Mapping the Essential Genes in the Pericentromeric Heterochromatin of the Second Chromosome's Left Arm in *Drosophila melanogaster*

Richard Yuditskiy

A Thesis Submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science

Graduate Program in Biology

York University

Toronto, Ontario

April 2022

Abstract

Heterochromatic regions perform essential roles in model organisms such as *Drosophila melanogaster*, yet not all have been genetically mapped. The current project focused on identifying and mapping the essential heterochromatic genes found in the lethal region of the left arm of the second chromosome (2Lh) in *D. melanogaster*, constructing a genetic map in the process. This was achieved by finding fatal mutations in *D. melanogaster* lines via the candidate gene approach, which compared genomic DNA sequences of potential mutant candidates with the gene's control line sequence. To indicate their importance, the mutations found were discussed and their effects on the genes are explained. Six out of the seven essential genes have been located: one which was previously found (*light)*, and five in this project: *CG42748*, *Ir40a*, *CG17493*, *Slmap*, and *Cht10*. The seventh and final gene, *l(2)40Fe*, has been deemed unfindable and is explained more thoroughly later in this paper.

Acknowledgments

First and foremost, I wish to thank you Arthur J. Hilliker for providing me with the wonderful opportunity of working in your laboratory. It has been a tremendous experience and an absolute privilege completing first an undergraduate and now a master's thesis in the fascinating world of Drosophila genetics. From the very beginning of our meeting, you were extremely welcoming and helpful to me, an approach that only grew with time. It has been truly an honor, especially since you have built the very foundations that made my and many others' works possible. I have gained a remarkable amount of knowledge and skills from you in these past three years, most of which I would never have found in any other place. The freedom you have given me when I was conducting my project can not be stressed enough. This freedom has given me motivation and the space required to keep pushing forward while keeping a healthy mindset throughout. Of course, I can not forget to acknowledge all the materials and lab spaces you provided me to make this project possible. I wish you nothing but the very best in terms of your health, luck, and happiness. Thank you Dr. Hilliker!

I wish to thank Alistair B. Coulthard for his unwavering willingness to help and mentor me throughout this project. Quite simply this would not have been possible whatsoever without the extensive and wonderful work done by you. Your expertise in the topic of Drosophila genetics and mapping is truly at an elite level and my current work is in many ways a continuation of your ground-breaking achievements. But more than your knowledge on these topics is your amazing character. Whenever I was in doubt, confused, or straight up lost, I could always count on you to clear everything up, even to the smallest details if necessary. I appreciate all the discussions we have had over these past years and hope I provided some help as well. I

wish you nothing but the very best in terms of your health, luck, and happiness. And may you find *chaser*!

I wish to also thank Alex Molnar and Harjot Deol for aiding me with the various workings of the laboratory and experimental procedures. Please know that I greatly appreciate your comradeships and incredible qualities which made me learn a lot in these years. Alex, beyond the help you provided me I thank you for your conversations, be they science related or not, for it had made the time pass cheerfully. Harjot, thank you for taking me under your wing and showing me the proper ways of experimenting when I was but a ridiculously inexperienced undergraduate student. Thank you both very much!

I would like to thank Gary Sweeney for the valuable time and insights provided to me during my master's degree. Truly, I appreciate it very much for you being part of my supervisory committee and helping me to become a better graduate student in numerous aspects. Thank you!

Thanks is also due to the countless wonderful people who have provided their own bit of help for this project and beyond it. Among them are Vladimir (Kyle) E. Belozerov, Spencer T. Mukai, Patricia Lakin-Thomas, and Joel S. Shore. Thank you all!

Finally, I want to thank my family who have provided me with so much unconditional love, warmth, and support throughout my life. Each one of you have and continue to push me to become a better person today than what I was yesterday. You have taught me place value in hard work and perseverance, and even more than that you showed me how to be a good person. I owe everything to you and this work is as much yours as it is mine. Thank you all so very much, I could not have asked for a better family!

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Chapter One: Introduction

Heterochromatin's Growing Significance

The recognition of differently packaged chromosomes is attributed to Emil Heitz back in 1928 while working on the model organism *Marchantia polymorpha*, the common liverwort. Heitz discovered that the nucleus of the cell contains a type of chromatin which condenses during mitosis before decondensing once the cell enters interphase, and another type of chromatin which spends the entire cell cycle condensed (Heitz 1928). These he named euchromatin and heterochromatin, respectively (Heitz 1928). In a later paper (Heitz 1929); Heitz noted the differentially condensed nature of certain heterochromatin segments after the cell undergoes telophase. He suggested that heterochromatin is more commonly found in the chromosome where the concentration of genes is either low or present in a passive state (Heitz 1929). These descriptions of heterochromatin by Heitz remain to this day the backbone of its current definition. Heterochromatin has always had an air of mystery about it, but over the past two centuries research has brought much light onto this topic to reveal its importance for genetics.

Later, in the year 1966, Spencer W. Brown separated heterochromatin into two distinct classes: facultative and constitutive. Facultative heterochromatin is dispersed throughout the length of the chromosome and, under certain conditions, can be induced to function like euchromatin (Brown 1966). Constitutive heterochromatin, on the other hand, is located mainly in the centromeric and telomeric regions of a chromosome and is always heterochromatic in appearance (Brown 1966). Facultative heterochromatin was thought to be akin to constitutive heterochromatin in the sense of being transcriptionally silent, yet it also has the significant ability to switch between heterochromatin and euchromatin (Brown 1966).

It was found that facultative heterochromatin plays a key role in regulating gene expression in cells, particularly, turning active when gene stimulation was needed for development and deactivating again once the gene's role was complete (Trojer and Reinberg 2007). The reason behind that facultative heterochromatin decondenses varies and is attributed to one of the following: temporal states such as cell-cycle stages; spatial states such as nuclear localization changes within the cell; or heritable reasons such as monoallelic gene expressions (Trojer and Reinberg 2007). Meanwhile, the stimuli that is responsible for changing facultative heterochromatin into euchromatin, and vice versa, falls under one of the following: addition of chromatin components; chromatin modulation; trans-acting factors; and subnuclear localization (Trojer and Reinberg 2007). These characteristics make facultative heterochromatin an extremely important component for all living organisms, but the focus of this paper will be on constitutive heterochromatin which will henceforth be referred to simply as "heterochromatin".

Another common feature of heterochromatic nature are the many simple highly repetitive sequences of satellite DNA (Lohe *et al.* 1993), which are the bane of sequencers worldwide. Sequencing is already an extremely difficult and detail-oriented task, but with the presence of excessive repeats the computational alignment and assembly steps lead to an increased number of errors presenting themselves during analysis (Treangen and Salzberg 2011). Such repeats even present challenges for recent, next-generation, sequencing technologies such as FINDER (Banerjee *et al.* 2021), which automatically annotates eukaryotic genes and transcript structures. FINDER is programmed to label all genes located in repeat regions as "low-confidence genes" due to the diverse problems they can lead to during analysis (Banerjee *et al.* 2021). One may even be so inclined as to ignore such repeat regions altogether. This however would come with a slew of problems since very important biological processes and interconnectedness of the

genome would be missing from the results (Treangen and Salzberg 2011). So, because both types of heterochromatins have this highly repeatable nature it is very difficult to sequence them accurately. It is therefore crucial to run duplicates of the important sequences if such repeat regions are unavoidable.

Position Effect Variegation

Work accomplished before Brown's distinction proved that constitutive heterochromatin silences euchromatic genes, as was shown by position effect variegation (PEV) of the *white* gene (Muller 1930). Under normal circumstances the *white* gene of *Drosophila melanogaster* is expressed in every single cell of the adult fruit fly, producing their distinctive red-eye colour phenotype (Muller 1930). However, a certain mutation called the *white-mottled-4* (w^{m4}) inversion, described by Hermann J. Muller in 1930, transforms the colour of the affected eye cell white. This was a truly peculiar finding which yearned to be further explored. The best-known example showing PEV in constitutive heterochromatin was Spofford's 1976 experiment. Spofford moved the *white* gene of *D. melanogaster* away from its native euchromatic position, near the telomere of the X chromosome, to a heterochromatic position (Spofford 1976). This gene translocation resulted in a heterochromatic inactivation of certain cells of the eyes, resulting in eyes with mosaic patterning of white and wildtype red colours (Spofford 1976). Interpretation of heterochromatin suppressing a euchromatic gene provided false support for the initial belief of the inactive nature of heterochromatin.

Other examples of the variegation effect were also observed when heterochromatic genes were moved from their native heterochromatic environment to a euchromatic one. The genes *rolled* (*rl*) and *light* (*lt*) are two heterochromatic genes for which the most amount of evidence

exists for such a phenomenon (Eberl *et al.* 1993; Howe *et al.* 1995). When the experimenters migrated these heterochromatic genes to euchromatic regions, the PEV effect mentioned above was observed. Both papers noticed that when moved to euchromatin, larger heterochromatic blocks with the gene resulted in lower levels of PEV while smaller blocks with the gene showed higher levels of PEV (Eberl *et al.* 1993; Howe *et al.* 1995). Also, there are several groups of PEV-modifying elements that affect euchromatin and heterochromatin very differently, and therefore leads to varying PEV effects in both chromatin regions (Locke *et al.* 1988). Therefore, both types of chromatins are highly dependent on the surroundings in which they find themselves in; losing at least some of its inherent function when present in an alien environment (Locke *et al.* 1988). As a result, instead of the long-standing belief of heterochromatin functioning as a silencer of euchromatic genes, they both negatively affect the genes found in the other chromatin type. It insinuates then that the nature of euchromatin and heterochromatin are not so wildly different as previously perceived.

Drosophila melanogaster Heterochromatin

Heterochromatin makes up a significant portion within many organisms' genomes, composing approximately 20% of the human genome, while being 34% of female and 46% of male *D. melanogaster* genomes (reviewed by Hoskins *et al.* 2007 and references therein; Adams *et al.* 2000). It comes as no surprise therefore that such an abundance in genomes led to the research of heterochromatin in multiple different organisms. The most thoroughly examined of these organisms is *D. melanogaster*, where several genes in its heterochromatic regions have been shown to be essential for its survival. One of the earliest discoveries of just such an essential heterochromatic gene was the previously stated *lt*. This gene was first identified by Jack Schultz back in 1936 who, when experimenting with thirteen different *D. melanogaster* mutants,

speculated the location of *lt* "...within the inert region of 2L" (Schultz 1936). Ever since then *lt* has grown to be one of the most extensively studied heterochromatic genes, revealing that it is responsible for a myriad of processes in *D. melanogaster*.

A major function *lt* is responsible for is its role in forming two vital complexes: the homotypic fusion and vacuolar protein sorting (HOPS) and class C core vacuole/endosome tethering (CORVET) complexes (Takáts *et al.* 2014). In *D. melanogaster*, the HOPS complex interacts with Syntaxin 17 to lead to maturation and clearance of autophagosomes, a vital part of any organism's immune system (Takáts *et al.* 2014). Also, when the HOPS complex was rendered dysfunctional by Takáts *et al.*, starvation-induced autophagy was prevented from working, indicating the importance of *lt* for cellular response during starvation. The CORVET complex has similarly important functions for endosome and lysosome fusion, in particular the maturation of early to late endosomes (Balderhaar and Ungermann 2013). These important functions were discovered many years after *lt* was located by Schultz in 1936, and the only reason it was revealed in the first place was due to a revolutionary breakthrough in heterochromatin a few decades ago.

Finding Essential Genes with Mutagenesis

In 1976, using the compound ethyl methanesulfonate (EMS), Arthur J. Hilliker induced mutagenesis in the second chromosome's heterochromatin of *D. melanogaster* (Hilliker 1976). With the goal of revealing and mapping the loci (genes) in this region, Hilliker used the heterochromatic deletions previously created by Hilliker and Holm (Hilliker and Holm 1975; Hilliker 1976). Hilliker used EMS to produce 118 recessive lethal and visible mutations which revealed five complementation groups in the right arm of chromosome 2 (2R) heterochromatin,

and four complementation groups in the left arm of chromosome 2 (2L) heterochromatin (Hilliker 1976). He noted the lack of evidence of deficiency in any of the EMS lethals, and so concluded that, although at a much lower density (about 1%) than euchromatin, heterochromatin indeed contains essential genetic loci (Hilliker 1976). Such a minute level of genes present in heterochromatin probably accounted for the earlier view of a genetically inert heterochromatin. It is no surprise therefore that previous researchers believed that the heterochromatin was a region devoid of genetic activity (reviewed in Yunis and Yasmineh 1971 and references therein). Hilliker's 1976 findings became extremely revolutionary for the field. The results put to question the long-standing view in the scientific community of labelling heterochromatin as a genetically insignificant portion of an organism, and therefore showcasing its importance.

Building directly upon Hilliker's previous work on the heterochromatic region in the 2R region of *D. melanogaster* were the two papers published by the Hilliker lab (Coulthard *et al.* 2003; Coulthard *et al.* 2010). The work done in 2003 presented a then up-to-date map of the second chromosome's centromeric heterochromatin (Coulthard *et al.* 2003). It also included analysis of the 21 genetic loci based on the cumulative research done on the topic in years past, along with their own unpublished findings (Coulthard *et al.* 2003). The team had located a brand new male sterile locus called *eunuch* and inferred the location of a new essential locus which they labelled *EMS 742*, adding it to the 19 previously found loci when forming the map (Coulthard *et al.* 2003). Overall, the map included the following elements: 16 essential loci, two parts of the Segregation Distorter system, a male sterile locus, a female sterile locus, and a *Minute* locus (Coulthard *et al.* 2003). The interallelic complementation map was constructed by following the system created by Hilliker, which separated the loci into nine deficiency

complementary groups, with the 2R heterochromatin loci assigned groups I-V and the 2L heterochromatin loci groups VI-IX (Hilliker 1976).

As the complementation map was created the researchers of the 2003 paper acknowledged that most of the essential loci or male/female sterility loci were found, even though other published papers suggested that they have found additional essential loci. Coulthard *et al*. in 2003 mentioned one paper (Dimitri *et al*. 1997) which found six potentially new vital loci. Coulthard and colleagues did not include these loci in their complementation map. They explained this choice by stating that the interallelic complementation complexity of alleles within even one locus would require multiple alleles to be used for a convincing allelism test, whereas the 1997 paper tested only one allele per locus (Coulthard *et al.* 2003; Dimitri *et al.* 1997). In the concluding remarks of the 2003 paper the researchers stressed the importance of running wider-ranging complementation analyses, which would ensure that any potentially unique essential loci are verified to be as such (Coulthard *et al.* 2003). These words were a foreshadowing, for seven years later another paper was published by Coulthard *et al*. which sought to resolve this exact dilemma.

Mapping and Identifying Essential Genes

The push behind the 2010 paper came from the disagreement in the number of essential loci found within the genetic region where *Df(2R)41A8* and *Df(2R)41A10* overlapped (Coulthard *et al.* 2010). Hilliker noted four essential loci within this region in his 1976 paper, however a newer mutagenesis screen conducted in 2004 by Myster *et al.* observed 15 essential loci while also using EMS as mutagen, which was too great of a difference to ignore. The reasoning behind this disagreement was because Hilliker based his research on previously found knowledge on EMS

(Lim and Snyder 1974) which stated that the mutagen causes mostly point mutations. This was different from the conclusion derived at by Myster *et al.*, because they considered EMS to cause not only smaller point mutations but also large abnormalities, like deletions, due to the sensitivity of the heterochromatin to EMS (Myster *et al.* 2004). As a result, this led them to interpret the complex interallelic complementation found by Hilliker in the *l(2)41Ae* locus as multiple genes, instead of one locus, due to large deletions present in the region (Myster *et al.* 2004). This dispute was resolved in the 2010 paper by Coulthard and colleagues.

To solve this dispute Coulthard *et al.* (2010) mapped *l(2)41Ae* and its surrounding regions. They have done this by cataloguing as many essential loci in the region as they could with a candidate gene method using the at the time most recent 2Rh annotations (Hoskins *et al.* 2007; Tweedie *et al.* 2009) and doing genetic and molecular mapping analyses (Coulthard *et al.* 2010). The map was constructed by assembling all mutant lines that were directly mapped to *l(2)41Ae* in previous research, then running them through an extensive *inter se* complementation analysis (Coulthard *et al.* 2010). Since *l(2)41Ae* was pinpointed beforehand to the h46 cytological band (Dimitri 1991; Coulthard *et al.* 2003; Dimitri *et al.* 2009), Coulthard and colleagues focused their primary attention on the region of genes from *CG42595* to *Nipped-B* found in the h45-h46 cytological bands (Coulthard *et al.* 2010). The researchers made sure that the newly created complementation map was comprehensive enough and so used 50 mutant lines relating to the genes under study to reveal four novel complementation groups, namely: *CG17665*, *CG40127*, *CG17683*, and *CG42595* (*unextended*) (Coulthard *et al.* 2010). These went along with two other complementation groups located in past research, *Rpl38* by Marygold *et al.* in 2005 and *Nipped-B* by Gause *et al.* in 2008, to bring the grand total of complementation groups in the map to six (Coulthard *et al.* 2010).

Mutation Types and Sequencing

Running sequences of candidate genes with the mutated lines of interest provided the answers when identifying the complementation groups, and hence the construction of the map. By comparing the gene sequences between the chosen mutant line and its control line counterpart the potential differences found between them would be of great interest to the researchers. These changes are mostly a result of different types of point mutations, some of which have no effect on the gene product whatsoever, and as such are referred to as silent mutations (Strachan and Read 1999). This occurs when a nucleotide base undergoes a mutation yet the amino acid it later produces remains the same due to the degenerate nature of the genetic code. It is possible because all 20 amino acids (except for tryptophan) can be formed by a few different types of codon combinations, and so even though a mutation is present it does not affect the final protein result (Strachan and Read 1999). However, certain mutations could be quite profound for the gene under study and may even lead to the fatality of the organism.

Among such dramatic sequence alterations are nonsense mutations, which replaces an amino acid for a premature stop codon (Strachan and Read 1999). The result is oftentimes devastating to the final protein product, which would become a truncated and dysfunctional version of its normal self (Strachan and Read 1999). Another form of point mutation is a missense mutation which turns the original amino acid into another, and as a result changes the structure and nature of the protein (Strachan and Read 1999). If a significant portion of the population under study contains such a missense mutation it is referred to as a polymorphism, which is crucial to assure that the sequencing results are consistent with the established wildtype sequences (Strachan and Read 1999). Yet, if a missense mutation that is present only in a small portion of the population (less than 1%) is found then it is deemed significant enough for further

study (Strachan and Read 1999). The resulting missense mutation could very well change the protein structure, and hence its function, perhaps to such a high degree that would lead to the organism's fatality (Strachan and Read 1999). The nonsense and missense mutations, along with deletions, are what Coulthard and his colleagues were searching for in 2010 to pinpoint the various locations of lethality in their genes of interest. All the identified mutations in the paper were clearly summarized to show the nucleotide change, amino acid change, and the location of the mutation to compare with the established gene polymorphisms (Coulthard *et al.* 2010).

Summary and Importance of Coulthard *et al.* 2010 Paper

To identify the four complementation groups found in the paper initial sequencing work focused on finding all complementation groups within the newly identified *Df(2R)34-2* heterochromatic deletion (Coulthard *et al.* 2010). Here the team begun with the *CG17665* gene due to both the *Df(2R)34-2* and *Df(2R)41A2* deletions breaking within the gene's boundaries and being lethal with the *EMS 45-34* complementation group (Coulthard *et al.* 2010). The group found a mutation in the third exon of the *CG17665* gene when sequenced with line *EMS 45-34* which turned a guanine nucleotide into adenine, and thereby changed a tryptophan into a premature stop codon (Coulthard *et al.* 2010). Four additional significant mutations were found by the groups when they used other lines in their sequencing of *CG17665*. Two of these were stop codons being formed from the wildtype arginine and tryptophan when analyzing lines *EMS 45-61* and *NC19,* respectively (Coulthard *et al.* 2010). Results of line *EMS 45-71* showed a glycine mutate into a glutamic acid, and line *NC28* mutating a leucine into a proline (Coulthard *et al.* 2010).

Next, the group sequenced gene *CG17683* with three lines and uncovered a significant mutation in each (Coulthard *et al.* 2010). Line *NC38* resulted in a premature stop codon from the

wildtype cysteine amino acid, line *EMS 34-13* contained a serine to phenylalanine change, while line *NC109* had two adjacent mutations which changed a valine into aspartic acid and a serine into alanine (Coulthard *et al.* 2010). The third gene was *CG42595* where the mutant line *L2* led the sequence to form a premature stop codon instead of the wildtype glutamine (Coulthard *et al.* 2010). When tested with line *NC1* the *CG42595* gene contained two small deletions, 13-bp and 4-bp, resulting in a frameshift and loss of a splice junction motif (Coulthard *et al.* 2010). Finally, gene *CG40127* had one mutation found in line *NC110* which mutated a leucine into a histidine (Coulthard *et al.* 2010).

The group from the 2010 paper also mapped two loci that were identified in the past: *unextended* (*uex*) and *l(2)41Ab*. Homozygous $uex^{34\text{-}7}$ DNA was tested with gene *CG42595* to reveal a methionine into isoleucine mutation, which meant that with the extensive genetic analysis conducted by the researchers on the *uex* gene it pointed to the two genes being interchangeable (Coulthard *et al.* 2010). In regards to the *l(2)41Ab* gene, the researchers used previous research (Coulthard *et al.* 2003) and the latest heterochromatin annotations at that time (Hoskins *et al.* 2007; Tweedie *et al.* 2009) to sequence all candidate genes. Sequencing of the predicted genes with line *EMS 45-10* failed to find any mutation, yet the polymorphisms used to validate the sequencing arrived back inverted (Coulthard *et al.* 2010). The researchers then tried to sequence the inverse position from the original candidate genes, which succeeded in finding a tryptophan to stop codon mutation within exon seven of the gene *CG41265* (Coulthard *et al.* 2010). Hence, with all the evidence provided the researchers concluded that the *l(2)41Ae* was not a single complex locus, but indeed a region of numerous loci which included: *Nipped-B*, *RpL38*, *CG17665*, *CG40127*, *CG17683*, and *CG42595* (*uex*).

Besides the above-mentioned mapping of essential loci that used transposable element mutagenesis and EMS, Coulthard and colleagues tried a new method for identification of vital genes in the heterochromatin of 2R: RNAi knock-down. The group tested the new RNAi method on 12 heterochromatic genes in the 2R region, some of which were used as controls because they have already been shown to be lethal in other studies, namely: *Nipped-A*, *Nipped-B*, *gus*, and *RpL38* (Rollins *et al.* 1999; Styhler *et al.* 2002; Marygold *et al.* 2005). These four genes were confirmed to be lethal using the RNAi knock-down method and so provided enough confidence of its merit to identify five additional lethal or semi-lethal genes that have never been identified as such previously (Coulthard *et al.* 2010). The other remaining three genes resulted in viable progeny when affected by RNAi knock-down, which the group summarized along with the above findings (Coulthard *et al.* 2010).

Therefore, once the researchers combined the results from their genetic analyses, sequencing, and RNAi knock-down they ended up with eight distinct vital loci (Coulthard *et al.* 2010). Two of these, *l(2)41Ab* and *uex*, were previously reported to be essential and in the paper were found to be synonymous with genes *CG41265* and *CG42595*, respectively (Coulthard *et al.* 2010). The other six essential loci found in this paper through the three methods mentioned were novel and included: *CG17684*, *CG17683*, *CG40127*, *CG17665*, *CG17883*, and *Atf6* (Coulthard *et al.* 2010). In the years following this success in identifying vital loci numerous other essential genes have been found. Some of these vital genes are presented in the current paper, with a focus on the ones found in the h35 cytological band of *D. melanogaster*'s second chromosome.

Chapter Two: Objectives of the Current Project

Identifying and mapping the essential heterochromatic genes in the lethal 2Lh region of *D. melanogaster* is the goal of this project. Up to now only one of the vital genes has been identified: *lt* as *l(2)40Fb*. So, using the candidate genes and mutant strains for the lethal 2Lh region, I will link the other essential genes with their respective location on the chromosome. Finding the entire set of these essential genes, *l(2)40Fa* to *l(2)40Fg*, would help create a map of the lethal 2Lh region and ease any future studies involving it. This project shares many similarities with the 2010 paper by Coulthard *et al.*, which identified genes in the heterochromatin of the right arm of the same chromosome. Therefore, this project is a humble continuation of that work, in the hopes that eventually the heterochromatin of chromosome 2 in *D. melanogaster* would be mapped genetically and molecularly.

Chapter Three: Materials and Methods

Fly Genetics

Fly culture: All fly stocks and crosses were kept on a 12-hour light-dark cycle at 25°C and fed a fly food recipe taken from the Hilliker lab. To create a 1 L volume of such fly food firstly a mixture was made in a 1 L Pyrex® Erlenmeyer flask containing the following ingredients: 800 mL water, 100 g sucrose, 22 g agar, 1 g potassium phosphate (KH_2PO_4) monobasic, 8 g potassium sodium (NaK) tartrate tetrahydrate, 0.5 g sodium chloride (NaCl), 0.5 g calcium chloride ($CaCl_2$) dihydrate, 0.5 g magnesium chloride ($MgCl_2$) hexahydrate, 0.5 g iron (III) sulphate ($Fe_2(SO_4)_3$). This mixture was heated to 80°C to melt the agar and then was left to cool in the fume hood. While the first mixture was cooling, a second mixture was made containing 200 mL water and 50 g of dry yeast (*Saccharomyces cerevisiae*). This mixture was heated until boiling to kill the yeast and once properly cooled for safe handling, combined with the first mixture. Lastly, once the temperature of the combined mixture reached 60°C or lower, 5 mL of propionic acid was added and thoroughly mixed. The final solution was then poured into either Drosophila mating bottles or vials, plugged with Styrofoam or cotton wool, and finally placed in 4°C storage for later use.

Fly strains: The chosen wildtype control *D. melanogaster* flies, namely *rosy +5* (ry^{+5}), and all mutant lines used for the experiments (**Table 1**) were taken from Dr. Hilliker's lab fly stocks located in York University, Ontario, Canada. Eight of the lines were created by Hilliker (1976), who produced them by inducing EMS mutations on an isogenic for chromosome 2 strain which was homozygous for the *Pin* mutation, called the *iso-2* strain. The other lines used, except ry^{+5}, were created by Sharp (1988) who induced EMS mutations in a *cinnabar* (*cn*) *brown* (*bw*) strain.

Table 1. The *D. melanogaster* mutant lines experimented upon in this project. All the flies were taken from Dr. Hilliker's fly stocks located in York University, Ontario, Canada. These mutant lines were created by either Hilliker (1976) or Sharp (1988), with EMS as the mutagen of choice to induce the mutations. The ry^{+5} wildtype flies were used as the control population.

Mutant Line	**Paper Reference**
Df(2L)PR31	Hilliker 1976
EMS 40-2	Hilliker 1976
EMS 40-5	Hilliker 1976
EMS 40-7	Hilliker 1976
EMS 40-18	Hilliker 1976
EMS 56-4	Hilliker 1976
EMS 56-24	Hilliker 1976
EMS 56-8	Hilliker 1976
532	Sharp 1988
534	Sharp 1988
540	Sharp 1988
591	Sharp 1988
621	Sharp 1988
652	Sharp 1988
PR12-5	Sharp 1988
I1	Sharp 1988

Chromosome balancing: All mutant lines used were balanced by Alistair Coulthard with the *CyO EGFP* balancer chromosome. This was achieved by individually crossing each mutant line to the *Df(2L)PR31/CyO EGFP* deletion which is deficient of all known essential genes in the heterochromatin of the second chromosome's left arm (Hilliker and Holm 1975; Hilliker 1975).

RNAi analyses: The RNAi knock-down work done by Alistair Coulthard provided a candidate list of the essential 2Lh genes, which was used during the sequencing stages of the paper.

Preliminary Work Involving the Lethal 2Lh Region

Mapping the essential genes in the lethal 2Lh region in *D. melanogaster* has been the ultimate goal of the current project. A growing number of genes have been identified in this region ever since it was first analyzed. The most recent gene sequence annotations reveal that currently there are 37 heterochromatic protein-coding genes in the 2Lh centromeric region (Hoskins *et al.* 2007; Larkin *et al.* 2011). These genes range from the 2L centromere to the 2L heterochromatin-euchromatin border, which was determined by Riddle *et al.* (2011) to be at the 22,001,009 base pair. This was done by observing the differences of histone markers between euchromatin and heterochromatin, which Riddle *et al.* then used to identify the borders for all the chromosomes in *D. melanogaster* (2011). The 37 protein-coding genes, shown in **Figure 1**, are therefore a list of all potential essential genes found in 2Lh, and so were the chosen candidates for the current work.

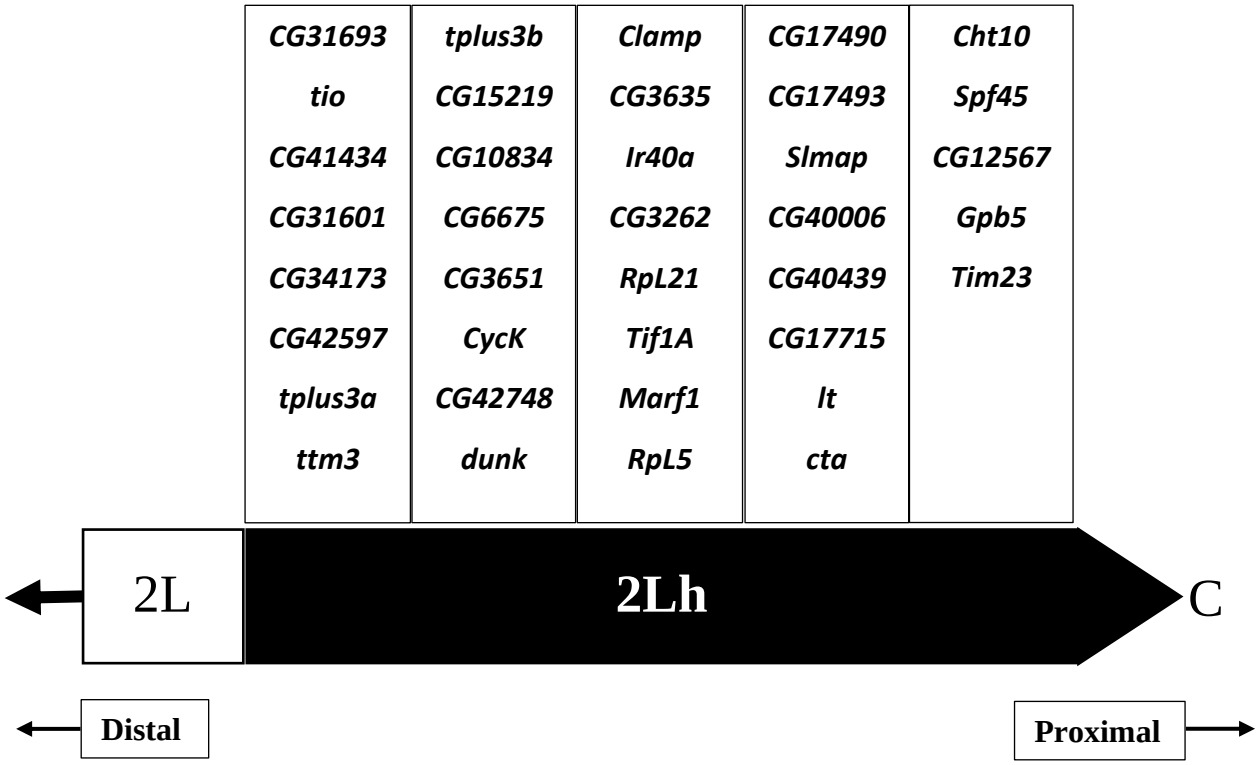


Figure 1. Map of the pericentromeric heterochromatic genes found in the left arm of the second chromosome in *D. melanogaster*. This map is a segment of the latest heterochromatic gene sequence annotations available (Hoskins *et al.* 2007; Larkin *et al.* 2021). The left arm of chromosome 2 is represented by the pentagon, within which the dark-shaded area is the 2L heterochromatic region situated between the light-shaded 2L euchromatic region and the centromere designated by the "C". List of genes within the boxes are all the protein-coding genes found between the 2L heterochromatin-euchromatin border found at the 22,001,009 base pair, as determined by Riddle *et al.* (2011), and the 2L centromere. More distal genes are found towards the top and left while more proximal genes are found towards the bottom and right, with the most distal one being *CG31693* and the most proximal being *Tim23*.

Procedures Prior to Polymerase Chain Reaction (PCR)

Embryo plating and collection: For each mutant line, approximately 50 male and 50 female adult *D. melanogaster* flies were placed together into a Drosophila mating bottle which contained air holes. The bottle was sealed with a tissue culture dish lid measuring 35 mm by 100 mm, which contained a modified agar solution. To create a 50 mL solution of this solution firstly 1.25 g of sucrose, 12.5 mL of grape juice, 1.125 g of agar, and 37.5 mL of water were added in a 250 mL Pyrex® Erlenmeyer flask. The solution was then heated to a boil with constant stirring and then removed from the heating block to cool to below 60°C. After this, 0.075 g of methyl 4-hydroxybenzoate, an anti-fungal agent also known as Tegosept or Nipagin provided by Sigma-Aldrich, was added and the combined solution was stirred well. The final mixture was then poured into culture dish lids and placed in a 4°C storage to solidify. To stimulate egg-laying, a small amount of active yeast was added on top of the collection plate, and then the flies were placed inside to lay eggs. After 16 hours all adult flies were removed from the bottle and the embryos on the collection plate were left to mature for an additional 2 hours.

Fly genomic DNA extraction and isolation: The gDNA of fly embryos was extracted by following the "Single Embryo PCR" protocol set by the Honda lab and modified by Christina Alm and Iulia Cealic (see Coulthard *et al.* 2010). The embryos were at first dechorinated, by rolling them on Scotch tape, and then each digested using 13 µL of Embryo Lysis Buffer and 2 µL of proteinase K (20 mg/mL). Following the crude gDNA extraction, the proteinase K inside the homogenate was activated and then deactivated. This was done using the MJ Research PTC-200 Peltier Thermal Cycler by incubating the homogenate first for 29 minutes at 39°C, and then for 10 minutes at 95°C.

For most fly lines the gDNA had to be extracted from embryos due to the flies failing to reach their adult phase. The only exceptions occurred for lines *540, 652,* and *I1*, all three of which shared the same genetic background and were created by Sharp (1988). The approximate 2% of surviving unbalanced adults in these three mutant lines allowed gDNA to be extracted and used as hemizygotes for the best *Cht10* gene candidate. In the case of the adults, each individual adult fly was digested using 46 µL of Embryo Lysis Buffer (10 mM Tris pH 8.2, 1 mM EDTA, 25 mm NaCl) and 4µL of 20 mg/mL of Proteinase K (provided by BioShop®). The proteinase K was activated and then deactivated following the same procedures as for the embryos.

Primer design: All primer pairs used, shown in **Appendix A**, were created using the 0.4.0 version of the "primer3" program (Untergasser *et al.* 2012; Koressaar and Remm 2007), which Bio Basic Inc. then synthesized. Each primer had the recommended amount of autoclaved water added to the primer pellet to form a 10 µM aqueous solution of the primer. The primers were created to have melting temperatures in the range of 60°C to 64°C, and to produce DNA products of approximately 750-bp in length.

Running PCR and *EGFP*-Diagnostic

PCR protocol: PCR reactions were mixed by following the recommended concentrations for the 2x MyTaq HS Mix provided by FroggaBio Inc (which contained Taq polymerase, 10x Taq DNA PCR reaction buffer, and dNTPs). Nuclease-free (NF) water was added in differing amounts depending on the number of primer pairs inside the reaction mix (see **Appendix B**). To avoid premature activation, the gDNA was the last component to be added. The PCR was run using the MJ Research PTC-200 Peltier Thermal Cycler. The program used when running the PCR started with a 4-minute incubation at 95°C, followed with 36 cycles of standard temperature control for

PCR amplification (95°C for 30 seconds, then 62°C for 30 seconds, then 72°C for 4 minutes), and a final 10-minute incubation for primer extension at 72°C.

EGFP-*diagnostic PCR:* Using the "Single Embryo PCR" protocol (see Coulthard *et al.* 2010), mutant embryos were run through *EGFP*-diagnostics to locate the ones which lacked the *EGFP* balancer chromosome, and therefore were homozygous for the mutation of interest. At the start, two primer pairs were combined with nuclease-free (NF) water, 2x MyTaq Polymerase Mix, and the mutant's gDNA as shown in **Appendix B**. This mixture was then inserted into the Thermal Cycler and the PCR protocol mentioned above was applied to get an amplification of the DNA product. Following this, the PCR products were separated by running an agarose gel electrophoresis. A 1.5% agarose gel was used here, composed of: 0.6 g agarose powder, 800 µL 50x Tris-acetate-EDTA (TAE) buffer, 39.2 mL reverse osmosis water (created using the Thermo Scientific™ Barnstead™ Easypure™ II UV/UF Ultrapure Water System), and 4 µL of iTaq Universal SYBR Green mix (provided by Bio-Rad Laboratories Ltd.). When loading the gel, each well contained 5 µL of PCR gDNA and 1 µL of 6x loading dye (0.25% bromophenol blue, 0.25% xylene cyanol FF, 30% glycerol) provided by New England BioLabs Inc. To obtain optimal separation the gel was run for 30 minutes at 75 V while immersed in 1x TAE buffer. The gel was then exposed to UV light using a Sigma Chemical Company™ T1201 Transilluminator and its image captured using a cellular phone camera.

Only homozygous embryos were used for sequencing, which were determined by utilizing the copper-zinc superoxide dismutase 1 (*SOD1*) and enhanced green fluorescent protein (*EGFP*) genes. *SOD1* is an abundant gene which creates a protein that protects the organism from oxidative damage by neutralizing superoxide radicals within the body (Pardo *et al.* 1995). Due to its abundance in beings such as *D. melanogaster* it can be was used as a positive control

for the presence of a DNA template. Meanwhile, the *EGFP* gene would only show in embryos that have been specifically bred to contain the *EGFP* balancer chromosome. Embryos that are homozygous for a mutation under study do not have this balancer, and so they likewise do not show the *EGFP* band on the gel. As a result, wells displaying two product bands in a gel indicated that the embryo being tested produced both the *SOD1* and *EGFP* product, and so is heterozygous for the mutation. On the other hand, if a well showed only one band in the gel it indicated that the embryo is missing the *EGFP* balancer and therefore the embryo is homozygous for the mutation. **Figure 2** below is one of the many examples of just such a diagnostic run in the current project, in this case testing line *EMS 40-5*.

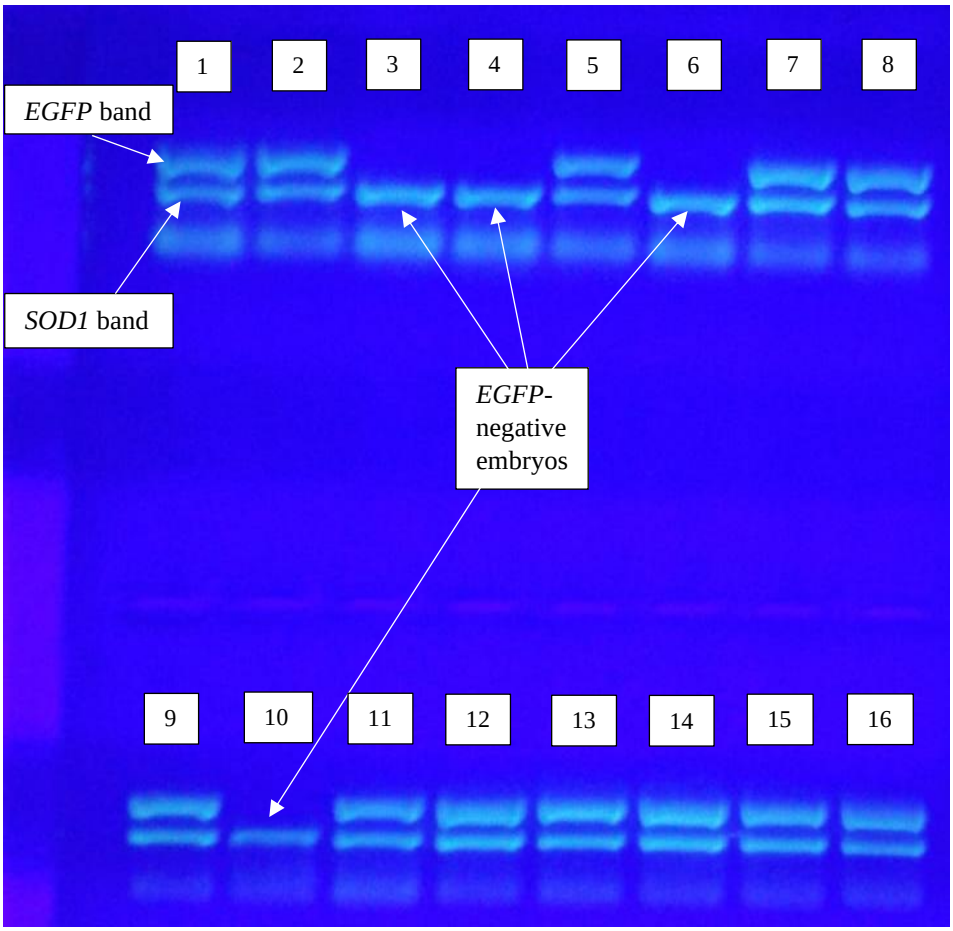


Figure 2. An *EGFP*-diagnostic testing homozygosity in embryos from the mutant line *EMS 40-5*. This test is run to determine which embryo samples are homozygous. *SOD1* primers (product size of 350-bp) act as a positive control for detecting the presence of template DNA, and the *EGFP* primers (product size of 683-bp) detect the presence of the *EGFP*-containing balancer. Homozygous embryos do not contain the *EGFP* balancer, therefore no *EGFP* band will show up for those embryos. In this particular gel there are a total of four homozygous embryos, found in wells 3, 4, 6, and 10. The bands seen in each well below the *SOD1* bands represents primer dimers and are irrelevant to the experiment.

Sequencing

Mutant DNA PCR: The DNA of homozygous mutant embryos was used to proceed to the sequencing step. Applying a similar PCR protocol as for the *EGFP*-diagnostic, the mutant DNA was mixed with the rest of the PCR components as mentioned in **Appendix B**. The one exception here was using exclusively the primer pair for the heterochromatic gene that was currently analyzed, instead of the *SOD1* and *EGFP* primer pairs.

Purification and sequencing of mutant DNA: After running through the PCR incubation program, the mutant DNA was purified following the recommended procedure by the QIAquick® PCR Purification Kit. Once purified, the DNA fragments and their respective primer aliquots were sent to Bio Basic Inc., who used Sanger sequencing for all DNA sequencings. Sequences which showed mutations of potential importance were run multiple times to confirm the finding. Genes in regions unproblematic for sequencing, such as exons and non-repetitive DNA, were duplicated two times. Sequences of genes in highly problematic regions, such as introns and repetitive DNA, were duplicated at least three times.

Analyzing Mutant Sequences

Comparing mutant sequences to FlyBase: With sequencing of the mutant genes finished and obtained, their exon sequences were compared with the latest release of the finished genomic sequence found at the gene annotation website FlyBase (www.flybase.org) (Larkin *et al.* 2021). The FlyBase sequences were generated by the enormous efforts of Hoskins *et al.*, who used whole-genome shotgun sequence (WGS3) to sequence and map the heterochromatin of *D. melanogaster* (2007). The researchers' used a strain generated by Brizuela *et al.* in 1994 which was isogenic for all four chromosomes, called the *iso-1* strain. The X chromosome of this strain

carried the *yellow* (*y*) mutation, while its chromosome 2 carried the *cinnabar* (*cn*), *brown* (*bw*), and *speck* (*sp*) mutations (Brizuela *et al.* 1994).

Comparisons between the FlyBase reference sequences and the obtained mutant sequences were then achieved by using the "EMBOSS water sequence alignment" program, which highlighted any nucleotide differences between the two sequences (Madeira *et al.* 2019). To best evade sequencing errors, only polymorphisms and nucleotide changes that showed up in the main body (about 90 nucleotides after a primer's start, but less than 90 nucleotides before a primer's end) of both forward and reverse primer sequences were noted as legitimate changes. How these changes affected the translated protein of the gene was determined by making use of the "ExPASy DNA to protein tool" (Duvaud *et al.* 2021). The two DNA sequences were first transcribed into their mRNA and then translated into their respective proteins. The differences between them would then be clear after running the "EMBOSS water sequence alignment" program for proteins (Madeira *et al.* 2019).

Detecting polymorphisms: Each of the sequences from the mutant lines of the candidate genes were searched through in hopes of uncovering potential significant mutations. However, an important note to remember is the presence of polymorphisms scattered throughout the many *D. melanogaster* lines. Such changes in the genome's sequence are considered background mutations since they are present in most of the fly population with the same genetic background. For example, the lines created by Hilliker in 1976 have a different genetic background (*iso-2 Pin*) compared to those used to produce the FlyBase sequences (*iso-1*), which in turn would also differ from Sharp's 1988 lines (*cn bw*). It is important therefore to compare the mutations found using lines of the same genetic background, as doing this would eliminate any irrelevant mutations from the analyses conducted later.

When sequencing the 2L heterochromatic genes with the lines indicated in this paper an enormous number of polymorphisms, seen in **Table 2** below, were revealed. The sheer quantity of these polymorphisms caused considerable trouble when attempting to locate unique mutations. Therefore, it was doubly important to identify such polymorphisms in the sequences of multiple fly lines from the same genetic background. Doing this greatly increased the confidence of the results and made certain the successful identification of polymorphisms and unique mutations.

Analyzing the predicted protein structure: A mutation can result in significant change to the protein it translates into. Locating a mutation is the first step of understanding the mutant studied, but the location of said mutation is equally important. The predicted protein structure and function helps indicate whether the mutation found is at a significant region of the protein. FlyBase annotations and NCBI's Conserved Domain Database (CDD) were to analyze the proteins of the genes under study to understand how significant the located mutations were (Marchler-Bauer *et al.* 2017; Larkin *et al.* 2021). Mutations present in important active domains would have a much more prominent effect on the protein structure and function than mutations in relatively trivial regions.

Table 2. Summary of all detected polymorphisms during sequencing. Lines where the polymorphisms were found and the lines where they were absent are both indicated. Data is based on the predicted translation start site, where +1 is the adenine of the first translation start codon. The location of the mutations is written as the number of amino acids downstream (+), or upstream (-), of this start site. The published sequences on FlyBase were used as a reference for all polymorphisms. The genes, ordered from distal to proximal, were chosen thanks to Alistair Coulthard's RNAi knock-down work, which supplied a list of essential 2Lh gene candidates.

Gene	**Polymorphism Type Found**	**Location of Polymorphism**	**Lines with Polymorphism**	**Lines without Polymorphism**
CG41434	C to T	+101 (exon 1)	*EMS 40-7, EMS 56-24, 540, 621*	
	C to T	+674 (exon 1)	*EMS 40-7, 621*	*EMS 56-24, 540*
	G to A	+792 (exon 1)	*EMS 40-7, 540, 621*	*EMS 56-24*
	3' UTR deletion	+1256 to +1267	*EMS 40-7, EMS 56-24, 621*	*540*
	3' UTR duplication	+1315 (16 nucleotides)	*EMS 40-7, EMS 56-24, 621*	*540*
	G to C	3' UTR	*EMS 40-7, EMS 56-24, 621*	*540*
CG31601	A to T	+68 (exon 1)	*EMS 40-7, EMS 56-4, 532, 534, 621, PR12-5*	*EMS 56-24*
	T to G	+71 (exon 1)	*EMS 40-7, EMS 56-4, 532, 534, 621, PR12-5*	*EMS 56-24*
	C to A	+91 (exon 1)	*EMS 40-7, EMS 56-4, 532, 534, 621, PR12-5*	*EMS 56-24*

Table 2 (Continued).

Gene	Polymorphism Type Found	Location of Polymorphism	Lines with Polymorphism	Lines without Polymorphism
CG31601 (Continued)	T to A	+425 (exon 5)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	
	A to T	+519 (exon 5)	*EMS 40-7, EMS 56-4, 532, 534, 621, PR12-5*	*EMS 56-24*
	G to A	+710 (exon 5)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	
	A to G	+809 (exon 5)	*EMS 40-7, EMS 56-4, 532, 534, 621, PR12-5*	*EMS 56-24*
	T to C	+1009 (exon 5)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	
	T to A	+1517 (exon 7)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	
	T to C	+1584 (exon 7)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	
	A to C	+1631 (exon 7)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	
	T to C	+2161 (exon 7)	*EMS 40-7, EMS 56-4, EMS 56-24, 532, 534, 621, PR12-5*	

Table 2 (Continued).

Gene	Polymorphism Type Found	Location of Polymorphism	Lines with Polymorphism	Lines without Polymorphism
CG42597	A to G	+129 (exon 1)	*EMS 40-7, EMS 56-24, 540, 621*	
	T to C	+204 (exon 1)	*EMS 40-7, EMS 56-24, 540, 621*	
	A to C	+341 (exon 1)	*EMS 40-7, EMS 56-24, 540, 621*	
	C to T	+354 (exon 1)	*EMS 40-7, 621*	*540, EMS 56-24*
	G to T	+486 (exon 1)	*EMS 40-7, EMS 56-24, 540, 621*	
ttm3	T to G	+54 (exon 1)	*532, PR12-5*	*EMS 56-4, EMS 56-24, 591*
CG6675	C to A	+261 (exon 2)	*EMS 56-4, 621*	*EMS 56-24*
	A to T	+1101 (exon 2)	*EMS 56-4, EMS 56-24, 621*	
CG42748	5’ UTR insertion	-103 (30 nucleotide)	*EMS 56-4, 534, 621*	*EMS 40-7*
	G to T	-101 (5’ UTR)	*EMS 56-4, 534, 621*	*EMS 40-7*
	T to C	-80 (5’ UTR)	*EMS 56-4, 534, 621*	*EMS 40-7*
	C to T	-31 (5’ UTR)	*EMS 56-4, 534, 621*	*EMS 40-7*
	G to A	+369 (exon 3)	*EMS 40-7, 534, 621*	*EMS 56-4*
	A to T	+682 (exon 3)	*EMS 40-7, 534, 621*	*EMS 56-4*
	T to C	+1953 (exon 3)	*EMS 40-7, 534, 621*	*EMS 56-4*
	A to G	+3506 (exon 4)	*EMS 40-7, EMS 56-4, 534, 621*	
	A to G	+4705 (exon 5)	*EMS 40-7, 534, 621*	*EMS 56-4*

Table 2 (Continued).

Gene	Polymorphism Type Found	Location of Polymorphism	Lines with Polymorphism	Lines without Polymorphism
Ir40a	G to T	+529 (exon 5)	*EMS 56-4, 532*	*EMS 56-24, 591*
	A to G	+1021 (exon 6)	*EMS 56-4, 532*	*EMS 56-24, 591*
Tif-IA	T to C	+485 (exon 2)	*EMS 56-4, 532*	*EMS 56-24, 591*
RpL5	G to T	+303 (exon 3)	*EMS 56-4, 591*	*EMS 56-24, 532*
	G to A	+324 (exon 3)	*EMS 56-4, EMS 56-24, 532*	*591*
Slmap	G to A	+301 (exon 1)	*EMS 40-2, EMS 40-18*	*EMS 56-18*
	T to C	+735 (exon 1)	*EMS 40-2, EMS 40-18, EMS 56-18*	
	A to C	+1956 (exon 3)	*EMS 40-2, EMS 40-18, EMS 56-18*	
	T to C	+2112 (exon 3)	*EMS 40-2, EMS 56-18*	*EMS 40-18*
Cht10	T to C	+644 (exon 1)	*540, 652, I1*	
	A to C	+4737 (exon 4)	*540, 652, I1*	
	T to C	+6753 (exon 6)	*540, 652, I1*	
Tim23	5’ UTR deletion	-52 to -40	*EMS 40-7, 532, PR12-5*	*EMS 56-4, EMS 56-24, 591*

Running Quantitative Reverse-Transcriptase PCR (RT-qPCR)

Adult fly RNA extraction and purification: Placed 2 males and a female of the *ry*$^{+5}$*/Df(2L)PR31* fly line in one 1.5 mL Eppendorf tube, and 5 females of the *PR12-5/Df(2L)PR31* fly line in another tube. Added 500 µL of TRIzol™ reagent (provided by Invitrogen™) to the two tubes, crushed the flies in the regent with a pestle, and incubated for 5 minutes in room temperature. Added 100 µL of chloroform into each tube and incubated for 2 minutes in room temperature. Then, centrifuged in a Micro High Speed Refrigerated Centrifuge Model VS-15000 CFN II (produced by Vision Scientific) for 15 minutes at 4°C and 12,000 RPM, after which three layers were formed (aqueous phase, interphase, and organic phase). Extracted 200 µL of the RNA-containing aqueous phase into new Eppendorf tubes. Added 250 µL of isopropyl alcohol to each tube and mixed well, before incubating at room temperature for 10 minutes. Next, centrifuged the two tubes for 10 minutes at 4°C and 12,000 RPM to form a pellet. Discarded the supernatant, leaving only the pellet in the tube. 500 µL of 75% ethanol was added to the pellet and the tubes were vortexed moderately for 15 seconds each. The tubes were then centrifuged for 5 minutes at 4°C and 7,500 RPM. The supernatant was discarded, and the pellet left to dry in the tube for 5 minutes. The pellets were resuspended in 20 µL of RNase-free water (DEPC-treated) and each was placed in new Eppendorf tubes. Finally, the tubes were placed on a dry bath heat-block incubator (produced by Boekel Scientific) for 10 minutes at 55-60°C.

Measuring RNA concentration: Before starting, the NanoDrop™ 2000 Spectrophotometer (produced by Thermo Scientific™) was cleaned of any potential foreign materials by adding onto the pedestal 5 µL of DEPC-treated water and closing the arm gently. Next, the spectrophotometer was blanked by adding 1 µL of DEPC-treated water, and then dried with a Kimwipe. Then, 1 µL of the RNA sample was added to the pedestal and its RNA concentration

was measured. After measuring, the sample was removed with a Kimwipe, and the pedestal cleaned with 1 µL of DEPC-treated water. The second sample was then measured likewise.

RT-qPCR protocol: A total of 33 reactions were prepared to compare the RNA concentrations between *ry*$^{+5}$*/Df(2L)PR31* and *PR12-5/Df(2L)PR31* in the following transcript variations of *Ir40a*: only transcript F; only transcript G; and the common exon near the 3' end for all three E, F, G transcripts. These three *Ir40a* transcript combinations had different initiation and termination locations. This therefore tested six unique experimental reactions, each of which had five duplicates for a total of 30 reactions. Additionally, three blank tubes corresponding to the three transcript variations were tested to gauge the basal RNA levels.

The 33 total reactions were prepared in 0.1 mL strip tubes (provided by QIAGEN), into which all the necessary components for the RT-qPCR were added (see **Table C** in the **Appendix**). For every applicable reaction, the RNA template was added at the very end to prevent any possible degradation. Once the strip tubes were capped, they were placed inside a QIAGEN Rotor-Gene Q and run through a RT-qPCR program. The program begun with a reverse transcriptase phase at 50°C for 10 minutes, after which the temperature was raised to 95°C for 1 minute during the polymerase activation and DNA denaturation phase. The 35-cycle amplification phase followed this where the reactions were first incubated at 95°C for 15 seconds, then annealed and extended at 60°C for 30 seconds. As the cycles were progressing the quantitative PCR curve was constructed on the computer and saved for further analyses.

Chapter Four: Results

The bulk of analyzation in the current research occurred after the sequencing stages. As described in the **Materials and Methods** section, utilization of identical genetic backgrounds was crucial to sift through the countless insignificant polymorphisms found in 2Lh. Reference sequences found on FlyBase aided in this when aligned with the mutant DNA sequences acquired from the sequencing ran in this project. It helped locate any changes found in the genes studied between the FlyBase and mutant sequences, which were then compared to the sequences of other lines sharing the same genetic background. Mutations present in multiple lines were identified as polymorphisms and were noted down to create a reference list of them that was presented in **Table 2** above.

Finally, the few mutations which did not match any of the polymorphisms, the truly unique and significant mutations, were the ones which were further examined. The sequences with potential mutations were verified by sequencing them again and, if confirmed, were added to a list of unique mutations. **Table 3** at the end of this **Results** section provides a summary of the confirmed significant mutations found, all of which are discussed later in this paper.

Sequencing of *Slmap*

A total of two mutations in *Slmap* were found in the sequences of two mutant lines. The first mutation was found in line *EMS 40-18*, being a guanine to adenine mutation in the 848th nucleotide. This changed the 283rd amino acid from a control "GGT" into a "GAT" codon, which therefore changed a non-polar glycine (G) into a negatively charged aspartic acid (D). The second mutation was found in line *EMS 56-8*. Within the second exon of *Slmap,* a thymine to

cytosine mutation in the 1837th nucleotide led to a codon change from a control "TCT" codon to a "CCT" codon. This made the 613th amino acid convert from a polar serine (S) into a non-polar proline (P).

Figure 3 below provides a visualization on the procedures for determining such changes, in this case for gene *Slmap* with mutant *EMS 40-18*. The two DNA sequences at the top of the figure are from the same 300-bp *Slmap* region, representing the nucleotide interval of 640 to 939 in the gene's first exon. They both contain a 20-bp yellow highlighted sequence which is a primer constructed to sequence this region. The respective amino acid products, in the gene's amino acid interval of 214 to 313, are shown below the DNA sequences. FlyBase sequences were used as the references for the genes studied in this project. These were produced by Hoskins *et al.* in 2007 who used the *iso-1* strain created by Brizuela *et al.* (1994). The reference DNA sequences of a particular gene region were compared with the sequencing of the same region completed in the current project using the chosen candidate mutant line.

If a nucleotide change was present in other lines of the same genetic background, it was a clear sign of a polymorphism. **Figure 3** shows one such example, where at the 735th nucleotide a thymine changed into a cysteine, which was also found in lines *EMS 40-2* and *EMS 56-18* (as shown in **Table 2**) and had no effect on the translation of the serine. Yet, if a nucleotide mutation was unique, and the amino acid change was significant, then it was a significant and unique mutation. **Figure 3** has a prime example of that, where a guanine into an adenine mutation also affected the translation of the respective amino acid. It turned a glycine amino acid into an aspartic acid and was not found in any other line. Such a procedure was used for all potential mutations located. It was done to determine the effect that the DNA change had on the amino acid sequence, and as a result on the viability of the protein which is discussed later in this paper.

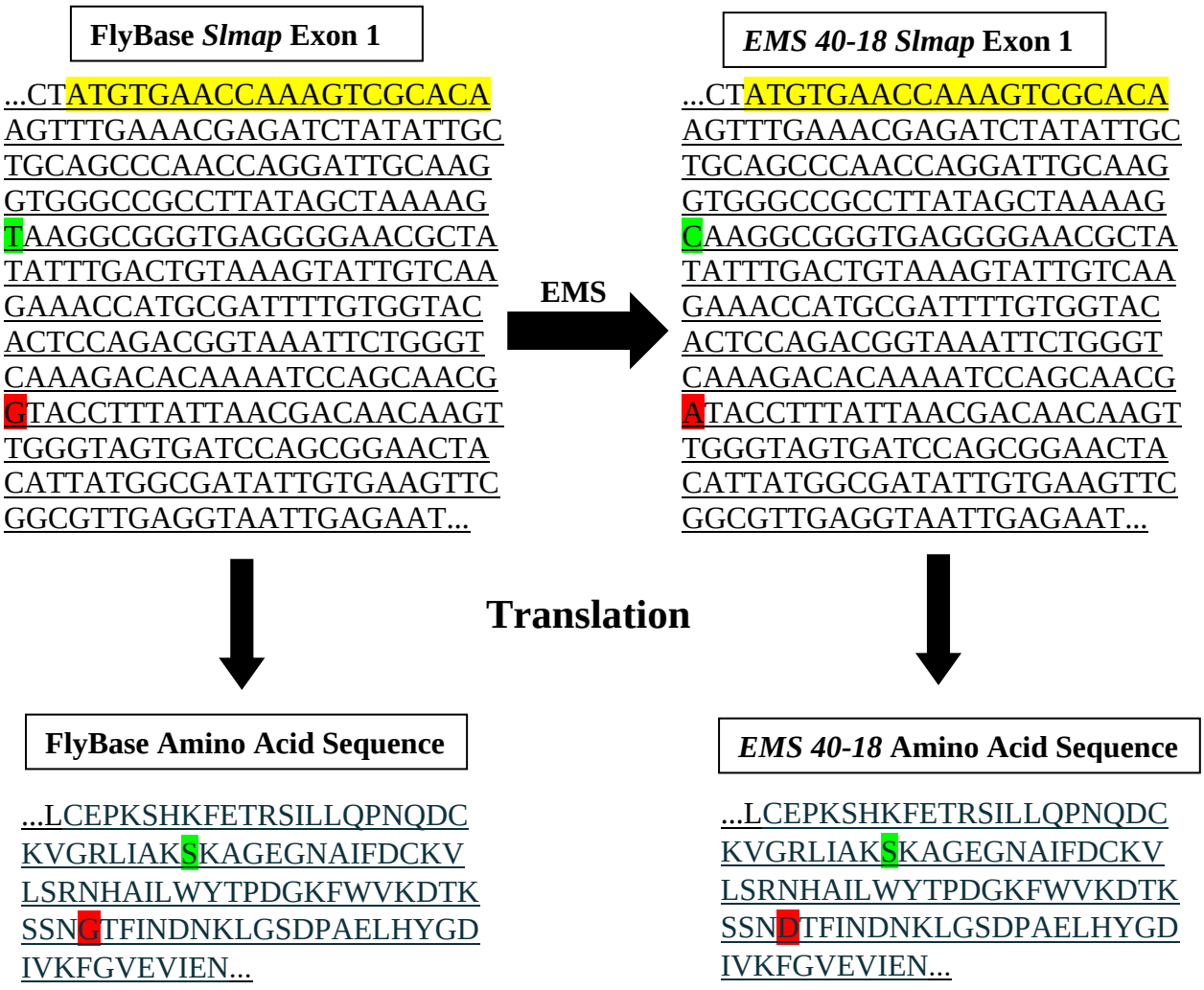


Figure 3. DNA and amino acid sequence comparisons of *Slmap* between FlyBase and *EMS 40-18*. The FlyBase DNA sequences were derived at by Hoskins *et al.*, using the *iso-1* strain (2007). Both DNA sequences represent the same 300-bp region of *Slmap*, being nucleotides 640 to 939 in the gene's first exon. Likewise, the two amino acid sequences are from the same region: amino acids 214 to 313. The 20-bp yellow highlighted sequences indicate a primer used to sequence this region. A polymorphism found here is shown by green highlighted thymine and cysteine, which did not change the translation of the serine. A unique mutation, highlighted in red, turned a guanine into an adenine, changing the glycine into aspartic acid.

Sequencing of *CG17493*

A significant mutation was found in the sole exon of *CG17493* in line *EMS 40-2*, at the 425th nucleotide. This affected the 142nd amino acid, where an adenine to guanine mutation changed a "GAA" codon found in the control sequence into a "GGA" codon. This transformed the negatively charged glutamic acid (E) found in the control to a non-polar glycine (G).

Sequencing of *Ir40a*

A very significant mutation was found when sequencing the *Ir40a* gene with the heterozygous line *PR12-5/Df(2L)PR31*. This mutation was revealed to be a large 157-bp deletion in the untranslated region (UTR) of the gene. Finding this deletion in the UTR of *Ir40a* lead to some issues, especially during the DNA sequence alignment with the FlyBase reference sequence. Firstly, since a gene's UTR is present in an intronic region of the gene it therefore contains many highly repetitive sequences of DNA, and the case with *Ir40a* is no different. Many errors could and did arise when sequencing such repetitive regions of the gene, requiring multiple duplicates to achieve somewhat confident results. *Ir40a* sequencings near the deletion was duplicated at least three times, yet even after repeating the for that many times the results could still be deemed unreliable. So, although alignment of the affected UTR region with the reference FlyBase sequence was successful to a degree it was not up to normal standards.

Another key finding was that the location of the UTR deletion of *Ir40a* contains an alternate splice site. This "ATGG" splice site could be used by the gene as an alternate pathway to create different transcripts of mRNA. Deleting a region used by *Ir40a* to generate some of its mRNA transcripts would be disastrous to the organism if it can not compensate such a loss.

RNA Concentration Measurement for *Ir40a* RT-qPCR

An RT-qPCR was run to confirm the above-mentioned *Ir40a* deletion. After extracting the RNA from the adult flies its concentration was measured to run a proportionate RT-qPCR. The RNA concentration for the *ry*$^{+5}$*/Df(2L)PR31* sample was 434.6 ng/µL. The RNA concentration for the *PR12-5/Df(2L)PR31* sample was 1217.6 ng/µL.

RT-qPCR of the *Ir40a* RNA Transcripts

Once the RNA was reverse transcribed into DNA a concentration graph was produced, onto which a fluorescence threshold line was added at the beginning of the graph's linear phase. It is from this fluorescence threshold line that all the cycle threshold (Ct) values of the 33 reactions was derived. Furthermore, the Ct values were used to calculate the DNA concentration for each reaction and their differences noted. The final DNA products from the six unique reaction types were then run through a 1.5% agarose gel and imaged, as can be seen in **Figure 4**. A DNA template volume of 5 µL was used in each well, unless otherwise stated. The wells found at the top of the gel includes all of the DNA products from the experimental line *PR12-5/Df(2L)PR31*, numbered 1 through 8. The bottom portion of the gel, numbered 9-16, contained the DNA products from the control wildtype line *ry*$^{+5}$*/Df(2L)PR31*. Imaging of the gel revealed four wells (1, 2, 9, 10) with clear solid bands, and therefore DNA products, which corresponded to the G and common EFG exon transcripts for both the experimental and control lines. There was an extremely faint band present in well 12, which was the F transcript from the experimental line with the highest obtained DNA concentration. Bands in the rest of the wells were absent and therefore no products were formed there. Well 3 contained a minute amount (1 µL) of DNA template from the common EFG exon transcript of the experimental line. Meanwhile, the F

transcripts from the experimental line were present in wells 4 and 5, which contained the lowest and highest obtained DNA concentrations, respectively. The three wells 6, 7, and 8 contained no DNA templates to act as controls for transcript G, common EFG exon, and transcript F. Well 11 contained the F transcript from the experimental line with the lowest obtained DNA concentration. Lastly, well 13 was completely empty and wells 14, 15, and 16 acted as controls in identical composition and order as wells 6, 7, and 8.

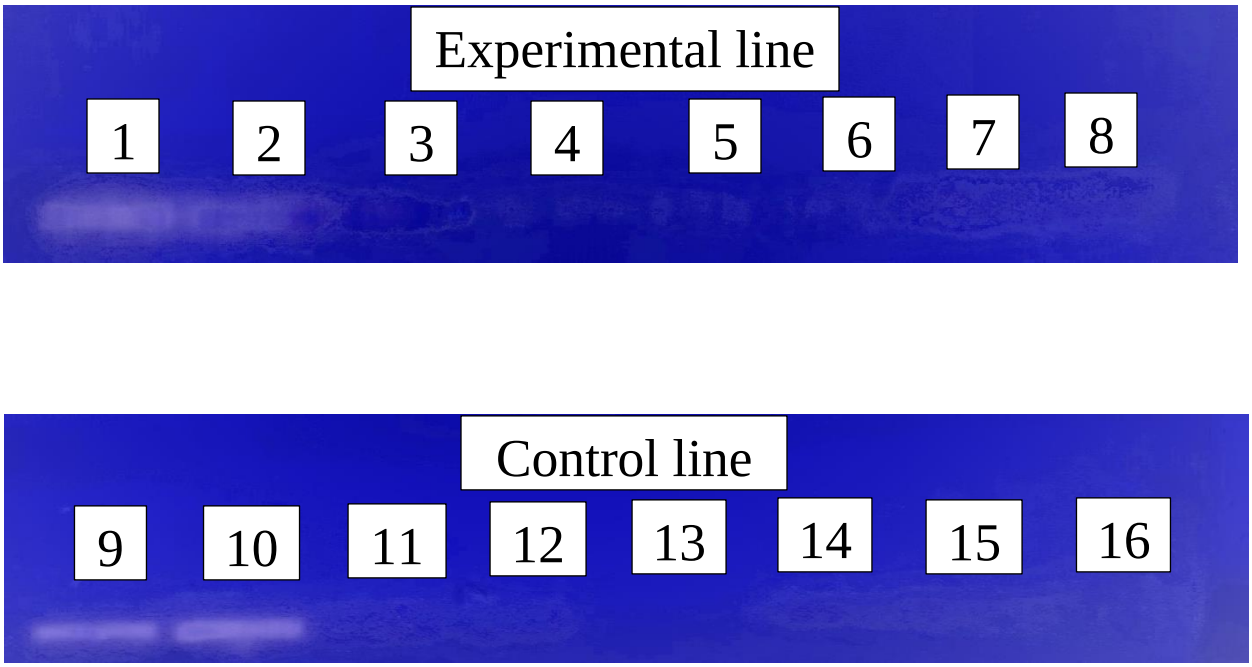


Figure 4. RT-qPCR DNA products of the three *Ir40a* transcripts E, F, and G for the experimental and control lines. The experimental line is *PR12-5/Df(2L)PR31*, which tested mutant line *PR12*-5. Line ry^{+5}*/Df(2L)PR31* represents the control; wildtype flies crossed with the *PR31* deletion. Each well contained 5 µL of DNA template. Lane 1 is the G transcript produced by the experimental line. Lanes 2 and 3 are the common EFG exon transcript from the experimental line, with lane 3 containing less DNA (1 µL). Lane 4 is the lowest DNA concentration obtained product of the F transcript from the experimental line. Lane 5 is the highest DNA concentration obtained product of the F transcript from the experimental line. Lanes 6, 7, and 8 contain no DNA product and act as controls for transcript G, common EFG exon, and transcript F, respectively. Lane 9 is the G transcript produced by the control line. Lane 10 is the common EFG exon transcript from the control line. Lane 11 is a low DNA concentration product of the F transcript from the control line. Lane 12 is a high DNA concentration product of the F transcript from the control line. Lane 13 is a blank well. Lanes 14, 15, and 16 acted as controls for transcript G, common EFG exon, and transcript F, respectively.

Sequencing of *CG42748*

The sequencing results for *CG42748* showed an important mutation in line *534*. This single nucleotide change, from an adenine to a guanine, affected four transcripts of the gene: G, H, K, and L. The control codon of "AAC" changed into "GAC" for all four transcripts, thus mutating a polar uncharged asparagine (N) into a negatively charged aspartic acid (D). These changes were confirmed by running duplicate sequences of the region to make certain that the single nucleotide change was not a polymorphism. To have convincing results three duplicates were run to verify this mutation, which all showed it occurring only in line *534*, being absent in lines *EMS 40-7*, *EMS 56-4*, and *621*.

Sequencing of *Cht10*

There was one mutation found using line *540* that occurred close to the 3' end of the *Cht10* gene at the 6455th nucleotide. This change is significant and was found on the 2,152nd amino acid from the 3' end of the gene, which occurred due to a conversion from a thymine to a cytosine nucleotide. This resulted in the "ATG" methionine (M) codon found in the control sequence turning into an "ACG" threonine (T) codon.

Table 3. Summary of all significant mutations found in the lethal 2Lh region. Data is based on the predicted translation start site, where +1 is the adenine of the first translation start codon, or first amino acid translated. The location of the mutations is written as the number of amino acids downstream (+), or upstream (-), of this start site. The published sequences on FlyBase were used as a reference for all mutations. The genes, ordered from distal to proximal, were chosen thanks to Alistair Coulthard's knock-down RNAi work which supplied a list of essential 2Lh gene candidates.

Gene	**Mutant Line**	**Nucleotide Change**	**Nucleotide Mutation Location**	**Amino Acid Change**	**Amino Acid Mutation Location**
CG42748	*534*	A to G A to G A to G A to G	+6097 +6430 +2014 +2146 *Single nucleotide change affecting 4 transcripts*	N to D N to D N to D N to D	+2033 (G transcript) +2084 (K transcript) +672 (H transcript) +716 (L transcript)
Ir40a	*PR12-5*	5' UTR deletion	5' UTR deletion from -244 to -69, limited to the E transcript		
CG17493	*EMS 40-2*	A to G	+425 (exon 1)	E to G	+142
Slmap	*EMS 40-18*	G to A	+848 (exon 1)	G to D	+283
	EMS 56-8	T to C	+1837 (exon 2)	S to P	+613
Cht10	*540*	T to C	+6455 (exon 6)	M to T	+2152

Chapter Five: Discussion

Various types of mutations can occur in sequenced DNA. They could range from insignificant polymorphisms or silent mutations all the way to very significant, protein-altering, mutations (Strachan and Read 1999). The insignificant polymorphisms and silent mutations have little to no impact on the protein complex because the bonding nature of the amino acid impacted by the mutation is preserved (Strachan and Read 1999). Oftentimes the exact same amino acid is retained and therefore the overall protein structure is completely the same, and so such mutations are of no importance for the goal of this study. The mutations that are of note, however, alter some part of the bonding structure of the protein in question, and usually in a critical location at that. Due to the highly repeated nature of heterochromatin and the oftentimes large number of primers required for sequencing the genes, all notable mutations found were sequenced multiple times to be confident. **Table 4**, presented at the end of this section, succinctly summarizes the important functions and homologs of the genes analyzed.

l(2)40Fa is *Slmap*

Based on previously completed RNAi analyses, by Alistair Coulthard, several different mutant lines were sequenced with the *Slmap* gene. Out of the tested mutant lines, the two mentioned in the results have revealed especially significant mutations. The result of the *Slmap* mutation in line *EMS 40-18* turned a hydrophobic non-polar glycine into a negatively charged aspartic acid. While the non-polar glycine would be most stable buried inside the protein core where it would avoid any surrounding water, the bonding nature of the aspartic acid would be entirely different. Instead of being concealed, the aspartic acid would very likely form a salt bridge between its negative end and an amino acid with a positively charged end, such as an arginine or a lysine.

Because such salt bridges are extremely stable, they stabilize the entire protein as a result, and this changes the way degradative enzymes or inhibitors access it and so could spell disaster regarding the protein's normal function. Hence, switching the glycine into an aspartic acid would lead to major tertiary structure reformation.

Besides the nature of the mutation, its location is even more important. The FlyBase annotations and NCBI's CDD for *Slmap* plainly shows that it is essential to *D. melanogaster* (Marchler-Bauer *et al.* 2017; Larkin *et al.* 2021). The gene forms part of the FAR/SIN/STRIPAK complex, allows for protein kinase binding activity, and participates in negative regulation of hippo signaling (Ashton-Beaucage *et al.* 2014; Zheng *et al.* 2017). The mutation found occurs in the 283rd amino acid, and so falls within the 216 to 311 amino acid interval responsible for a forkhead-associated domain (FHA) (Marchler-Bauer *et al.* 2017; Larkin *et al.* 2021). This domain is broadly found in many regulatory proteins because it recognizes various kinds of phosphopeptides (Hofmann and Bucher 1995). In eukaryotes, such as *D. melanogaster*, proteins containing the FHA domain localize to the nucleus, where they establish or maintain cell cycle checkpoints, repair DNA, and regulate transcription (Hofmann and Bucher 1995). Additionally, this particular FHA domain contains several phosphopeptide binding sites, seven to be exact, which bind directly to the ligand backbone and phosphate group. The mutation found in the 283rd amino acid occurs precisely in one of these phosphopeptide binding sites. A mutation of such a caliber, in this vital domain, will most definitely lead to the organism's fatality, like in the *EMS 40-18* line studied. Hence, *Slmap* can now be identified as *l(2)40Fa[EMS 40-18]*.

Similarly, the mutation in line *EMS 56-8* is just as significant in terms of *Slmap*'s protein structure. Here, a hydrophilic polar serine turned into a hydrophobic non-polar proline. The serine in the control line protein would be most stable at the periphery of the protein complex,

where it could interact with polar molecules such as water. On the other hand, the proline would act as any hydrophobic molecule would, by concealing itself deep inside the protein's core away from polar molecules like water. Such a change would rearrange the tertiary structure of the protein. A more enclosed protein complex could help protect it from degradative enzymes but could also hinder its innate functions.

Using NCBI's CDD, it was found that *Slmap* produces a protein at the amino acid interval of 355 to 615 which translates into a chromosomal segregation protein SMC (structural maintenance of chromosomes) (Marchler-Bauer *et al.* 2017). The mutation located in line *EMS 56-8*, at the 613th amino acid, is within this interval. SMCs are a highly conserved family of chromosomal ATPases whose purpose is to bind DNA and organize and segregate chromosomes for separation (Harvey *et al.* 2002). Curiously, the size of this SMC protein shrinks when the mutation found is considered. The first amino acid of this protein remains the 355th, yet the terminal one is changed to the 558th amino acid. This shortening of the protein would expose the affected end, which could end up detrimental for its functionality. As can be deduced from its functionality, this protein is essential for an organism to survive, which would otherwise die as seen in mutant line *EMS 56-8*. Hence, *Slmap* can also be identified as *l(2)40Fa[EMS 56-8]*.

l(2)40Fc is *CG17493*

Similar forms of proofs using RNAi analyses (Alistair Coulthard) and sequencing results helped identify the *l(2)40Fc* gene. The significant mutation found in the small *CG17493* gene occurred with mutant line *EMS 40-2* and turned a negatively charged glutamic acid into a hydrophobic non-polar glycine. Such a change would eliminate any potential salt bridge existing between the negatively charged side chain of the glutamic acid and a positively charged side chain of a basic

amino acid, like lysine or arginine. The glycine formed instead of the aspartic acid would stabilize itself with other hydrophobic interactions, finding them deep within the protein complex. Therefore, this mutation would lead to major tertiary structure changes for the protein.

According NCBI's CDD, *CG17493* translates one protein, a centrin, in the amino acid interval of 26 to 182 (Marchler-Bauer *et al.* 2017). Centrins are a universal family of calcium-binding phosphoproteins that act as calcium sensors with numerous proteins as their biological target (Satisbury 1995). They are found in the centrosome of practically all living organisms where they are vital for centriole duplication and, in extension, miotic division (Satisbury 1995). Since the mutation found in these results occurs in the 142nd amino acid, it directly affects the translation of this centrin protein. Once again, such a protein is very important for an organism and without it, such as seen in line *EMS 40-2*, fatality is certain. Therefore, *CG17493* can now be identified as *l(2)40Fc[EMS 40-2]*.

l(2)40Fd is *Ir40a*

Ir40a, representing ionotropic receptor 40a, is an extremely important gene because it plays a vital role in the hygrosensitivity of *D. melanogaster* (Enjin *et al.* 2016). Found in the fly's antenna, it is a crucial protein for the fruit fly to pick up signals of changing humidity levels in the surrounding environment (Enjin *et al.* 2016). It works in concert with *Ir25a* and *Ir93a* to sense changes in humidity and in dry detection behaviour, by mediating the response of the hygrosensory sacculus neurons (Enjin *et al.* 2016). These functions help the fly locate areas of proper humidity, preventing it from dying via desiccation.

Three mRNA transcripts exist for *Ir40a*: E, F, and G (Larkin *et al.* 2021). Transcript F is the only one that encodes all seven exons found in the gene, while the E and G transcripts encode

two unique variations of the three exons closest to the 3' end of the gene. Also, the three *Ir40a* transcripts share a common exon towards the 3' end of the protein which was used in the RT-qPCR experiment. As mentioned in the **Results** section above, the sequencing data revealed a 157-bp deletion in the 5' UTR *Ir40a*. By examining how this deletion affected each of the three *Ir40a* transcripts the deletion's impact on them could be identified.

Using the FlyBase gene annotations for the *Ir40a* transcript sequences, it was found that the deletion had no impact whatsoever on transcript G or transcript F. This is because the exons of these two transcripts are not affected by the deleted region of the gene, and so its absence would not impact their viability. However, the deletion did have a direct impact on transcript E. The 157 nucleotides that were removed happened to be a major part of the 5' UTR of transcript E, affecting nucleotides -244 to -69. The important role of the 5' UTR for the regulation of the gene's translation can not be understated for proper *Ir40a* gene function. Removing such an important region would disrupt transcription of *Ir40a*, rendering it unusable for the affected organism. This could explain the completely lethal nature observed in flies homozygous for line *PR12-5*.

Based on the sheer significance that stems from such a deletion, an RT-qPCR was decided to be used. The purpose of running this RT-qPCR was to find if *PR12-5/Df(2L)PR31* flies, the proposed mutant strain for *Ir40a*, produced lower transcript E DNA than the control ry^{+5}*/Df(2L)PR31* flies. A few variations of the three DNA transcripts near the location of deletion were tested for each fly line, and they were compared to determine any differences. Transcript E could of course not be used because its only unique region was the deleted sequence under question. Since the bands found in the gel following the RT-qPCR revealed definite presence of transcript G and common exon transcript EFG DNA in both fly lines, they are hence

unaffected by the deletion. The story is different when it comes to the results of the F transcript. The sample containing the highest attained DNA concentration of the F transcript in the control line shows a very faint band, compared to a complete absence of a band in the experimental line. Such a result could indicate, however slightly, that the deletion affects the production of the F transcript. Indeed, it is peculiar to observe that the control band for the F transcript is so very faint. One would assume that the DNA levels for all three transcript variations using the control line would show clear bands, yet only two out of them do. The cause could be due to human error when running the numerous RT-qPCR procedures for the F transcript samples. As a result, the F transcript bands mentioned can not be used as proof for determining the deletion's effect on the F transcript. Since only the production of transcripts G and common exon EFG can be deemed to be unaffected, the mutation present in line *PR12-5/Df(2L)PR31* does not impact them. This leaves transcripts E and F as potential candidates affected by the mutation.

The proof used for identifying *l(2)40Fd* was RNAi analyses (Alistair Coulthard) from line *PR12-5*, sequences of the *Ir40a* gene, and the **Figure 4** gel containing DNA products from the RT-qPCR of the heterozygous line *PR12-5/Df(2L)PR31*. Firstly, sequencing showed that the deletion affected solely the 5' UTR of the E transcript of the gene, leaving the F and G transcripts intact. Secondly, the RT-qPCR results were used to eliminate some of the transcripts as candidates affected by the mutation, narrowing them down to either transcript E or F. So, the two methods complemented each other and could both prove the existence of the 5' UTR transcript E deletion. One transcript deleted out of three would have a very substantial impact on the survival of the flies, however, there are two other possible transcripts. These two unaltered transcripts are possibly sufficient for the heterozygous line *PR12-5/Df(2L)PR31* flies to survive at the very low rate observed, but insufficient for homozygous *PR12-5* flies to live. The resulting

lethality, alongside the data analyzed, can be used as proof to confirm that *Ir40a* be identified as *l(2)40Fd[PR12-5]*.

l(2)40Ff is *CG42748*

Based on the latest FlyBase annotations (Larkin *et al.* 2021), the *CG42748* gene is described as playing a role in protein phosphatase 1 (PP-1) binding, which contributes to several important functions (Bollen and Stalmans 1992). Found in all eukaryotes, its role is to dephosphorylate the serine and threonine of many phosphoproteins (Bollen and Stalmans 1992). The resulting impact of PP-1 is tremendous, with essential roles in the following: muscle contraction, glycogen metabolism, intracellular and calcium transport, protein synthesis, and cell division (Bollen and Stalmans 1992). Critically mutating *CG42748* would leave the fly unable to develop beyond its embryonic stage, making the gene crucial for the fly's survival.

In the case of finding the *l(2)40Ff* gene, RNAi analyses (Alistair Coulthard) of line *534* and sequencing results of gene *CG42748* were used. Here the result of a single nucleotide change, which was confirmed three times using duplicate sequences, affected four of the gene's transcripts: G, H, K, and L. The mutation changed an asparagine into aspartic acid, replacing an amino acid that would usually be found at the protein's surface into one that is more likely to stabilize itself by forming a salt bridge. Adding an additional salt bridge, via the ionic bond created between the negative side chain of aspartic acid and the positive side chain of either arginine, histidine, or lysine would add stability to the protein. A change like this would lead to tertiary structure reformation of the protein due to the new, highly favoured, ionic bond produced by the aspartic acid. Such an ionic bond formation would rearrange the protein structure

completely, including the protein's active site, and so will likely lead to the organism's death, as seen in line *534*.

The *CG42748* gene creates 11 potential mRNA transcripts, from various combinations of its 10 exons (Larkin *et al.* 2021). Out of the four transcripts that were affected by the single nucleotide mutation located the K transcript contained the most exons, at 10. This was followed closely by transcript G with nine exons, then transcript L with six exons, and finally transcript H with five exons. Affecting four out of the potential 11 mRNA transcripts could spell disaster for the proper function of *CG42748*. This single nucleotide mutation affecting four transcripts could explain the reason behind line *534* flies failing to reach maturity. Hence, using the data above, alongside the completely lethal nature of flies with mutated *CG42748,* it can be confirmed that *CG42748* be identified as *l(2)40Ff[534]*.

l(2)40Fg is *Cht10*

Locating the *l(2)40Fg* gene was tricky because the most likely gene candidate, *Cht10*, is enormous compared to the other genes studied. This naturally increased the number of primers required for sequencing this gene which led to more obstacles appearing. Along the journey of analyzing *Cht10,* several primers failed to sequence and so had to be broken up into smaller primers or rebuilt altogether. Eventually, a very significant mutation was found in the 2,152nd amino acid of *Cht10* when analyzing line *540*. Here, a non-polar hydrophobic methionine was mutated into a hydrophilic polar threonine. The hydrophobicity of methionine forces it to avoid water by translocating into the interior of the protein complex it is a part of. This evasion of the protein's exterior changes the tertiary structure of the overall protein as methionine seeks hydrophobic interactions with similar amino acids at the protein's core, which results in a more

compact protein confirmation. On the other hand, the hydrophilicity of the threonine amino acid forces it to preferentially interact with polar molecules, such as water. Compared to methionine, this causes a completely opposite type of tertiary structure for the protein complex where, as the threonine moves towards the exterior of the protein complex, it unfolds the protein structure. Overall, this sort of interaction will lead to a less compact protein confirmation than the one seen with the methionine leading to it being more easily accessible to degradative enzymes or inhibitors, which could spell disaster for *Cht10*.

This mutation becomes even more significant when considering where it occurs. Based on NCBI's CDD results for *Cht10*, the mutated 2,152nd amino acid is located at one of four catalytically active chitotriosidase domains closest to the 3' terminal end, which spans the amino acid interval of 1,910 to 2,274 (Marchler-Bauer *et al.* 2017). These four domains are very important for the *Cht10* to function since chitotriosidases hydrolyze glycosidic bonds in the chitin of many animals, including crustaceans and insects, such as our beloved *D. melanogaster* (Kramer and Koga 1986). They are essential for these animals since they help reshape the chitin protecting the organism and help digest the chitin consumed from other animals or fungi (Reynolds and Samuels 1996). Because the mutation would reshape the protein's tertiary structure it could disrupt one of these four catalytic domains. This would then cause *Cht10* to lose its inherent functionality, proving lethal for the organism in question as seen in line *540*. Hence, *Cht10* can now be identified as *l(2)40Fg[540]*.

Search for *l(2)40Fe*

Previously, the only gene that has been identified in the lethal 2Lh heterochromatic region was *lt* which is synonymous with *l(2)40Fb*. The present project focused on classifying the other six essential 2Lh heterochromatic genes that were up until now unidentified: *a*, *c*, *d*, *e*, *f*, and *g*. Five out of these six genes have been successfully identified in this project, as discussed extensively in the previous sections. However, one elusive gene that remains unknown is *l(2)40Fe*. No success was found in locating any unique mutations in line *EMS 56-24* flies, which is the sole candidate line available attributed to this gene based on RNAi knock-down data (Alistair Coulthard). Since this line was tested numerous times with multiple potential gene candidates to no avail, and the rest of the genes between *l(2)40Fd* and *l(2)40Ff* have been shown to be non-essential, it is therefore deemed unfindable.

Here are some of the potential reasons behind the lack of candidates in this line, thanks to discussions with Alistair Coulthard. Firstly, the large intron sizes naturally found in all heterochromatic regions makes the process of finding multiple candidate lines that hit a gene of interest extremely difficult, as could be the case with *l(2)40Fe*. Secondly, this gene could be haploinsufficient. In this case a single wildtype copy present is not enough for normal gene function and would require a special screen to be conducted to recover a mutant. And thirdly, the gene might map to a completely different region in the *D. melanogaster* genome due to various position effects that would obscure mapping of this gene.

Additionally, a peculiar finding was uncovered when sequencing some of the gene candidates with line *EMS 56-24*. It was revealed that its background polymorphisms did not match the sequenced polymorphisms of other lines from the same genetic background. This was

strange because it was not the case in any of the other lines sequenced, and it could possibly mean that the line *EMS 56-24* flies used in this project are not from the true line. This is also supported by previous research using the same line. During his work in the 1980's, Sharp noted that line *EMS 56-24* was lost. In Figure 26 of his 1988 PhD thesis, which presented 2Lh mutants and the mutations used to localize them, he highlighted the lost lines: *EMS 40-6* and *EMS 56-24* (Sharp 1988). He chose to include these lines in his diagram because they have been used in the past to identify their complementation groups (Sharp 1988). Therefore, the *EMS 56-24* stock copy flies used in this project could very well be different than those initially created by Hilliker in 1976. Perhaps the original line was lost, or the flies reverted to wildtype, either way the current line in stock is most likely not the original *EMS 56-24* line.

Mapping the Lethal 2Lh Region

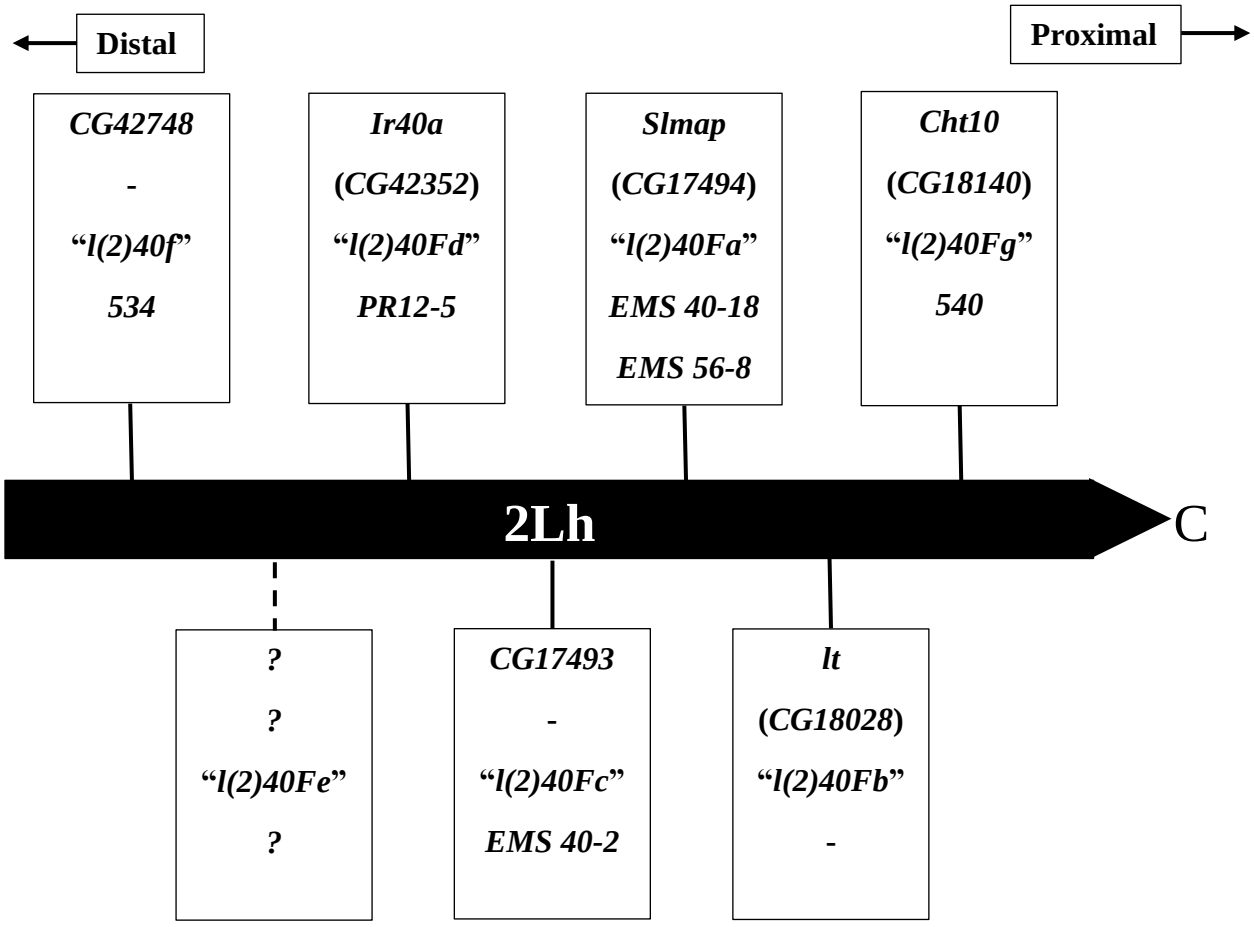
With the amount of evidence presented for each of the genes it is therefore possible to finally create a map of the vital genes found in the lethal 2Lh region. The map is presented below after **Table 4** and is labelled as **Figure 5**. Included in the map are the five essential genes successfully mapped in the current paper, and additionally the previously discovered essential gene *lt* and the unidentified gene *l(2)40Fe*. Chromosome 2 is represented by the large horizontal line running through the middle and the thinner vertical lines protruding from it are the locations of the genes mapped. The genes are separated by placing them in boxes. Within them, at the top, is written the gene's name as it is widely known and below it in brackets is the synonym, if any, it is referred to. The gene's complementation group within the lethal 2Lh region, ranging from *l(2)40Fa* to *l(2)40Fg*, is then written below that and presented in quotation marks. Finally, the mutant line where the mutation was found in is stated at the very bottom, which in the case of *Slmap* is two mutant lines. Distal genes are listed towards the left of the map, while the more

proximal ones are listed towards the right. The question marks in the *l(2)40Fe* box are due to the failed attempts at finding it during sequencing, rendering mapping of it currently impossible. A dashed line is used to represent the theoretical location of *l(2)40Fe* in chromosome 2 with respect to the other vital genes, yet its exact location can not be certain. The construction of this map would not have been possible without Alistair Coulthard, who used the RNAi knock-down method to create a list of 2Lh candidate genes which was used in this project.

Table 4. Assigned functions and homologs of the essential genes mapped. The homologs were found using the DIOPT version 8.0 ortholog mapping online resource (Hu *et al.* 2011). Only the closest gene homolog was presented, unless two genes showed equal similarities such as the *CG17493* and *Cht10* homologs.

Gene	**Function in *Drosophila melanogaster***	**Homologs**
CG42748	Plays a role in cell-cell junction organization. Enables protein phosphatase 1 binding activity (Bollen and Stalmans 1992).	*AJM1* (Apical Junction Component 1 Homolog) (*Homo sapiens*)
Ir40a	Essential role for hygrosensation (response to environmental changes in humidity) found in the antennal neurons (Enjin *et al.* 2016).	*GD24353* (*Drosophila simulans*)
CG17493	Allows calcium ion binding activity. Important for centriole replication and mitotic cell cycle (Satisbury 1995).	*CETN1* (centrin 1) and *CETN2* (centrin 2) (*Homo sapiens*)
Slmap	Forms part of FAR/SIN/STRIPAK complex. Enables protein kinase binding activity. Participates in negative regulation of hippo signaling (Ashton-Beaucage *et al.* 2014; Zheng *et al.* 2017).	*SLMAP* (sarcolemma associated protein) (*Homo sapiens*)
lt	Fusion of autophagosome with lysosomes, endocytic down-regulation, and eye pigment biogenesis. Forms part of the HOPS and CORVET complexes (Balderhaar and Ungermann 2013; Takáts *et al.* 2014).	*VPS41* (Vacuolar protein sorting-associated protein 41 homolog) (*Homo sapiens*)
Cht10	Chitin binding activity, chitin catabolic process, and chitinase activity (Kramer and Koga 1986; Reynolds and Samuels 1996).	*CHI3L1* (chitinase 3 like 1) and *CHI3L2* (chitinase 3 like 2) (*Homo sapiens*)

Figure 5. A simple genetic map of the essential genes in the lethal 2Lh region. The map presents the five essential genes successfully mapped in the current paper, alongside the previously discovered essential gene *lt*, and the unidentified gene *l(2)40Fe*. Thick horizontal line running through the middle signifies chromosome 2, while the thinner vertical lines protruding from it are the locations of the genes mapped. Each gene has its own box, with a synonym (if any) in brackets below, followed by its complementation group within the lethal 2Lh region ranging from *l(2)40Fa* to *l(2)40Fg* in quotation marks, and finally the mutant line(s) where the mutation was found (two for *Slmap*). Genes run distal (left) to proximal (right), and the centromere is represented by "C". Vital locus *l(2)40Fe* is unfindable and so is represented by question marks in its box. The genes were chosen thanks to Alistair Coulthard's RNAi knock-down work which supplied a list of essential 2Lh gene candidates.



Contribution to New Knowledge

Heterochromatin has been labelled as an unimportant part of the genome for many decades. It was only after numerous revolutionary contributions were completed, by countless scientists, that its significance for living organisms was revealed. Using the accomplishments of the past, the goal of the current project was the identification and mapping of the essential heterochromatic genes in the lethal 2Lh region of *D. melanogaster*. In particular, the unidentified essential complementation groups *l(2)40Fa* to *l(2)40Fg*. Before this project, out of these seven complementation groups, only *l(2)40Fb* was mapped to the gene *lt*. Utilizing the candidate gene approach I have used the mutant strains for the lethal 2Lh region to successfully map five out of the missing six essential genes. They are as follows: *l(2)40Fa* is *Slmap*, *l(2)40Fc* is *CG17493*, *l(2)40Fd* is *Ir40a*, *l(2)40Ff* is *CG42748*, and *l(2)40Fg* is *Cht10*. These genes can now act as effective targets for a multitude of applications, ranging from gene linkage to drug development against diseases.

Besides this success, the work accomplished here provides a possible clue for the discrepancy seen between the genetic mapping completed by Hilliker in 1976 and the candidate genes for 2Lh. Based on the available data, there is simply not enough candidate genes present that are proximal of *lt* to match Hilliker's 2Lh complementation map (1976). The most likely reason is that the original genetic backgrounds of the strains used to map the 2Lh region contained multiple inversions. If true, this could explain the extreme difficulties faced in this project, and others before it, when mapping the lethal 2Lh region. This degree of difficulty could clearly be seen in the mapping of the 2Lh locus *E(Sd)*, which was mapped to two different locations of the chromosome by separate researchers (Brittnacher and Ganetzky 1984; Sharp *et al.* 1985).

Using the results from the identification and mapping of the genes *l(2)40Fa* to *l(2)40Fg*, my hypothesis is that at least two inversions occurred in this region. A very possible scenario could have begun with the seven genes in original alphabetical order of *a* to *g*. They would run distal to proximal and would assume that the unidentified gene *e* is present between *d* and *f.* The first inversion would include strictly the region's *a* to *f* segment, changing the gene order to "*f, e, d, c, b, a, g*". A second inversion could have then occurred, involving only the *a* to *b* segment, which would have changed the gene order to "*f, e, d, c, a, b, g*". This result would then agree with the one derived at using the extensive analyses shown in this project and is more clearly illustrated above in **Figure 5**. Of course, many more combinations of inversions are possible which lead to the same gene order, but the one presented here is the simplest and hence most probable solution.

Combining the five essential genes identified in the current research with the previously known *l(2)40Fb* results in six out of the seven vital genes of the lethal 2Lh region. Because the above-mentioned genes have been identified as essential, they are now all viable candidates for further examination. This could include searching for the genes' specific effects on the organism which will help any future studies involving *D. melanogaster*. For instance, potential research could focus on linkage mapping by determining the genes responsible for certain traits or diseases. It is known that the absence of these vital genes leads to either complete or partial fatality of the affected organism, yet it is uncertain of the specific disorders that could occur as a result. Studies related to this could make it possible to pinpoint traits and diseases that were not possible before and therefore conduct experiments to understand them better. The new knowledge could then be further extended to other organisms, for example humans, by studying their respective homologs. Identifying the homologs with direct impact on diseases, such as

cancer, could eventually provide working models on them and lead to the development of effective drugs to combat such terrible conditions.

In conclusion, the goals described in the beginning of the paper have been successfully accomplished. Six out of the seven essential genes are now tied to actual genes and mutant lines. There was no success identifying and mapping one of the genes, *l(2)40Fe*, even though all available methods have been exhausted in its search. It will have to remain unidentified for the time being, at least until new mutagenesis screens or RNAi knock-down or other methods reveal information yet unknown. Barring this, the lethal pericentromeric 2Lh region is for the most part mapped genetically and molecularly. Therefore, although being a modest contribution, this project was conducted in the hopes that in the future the entire chromosome 2 heterochromatin is elucidated, and eventually the whole *D. melanogaster* genome as well.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al*. 2000. The genome sequence of *Drosophila melanogaster*. *Science*. **287**:2185–2195.

Ashton-Beaucage D, Udell CM, Gendron P, Sahmi M, Lefrançois M, Baril C, Guenier AS, Duchaine J, Lamarre D, Lemieux S, *et al*. 2014. A Functional Screen Reveals an Extensive Layer of Transcriptional and Splicing Control Underlying RAS/MAPK Signaling in Drosophila. *PLoS Biol*. **12**(3):e1001089.

Balderhaar HJK, Ungermann C. 2013. CORVET and HOPS tethering complexes - coordinators of endosome and lysosome fusion. *J Cell Sci*. **126**(Pt 6):1307-1316.

Banerjee S, Bhandary P, Woodhouse M, Sen TZ, Wise RP, Andorf CM. 2021. FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics*. **22**(205).

Bollen M, Stalmans W. 1992. The Structure, Role, and Regulation of Type 1 Protein Phosphatases. *Crit Rev Biochem Mol Biol*. **27**(3):227-281.

Brittnacher JG, Ganetzky B. 1984. On the components of segregation distortion in *Drosophila melanogaster*. III. Nature of enhancer of SD. *Genetics*. **107**(3):423–434.

Brizuela BJ, Elfring L, Ballard J, Tamkun JW, Kennison JA. 1994. Genetic Analysis of the *brahma* Gene of *Drosophila melanogaster* and Polytene Chromosome Subdivisions 72AB. *Genetics*. **137**(3):803–813.

Brown SW. 1966. Heterochromatin. *Science.* **151**(3709):417-425.

Coulthard AB, Eberl DF, Sharp CB, Hilliker AJ. 2003. Genetic analysis of the second chromosome centromeric heterochromatin of *Drosophila melanogaster*. *Genome*. **46**(3):343-352.

Coulthard AB, Alm C, Cealiac I, Sinclair DA, Honda BM, Rossi F, Dimitri P, Hilliker AJ. 2010. Essential Loci in Centromeric Heterochromatin of *Drosophila melanogaster*. I: The Right Arm of Chromosome 2. *Genetics*. **185**(2):479–495.

Dimitri P. 1991. Cytogenetic analysis of the second chromosome heterochromatin of *Drosophila melanogaster*. *Genetics*. **127**(3):553-564.

Dimitri P, Arcà B, Berghella L, Mei E. 1997. High genetic instability of heterochromatin after transposition of the LINE-like I factor in *Drosophila melanogaster*. *PNAS*. **94**(15):8052–8057.

Dimitri P, Caizzi R, Giordano E, Accardo MC, Lattanzi G, Biamonti G. 2009. Constitutive heterochromatin: a surprising variety of expressed sequences. *Chromosoma*. **118**(4):419-435.

Duvaud S, Gabella C, Lisacek F, Stockinger H, Ioannidis V, Durinx C. 2021. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* **49**(W1):W216-W227.

Eberl DF, Duyf BJ, Hilliker AJ. 1993. The role of heterochromatin in the expression of a heterochromatic gene, the *rolled* locus of *Drosophila melanogaster*. *Genetics*. **134**(1):277–292.

Enjin A, Zaharieva EE, Frank DD, Mansourian S, Suh GS, Gallio M, Stensmyr MC. 2016. Humidity Sensing in *Drosophila*. *Curr Biol*. **26**(10):1352-1358.

Gause M, Webber HA, Misulovin Z, Haller G, Rollins RA, Eissenberg JC, Bickel SE, Dorsett D. 2008. Functional links between *Drosophila* Nipped-B and cohesin in somatic and meiotic cells. *Chromosoma.* **117**(1):51–66.

Harvey SH, Krien MJE, O'Connell MJ. 2002. Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. *Genomes Biol*. **3**(2):reviews3003.1–reviews3003.5.

Heitz E. 1928. Das heterochromatin der moose. I. *Jahrb. Wiss. Bot.* **69**:762-818.

Heitz E. 1929. Heterchromatin, Chromocentren, Chromomenen. *Ber Bot Ges.* **47**(4):274-284.

Hilliker AJ, Holm DG. 1975. Genetic analysis of the proximal region of chromosome 2 of *Drosophila melanogaster*. I. Detachment products of compound autosomes. *Genetics.* **81**(4):705-721.

Hilliker AJ. 1976. Genetic analysis of the centromeric heterochromatin of chromosome 2 of *Drosophila melanogaster*: deficiency mapping of EMS-induced lethal complementation groups. *Genetics.* **83**(4):765-782.

Hofmann K, Bucher P. 1995. The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem Sci*. **20**(9):347-349.

Hoskins RA, Carlson JW, Kennedy C, Acevedo D, Evans-Holm M, Frise E, Wan KH, Park S, Mendez-Lago M, Rossi F, *et al*. 2007. Sequence Finishing and Mapping of Drosophila melanogaster Heterochromatin. *Science.* **316**(5831):1625-1628.

Howe M, Dimitri P, Berloco M, Wakimoto BT. 1995. *Cis*-effects of heterochromatin on euchromatic and heterochromatic gene expression in *Drosophila melanogaster*. *Genetics*. **140**(3):1033–1045.

Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. 2011. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*. **12**(357).

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics*. **23**(10):1289-1291.

Kramer KJ, Koga D. 1986. Insect chitin: Physical state, synthesis, degradation and metabolic regulation. *Insect Biochem*. **16**(6):851-877.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, *et al*. and the FlyBase Consortium. 2021. FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res*. **49**(D1):D899–D907.

Lim JK, Snyder LA. 1974. Cytogenetic and complementation analyses of recessive lethal mutations induced in the X chromosome of *Drosophila* by three alkylating agents. *Genet Res*. **24**(1):1-10.

Locke J, Kotarski MA, Tartof KD. 1988. Dosage-dependent modifiers of position effect variegation in *Drosophila* and a mass action model that explains their effect. *Genetics*. **120**(1):181-198.

Lohe AR, Hilliker AJ, Roberts PA. 1993. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics*. **134**(4):1149-1174.

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, *et al*. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. **47**(W1):W636-W641.

Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, *et al*. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res*. **45**(D1):D200-D203.

Marygold SJ, Coelho CM, Leevers SJ. 2005. Genetic analysis of RpL38 and RpL5, two Minute genes located in the centric heterochromatin of chromosome 2 of *Drosophila melanogaster*. *Genetics*. **169**(2):683–695.

Moschetti R, Celauro E, Cruciani F, Caizzi R, Dimitri P. 2014. On the Evolution of *Yeti*, a *Drosophila melanogaster* Heterochromatin Gene. *PLoS ONE*. **9**(11):e113010.

Muller HJ. 1930. Types of visible variations induced by X-rays in *Drosophila*. *J. Genet*. **22**(3):299–334.

Myster SH, Wang F, Cavallo R, Christian W, Bhotika S, Anderson CT, Peifer M. 2004. Genetic and bioinformatic analysis of 41C and the 2R heterochromatin of *Drosophila melanogaster*: a window on the heterochromatin-euchromatin junction. *Genetics*. **166**(2):807-822.

Pardo CA, Xu Z, Borchelt DR, Price DL, Sisodia SS, Cleveland DW. 1995. Superoxide dismutase is an abundant component in cell bodies, dendrites, and axons of motor neurons and in a subset of other neurons. *Proc Natl Acad Sci U S A.* **92**(4):954-958.

Reynolds SE, Samuels RI. 1996. Physiology and Biochemistry of Insect Moulting Fluid. *Adv Insect Physiol*. **26**:157-232.

Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, *et al*. 2011. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res.* **21**(2):147-163.

Rollins RA, Morcillo P, Dorsett D. 1999. Nipped-B, a *Drosophila* homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. *Genetics*. **152**(2):577-593.

Satisbury JL. 1995. Centrins, centrosomes, and mitotic spindle poles. *Curr Opin Cell Biol*. **7**(1):39-45.

Schultz J. 1936. Variegation in *Drosophila* and the Inert Chromosome Regions. *PNAS*. **22**(1):27-33.

Sharp CB, Hilliker AJ, Holm DG. 1985. Further characterization of genetic elements associated with the segregation distorter phenomenon in *Drosophila melanogaster*. *Genetics*. **110**(4):671–688.

Sharp CB. 1988. Biometrical and genetic studies of segregation distortion in *Drosophila melanogaster*. Ph.D. Thesis, University of Guelph, Guelph, ON, Canada.

Spofford JB. 1976. Position-effect variegation in *Drosophila*. In *The genetics and biology of* Drosophila (ed. M. Ashburner and E. Novitski), vol. 1, pp. 955-1018. New York (NY): Academic Press.

Strachan T, Read AP. 1999. *Human Molecular Genetics 2*. 2nd ed. New York (NY): Wiley-Liss.

Styhler S, Nakamura A, Lasko P. 2002. VASA localization requires the SPRY-domain and SOCS-box containing protein, GUSTAVUS. *Dev Cell*. **3**(6):865–876.

Takáts S, Pircs K, Nagy P, Varga A, Kárpáti M, Hegedűs K, Kramer H, Kovács AL, Sass M, Juhász G. 2014. Interaction of the HOPS complex with Syntaxin 17 mediates autophagosome clearance in *Drosophila*. *Mol Biol Cell*. **25**(8):1338-1354.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13(1):36-46.

Trojer P, Reinberg D. 2007. Facultative heterochromatin: is there a distinctive molecular signature? *Mol Cell.* **28**(1):1–13.

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, *et al*. 2009. FlyBase Consortium. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.* **37**:D555–D559.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3 - new capabilities and interfaces. *Nucleic Acids Res*. **40**(15):e115.

Yunis JJ, Yasmineh WG. 1971. Heterochromatin, satellite DNA and cell function. *Science.* **174**(4015):1200-1209.

Zheng Y, Liu B, Wang L, Lei H, Prieto KDP, Pan D. 2017. Homeostatic Control of Hpo/MST Kinase Activity through Autophosphorylation-Dependent Recruitment of the STRIPAK PP2A Phosphatase Complex. *Cell Rep*. **21**(12):3612-3623.

Appendices

Appendix A. The primer pairs used for the sequencings and diagnostic tests. They were created using the 0.4.0 version of the "primer3" program (Untergasser *et al.* 2012; Koressaar and Remm 2007), which Bio Basic Inc. then synthesized. Melting temperatures of approximately 60°C to 64°C were used when constructing the primer pairs.

Gene	**Note about Primer Pair**	**Forward Primer**	**Reverse Primer**
EGFP		ctggtcgagctggacggcgacg	cacgaactccagcaggaccatg
EGFP diagnostic control	*SOD1*	gagcagcggtacgcccgtga	gcaccagcgttgcccgttga
CG41434	5' end	tgttcgtcgcaactcgatag	agatgcggcaagaatagagc
	Exons 1-2	gaagccttggcctacctttt	gtttcgccgctgagagatac
	3' end	ggcctgatggttgttttgtt	ctgtttcaatctggcaagca
CG31601	5' end	cataaagttgctcggttctgc	tggattgtggggaagatgat
	Exons 4-5	gtctgttcattccgcctttc	cgattcggaggtgtcatctt
	Exons 5-6	tggacaagggcttttgtgt	ccctttgttaacacggcatt
	Exon 7	gtatccaacccgacaaggaa	gggccgtacaattattccag
	3' end	aacgtccttcttgacgtgct	acaccccgaataagtgcaac
CG34173		tttttcgccagaacactcag	cggaagtacacaatcgcatc
CG42597		cgcaggcaagtcacttctaa	aaggaaccgaaaacgatgtg
ttm3	5' end	gtttctggttgggatcgtca	tgcggtcataccttgttcag
	3' end	ttcactggtgctggaaatca	tcgaactgccaaaacaatga
CG6675	5' end	gctttcctacctccgctgta	tccaggcttgccaaaatatc
	Exons 1-2	acttagcccgatcgaacaga	ctgagctcccaaggagtgac
	Exon 2b	aatcgggatgtgaagaatgc	atggatcactgctggaggtc
	3' end	catccgagacaatggaaggt	ccgcaagcactcaaatgtta

Appendix A (Continued).

Gene	Note about Primer Pair	Forward Primer	Reverse Primer
CG42748	5’ end	cgaacgttctgttggcacta	cgccgcaatttctaaacagt
	Exons 2-3a	ttagaaattgcggcgctaac	ctcgggatcgtaaggcatag
	Exon 3b	gccagatttcagaggattgc	tttcccgctttagggttctt
	Exon 3c	ctctagctccagctcgtcca	cagctcctccagttcctcac
	Exon 3d	cattcactcaacgagctcca	tgggatcctggtcaaatgtt
	Exon 3e	gataagccggtgctggatac	aatgtggggtccatcagttc
	Exon 3f	gccggaatccgagtacgata	tcttcaagcaagcacaaaca
	Exon 4a	ttcattctgcaaatcgcaag	tggagaagaggcttctggag
	Exon 4b	agaaactccgagcgatgcta	gcttgaaccttagccaccaa
	Exon 4c	gaaacgaccggtgaaactgt	gcaacgaggcaaaacaaatc
	Dunk 5’ end	attctgagtaggcccccttg	gattccgccacagctacagt
	Dunk 3’ end	cgagccgctcttatatctgg	caactggacaacttgcatgg
	Alternate splice	cccaccgatagttaggcact	gttaacttcggacggggtgt
	Exon 6	cagtgttatggtttttcggtttg	cgagctcgatccactcaata
	Exon 7	ttcgtcctaaacccaataacg	ccgaaacgcgttaaagttgt
	Exon 8	tgccatcagtcaggaagaaa	tttttccgcggatctattt
	Exon 9	ctcgagtaagcgttaaggttgt	tgttttggactgccttttga
	Exon 10a	ggggaatgctctttgttgtt	cctttgcgggctaaaagact
	Exon 10b	acacggtcggacataggaag	tcgaagacctccggatagtg
	3’ end	tgtgccaagctgaaactctg	cagatcgacgcggattatg
Ir40a	5’ end	ccgtacatcgaaatgctcct	gttatcgccgtcgtaccaat
	Exons 2-4	cagctcccgtgtaattcgtt	cctgaatcagccagtgtgaa
	Exons 3-4	gaggtcagaagatggcaagg	cactgtgtgacgggaactgt
	Alternate splice	actgcttctggcaagtccac	tctgatgacgcaattttgact
	Exon 5	tggactttgattactggaccaa	cgattttcgttacgatggaga
	Transcript G	gatcgttcctttccgttgac	ttgaatccatcagccacgta
	Transcript EFG	gcgagccaccaatcaatact	tgtggaagcgggaattaaac

Appendix A (Continued).

Gene	Note about Primer Pair	Forward Primer	Reverse Primer
RpL21		tagccttgaggacggatttg	tcagtggcaacgctattcaa
Tif-1A	5’ end	tcaatgagcacgttaaaaagc	gcctcgacaacggttacaaa
	Exons 1-2	cgaaatttccatctggtcaag	tcacatggtacggctttttg
	Exon 2b	atcttaactgccgtgccaat	aaatcgtggcgaaaaacg
	Exon 3	agatcctacttttcgacttcgga	ctagccatgtatcccaccgc
	Exons 3b-4	gacgatgtatgcgccgaaaa	ggcctgtcaacgccatatga
	Exons 5-6	aagtgaacctcctgtgagcg	atcatggcgaagatccgaaa
	3’ end	cgatgagacgctggaatcgt	taagccaaaggtaccagcgc
RpL5	5’ end	agggagtccagtcccaactt	ccggcgaggtatgtgtaaaa
	Exons 3-4	gatcagcccttaaggcaaca	tggcgtactcaacttgatgg
	3’ end	gccgcacactacagattgaa	ttgcgtatcgactgtgttagc
CG17493		taggaatagggtgggctgtg	aggaacgcacgaagtgtttt
Slmap	5’ end	ttgtcaaacagcagccactc	tgtgcgactttggttcacat
	Exon 1	ttacaaagggaggtggcatc	cggtcgctaatgggacttac
	Exon 2	aggcaagcaaattgagacca	gataaatcgcgatggattcg
	Exons 2-3	cgagcttaaggagcttcgtc	gatcaacgcctttctcaagc
	Exon 3	ccgaaatcgtggatcacttt	cgcgagttttcctcattagc
	3’ end	tgggatacgaatacgccata	ttaaccggaccaccacaaat

Appendix A (Continued).

Gene	Note about Primer Pair	Forward Primer	Reverse Primer
Cht10	5’ end	acgcccactctaacagactt	tgagtttccgcgggcaataa
	Exon 1a	atcagagcaacacgtctagg	cagtcacaggtccccttgag
	Exon 1b	tacgtcatggattcagcggg	gcaaagttgttgttagccgc
	Exon 1c	tatgaagaccccaccagtgc	accgtcccaatgcaagtct
	Exon 1d	gcacaaccaactcacaccac	caatcgcagttcccaattct
	Exon 1e	tgcaaaggcgaaaagagagt	gccctcaaagccgtattttt
	Exons 1f-2	gctactttacaaactgggcct	aaagagtgaggccgtatcag
	Exons 3a-4a	ctcacatgccgcaaaccttt	ccctccccattagactcatgt
	Exons 3b-4b	tgtggaaatggcgtgcatc	tgccgtccctttcttgcaa
	Exon 4c	ggcgataagtactcccgact	ccccgaaacctctgagtaca
	Exons 4d-5	tcatggggatcagtgggttt	gggctaagtcttgtgcgttt
	Exon 6a	ggctttatattgggcgtggt	tcgctcatagaaacggttg
	Exon 6b	tggtaatatgggtcgacaaac	tcgctcatagaaacggttg
	Exon 6c	gcctaccaatgttcagtcctc	attttactgggcgagactgc
	Exon 6d	tggacttagcctgggagttt	ctgagtaatgggtgcttgcc
	Exon 6e-3’ end	ctagagcggatggattcttg	tatggttcaagccccgttac
	Exon 6f-3’ end	ctagagcggatggattcttg	caatatcggagaacccgttg
Tim23	5’ end	ttttgctaatagagcgcctga	cgaatacagcaccgtcaatg
	3’ end	aagcaattggcggaattaaa	aactgtcactgccagacgtg
Gpb5	5’ end	tcccccacgcaatttagtta	ccttttttattaatgcggcg
	5’ endb	cccccagccatttcattat	cgtagcgcatgctcagttta
	3’ end	aagcaattggcggaattaaa	aactgtcactgccagacgtg

Appendix B. Volume of individual component added for the PCR reaction. Note the change in nuclease-free (NF) water depending on whether one or two primer pairs are used. Two primer pairs (*SOD1* and *EGFP*) were used when preparing the PCR reaction for the *EGFP*-diagnostic. All PCR reactions destined for sequencing used only one primer pair.

Component	**Reaction Volume with a Single Primer Pair**	**Reaction Volume with Two Primer Pairs**
NF water	8.5 µL	6.5 µL
Primer forward (10 µM)	1.0 µL	1.0 µL
Primer reverse (10 µM)	1.0 µL	1.0 µL
SOD1 forward (10 µM)	-	1.0 µL
SOD1 reverse (10 µM)	-	1.0 µL
2X MyTaq HS Mix	12.5 µL	12.5 µL
Genomic DNA	2.0 µL	2.0 µL
Total:	**25 µL**	**25 µL**

Appendix C. Reaction setup for the RT-qPCR reaction. Note the change in NF water depending in whether the reaction contains the RNA template or not. The RNA template, if applicable, was always added at the very end.

Component	**Volume for RNA Template Reaction**	**Volume for Blank Reaction**
iTaq universal SYBR® Green reaction mix (2x)	10.0 µL	10.0 µL
iScript reverse transcriptase	0.25 µL	0.25 µL
Forward primer (10 µM)	0.6 µL	0.6 µL
Reverse primer (10 µM)	0.6 µL	0.6 µL
RNA (100 ng/µL)	2.0 µL	-
NF water	6.55 µL	8.55 µL
Total:	**20 µL**	**20 µL**