Mapping Modifiers of Protein Aggregation

A Thesis

Submitted to the Faculty

of

Drexel University

by

Anna Lysenko

in partial fulfillment of the

requirements for the degree

of

Master of Science

June 2014

This thesis is dedicated to

My husband and my parents, who have supported me throughout this process and without whom I would not be where I am today

I would like to Acknowledge

My supervisor, Dr.Tali Gidalevitz, and my committee members Dr.Aleister Saunders

And Dr.Joseph Bentz, who have guided me through my time at Drexel University; as well as lab members Jasmine Alexander-Floyd and Gowri Gouda who contributed to this project.

Table of contents

Figures

Mapping Modifiers of protein aggregation

Anna Lysenko

June 2014

Abstract

Polyglutamine expansions that go beyond 35 CAG repeats in disease causing genes such as the Huntington gene, cause misfolding of proteins and can lead to aggregation and disease. Disruption of proteostasis and aggregation it thought to play a significant role in conformational diseases. In addition, there are other aspects that influence onset, progression and severity of disease. In this study, we look at how natural variants can contribute to increased protein aggregation as well as suppress aggregation. Specifically, we try to map a single recessive locus that contributes to increased aggregation in head muscle cells in *Caenorhabditis elegans*, as well as a network of modifiers contributing to suppression of aggregation. Here we describe how these strains were created, the process of determining whether these phenotypes are a result of a single modifier, or network of genes as well as mapping the modifier responsible to high aggregation of head muscle cells.

Protein folding and Proteostasis

Protein folding is an essential function in any living organism. Found in the coding regions of genes, amino acid sequences define the protein structure. When amino acids come together they form a long unfolded chain, which then needs to be folded correctly into its tertiary structure to become a functional protein. In the process of forming a tertiary structure, proteins go through a process called folding intermediates (Gianni et al.,2007). This is a critical point for a protein, and one where the process can go awry (Horwich et al.,2002).Proteostasis, or protein-folding homeostasis is when there is harmonious function between the biological pathways that control protein synthesis, folding, trafficking and degradation.. Disruption in proteostasis is thought to contribute to the development of disease; proteostasis disruption can results in off pathway intermediates and the formation of toxic oligomers; these are structures made of monomers created by aberrant or mutant proteins that can cause aggregation or toxicity to occur (Winner et al, 2011, Wetzel et al.,1994) Protein quality control mechanisms, such as the ubiquitin-proteasome system (UPS), Unfolded protein response (UPR), Endoplasmic Reticulum-associated degradation (ERAD), aggregate clearance via autophagy, Heat shock response (HSR) and chaperone systems are critical in maintaining stability of protein folding. These are essential cellular functions, and without molecular chaperone assistance or proper degradation (Roth et al.,2008), aberrant proteins are more prone to misfolding (Bence et al.,2001, Bukau et al.,2006).

Two mechanisms that are used to alleviate the stress of misfolded proteins and attempt to prevent aggregation, are folding the proteins with the assistance of molecular

chaperones (Hartl, et al.2011, Bar-Lavan et al.,2012) or if this is not possible, target the misfolded proteins for degradation (Powers et al.,2009). Misfolded proteins that are not refolded to their native state, or have been targeted for degradation prematurely and remain functional but abarrent, are a contributing factor in conformational neurodegenerative diseases. Protein quality control is an essential process, that if disrupted can have detrimental effects

Molecular chaperone

Molecular chaperones, such as Hsp70, are able to convert toxic aggregates and misfolded proteins to their native forms (Hinault et al.,2010, Bukau et al.,2006) In an ideal situation, molecular chaperones would assist in folding proteins to their proper form. In situations where there is an overproduction of mutant protein, such as a mutant huntingtin , molecular chaperones attempt to refold them, and failure in these mechanisms may in part be responsible for disease. Organisms can delay disease onset and show the ability to protect against aggregation by increasing chaperone activity through gene overexpression such as hsf-1 which a transcription factor activated in a cellular stress response, suggesting chaperones have to ability to mediate aggregation (Morley et al., 2004, Cohen et al.,2006). Molecular chaperones are also critical in tagging mutant proteins for degradation, and proper disposal can prevent disease causing aggregates. Failure of these mechanisms may be the cause or a contributing factor in neurodegenerative diseases (Bennett et al.,2002, Bukau et al.,2006). Stable oligomers

have been shown to hinder chaperone activity, leading to buildup of these mutant protein oligomers, and as a result disease causing aggregation (Hinault et al.,2010).

Polyglutamine

Conformational neurodegenerative diseases such as Parkinson's, Alzheimer's, Amyotrophic Lateral Sclerosis and Huntington's disease, are diseases associated with buildup of plaques and aggregates (Kakizuk et al.,1998, Chen et al.,2002). For example Huntington's disease is caused by polyglutamine (polyQ) expansions (Kopito and Ron, 2000), or CAG repeats, that results in detrimental muscular, neurological and cognitive effects (Orr, 2001). Polyglutamine expansions are long stretches of CAG repeats that code for the amino acid Glutamine. Studies performed on brain tissue of postmortem patients displaying symptoms of Huntington's disease, have identified aggregation of mutated huntingtin protein (Kakizuka et al.,1998, Chen et al.,2002). A number of GWAS and cohort studies performed, such as a study done on Venezuelan kindreds, reveals strong correlations between length of CAG repeat in an individual and their age of onset and severity of disease (Wexler et al.,2004, Brinkman et al.,1997).In many cases, we still do not understand the mechanisms that cause the aggregation of misfolded proteins, and why they are toxic, which contributes to our inability to find effective treatments.

Conformational neurodegenerative disease

Huntington's disease, as well as other trinucleotide repeat disorders and conformational neurodegenerative diseases, such as X-linked spinal and bulbar muscular atrophy, fragile X syndromes, and spinocerebellar ataxia (La Spara et al.,1994), are caused by a gain-of-function mutation that results in unstable proteins.. These mutated proteins can result in a disruption of the global balance of protein folding, and not only influence aggregation of one mutant protein, but result in misfolding of other proteins necessary for cellular function; this imbalance results in quicker aggregation of disease proteins such an huntingtin (Gidalevitz et.al.,2006,). Hindering chaperone activity (Hunault et al.,2010) as well as the activity of other proteins required for cellular function (Olzscha et al.,2011), these misfolded proteins lead to a dysfunction of protein folding and degradation pathways; as an example, studies have been done that show aggregation of human amyloidogenic proteins interfere with proteostasis in numerous models such as *C.elegans*, (Gidalevitz et al.,2006), and yeast (Eremenko et al.,2012) . Polyglutamine expansions, beyond the acceptable threshold of 35 CAG repeats, lead to the conformational changes in protein and result in misfolded protein, and ultimately aggregation (Stine et al.,1993, Soto CJ.,2003, Dobson et al.,2003, Kakizuka, 1998) (figure 1 and 2).

CAG repeats can vary from 6 to 39 repeats in an individual unaffected by the disease, and can range from 36 to 180 repeats in an individual showing symptoms of Huntington's (Rubunsztein et al.,1996, Satbasivam et al.,1997). The threshold for the disease is very abrupt, a person without Huntington's can have 34 repeats present, while one with the

disease can have 36 (Duyao et al, 1993). Persons with repeats slightly above the disease threshold tend to show disease symptoms in old age, beyond their 50's, while juvenile onset Huntington's is seen in patients with 50 and higher repeats (Brinkman et al.,1997). Age and CAG repeat length are two major contributing factors in diseases such as Huntington's. The aggregation of misfolded proteins and subsequent increase in toxicity is intensified as an individual with one of these diseases ages, but this is also tied to the number of CAG repeats (Morley et al., 2002, Brinckman et al.,1997). The efficiency of proteostasis declines with age either due to environmental stressors or intrinsic variation resulting from polymorphisms and disease causing mutations (Kikis et al.,2012). Using *C.elegans* as a model, it has been shown that there are several hundred proteins that become insoluble as the organism ages, suggesting protein folding mechanisms might decline in function and can have detrimental disease consequences (David et al.,2010

It has been shown that there are early misfolded proteins present prior to the appearance of disease symptoms. It is now possible to detect trace amounts of misfolded proteins in early age mice, whereas large inclusion bodies are not found until much later in life (Gupta et al.,2011) This early misfolding and aggregation formed by an array of proteins other than the disease associated protein, such as HypF protein in *E.coli* and P13-SH3 from bovine, both which are small globular proteins that are capable of forming fibrillar aggregates *in vitro* but are non pathogenic (Bucciantini et al., 2002) These aggregates have been shown to be cytotoxic and causes cellular dysfunction that leads to quicker aggregation and buildup of plaques of pathogenic mutant proteins(Bucciantini et al.,2002) This demonstrates that avoiding aggregation of any protein is crucial in preserving proper biological and cellular functions.

Natural variation and disease

Natural variation in genetic background that occur due to random mutations can assist or prevent organisms from successfully adapting to unfavorable internal conditions, such as aberrant proteins (Gidalevitz et al.,2013). It is well know that genetic variation in individuals genetic background influence phenotypic expression of conformational disease (Genin et al.,2008, Doetschamn et al.,2009). Conformational diseases tend to have a large variability in the age of onset (Wexler et al.,2004). Even though mutations that are causative of these diseases are expressed from birth, onset is typically seen later in life. In the example of Huntington's diseases, the length of polyglutamine repeats will be in part the determining factor of the age of onset of the disease (figure 3)(Wexler et al.,2003), In individuals with 40 to 50 CAG repeats, polyQ length accounts for about 44% of variation in the age of onset as well as the severity of the disease, while in individuals with greater than 50 CAG repeats, it accounts for 72% of the variation. Of the residual variation, the remaining difference in age of onset and severity of disease , about 40% is attributed to genes other than the Huntington gene, while 60% is attributed to environmental factors (Wexler et al.2003). This variation in onset of the disease even in individuals with the same number of polyQ repeats, is indicative of other natural variants that may be present and influencing the disease progression. This brings us to the question of what are the other genetic modifiers cause this variation in disease onset and severity, in patients with the same length of polyQ expansions.

Natural genetic variation determines susceptibility to mutation-induced protein aggregation (Gidalevitz et al., 2013). It is as important to study natural variants present in an individual's background as important as it is to study disease genes themselves. Although genetic manipulations such as knockout/knock-in organisms have provided a wealth of information on stress response pathways and adaptation mechanisms, these gene modifications that are not naturally produced can have adverse effects on an organism and its development and fitness(Jenkins et.al.,2004). Looking at how natural variants affect disease phenotype allows us to use modifiers to our advantage that are compatible with normal development of an organism (Hamilton et.al.,2012). Natural variants in genetic background can dampen or enhance the aggregation of mutant proteins, by enhancing or suppressing pathways involved in the process. Using a model organism such as *C.elegans* allows us to identify these pathways and provides a valuable tool for understanding how variants in genetic background influence disease outcome

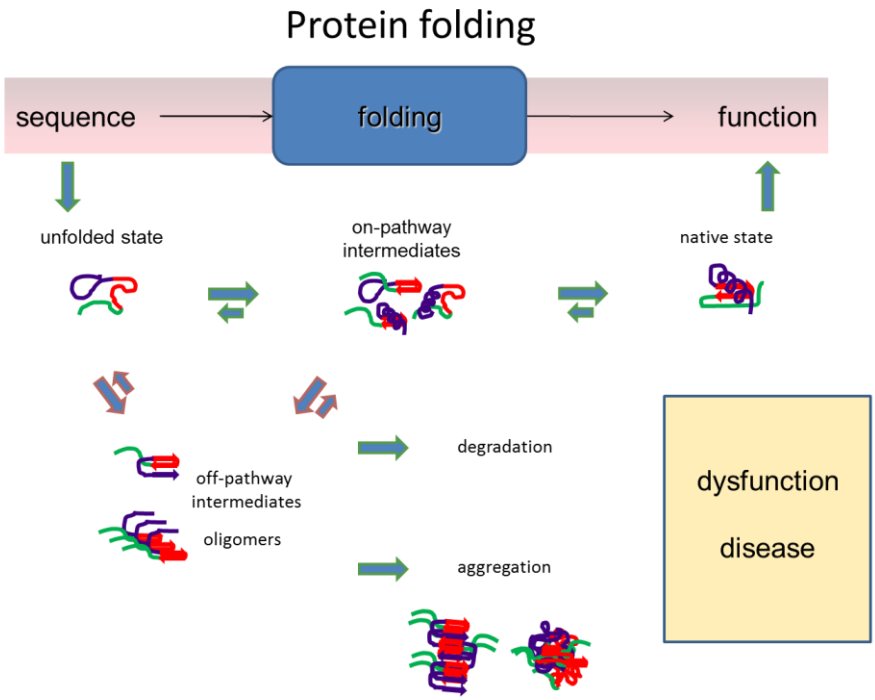

Figure 1: Protein folding. When protein homeostasis is maintained, proteins go from a nacent polypeptide, to a folding intermediate and finally a native protein. Different factors can contribute to disruption of this process, such as environmental stressors, mutations or translational errors. This can lead to misfolded proteins. In ideal circumstances, proteins will either be assisted by chaperones and refolded to their correct state, or degraded. In some situations, proteins can stay misfolded, and in turn form aggregates. This can lead to disease.

(Adapted from Gidalevtiz et al., 2013)

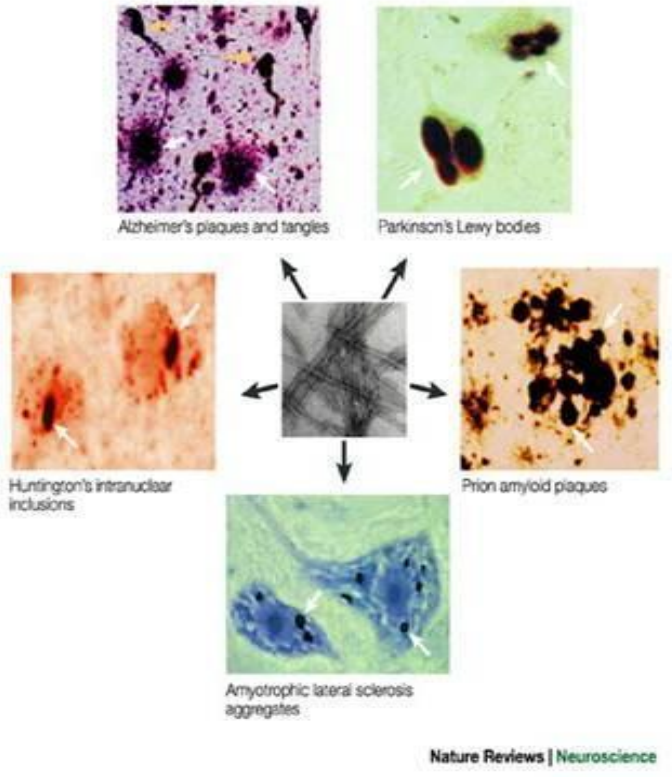

Figure 2: Types of aggregates formed in conformational diseases. Protein aggregation can lead to diseases such as Alzheimer's, Parkinson's, Huntington's, Prion and Amyotrophic later sclerosis. Although each disease has its own associated protein that misfolds and aggregates, the structures of the plaques and aggregates is the same between diseases. These diseases vary highly in their age of onset and severity. They can be sporadic or inherited. (Adapted from Soto CJ, 2003)

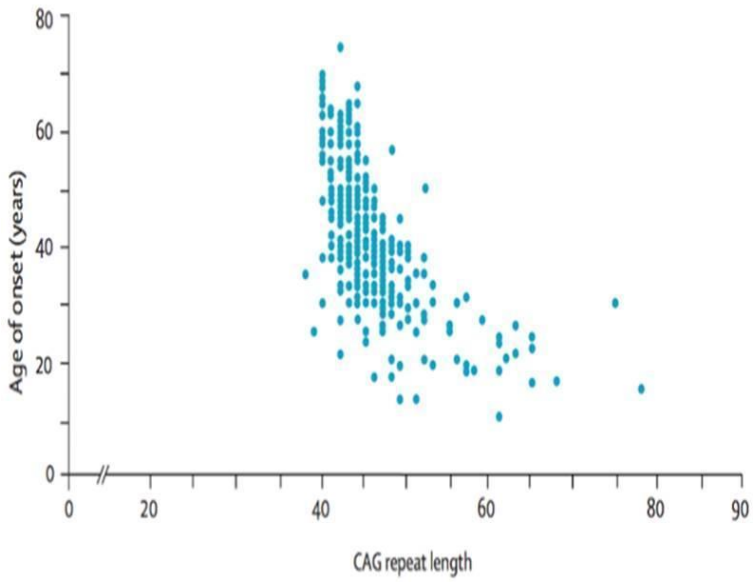

Figure 3:Variation in Huntington's disease. Conformational diseases such Huntington's disease, which is shown here, display high variability in their age of onset, severity and progression. In part this is due to the length of polyglutamine repeats. There are also other factors that are involved in this variability, including other natural genetics variations found in different patients. These naturally occurring genetic variations contribute to the variability in onset shown here, even in patients with the same number of CAG repeats. (Adapted from Wexler et al.,2004)

Caenorhabditis elegans as a model organism

When studying conformational diseases and the influence natural variants in genetic background have on its outcome, it is important to choose a model organism that will allow us to study multiple aspects of the disease. This includes aggregation of proteins, toxicity that is caused by protein aggregation, longevity, and motor abilities, all of which are symptoms associated with protein conformational diseases. *C.elegans* is a very effective model organism for studying conformational diseases (figure 4). Their short generation time, which is approximately 3.5 days, makes it easy to create recombinant inbred lines, allowing you to study effects of recombination and different genetic background on disease. Short life span, with a maximum survival of 30 days, makes the study of lifespan in this system easier than in other systems, such as rodents (Wormbase). The ability to cultivate *C. elegans* in very large colonies allows for easy development of experiments that have statistical significance. Important to the study of protein aggregation, the transparent body wall of *C. elegans* allows for the observation of protein aggregation with simple microscopy techniques in live organism's using GFP-fusion protein to visualize the aggregation process (Markaki and Tavernarakis, 2010), while some other models used for the study of Huntington's disease such as mice and rats, necessitate euthanization to study protein aggregation.

C.elegans are also a valuable tool due to their high genetic variation in different strains found in different parts of the world, as an example Hawaiian genetic distance to N2 being 1.2-2.7 10^{-3} per nucleotide (figure 5). This allows for the creation of genetically

diverse populations, which are key in studying natural variation in genetic backgrounds, and how they affect disease. The *C. elegans* genome has been completely sequenced, and in addition to their genetic diversity, this provides a valuable tool for studying mechanisms that control polyglutamine aggregation Due to this genetic diversity, it is possible to study how natural variants in genetic background contribute to the outcome of disease; and *C.elegans* can be used as a tool to study these natural variants, and their influence on mutant proteins and disease.

The model used to study the influence of natural variants in genetic background have on disease is a Bristol laboratory strain with a 40 polyglutamine expansion. The 40 polyQ expansion was fused to YFP for visualization, with a myosin heavy chain promoter and integrated into a location on the X chromosome. This model has Q40 expressed in the muscle cells.

C.elegans display a threshold for polyQ repeats, much like humans do (figure 6). It has been previously shown that genetic variation among wild isolates of *C. elegans* influences aggregation of polyQ (Gidalevitz et al, 2013) (figure 7).

Figure 4: *C.elegans.C.elegans* are a valuable tool in studying conformation diseases. Their small size and fast generation time makes it easy to study genetic variation influences on disease as well as perform numerous replicates. They are also transparent, which allows for the study of protein aggregation diseases in real time, without euthanization of the animal.

Figure 5: *C.elegans* wild isolates. Different wild isolates of *C.elegans* found in different parts of the world display significant variance in their genetic backgrounds. (Adapted from Wormbook).

Recombinant inbred lines in the study of natural variation in disease

A panel of 21 recombinant inbred lines (RILs)- which are organisms from the same parental strains, but through independent recombination followed by self reproduction, result in a panel of strains with different recombinant genetic backgrounds- were created by random selection of random F2 progeny from a cross of a wild isolate DR1350 and a laboratory strain Q40Bristol, also displayed a large variation in polyQ aggregation (Gidalevitz et al, 2013)(figure 8 and 9). This shows that genetic background has a large influence on the aggregation patterns in organisms containing the same 40 polyglutamine expansion. Independent recombination events can lead to different genes being inherited from either parent, resulting in mixed genetic background, and different aggregation phenotypes. Of particular interest to this project came a set of RIL's that display high head aggregation and high body aggregation, as well as a line that had complete suppression of aggregation, neither of which is consistent with any parental aggregation patterns (figure 9 and 10). RIL 2 was chosen as a representative to study this phenomenon of high aggregation in the head muscle cells (figure 9). These two strains, one with very high aggregation in head muscle cells and one with very suppressed aggregation, indicate that there are natural variants that can occur in genetic background that influence disease progression in organisms with the same length of polyglutamine repeats.

Natural variation in genetic background plays an important role in the onset and severity of disease. It is important to understand how an individual's genetic makeup can influence disease progression. In this study, we look at how genetic background can influence protein aggregation and how to determine whether single modifiers or networks of genes are responsible.

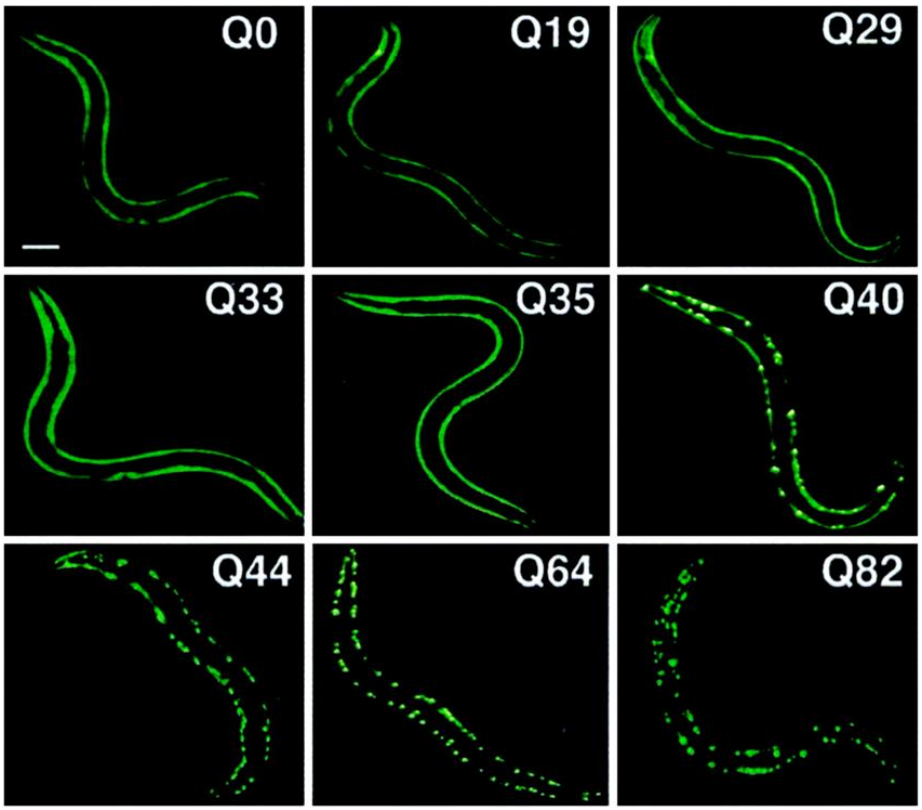

Figure 6: Threshold for polyglutamine aggregation. The threshold for polyglutamine aggregation as shown here is 40 CAG repeats. Lower than 40 repeats, we only see soluble protein. (Adapted from Morley et al.,2002)

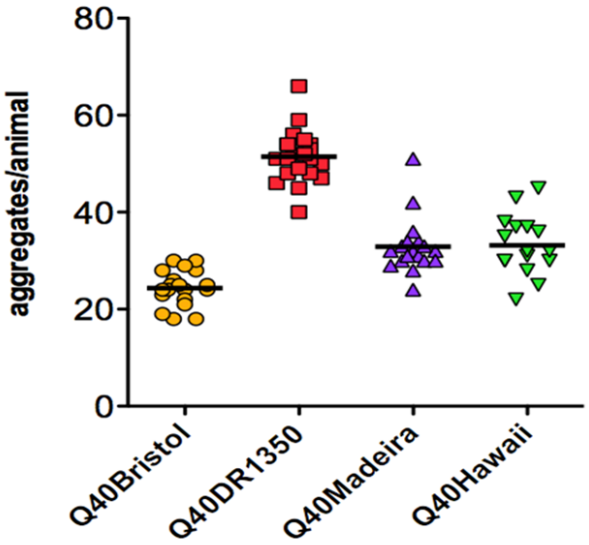

Figure 7: Different wild isolates display different levels of aggregation even with the same number of CAG repeats. This is indicative of other natural variants present in their genetic backgrounds contributing to the effects of long polyglutamine expansions. (Adapted from Gidalevitz et al.,2013)

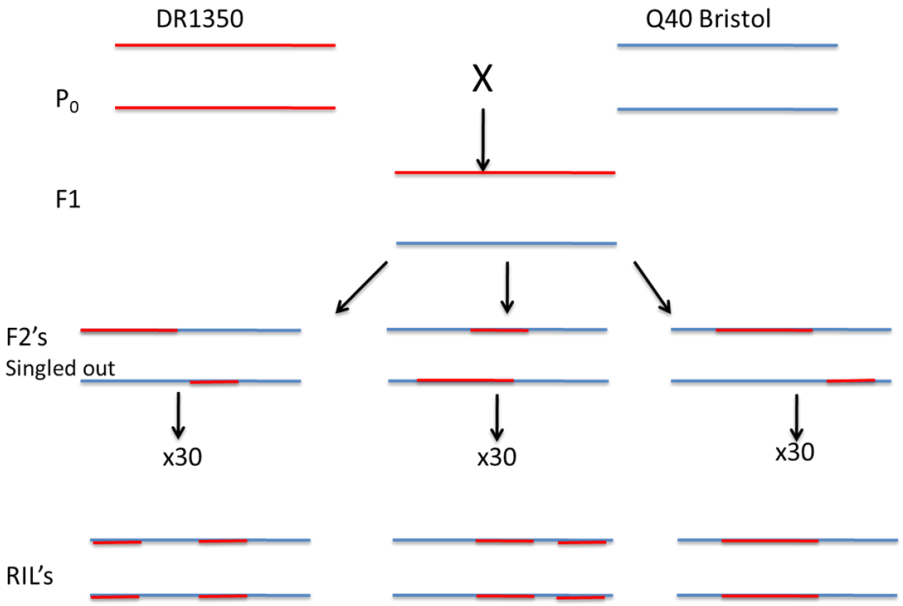

Figure 8: Wild isolate DR1350 was crossed to laboratory strain Q40 Bristol. Random F2's were singled out to create 21 lines, each would display a different assortment of parental loci. Lines were allowed to go through 30 cycles of self reproduction to homozygose at each locus, creating a panel of 21 Recombinant inbred lines (RILs).

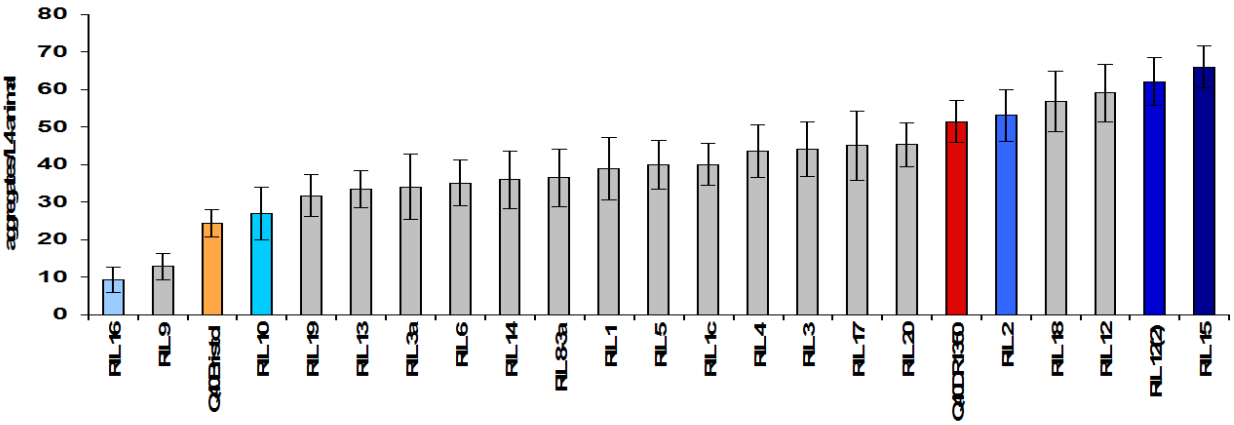

Figure 9: 21 RIL's show variation in protein aggregation. Some show aggregation patterns well below their parental phenotype, some well above. Of particular interest are RIL's 2, 18, 12, 12(2) and 15, which fall on the high end of the spectrum. These 5 RIL's display early onset of aggregation in head muscle cells, not found in any of the other lines. 5/20 displaying this phenotype is suggestive of a single recessive locus being responsible for this phenotype. (Adapted from Gidalevitz et al.,2013)

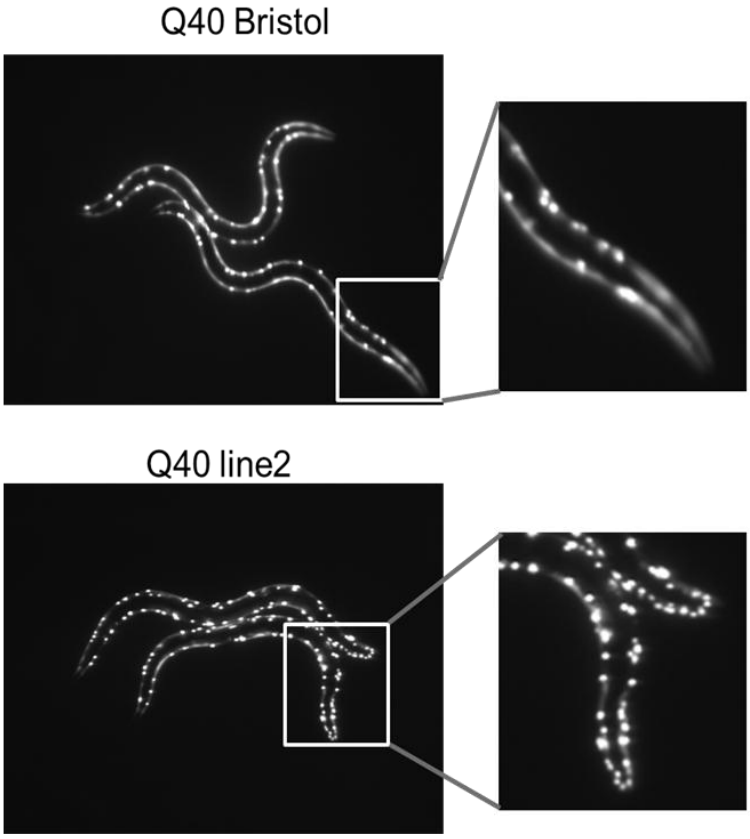

Figure 10: RIL 2 and Q40Bristol. RIL 2 displays high aggregation of head muscle cells with early onset, this phenotype is not present in its parent Q40 Bristol. (Gidalevitz et al.,2013)

Results

Previous work

Results described in this section are contributed by Jasmine Alexander-Floyd and Gowri Gouda

It was initially important to identify whether the presence of high polyglutamine aggregation in the head muscle cells was caused by genetic variation at a single locus, or by a network of genes. This is important because mapping a single locus can be done with mapping strains and SNP mapping, while mapping a network or genes would involve different methods to determine how multiple genes work together to produce one phenotype. In the initial panel of recombinant inbred lines, 5 out of 20 RIL;s displayed high aggregation in head muscle cells; since all of these were a result of independent recombination events, and the expectation would be that different recombination events would lead to differing phenotypes, it was suggestive of a single recessive locus being responsible for the high aggregation in head muscle cells.-To confirm this initial observation, RIL 2 was backcrossed to its parental strain, Q40 Bristol, and by looking at patterns of inheritance, we could see a 3:1 inheritance ratio in F2 progeny; where 75% showed Q40Bristol like phenotypes (1,629), and 25% showed high aggregation in the head muscle cells (528) (Jasmine Alexander-Floyd, unpublished observations) (figure 11). This again indicated the presence of a single recessive locus responsible for the early onset of aggregation in head muscle cells.

Next the location of this locus was mapped to roughly the center of chromosome I using mapping strains. This was done by using a strain with a single marker on each chromosome to identify what chromosome the modifier lies on and roughly where on the chromosome. Then, using two point mapping by crossing RIL 2 to a double mutant mapping strain with phenotypic markers in different areas of chromosome I, it was attempted to map the modifier further. Markers unc-11, at 3798566bp, and dpy-5, at 5432158bp, were used for the left are and markers dpy-5, at 5432158bp, and unc-13, at 7422785bp, were used for the right arm of the chromosome. The distribution of phenotypes was scored in F2 progeny after crossing RIL 2 into one of these strains. F1's were selected and then the F2 progeny were scored for unc non-dpy and dpy non-unc phenotypes. These phenotypes were scored because they occur based on recombination events that would help determine if the modifier was in between the two markers. Inconsistent results suggested there were genetic factors in this region affecting recombination frequencies. This could be explained, for example, by the presence of small deletions or duplications in the genomes of the parental strains. Some results were suggestive of it possibly lying toward the left side of the chromosome, but because this was not consistent, the strain was sent for sequencing.

Prior to sequencing, to isolate the locus that contributes to head muscle cell aggregation, and filter out variants that do not contribute to this phenotype, RIL 2 was backcrossed to its parental strain Bristol 23 times. This was done with selection for the head muscle cell aggregation phenotype at each F2 generation, using the following crossing scheme: RIL 2 hermaphrodites were set up to mate with Q40 Bristol males; because we have previously determined that RIL2 phenotype was recessive, F1 cross

progeny was identified by having a Q40Bristol like phenotype. F2 progeny were selected based on the RIL2-like phenotype - high aggregation in the head muscle cells. This was repeated 23 times. The goal was to isolate as small of a donor fragment containing causative variants, in a predominantly wild type Bristol background (figure 12) (Gidalevitz et al.,2013). RIL 2 (23 times backcrossed) was sequenced using whole genome high throughput sequencing at the Wistar institute. After sequencing, a 1.5Mb region of interest between 832kb and 2.22Mb on the left arm of chromosome I was identified using Cloudmap through Galaxy (Minevich et al.,2012). This region was found to contain high density of DR1350-derived SNPs (approximately 300 changes/100kb) (figure 13), and was located on the left arm of chromosome I. Since this location was consistent with the three-point mapping results, it was chosen as a candidate locus. A list of potential genes found in this area was created by choosing SNPs that are in regulatory regions and thus could control the expression of genes, as well as non-synonymous SNPs, resulting in approximately 100 genes. Some genes are of particular interest such as *moag-4*, which specifically modifies amyloid aggregation, including in Q40 *C. elegans* model, by interaction with early amyloid precursor (TJ van Ham - 2010 Cell Falsone et al.,2012). A few others of particular interest in this region are SHC1 (Src Homology 2 Domain Containing transforming protein 1) that is important for regulation of apoptosis, and interacts with the insulin-like signaling pathway to influence stress response and longevity; it encodes three isoforms, one of which may be involved in lifespan regulation and stress response to reactive oxygen species, while the other two act on the Ras pathway, which activates genes responsible for cell growth, differentiation and survival. *atg-5* (Autophagy related 5) is another gene of importance and plays a role in building

autophagosomes, necessary for autophagy and aberrant protein disposal; ubc-3 encodes for an E2 ubiquitin ligase (homologue of yeast and human CDC34) and is also important in protein degradation and ubiquitination.

After the region of interest was identified, further backcrossing was performed to attempt to decrease the size of the donor fragment. During this backcrossing, a series of lines were isolated based on differing phenotypes. These were given their own names and used as independent lines for further mapping (figure 14). This was important because different phenotypes that were rare, could indicate rare recombination events that could possibly have occurred in the donor fragment during backcrossing. These strains could potentially play valuable roles in shortening the donor fragment and decreasing the list of potential genes contributing to the high aggregation in the head muscle cells. Among those isolated out were two strains that were smaller, less reproductive and had much higher body aggregation. These were named 8a and 8b. 8a maintained the high head aggregation phenotype, but was more highly aggregated than the original RIL 2, while 8b lost the high head aggregation phenotype, but maintained the high body aggregation. This was of great value because any overlap in the donor fragment found in 8b and other RIL's could be eliminated from the gene list. This is because 8b was isolated from RIL 2, but lost its early onset of aggregation in the head muscle cells, in turn losing the modifier responsible for this phenotype. Line #7 is RIL 2 continuously backcrossed and currently on backcross 33. Line #9 was isolated out for being slightly more aggregated, and slightly smaller than line #7, but still maintaining the phenotype.

Current work

Following identification of this 1.5Mb region, we attempted to narrow down the size of the donor fragment by SNP mapping protocol (Davis et al.,2005). It was necessary to narrow down the list of potential target genes to be able to take a deeper approach to study candidate genes and their effect on the phenotype and pathways they are involved in.

The SNP mapping approach was used to try to find the breakpoints of the donor fragment based on data compared from a number of strains. RIL's 12, 12(2), 15 and 18 were tested to determine if their donor fragments were smaller in size compared to RIL 2. This was done because all these lines display increased head muscle cell aggregation pattern, but acquired this phenotype through independent recombination events. In addition to RILs, four strains generated during backcrossing were also used – designated as lines 7, 8a, 8b, and 9. A combination of data from all of the RIL's and lines created through independent recombination events may allow us to narrow down the size of the donor fragment, and narrow down our list of target genes.

Primers were designed to cover the candidate region between 832kb and 2.22Mb on the left arm of chromosome I. We selected 8 primer pairs, each designed to amplify a small region containing a SNP within the donor fragment that would indicate what background is present in this location, either DR1350 or Bristol (figure 15). RIL 2 (23x backcrossed) and Q40 Bristol were used to test these primers and optimize PCR and restriction digest conditions. For example, primer pair 1 was designed to amplify genomic fragment from 832228bp to 832907bp that contained a SNP at position 832671. This SNP results in a loss of the recognition site for ApoI restriction nuclease in DR1350

background; a resulting PCR fragment will be susceptible to the digest with ApoI if amplified from Bristol background, and resistant if amplified from DR1350 background (figure 16). Genotyping performed using primer pair 1 revealed that all lines - RIL's 12, 12(2), 15 and 18 as well as lines 7, 8a and 8b, have a DR1350 background present in this location. This can be seen by the generation of bands consistent with the sizes expected from the loss of a restriction site - 331bp and 175bp (figure 17). This is also the case for primer pair 3, where a loss of a restriction site, resulting in a 603bp band, indicates a DR1350 background present in this position (figure 18). This shows that none of the RIL's or the strains isolated during backcrossing have a shorter donor fragment 5' side of the donor fragment since all show genotyping results consistent with DR1350 genetic background.

Genotyping performed using primer pair 8 revealed digested PCR products at around 278bp and 206bp (figure 19), consistent with DR1350 background, for all lines tested except RIL 15 and line 7. RIL 15 and line 7 appeared to have a Bristol background at this position (figure 19, lanes 7 and 11, respectively), since there is a loss of a restriction site in the Bristol background, resulting in no digest of the PCR product. This was an important result because it indicated further genotyping should only be performed on RIL 15 and line 7. It also helped conclude that although all RILs were the result of independent recombination, they inherited atleast the same size donor fragment as RIL 2,indicating the importance of the genes within this region.

Since the interpretation that RIL 15 and line 7 have shorter donor fragments is based on the lack of restriction digest, it was treated as a suggestive result, because if conditions are not optimal, a failure of digest can occur and give a false result in this

situation. Therefore, primer pair 7 was designed, 240 Kb 5' to the primer pair 8 location. This would not only confirm the result, but it would also show whether the region is even smaller, allowing us to narrow down the list of possible gene targets further. After this genotyping, it was evident that RIL 15, line 8b, and line 7,all had a shorter donor fragments (figure 20, lanes 7, 8, and 11, respectively). This was based on the digest of the PCR product and products around 227bp and 145bp, consistent with Bristol background.

To map the 5' boundary between the DR1350 and Bristol backgrounds in the lines with the shorter donor fragment, the RIL's and line #7 were tested for primer pair 5, which is roughly in the center of the donor fragment. This genotyping showed no digest of the 743bp PCR products, indicating that all RIL's have a DR1350 background present in this location (figure 21). This suggests that the breakpoint of the donor fragment is somewhere between the locations of primer pairs 5 and 7.

After this result, a new primer pair was designed that falls roughly in between primer pairs 5 and 7, at 1.81Mb. This primer pair was used to test RIL 15 and line 7, because they have been shown to have shorter donor fragments, as well as line 8b to determine if there was any overlap. The results indicate both the lines have a DR1350 background in this location, now putting our donor fragment breakpoint between 1.81Mb and 1.98Mb. (Figure 22). For line 8b, the results indicated that there was Bristol background in this location, resulting in an even shorter donor fragment line. These results eliminated approximately 20 genes from the 3' side of the donor fragment, After determining where the breakpoint of the donor fragment is for line 8b, this will allow us to cut out even more genes. Ideally the breakpoint will fall between closer to 1.64Mb

primer, with a large overlap between RIL 15, line 7 and line 8b. If the breakpoint falls closer to the primer pair at 1.17Mb, this will result in a smaller overlap, with a larger number of genes to test further.

Is suppression of aggregation a result of a single modifier or a network of genes?

An interesting question is whether similar modifiers and mechanisms are used in suppressing aggregation, as in causing high aggregation. To look at this, in addition to mapping the single recessive locus responsible for high aggregation in head muscle cells, we took a brief look at the suppression of aggregation in RIL's and focused on RIL 16 (Figure 9) to determine whether it was controlled by a single recessive locus as well, or if there were more players. RIL 16 has suppression of aggregation and has the lowest aggregation of all the RIL's, with approximately 1 aggregate present at late L4 stage, while Q40 Bristol will have over 10 (figure 23). It is interesting to look at whether there are any similar underlying genes controlling both high aggregation of head muscle cells, and suppression of aggregation. These experiments were performed to determine this.

We first wanted to test whether the suppression of aggregation is a result of a single modifier or a network of genes. RIL 16 was crossed to Q40 Bristol to determine how the phenotype is inherited, much like RIL2 was. In the F1 generation we found there was moderate suppression of aggregation, and no progeny showed high aggregation like the Q40 Bristol parent. This indicates a dominant inheritance, which was not seen in RIL 2. F2 progeny were then singled out and allowed to self reproduce. In the F2 progeny we detected large variability in the aggregation phenotypes. There were 7 out of 20

progeny in F2 generation that showed RIL 16 like suppression of aggregation, which initially indicated a potential recessive locus involved in suppression of aggregation (figure 24). But when F2's were allowed to self reproduce, we saw again a large variation in their phenotypes. Suppression of aggregation in F2 was not an indication that there would be similar suppression in their self progeny (figure 25). This is suggestive of a network of genes involved in regulating suppression of aggregation, and that multiple variants need to be inherited in order to see this protective phenotype. These results indicate there probably is not much overlap between the modifiers controlling suppression of aggregation and the modifier controlling high aggregation of head muscle cells.

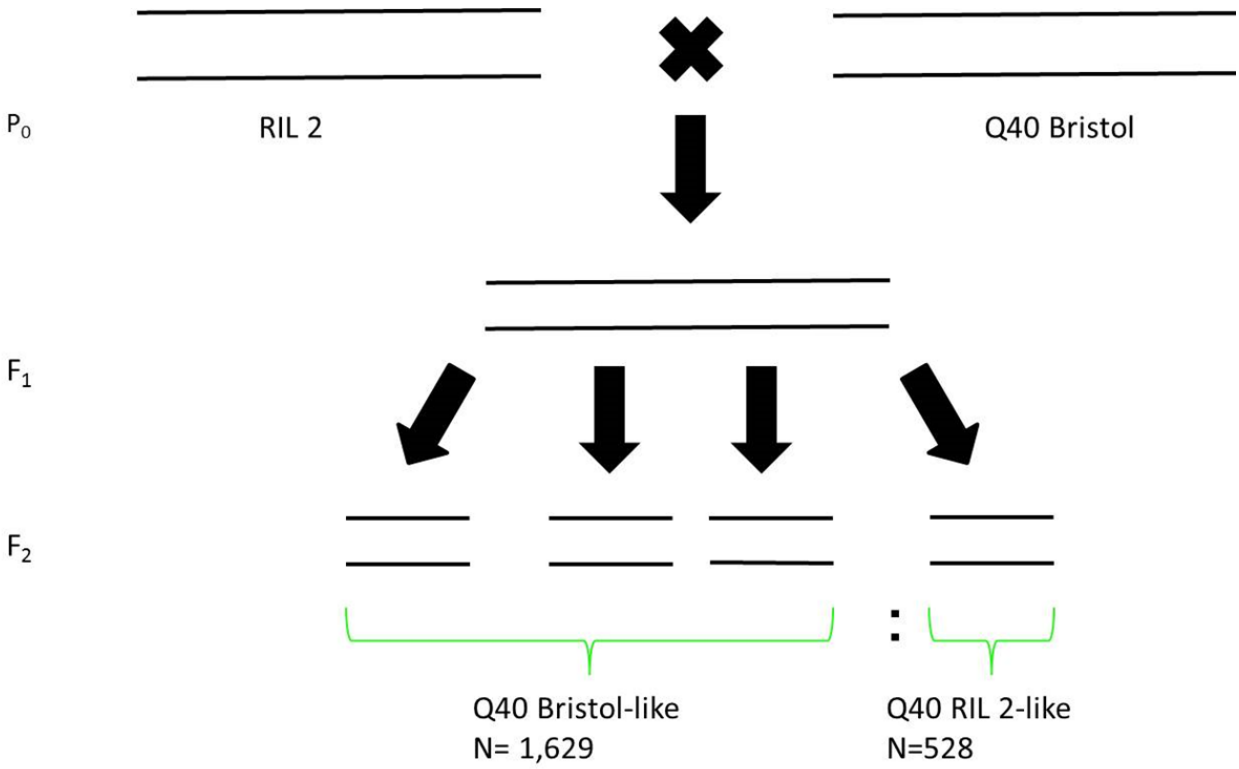

Figure 11: High aggregation of head muscle cells is controlled by a single recessive locus. RIL 2 was crossed back to its parental strain, Q40 Bristol, to determine what the inheritance was for the RIL 2 phenotype. F1 progeny that were heterozygous were allowed to self-reproduce, F2's were counted. With the resulting number, it is evident we have a 3:1 segregation, much like it was seen in the panel of 20 RIL's. This is suggestive of high aggregation in head muscle cells being controlled by a single recessive locus. (Jasmine Alexander-Floyd)

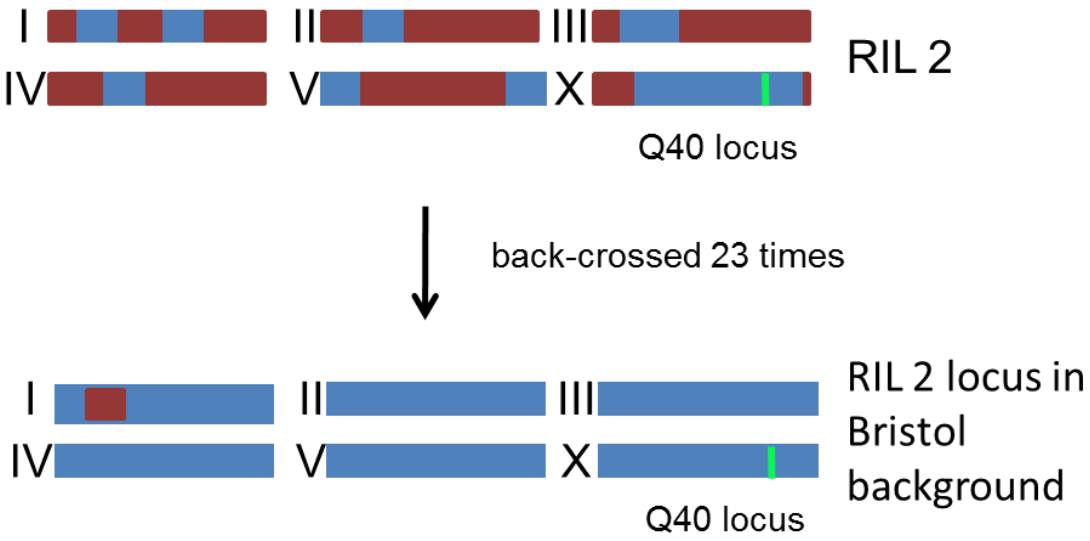

Figure 12: Backcrossing RIL 2. RIL 2 was backcrossed 23 times to filter out natural variants not contributing to its phenotype and isolate the locus responsible for the early onset aggregation in the head muscle cells in a predominantly Bristol background.

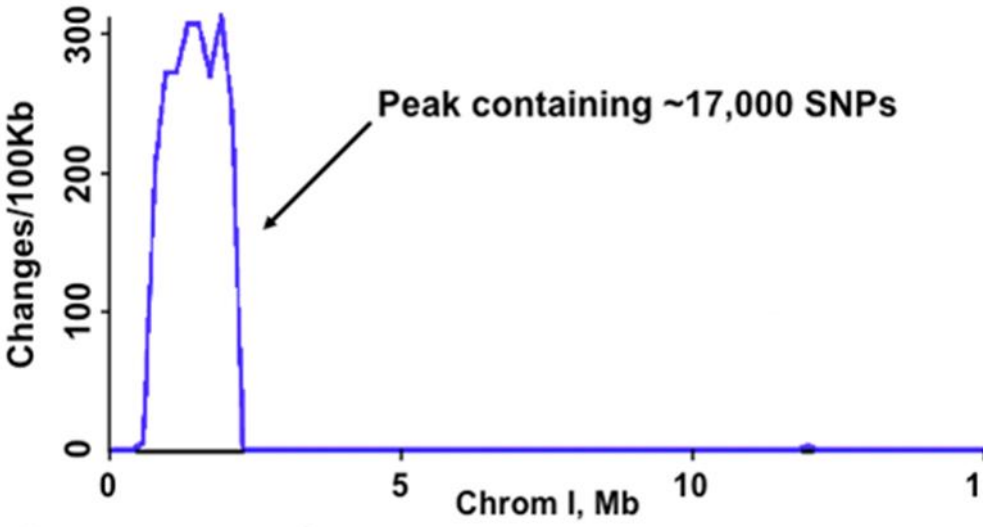

Figure 13: Sequencing results of RIL 2. Through sequencing and analysis using Cloudmap in Galaxy, a 1.5Mb region of interest was identified on the left arm of chromosome I. This region contains ~300 changes/100kb. (Adapted from Jasmine Alexander-Floyd)

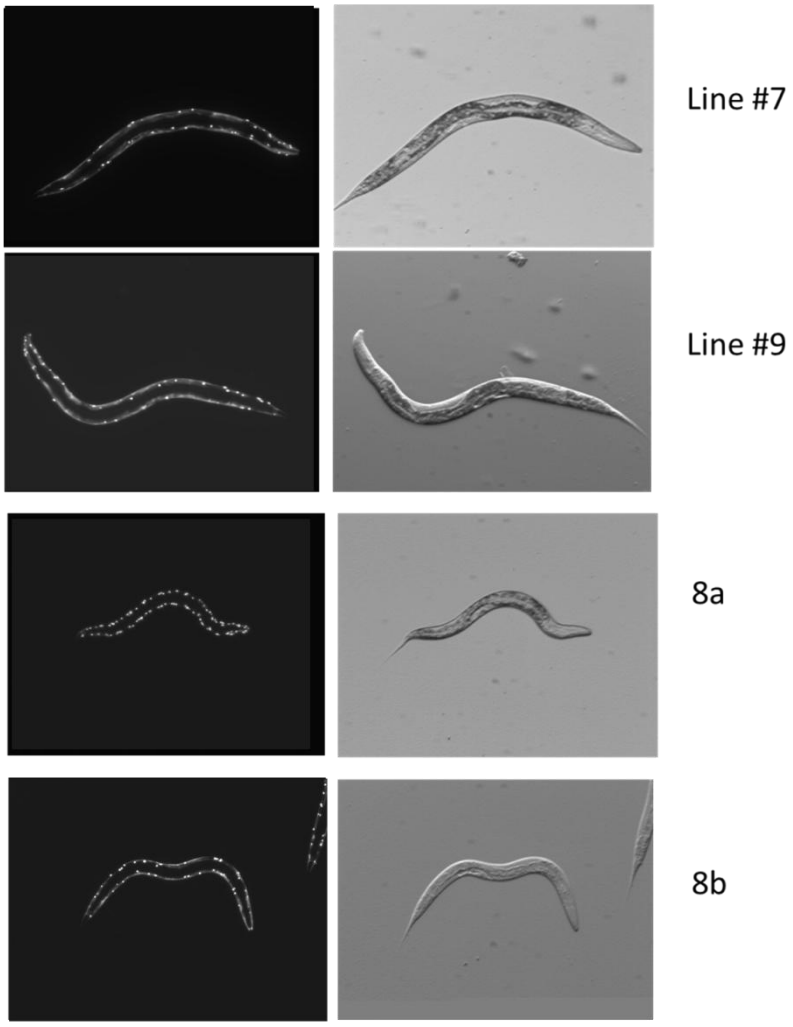

Figure 14: Independent lines isolated during backcrossing RIL2.
Independent lines were isolated during further backcrossing of RIL 2 (23xbackcrossed) in an attempt to decrease the size of the donor fragment.

	WBVar	Position-wormbase	product size	enzyme	PCR temp	digest temp	products N2	products DR
1	WBVar00004591	832671	679	ApoI	68	50	263,175,88,68,53,32	331,175,88,53,32
2	WBVar01529903	846351	612	BmpI	60	37	233,139,60	293,139,
3	WBVar01529977	944150	603	AlwNI	60	37	405,198,	no digest
4	WBVar0014396	1176473	158	BsrI	60	65	86,72,	no digest
5	WBVar00016276	1647221	743	SalI	60	37	432,311,	no digest
6	WBVar00017051	1850249	132	AvaII	60	37	no digest	98, 32
7	WBVar00017401	1981603	372	ApoI	60	50	227,145,	152,145,72,
8	WBVar00023136	2222271	484	AlwNI	60	37	no digest	278,206,

Figure 15: Primers spanning the 1.5Mb region of interest in RIL 2. A series of primers were designed to span the region of interest in RIL 2 that contains the recessive modifier contributing to its phenotype. Each primer pair amplifies a region that gains or loses a restriction site, creating a different pattern of digest, or complete loss of digest. This allows us to determine whether a DR1350 or Bristol background is present in that location.

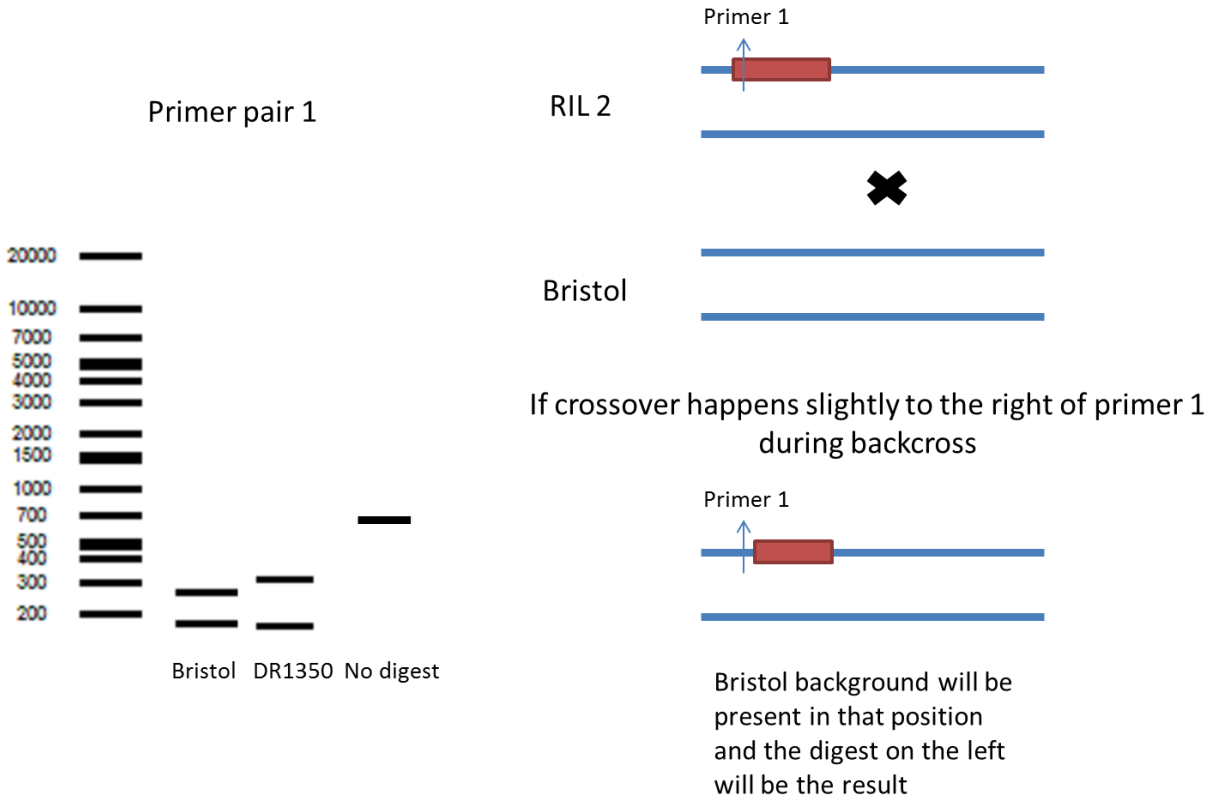

Figure 16: An example of the results for genotyping using primer pair 1. If crossover happens slightly to the right of primer pair 1 during backcrossing, you will have the Bristol genotyping result. If it does not, you will have a DR1350 genotyping result.

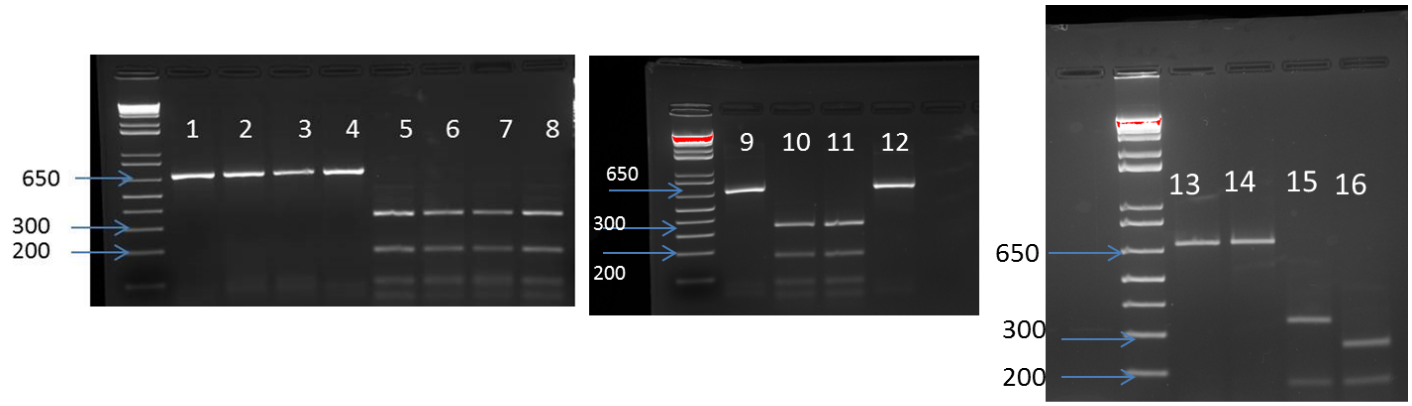

Figure17: Genotyping data for primer pair 1. Lanes 1-4 show undigested PCR products for RIL's 12, 12(2), 15 and 18. Lanes 5-8 are digested PCR products with bands consistent with DR1350 background. Lanes 9 and 12 are undigested PCR products for line #7 and 8a, and lanes 10-11 are digested PCR products with bands consistent with DR1350 background as well. Lanes 13-14 are undigested PCR products for Q40Bristol and 8b. Lanes 15-16 are digested PCR products, Q40Bristol consistent with Bristol digest patterns and 16 consistent with DR1350 background.

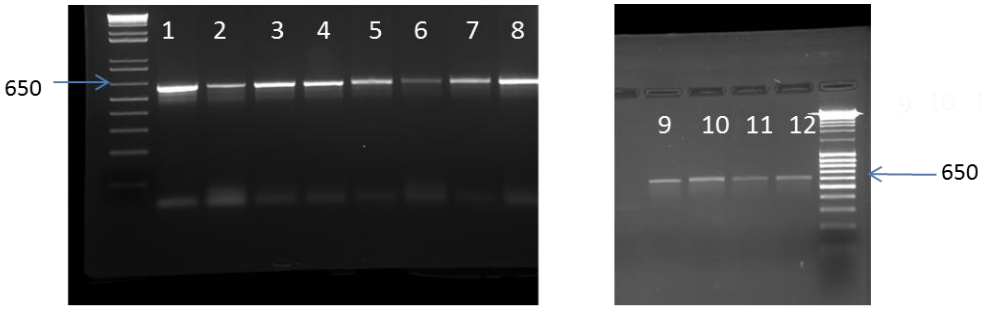

Figure 18: Genotyping data for primer pair 3. Lanes 1-4 show RILs 12, 12(2), 15 and 18 undigested PCR products, and lanes 5-8 and 11-12 show digested PCR products. A loss of a restriction site, resulting in no digest of the PCR product is suggestive of a DR1350 background present in this location

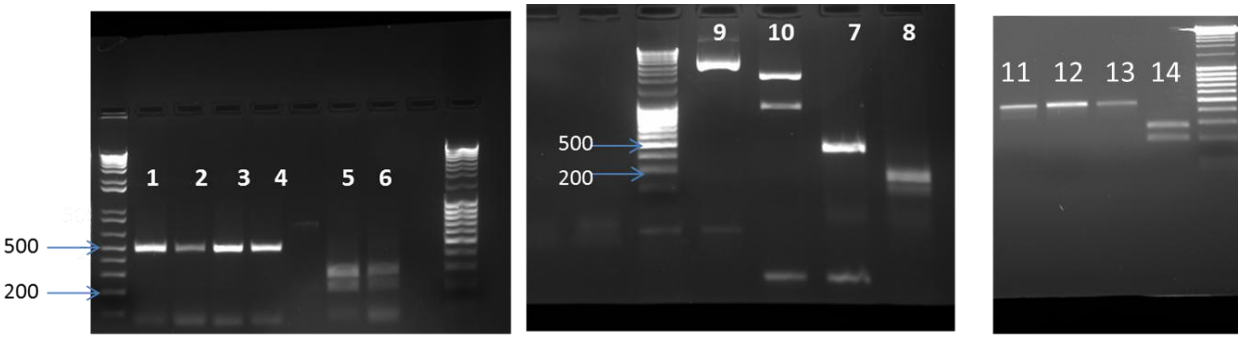

Figure 19: Genotyping for primer pair 8. Lanes 1-4 and 11-12 show undigested PCR products for RILs 12, 12(2), 15 and 18. Lanes 5-8 and 13-14 show digested PCR products. Of particular interest are lanes number 7 and 13, which show no digest of the PCR product. A loss of a digest site at this location is suggestive of a Bristol background at this location

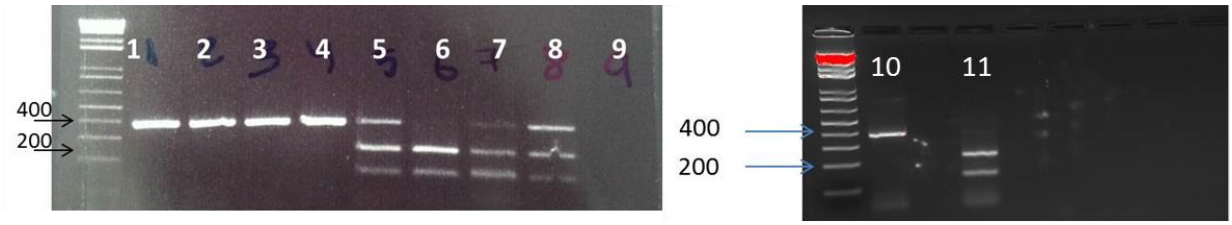

Figure 20: Genotyping for primer pair 7. Lanes 1-4 represent undigested PCR products for primer pair 7, using Q40 Bristol in lanes 1 and 2 with different DNA concentrationsRIL 15 in lane 3 and line 8b in lane 4. Lanes 5-8 represent digested samples in the same order, Lane 9 is a no template control. Partial digests can be seen in 3 of the samples, where the PCR product was not entirely digested by the enzyme. Lanes 10 is undigested PCR product for line #7 and lane 11 is digested PCR product for line #7,these show line 7 also has a Bristol background present in this position

Figure 21: Genotyping for primer pair 5. RIL's 12, 12(2), 15 and 18 were genotyped for primer pair 5. Lanes 1-4 show undigested PCR products. Lanes 5-8 show digested PCR products. Loss of a restriction site indicates a DR1350 background present in this location for all RILs.

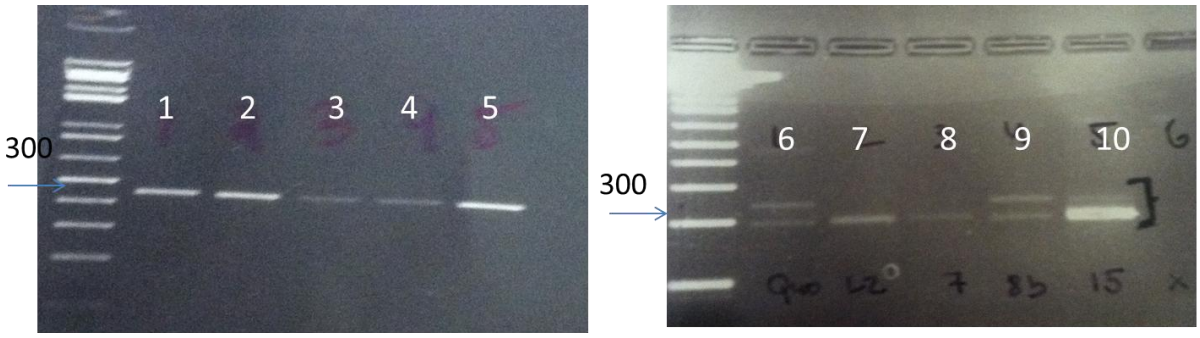

Figure 22: Primer pair 1.81 genotyping. Q40 Bristol, RIL 2, line 7, 8b and RIL 15 were genotyped for the new primer pair at position 1.81Mb. Lanes 1-5 are undigested PCR products of these lines .Lanes 6-10 are digested products. Digests in lanes 7,8 and 10 indicate a DR1350 background in these positions, while digests in lanes 1 and 4 indicate Bristol background in this position.

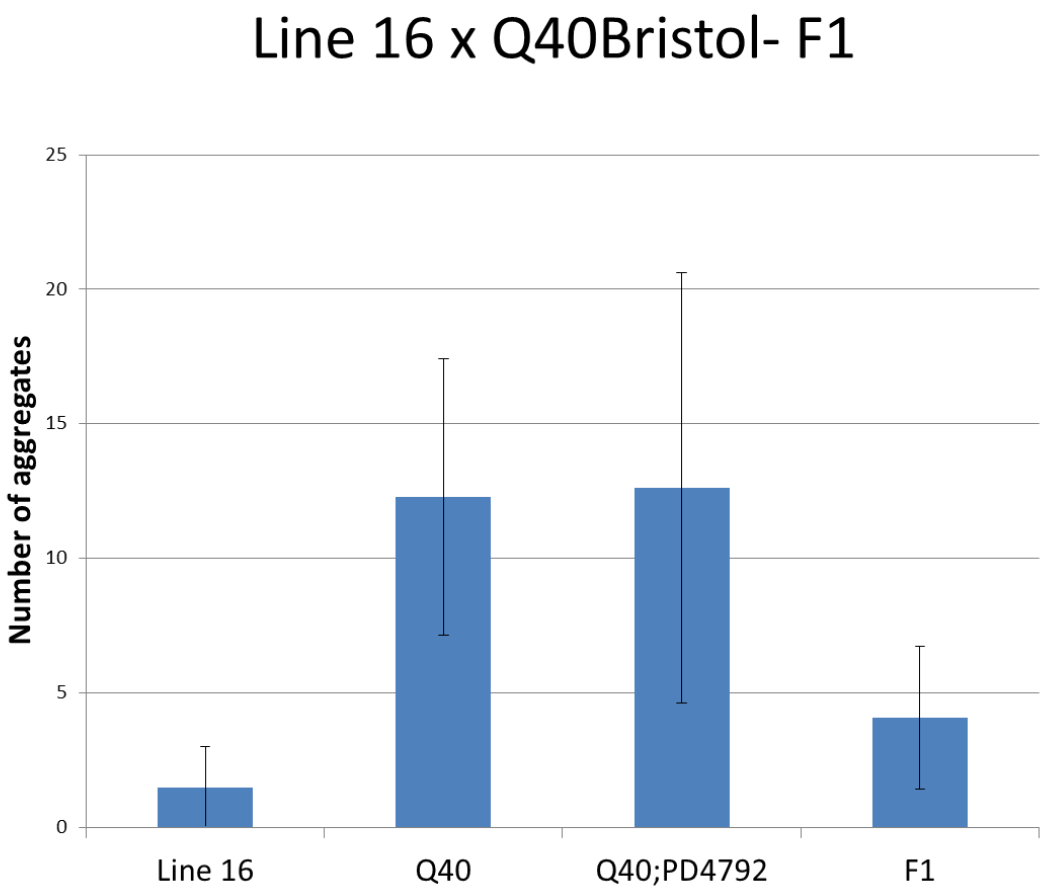

Figure 23: F1 generation in a cross between Q40 Bristol and RIL 16 show moderate suppression, with no progeny displaying high aggregation phenotype seen in Q40 Bristol. Q40;PD4792 was a strain used with a neuronal marker to help identify cross progeny.

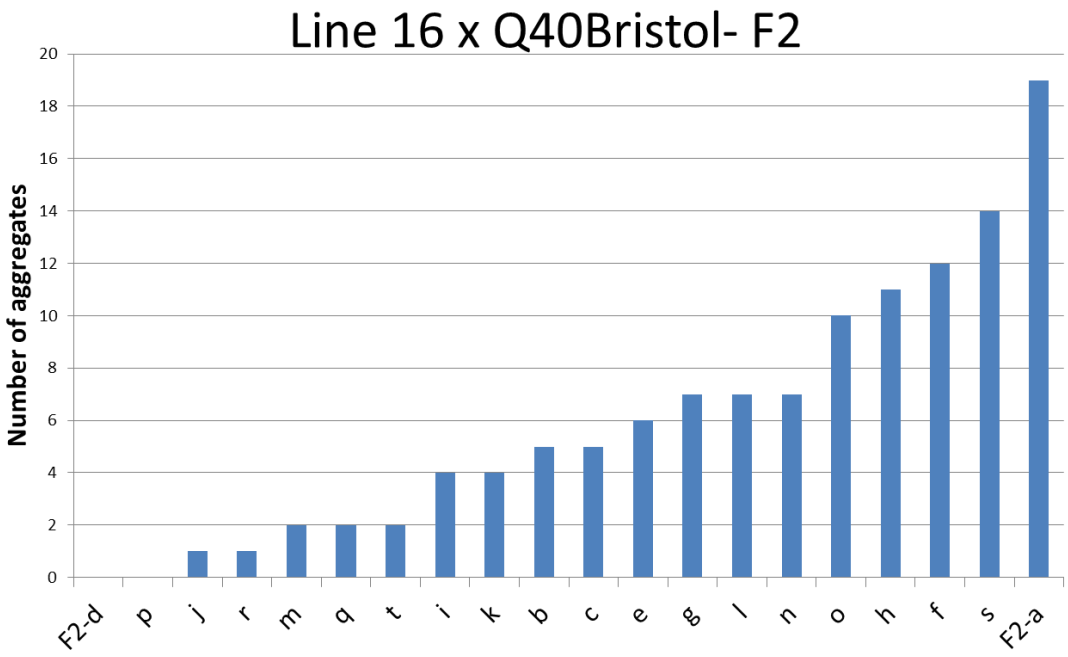

Figure 24: F2 progeny from experiment in Figure 23 were singled out and scored for aggregation. There is a wide range of phenotypes seen from complete suppression to very high aggregation. 7 out of 20 nematodes had suppression of aggregation similar to that seen in RIL 16.

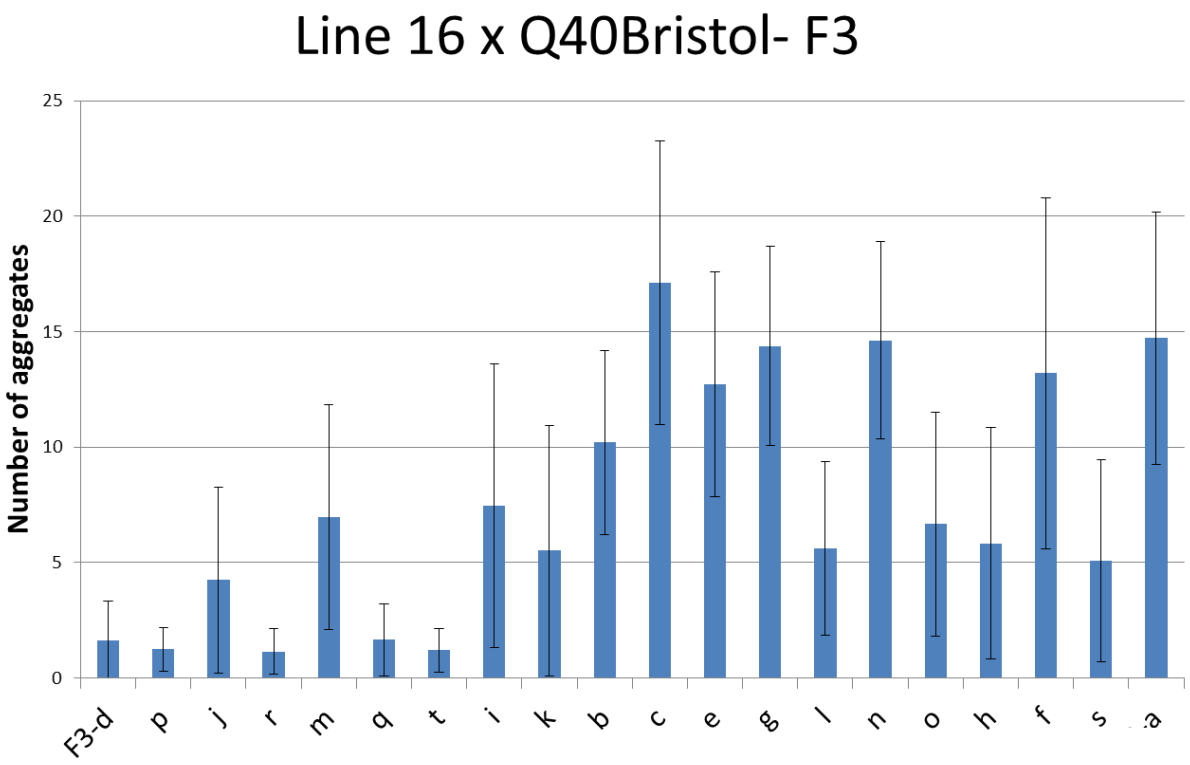

Figure 25: F2 animals from Figure 24 were allowed to self reproduce and F3 progeny were scored for aggregation. Data are arranged as in Figure 24, from F3 progeny of F2 parents with lowest aggregation on the left to those from F2 parents with highest aggregation on the right. A wide range of phenotypes is seen again, and suppression in F2 is not an indication of suppression in F3. This is suggestive of a network of genes that need to be inherited together to see suppression of aggregation.

Summary and future experiments

As a result of SNP mapping and attempting to find the location of the single recessive locus that is believed to be responsible for early onset of aggregation in head muscle cells, we have narrowed down the region of interest to a smaller donor fragment, and cut the list of genes to between 25 and 53 that are of interest, based on determining a small region where the breakpoint of the donor fragment lies for RIL15 and line 7, between 1.81Mb and 1.98Kb, and where the donor fragment breakpoint for line 8b lies between 1.17 Mb and 1.64Mb.

Currently the lines that are of particular interest are RIL 15 and lines 7 and 8b. These are of interest because RIL 15 and line 7 both have shortened donor fragments, but maintain the high aggregation of head muscle cells phenotype. While line 8b was a result of backcrossing and has lost its high head aggregation, but still has at least part of the donor fragment present (figure 26). This allows us to eliminate any genes present in overlap with line 8b. Depending where the breakpoint of the donor fragment lies, this allows us to cut our list of genes down to between 27 and 55 genes, bringing the list of potential genes that to between 53 and 25. This brings us to a much more manageable list of genes. Further fine tune mapping by sequencing will be necessary to find the true breakpoints of these fragments, and determine what genes are left that are of particular interest. After identification of a shorter list of genes, other experiments can be carried out to determine which gene in the list is responsible for high aggregation of head muscle cells. This can be done through a series of different experiments such as RNAi to

downregulate target genes. This would allow us to see if downregulation of a specific gene from the narrowed down gene list would cause the high head aggregation phenotype in Q40Bristol. By using Q40Bristol in RNAi experiments we would be able to determine if certain genes, from our target gene list, with a loss of function are the reason for high aggregation in head muscle cells. Alternatively, experiments can be performed using cosmid injections into the RIL 2 with DR1350 background to see if it is possible to rescue the high aggregation phenotype. These approaches assume that the modifier is not part of a larger network of genes within a single locus. If it turns out a single modifier does not influence the phenotype on its own, it will be necessary to further test pathways it can be involved in, and what other genes are present in this pathway that work together to produce this phenotype.

Following identification of the modifier, experiments need to be performed to determine whether the effects of this modifier are unique to polyQ, or whether it will have the same effects on other proteins that cause aggregation in other conformational neurodegenerative diseases, such as Amyloid-β in an Alzheimer's model. Identification of this modifier would lead us to a better understanding of how networks of genes present in an individual's background work together to influence the phenotypic outcome of a disease. This is valuable information to understand why individuals that have the same number of CAG repeats, but different age of onset and severity of disease.

For RIL 16, now that we are aware of a complex inheritance being responsible for the suppression of aggregation, further mapping needs to be performed to determine what chromosome/chromosomes these genes lie on, and perform linkage analysis to get further in identifying them. This is of great importance because this shows that we can model

diseases with complex inheritance in *C.elegans.* A large number of diseases have a complex genetic inheritance or, like in the example of Huntington's disease, show significant influence of the individual's genetic background, despite in itself being a monogenic disease. This is critical in moving research in this area forward and we can take this model system beyond identifying single target genes. Further research using this line would allow us to identify methods we can use to map complex inheritance in diseases and determine how networks of genes work together to create a disease phenotype.

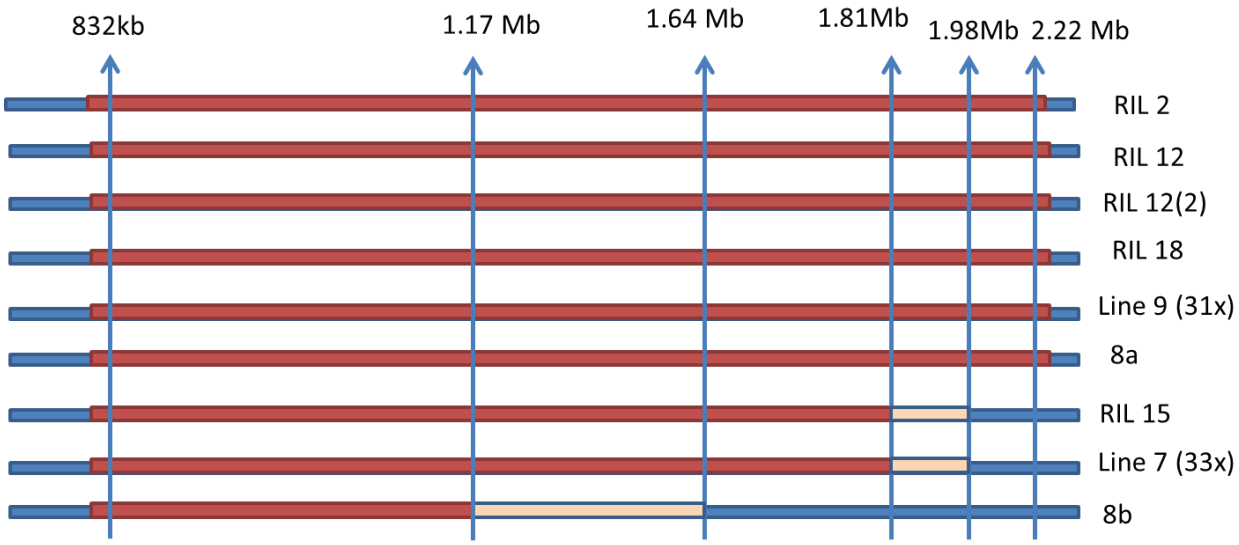

Figure 26: Results of donor fragment sizes for all RILs and lines isolated during backcrossing. Blue indicates Bristol background; red indicated the sizes of the donor fragments. Yellow? indicates regions of uncertainty that need to be mapped further to determine exact breakpoints for these lines.

References

Bennett EJ, Bence NF, Jayakumar R, Kopito RR. Global impairment of the ubiquitin-proteasome system by nuclear or cytoplasmic protein aggregates precedes inclusion body formation. *Mol Cell*. 2005;17(3):351-365.

Brenner S. The genetics of *caenorhabditis elegans*. *Genetics*. 1973:71-94.

Brignull HR, Morley JF, Garcia SM, Morimoto RI. Modeling polyglutamine pathogenesis in C. elegans. *Methods Enzymol*. 2006;412:256-282.\

Brinkman RR, Mezei MM, Theilmann J, Almqvist E, Hayden MR. The likelihood of being affected with huntington disease by a particular age, for a specific CAG size. *Am J Hum Genet*. 1997;60(5):1202-1210.

Bucciantini M, Giannoni E, Chiti F, Baroni F, Formigli L, Zurdo J, Taddei N, Ramponi G, Dobson CM, Stefani M. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*. 2002;11:507–511.

Bukau B, Weissman J, Horwich A. Molecular chaperones and protein quality control. *Cell*. 2006;125(3):443-451.

Chen S, Ferrone FA, Wetzel R. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. *Proc Natl Acad Sci* USA. 2002;11:11884–11889.

Davis, M. Wayne, et al. "Rapid single nucleotide polymorphism mapping in C. elegans." *BMC genomics* 6.1 (2005): 118.

Djousse L, Knowlton B, Hayden M, et al. Interaction of normal and expanded CAG repeat sizes influences age at onset of huntington disease. *Am J Med Genet A*. 2003;119A(3):279-282.

Dobson CM. Protein folding and misfolding. *Nature*. 2003;426(6968):884-90.

Duyao M, Ambrose C, Myers R, et al. Trinucleotide repeat length instability and age of onset in huntington's disease. *Nat Genet*. 1993;4(4):387-392.

Eremenko E, Ben-Zvi A, Morozova-Roche LA, Raveh D. Aggregation of human S100A8 and S100A9 amyloidogenic proteins perturbs proteostasis in a yeast model. *PLoS One*. 2013;8(3):e58218.

Gianni S, Ivarsson Y, Jemth P, Brunori M, Travaglini-Allocatelli C. Identification and characterization of protein folding intermediates. *Biophys Chem*. 2007;128(2-3):105-113.

Gidalevitz T, Ben-Zvi A, Ho K, Brignull H, and Morimoto R. Progressive disruption of cellular protein folding in models of polyglutamine diseases. *Science*. 2006;311(5766):1471-1474.

Gidalevitz,T., Kikis,EA., Morimoto,RI. A cellular perspective on conformational disease: The role of genetic background and proteostasis networks. *Current Opinion in Structural Biology*. 2010;20:23-32.

Gidalevitz,T., Wang, N., Deravaj, T., and Morimoto, RI. Genetic variation unmasks distinct states of resistance and tolerance of polyglutamine aggregation. *Unpublished*.

Gidalevitz T, Prahlad C, Morimoto R. The stress of protein misfolding: From single cells to multicellular organisms. *Cold Spring Harbor Perspectives in Biology*. 2011.

Gidalevitz T, Wang N, Deravaj T, Alexander-Floyd J, Morimoto RI. Natural genetic variation determines susceptibility to aggregation or toxicity in a C. elegans model for polyglutamine disease. *BMC Biol*. 2013;11:100-7007-11-100.

Gupta S, Jie S, Colby DW. Protein misfolding detected early in pathogenesis of transgenic mouse model of huntington disease using amyloid seeding assay. *J Biol Chem*. 2012;287(13):9982-9989.

Gusella JF, MacDonald ME. Huntington's disease: The case for genetic modifiers. *Genome Med*. 2009;1(8):80.

Hinault MP, Cuendet AF, Mattoo RU, et al. Stable alpha-synuclein oligomers strongly inhibit chaperone activity of the Hsp70 system by weak interactions with J-domain co-chaperones. *J Biol Chem*. 2010;285(49):38173-38182.

Hsu AL, Murphy CT, Kenyon C. Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *Science*. 2003;11:1142–1145.

Ho LW, Brown R, Maxwell M, Wyttenbach A, Rubinsztein DC. Wild type huntingtin reduces the cellular toxicity of mutant huntingtin in mammalian cell models of huntington's disease. *J Med Genet*. 2001;38(7):450-452.

Horwich A. Protein aggregation in disease: A role for folding intermediates forming specific multimeric interactions. *J Clin Invest*. 2002;110(9):1221-1232.

Kakizuka A. Protein precipitation: A common etiology in neurogenerative disorders? *Trends in Genetics*. 1998;14(10):396-402.

Kikis EA.,Gidalevitz T.,Morimoto, RI: Protein homeostasis in models of aging and age-related conformational disease. *Adv.Exp Med Biol,* 2012:694:138-159.

Kopito,RR., Ron, D. Conformational disease. *Nature Cell Biology*. 2000;2:E207-E209.

Markaki M, Tavernarakis N. Modeling human diseases in caenorhabditis elegans. *Biotechnol J*. 2010;5(12):1261-1276.

Morley JF, Brignull HR, Weyers JJ, Morimoto RI. The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in caenorhabditis elegans. *Proc Natl Acad Sci U S A*. 2002;99(16):10417-10422.

Nagai Y, Inui T, Popiel HA, et al. A toxic monomeric conformer of the polyglutamine protein. *Nature Structural & Molecular Biology*. 2007;14(4):332-40.

Orr, Harry T. "Beyond the Qs in the polyglutamine diseases. *Genes & development* 15.8.2001; 925-932.

Powers ET, Morimoto RI, Dillin A, Kelly JW, Balch WE. Biological and chemical approaches to diseases of proteostasis deficiency. *Annu Rev Biochem*. 2009;78:959-991.

Rubinsztein DC, Leggo J, Coles R, et al. Phenotypic characterization of individuals with 30-40 CAG repeats in the huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am J Hum Genet*. 1996;59(1):16-22.

Satyal SH, Schmidt E, Kitagawa K, et al. Polyglutamine aggregates alter protein folding homeostasis in caenorhabditis elegans. *Proc Natl Acad Sci U S A*. 2000;97(11):5750-5755.

Soto C. Unfolding the role of protein misfolding in neurodegenerative diseases. *Nat Rev Neurosci*. 2003;4(1):49-60.

Stine OC, Pleasant N, Franz ML, Abbott MH, Folstein SE, Ross CA. Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. *Hum Mol Genet.*1993;11:1547–1549.

Tyedmers, Jens, Axel Mogk, and Bernd Bukau. "Cellular strategies for controlling protein aggregation." *Nature reviews Molecular cell biology* 11.11. 2010; 777-788.

Wetzel R. Mutations and off-pathway aggregation of proteins. *Trends Biotechnol*. 1994;12(5):193-198.

Wexler NS, Lorimer J, Porter J, et al. Venezuelan kindreds reveal that genetic and environmental factors modulate huntington's disease age of onset. *Proc Natl Acad Sci U S A*. 2004;101(10):3498-3503.

Winner B, Jappelli R, Maji SK, et al. In vivo demonstration that alpha-synuclein oligomers are toxic. *Proc Natl Acad Sci U S A*. 2011;108(10):4194-4199.