elegans. *Nature* 1999, 401:371-375.
51. Shen K, Bargmann CI. The immunoglobulin superfamily protein SYG-1 determines the location of specific synapses in C. elegans. *Cell* 2003, 112:619-630.
52. Wu YE, Huo L, Maeder CI, Feng W, Shen K. The balance between capture and dissociation of presynaptic proteins controls the spatial distribution of synapses. *Neuron* 2013, 78:994-1011.
53. San-Miguel A, Lu H. Microfluidics as a tool for C. elegans research. *WormBook* 2013:1-19.
54. Whitesides GM. The origins and the future of microfluidics. *Nature* 2006, 442:368-373.
55. Albrecht DR, Bargmann CI. High-content behavioral analysis of Caenorhabditis elegans in precise spatiotemporal chemical environments. *Nat Methods* 2011, 8:599-605.
56. Chung K, Rivet CA, Kemp ML, Lu H. Imaging single-cell signaling dynamics with a deterministic high-density single-cell trap array. *Anal Chem* 2011, 83:7044-7052.
57. Chung K, Kim Y, Kanodia JS, Gong E, Shvartsman SY, Lu H. A microfluidic array for large-scale ordering and orientation of embryos. *Nat Methods* 2011, 8:171-176.

58. Chronis N, Zimmer M, Bargmann CI. Microfluidics for in vivo imaging of neuronal and behavioral activity in Caenorhabditis elegans. *Nat Methods* 2007, 4:727-731.

59. McCormick KE, Gaertner BE, Sottile M, Phillips PC, Lockery SR. Microfluidic devices for analysis of spatial orientation behaviors in semi-restrained Caenorhabditis elegans. *PLoS One* 2011, 6:e25710.

60. Zhang Y, Lu H, Bargmann CI. Pathogenic bacteria induce aversive olfactory learning in Caenorhabditis elegans. *Nature* 2005, 438:179-184.

61. Duffy DC, McDonald JC, Schueller OJ, Whitesides GM. Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane). *Anal Chem* 1998, 70:4974-4984.

62. San-Miguel A, Crane MM, Kurshan P, Shen K, Lu H. Unsupervised Identification of Subtle Synaptic Pattern Phenotype C. elegans Mutants and their Characterization by Whole-Population Phenotypic Profiling. (Accepted). *Nature: Communications* 2016.

63. Chung K, Crane MM, Lu H. Automated on-chip rapid microscopy, phenotyping and sorting of C. elegans. *Nat Methods* 2008, 5:637-643.

64. Crane MM, Chung K, Stirman J, Lu H. Microfluidics-enabled phenotyping, imaging, and screening of multicellular organisms. *Lab Chip* 2010, 10:1509-1517.

65. Larsch J, Ventimiglia D, Bargmann CI, Albrecht DR. High-throughput imaging of neuronal activity in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America* 2013, 110:E4266-4273.

66. Albrecht DR, Bargmann CI. High-content behavioral analysis of Caenorhabditis elegans in precise spatiotemporal chemical environments. *Nature methods* 2011, 8:599-605.

67. Stirman JN, Brauner M, Gottschalk A, Lu H. High-throughput study of synaptic transmission at the neuromuscular junction enabled by optogenetics and microfluidics. *J Neurosci Methods* 2010, 191:90-93.

68. Ben-Yakar A, Chronis N, Lu H. Microfluidics for the analysis of behavior, nerve regeneration, and neural cell biology in C. elegans. *Curr Opin Neurobiol* 2009, 19:561-567.

69. Caceres Ide C, Valmas N, Hilliard MA, Lu H. Laterally orienting C. elegans using geometry at microscale for high-throughput visual screens in neurodegeneration and neuronal development studies. *PLoS One* 2012, 7:e35037.

70. Kim E, Sun L, Gabel CV, Fang-Yen C. Long-term imaging of Caenorhabditis elegans using nanoparticle-mediated immobilization. *PLoS One* 2013, 8:e53419.

71. Chokshi TV, Ben-Yakar A, Chronis N. CO2 and compressive immobilization of C. elegans on-chip. *Lab Chip* 2009, 9:151-157.

72. Nakai J, Ohkura M, Imoto K. A high signal-to-noise Ca(2+) probe composed of a single green fluorescent protein. *Nat Biotechnol* 2001, 19:137-141.

73. Miyawaki A, Llopis J, Heim R, McCaffery JM, Adams JA, Ikura M, Tsien RY. Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature* 1997, 388:882-887.

74. Crane MM, Stirman JN, Ou CY, Kurshan PT, Rehg JM, Shen K, Lu H. Autonomous screening of C. elegans identifies genes implicated in synaptogenesis. *Nat Methods* 2012, 9:977-980.

75. Yang J, Chen Z, Yang F, Wang S, Hou F. A microfluidic device for rapid screening of chemotaxis-defective Caenorhabditis elegans mutants. *Biomed Microdevices* 2013, 15:211-220.

76. Hendricks M, Ha H, Maffey N, Zhang Y. Compartmentalized calcium dynamics in a C. elegans interneuron encode head movement. *Nature* 2012, 487:99-103.

77. Schafer WR. Neurophysiological methods in C. elegans: an introduction. *WormBook* 2006:1-4.

78. Kerr RA. Imaging the activity of neurons and muscles. *WormBook* 2006:1-13.

79. Oreopoulos J, Berman R, Browne M. Spinning-disk confocal microscopy: present technology and future trends. *Methods Cell Biol* 2014, 123:153-175.

80. Tokunaga T, Hirose O, Kawaguchi S, Toyoshima Y, Teramoto T, Ikebata H, Kuge S, Ishihara T, Iino Y, Yoshida R. Automated detection and tracking of many cells by using 4D live-cell imaging data. *Bioinformatics* 2014, 30:i43-51.

81. Ben Arous J, Tanizawa Y, Rabinowitch I, Chatenay D, Schafer WR. Automated imaging of neuronal activity in freely behaving Caenorhabditis elegans. *Journal of neuroscience methods* 2010, 187:229-234.

82. Greenwald IS, Sternberg PW, Horvitz HR. The lin-12 locus specifies cell fates in Caenorhabditis elegans. *Cell* 1983, 34:435-444.

83. Sundaram MV. RTK/Ras/MAPK signaling. *WormBook* 2006:1-19.

84. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995, 20:273-297.

85. Kniss A. Computational and Microfluidic Platforms to Investigate the Role of Ca2+ and ROS in T Cell Activation with Single-Cell Resolution. Ph.D Thesis. *Georgia Institute of Technology* 2016.

86. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1979, 28:100-108.

87. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96 Proceedings* 1996:226-231.

88. Hamerly G, Elkan C. Learning the K in K-Means. *Neural Information Processing Systems* 2003.

89. Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 2015, 16:275-284.

90. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008, 9:356-369.

91. Consortium CeS. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* 1998, 282:2012-2018.

92. Ayyadevara S, Ayyadevara R, Hou S, Thaden JJ, Shmookler Reis RJ. Genetic mapping of quantitative trait loci governing longevity of Caenorhabditis elegans in recombinant-inbred progeny of a Bergerac-BO x RC301 interstrain cross. *Genetics* 2001, 157:655-666.

93. Shook DR, Johnson TE. Quantitative Trait Loci Affecting Survival and Fertility-Related Traits in Caenorhabditis elegans Show Genotype-Environment Interactions, Pleiotropy and Epistasis. *Genetics* 1999, 153:1233-1243.

94. Vertino A, Ayyadevara S, Thaden JJ, Shmookler Reis RJ. A narrow quantitative trait locus in C. elegans coordinately affects longevity, thermotolerance, and resistance to paraquat. *Front Genet* 2011, 2:63.

95. Doroszuk A, Snoek LB, Fradin E, Riksen J, Kammenga J. A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans. *Nucleic Acids Res* 2009, 37:e110.

96. Gaertner BE, Phillips PC. Caenorhabditis elegans as a platform for molecular quantitative genetics and the systems biology of natural variation. *Genet Res (Camb)* 2010, 92:331-348.

97. Shinoda Y, Sadakata T, Furuichi T. Animal models of autism spectrum disorder (ASD): a synaptic-level approach to autistic-like behavior in mice. *Exp Anim* 2013, 62:71-78.

98. Penzes P, Buonanno A, Passafaro M, Sala C, Sweet RA. Developmental vulnerability of synapses and circuits associated with neuropsychiatric disorders. *J Neurochem* 2013, 126:165-182.

99. Chen J, Yu S, Fu Y, Li X. Synaptic proteins and receptors defects in autism spectrum disorders. *Front Cell Neurosci* 2014, 8:276.

100. Grant SG. Synaptopathies: diseases of the synaptome. *Curr Opin Neurobiol* 2012, 22:522-529.

101. Kleinman JE, Law AJ, Lipska BK, Hyde TM, Ellis JK, Harrison PJ, Weinberger DR. Genetic neuropathology of schizophrenia: new approaches to an old question and new uses for postmortem human brains. *Biol Psychiatry* 2011, 69:140-145.

102. Pollard DA. Design and construction of recombinant inbred lines. *Methods Mol Biol* 2012, 871:31-39.

103. Crow JF. Haldane, Bailey, Taylor and Recombinant-Inbred Lines. *Genetics* 2007, 176:729-732.

104. Rockman MV, Kruglyak L. Recombinational landscape and population genomics of Caenorhabditis elegans. *PLoS Genet* 2009, 5:e1000419.

105. Andersen EC, Shimko TC, Crissman JR, Ghosh R, Bloom JS, Seidel HS, Gerke JP, Kruglyak L. A Powerful New Quantitative Genetics Platform, Combining Caenorhabditis elegans High-Throughput Fitness Assays with a Large Collection of Recombinant Strains. *G3 (Bethesda)* 2015, 5:911-920.

106. Rockman MV, Kruglyak L. Breeding designs for recombinant inbred advanced intercross lines. *Genetics* 2008, 179:1069-1078.

107. Li H, Ribaut J-M, Li Z, Wang J. Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theoretical and Applied Genetics* 2008, 116:243-260.

108. Li H, Ye G, Wang J. A modified algorithm for the improvement of composite interval mapping. *Genetics* 2007, 175:361-374.

109. McGrath PT, Rockman MV, Zimmer M, Jang H, Macosko EZ, Kruglyak L, Bargmann CI. Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors. *Neuron* 2009, 61:692-699.

110. Glater EE, Rockman MV, Bargmann CI. Multigenic natural variation underlies Caenorhabditis elegans olfactory preference for the bacterial pathogen Serratia marcescens. *G3 (Bethesda)* 2014, 4:265-276.

111. Seidel HS, Ailion M, Li J, van Oudenaarden A, Rockman MV, Kruglyak L. A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in C. elegans. *PLoS Biol* 2011, 9:e1001115.
112. Fay D. Genetic mapping and manipulation: chapter 1--Introduction and basics. *WormBook* 2006:1-12.
113. Stiernagle T. Maintenance of C. elegans. *WormBook* 2006:1-11.
114. Muschiol D, Schroeder F, Traunspurger W. Life cycle and population growth rate of Caenorhabditis elegans studied by a new method. *BMC Ecology* 2009, 9:1-13.
115. Cand EJ, #232, Li X, Ma Y, Wright J. Robust principal component analysis? *J. ACM* 2011, 58:1-37.
116. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 2003, 19:889-890.
117. Vinuela A, Snoek LB, Riksen JA, Kammenga JE. Genome-wide gene expression regulation as a function of genotype and age in C. elegans. *Genome Res* 2010, 20:929-937.
118. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JA, Hazendonk E, Prins P, Plasterk RH, Jansen RC, et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet* 2006, 2:e222.
119. Kato S, Kaplan HS, Schrodel T, Skora S, Lindsay TH, Yemini E, Lockery S, Zimmer M. Global brain dynamics embed the motor command sequence of Caenorhabditis elegans. *Cell* 2015, 163:656-669.
120. Richmond JE. Electrophysiological recordings from the neuromuscular junction of C. elegans. *WormBook* 2006:1-8.
121. Chalfie M, Hart AC, Rankin CH, Goodman MB. Assaying mechanosensation. *WormBook* 2014.
122. Goodman MB, Hall DH, Avery L, Lockery SR. Active currents regulate sensitivity and dynamic range in C. elegans neurons. *Neuron* 1998, 20:763-772.
123. Cook A, Franks CJ, Holden-Dye L. Electrophysiological recordings from the pharynx. *WormBook* 2006:1-7.
124. Miyawaki A, Griesbeck O, Heim R, Tsien RY. Dynamic and quantitative Ca2+ measurements using improved cameleons. *Proc Natl Acad Sci U S A* 1999, 96:2135-2140.
125. Chen TW, Wardill TJ, Sun Y, Pulver SR, Renninger SL, Baohan A, Schreiter ER, Kerr RA, Orger MB, Jayaraman V, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* 2013, 499:295-300.
126. Faumont S, Lindsay TH, Lockery SR. Neuronal microcircuits for decision making in C. elegans. *Curr Opin Neurobiol* 2012, 22:580-591.
127. Schafer WR. Egg-laying. *WormBook* 2005:1-7.
128. Zhen M, Samuel AD. C. elegans locomotion: small circuits, complex functions. *Curr Opin Neurobiol* 2015, 33:117-126.
129. Shipley FB, Clark CM, Alkema MJ, Leifer AM. Simultaneous optogenetic manipulation and calcium imaging in freely moving C. elegans. *Front Neural Circuits* 2014, 8:28.
130. Husson SJ, Gottschalk A, Leifer AM. Optogenetic manipulation of neural activity in C. elegans: from synapse to circuits and behaviour. *Biol Cell* 2013, 105:235-250.

131. Liu J, Ward A, Gao J, Dong Y, Nishio N, Inada H, Kang L, Yu Y, Ma D, Xu T, et al. C. elegans phototransduction requires a G protein-dependent cGMP pathway and a taste receptor homolog. *Nat Neurosci* 2010, 13:715-722.
132. Basca CA, Talos M, Brad R. Randomized Hough Transform for Ellipse Detection with Result Clustering. In: *EUROCON 2005 - The International Conference on "Computer as a Tool"*; 2005.
133. Yonghong X, Qiang J. A new efficient ellipse detection method. In: *Pattern Recognition, 2002. Proceedings. 16th International Conference on*; 2002.
134. Kuhn HW. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 1955, 2:83-97.
135. Bertsekas DP. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of Operations Research* 1988, 14:105-123.
136. Myronenko A, Song X. Point Set Registration: Coherent Point Drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2010, 32:2262-2275.
137. Besl PJ, McKay HD. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1992, 14:239-256.
138. Fitzgibbon AW. Robust registration of 2D and 3D point sets. *Image and Vision Computing* 2003, 21:1145-1153.
139. Gold S, Rangarajan A, Lu C-P, Pappu S, Mjolsness E. New algorithms for 2D and 3D point matching: pose estimation and correspondence. *Pattern Recognition* 1998, 31:1019-1031.
140. Tsin Y, Kanade T. A Correlation-Based Approach to Robust Point Set Registration. In: Pajdla T, Matas J, eds. *Computer Vision - ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004, 558-569.
141. Chui H, Rangarajan A. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding* 2003, 89:114-141.
142. Gale D, Shapley LS. College Admissions and the Stability of Marriage. *The American Mathematical Monthly* 1962, 69:9-15.
143. Chen T, Martin E, Montague G. Robust probabilistic PCA with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis* 2009, 53:3706-3716.
144. Pope G, Baumann M, Studer C, Durisi G. Real-time principal component pursuit. In: *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*; 2011.
145. Rodriguez P, Wohlberg B. Incremental Principal Component Pursuit for Video Background Modeling. *J. Math. Imaging Vis.* 2016, 55:1-18.
146. Chartrand R. Numerical Differentiation of Noisy, Nonsmooth Data. *ISRN Applied Mathematics* 2011, 2011:11.
147. Toyoshima Y, Tokunaga T, Hirose O, Kanamori M, Teramoto T, Jang MS, Kuge S, Ishihara T, Yoshida R, Iino Y. Accurate Automatic Detection of Densely Distributed Cell Nuclei in 3D Space. *PLoS Comput Biol* 2016, 12:e1004970.
148. Brenner S. The genetics of Caenorhabditis elegans. *Genetics* 1974, 77.
149. Boulin T, Bessereau JL. Mos1-mediated insertional mutagenesis in Caenorhabditis elegans. *Nat Protoc* 2007, 2:1276-1287.

150. Stewart HI, Rosenbluth RE, Baillie DL. Most ultraviolet irradiation induced mutations in the nematode Caenorhabditis elegans are chromosomal rearrangements. *Mutat Res* 1991, 249:37-54.

151. Martin E, Laloux H, Couette G, Alvarez T, Bessou C, Hauser O, Sookhareea S, Labouesse M, Segalat L. Identification of 1088 new transposon insertions of Caenorhabditis elegans: a pilot study toward large-scale screens. *Genetics* 2002, 162:521-524.

152. He L, Kniss A, San-Miguel A, Rouse T, Kemp ML, Lu H. An automated programmable platform enabling multiplex dynamic stimuli delivery and cellular response monitoring for high-throughput suspension single-cell signaling studies. *Lab Chip* 2015, 15:1497-1507.

153. Levario TJ, Zhan M, Lim B, Shvartsman SY, Lu H. Microfluidic trap array for massively parallel imaging of Drosophila embryos. *Nat Protoc* 2013, 8:721-736.

154. Daszykowski M, Walczak B, Massart DL. Looking for natural patterns in analytical data. 2. Tracing local density with OPTICS. *J Chem Inf Comput Sci* 2002, 42:500-507.

VITA

CHARLES L. ZHAO

ZHAO was born in Stillwater, Oklahoma. He attended public schools in Portales, New Mexico and Fremont, California, received a B.S. in Bioengineering from University of California, Berkeley in 2009 and a M.S. in Translational Medicine (and Bioengineering) also from University of California, Berkeley in 2011, before coming to Georgia Tech to pursue a doctorate in Biomedical Engineering. When not working on his research, he writes creative fiction, plays and mods video games, runs tabletop games, and blogs about all three of those, and also occasionally anime.