Stress, Aging and Gene Expression in *Drosophila melanogaster*

By

Michael P. Antosh

B.S. Physics, University of New Hampshire, 2007

Dissertation

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Physics at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2012
© Copyright 2012 by Michael P. Antosh
This dissertation by Michael P. Antosh is accepted in its present form
by the Department of Physics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date______________  _______________________________
                    Leon Cooper, Advisor

Recommended to the Graduate Council

Date______________  _______________________________
                    Nicola Neretti, Co-Advisor

Date______________  _______________________________
                    Derek Stein, Reader

Approved by the Graduate Council

Date______________  _______________________________
                    Peter Weber, Dean of the Graduate School
Curriculum Vitae:

Michael Antosh was born in Wakefield, Rhode Island in 1984. Previous degrees earned include Bachelor of Science in Physics from the University of New Hampshire in 2007 and Masters of Science from Brown University in 2009.

Positions and Employment

- 2011-Present: Teaching Assistant, Department of Physics, Brown University, Providence RI
- 2008-Present: Research Assistant, Department of Physics, Brown University
- 2006: Lung MRI Project, Department of Physics, University of New Hampshire, Durham, NH
- 2006: Research Experience for Undergraduates Program, Laboratory for Elementary Particle Physics, Cornell University, Ithaca, NY
- 2005-2006: Academic Mentor, Center for Academic Resources, University of New Hampshire,
- 2005: Environmental Engineering Intern, PARE Corporation, Lincoln, RI
- 2005: Asphalt Design Group, Department of Civil Engineering, University of New Hampshire,

Honors:

- 2007 Brown University Graduate Fellow
- 2006 Barry M. Goldwater Scholar
- 2006 University of New Hampshire Edward Stolworthy Scholarship
- 2005 Rhode Island Consulting Engineer’s Scholarship
- 2004, 2005 University of New Hampshire Civil Engineering Achievement Award
- 2003 National Merit Scholarship

Peer Reviewed Publications

Acknowledgements:

Mom, Dad, Kate and Heidi – for everything I needed outside of the classroom

Leon Cooper, Nicola Neretti, David Fox, Stephen Helfand, Rachel Whitaker – for everything I needed inside the classroom
Table of Contents:

Chapter 1: Gene Expression Analysis

- 1.1 Introduction 1
- 1.2 Single Gene Analysis 3
- 1.3 Gene Set Analysis 10
- 1.4 Discovery of takeout as a Gene That Affects Aging 14
- 1.5 Comparison of Biological Pathway Behavior Between Lifespan Extending Treatments 22
- Chapter 1 References 30

Chapter 2:

- 2.1 Introduction 33
- 2.2 Literature Survey 34
- 2.3 Methods 35
- 2.4 Simulations 49
- 2.5 Real Data 53
- 2.6 Discussion 58
- 2.7 Conclusion 62
- Chapter 2 References 63

Chapter 3: Response to Low-Dose Radiation in *Drosophila melanogaster*

- 3.1 Introduction 64
- 3.2 Literature Search 69
- 3.3 Experimental Methods 89
- 3.4 Results 92
• 3.5 Discussion 104
• 3.6 Conclusion 106
• 3.7 Calculation of Compton Scattering Interaction Probability for a Photon in Water 107
• Chapter 3 References 110
List of Tables

- Table 2.3.1: 3x3 Contingency Table Applied to Ranked List Steps 37
- Table 2.3.2 Example of Set Size Selection 47
- Table 3.2.1: Baxter and Blair’s Low-Dose Radiation Lifespans 77
- Table 3.2.2: Formal Low-Dose Radiation Experiments by Baxter and Blair 78
- Table 3.2.3: Low and High-Dose Response, Moskalev’s Control Flies 86
- Table 3.4.1 Summary of Pilot Gene Expression Results 99
- Table 3.4.2 GOStat Analysis Results for Available Full Experiment Data 100
List of Illustrations

- Figure 1.4.1 Verification of microarray data using qPCR 16
- Figure 1.4.2 Comparison of genes upregulated and downregulated with DR in flies of genetic background yw/w^{1118} and Canton-S 17
- Figure 1.4.3 Comparison of genes upregulated and down-regulated in DR, sir2 and p53 expressing long-lived flies at Day 10 18
- Figure 1.5.1 Heat map of KEGG gene sets from sir2 and p53 long-lived flies show similarity with a DR signature gene set (significant in both genetic backgrounds) 23
- Figure 1.5.2 Heat map of KEGG gene sets from DR of female whole body and female head and thorax appear very different 26
- Figure 1.5.3 Heat map of KEGG gene sets from DR and resveratrol female head and thorax appear very similar 27
- Figure 2.3.1 Illustration of a step 36
- Figure 2.3.2: Sections of Ranked Lists 37
- Figure 2.3.3: Significant Step Diagram 43
- Figure 2.3.4 Illustration of different step size choices 50
- Figure 2.4.1 Simulation Results for m=200, n=400, k=200, N=1200, \( \Delta=50 \) 50
- Figure 2.4.2 Simulation Results for m=400, n=400, k=200, N=1200, \( \Delta=50 \) 51
- Figure 2.4.3 Simulated Data With Decreasing Correlation 52
• Figure 3.1.1 Theoretical Illustration of Threshold or Hormetic Mechanism 66
• Figure 3.1.2: Excess Solid Cancer Risk vs. Radiation Dose (Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation 2006) 67
• Figure 3.2.1: Sacher’s Radiation Lifespan Data 72
• Figure 3.2.2: Sacher’s Survivorship Curves, Flies Given a Dose of Radiation Daily 73
• Figure 3.2.3: Sacher’s Model of Repair 74
• Figure 3.2.4: Survival for Drosophila females in background radiation (solid curve) and shielded (dashed) 79
• Figure 3.2.5: Carey’s Lifespan Mortality Rates with Irradiation and Diet 81
• Figure 3.2.6: Eye Color Mutation Frequency with Radiation Dose in Ogura et al (2009) 85
• Figure 3.4.1 Pilot Experiment Survivorship 93
• Figure 3.4.2 Full Experiment Survivorship 94
• Figure 3.4.3 Full Experiment Mortality Rate 94
• Figure 3.4.4 Median Lifespan Versus Dose for Baxter and Blair Data and Antosh Pilot Experiment 95
• Figure 3.4.5 Comparison with Control Lifespan versus Number of Incident Photons 97
• Figure 3.4.6 Data Comparison with Full Experiment Added 98
• Figure 3.4.7 Data Comparison in Log Scale  
  
  Figure 3.4.8. Dendrogram of Low Dose Radiation Treatments  

98  
103
Chapter 1: Gene Expression Analysis

1.1 Introduction

An organism has the ability to produce proteins that are specific to each gene in its genome. The interactions of these proteins with each other determine biological function. Gene expression is the process where the organism produces these proteins. Genes code for an intermediate step, messenger RNA (mRNA), through a process called transcription. Messenger RNA is used to produce proteins through a process called translation.

In the past decade, experiments have evolved with the ability to test gene expression for every gene at the same time. A major difficulty in this type of experiment is that more than one protein at a time cannot be measured easily because of their complex folding structures. Instead, most efforts are based on the idea of measuring messenger RNA, which is an intermediate step in the process of making proteins. The result for each experiment is a number corresponding to the amount of messenger RNA for each gene.

Analysis of gene expression data is complex and interesting in part because of the large amount of data produced in these experiments. For example, a fruit fly \textit{(Drosophila melanogaster)} has approximately 15,000 genes in its genome. If 3 experiments are done together, each with 3 replicates for an experimental condition
as well as for a separate control condition, the number of data points is roughly 270,000 (15,000*3*3*2). Gene expression analysis methods can be broken into two distinct groups: a gene-by-gene approach, referred to here as single gene analysis, and a biological pathway approach that looks at the behavior of groups of genes that have a known common function, referred to here as gene set analysis.

This chapter will introduce methods of single gene analysis as well as methods of gene set analysis. It will describe my analysis of three gene expression experiments done in the Helfand laboratory at Brown University. The following two chapters will describe a new gene expression analysis method that I developed and a gene expression experiment that I designed and implemented.
1.2 Single Gene Analysis

The goal of single gene analysis is to determine whether or not the expression of a given gene changed from the control samples to the experimental samples. Three statistics help to determine this:

- P Value
- Fold Change
- Present/Absent

Each of these statistics requires setting a cutoff, and genes will be labeled as having significantly changed between experiment and control if the gene has a better than cutoff value in all 3 statistics. The statistics are described in detail below.

P Value

*P Value* is the statistical probability that the difference between the expression levels in control and experimental samples could happen randomly. Generally, a gene will be considered significant if the p value is less than 5 percent (or a similar cutoff). It will depend on the average expression values for control and experimental samples as well as standard deviations between replicates.

The statistical test that gives the p value differs depending on the experimental method. One experiment type, microarray, measures the fluorescence of tags bonded to the messenger RNA. This results in a continuous distribution of values. Another experiment type, sequencing, measures the number of messenger RNA
fragments in the sample that match each gene. This results in a discrete distribution of values, since each gene will have an integer number of counts. A standard method for microarray data is to assume that the logarithm of expression is normally distributed and uses a t test. A standard method for discrete distributions is to assume that the data is distributed by the negative binomial distribution (Anders and Huber 2010). One sequencing analysis algorithm, Cuffdiff (Trapnell et al. 2010), uses the beta negative binomial distribution and a two-tailed student’s t test. Another algorithm, edgeR (Robinson, McCarthy and Smyth 2009), uses a Poisson model that generalizes to a negative binomial distribution with a term for biological variability and a modified exact test. A third algorithm, DESeq (Anders and Huber 2010), uses a similar technique to edgeR but scales variance using a model that adds a random variable to account for variations in the concentration of gene fragments during sequencing. EdgeR scales the variance with a single linear factor. Garber et al. (2011) discusses other sequencing analysis algorithms in a review article.

If many p value tests are done at once, the instance of false positives (type 1 error) becomes large enough that a correction is necessary. This problem is referred to as Multiple Hypothesis Testing. In this problem, the goal is to maximize the statistical power (1 minus the false negative rate) while controlling the false positive rate. For example, a p value of 0.05 means a 5 percent chance that the result occurred randomly. If 15,000 genes are tested in a given experiment and 2000 have a p value of 0.05 or less, roughly 5 percent of those 2000 genes (100 genes) probably only were significant by random chance (meaning that the gene is not actually
significant). Many strategies have been developed to solve this problem, but three common methods for correcting for this problem are as follows:

- The **Bonferroni Correction** is a change of the p value cutoff used to label significance by dividing by the number of total tests. In the example above, the p value cutoff for significance would become 0.05/15,000. This is the most stringent correction. The goal of this correction is to ensure that the probability of having even one false positive rate is equal to the significance cutoff probability (0.05 in the example above).

- The **Benjamini-Hochberg Method** (Benjamini and Hochberg 1995) ranks the genes by their p value, then takes as a cutoff the i^{th} p value in the list such that:

\[
\text{maximum } i \text{ such that } p_i \leq \frac{\text{cutoff}}{i} \quad (1.2.1)
\]

The goal of this test is to keep the false positive rate below the cutoff value, but unlike the Bonferroni correction the focus is on the rate of false positives and not the rate of having even one false positive. The allowance of some false positives increases the statistical power by decreasing the rate of false negatives.

- **Q Value** (Storey and Tibshirani, 2003) estimates the proportion of false positives for every possible value of a p value cutoff. This is done by assuming that non-significant genes have p values randomly distributed between 0 and 1. If the p values for each gene are viewed in a histogram, the height of the histogram near p=1 is considered the height of the random
genes’ contribution. Comparing this height to the total height of the histogram near p=0 gives an estimate of the number of false positives. The goal of this correction is the same as the Benjamini-Hochberg; however it uses an approach based on the assumption of random p values for non-significant genes while the Benjamini-Hochberg method ranks the p values. The choice of correction method depends on the form of the data and the desired solution. If it is not acceptable to have any false positives, the Bonferroni correction is best. If a histogram of the p values is consistent with Storey’s model (the height is consistent across non-significant p values), Storey’s model is likely best. In the algorithm discussed in chapter 2, we want the correction to be consistent across experiments with a different number of p values. We chose the Benjamini-Hochberg correction because the rank approach explicitly takes the number of p values into account.

**Fold Change**

The fold change is a measure of the change in average gene expression between experiment and control:

\[
\text{Fold Change} = \frac{\text{Average Experimental Sample Gene Expression}}{\text{Average Control Sample Gene Expression}}
\]  

(1.2.2)

Fold change is often expressed in terms of logarithms. This is beneficial mathematically because the sign of the fold change will indicate whether experimental or control samples have more expression. Cutoff values vary for this statistic.
Present/Absent

There is experimental noise in gene expression experiments, and genes with expression that is measured to be at or below the noise level are unreliable. Present/Absent is a measure of whether or not a gene was reliably measured. It has two possible values: 1 (“present”) or 0 (“absent”). For a given cutoff value that represents accurate expression measurement, a gene is labeled present if it has average expression greater than the cutoff in either the experimental samples or the control samples. It is labeled absent if it has average expression less than the cutoff in both experiment and control samples. The cutoff varies, but one of the many approaches in microarray experiments is to make the cutoff the 25th percentile of all expression values in the experiment(s).

Combining the Statistics

The combination of all three statistics is important. P values give statistical significance, but the effect may not be biologically relevant if the fold change isn’t large enough. And neither statistic matters if the gene is not reliably measured (absent).

Uses of Single Gene Analysis Results

Single gene analysis can be used to evaluate the significance of a given gene in an experiment. For example, the experiment discussed in section 1.4 details the discovery of the gene *takeout* as an aging influence, finding that it changes
significantly from control in many different experiments that extend lifespan in *Drosophila*. The result of single gene analysis can also be used to investigate the significance of biological pathways.

Biological pathways are groups of genes with a shared function. Two common repositories of pathway information are Gene Ontology and the Kyoto Encyclopedia of Genes and Genomes (The Gene Ontology Consortium, 2000; Kanehisa 2000). Gene Ontology contains many pathways (more than 3,000), organizing the pathways into 3 categories: cellular component, biological process and molecular function. It includes genes in pathways when the connection to that function is inferred but not necessarily shown in an experiment. The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains many fewer pathways, but these pathways are manually curated, meaning that they are only genes shown to be associated with a function in an experiment.

The most common pathway analysis algorithm that uses single gene analysis results is GOStat (Beissbarth and Speed, 2004). The GOStat algorithm works as follows: for a given biological pathway, compute the number of genes in the pathway that are both changing significantly from control (call it $k_1$). Then, use Fisher’s Exact Test (equation 1.2.3) on $k_1$, the total number of significant genes (call it $m$), the total number of genes in the pathway ($n$) and the total number of genes in the genome ($N$) to compute the statistical significance of $k_1$. The parentheses in the equation represent binomial coefficients.
P Value of Pathway = \sum_{k=\min(m,n)}^{\min(m,n)} \frac{m \choose k \frac{N-m}{n-k} \frac{N}{n} \choose k}{n-k} \text{ (1.2.3)}
1.3 Gene Set Analysis

Gene Set Analysis is a class of algorithms where the focus is on entire groups of genes together instead of single genes. Usually, these groups will represent biological function pathways (for example, metabolism or reproduction). This focus comes from the fact that most biological functions are achieved by groups of genes working together instead of just one gene controlling a function by itself. This section introduces the most commonly used algorithm, Gene Set Enrichment Analysis (GSEA) as well as two alternative methods (Parametric Gene Set Enrichment Analysis (PGSEA) and Gene Set Analysis (GSA)).

**Gene Set Enrichment Analysis**

Gene Set Enrichment Analysis (GSEA) (Mootha et al. 2003; Subramanian et al. 2005) uses a non-parametric ranked list test to determine the statistical significance of a group of genes (also called a gene set). The set of genes in the experiment (usually the set of genes labeled as present in single gene analysis) are ranked by a statistic (usually the logarithm of fold change). Then, the ranks of the genes that are in the gene set are examined to see if most of the genes in the gene set tend to be towards either the top or the bottom of the ranked list. The statistical test used is a variant of the Kolmogorov-Smirnov test – equation 1.2.4 expresses the test statistic (enrichment score) in terms of two functions $P_{in\ set}$ and $P_{not\ in\ set}$ (unrelated to p values). $P_{in\ set} - P_{not\ in\ set}$ is calculated for each rank and the maximum difference is chosen as the value of the enrichment score.
Enrichment score = max change from zero of \((P_{\text{in set}} - P_{\text{not in set}})\) where

\[
P_{\text{in set}} (\text{gene set } S, \text{rank } i) = \sum_{\text{genes in } S \text{ with rank } \leq i} \frac{|\text{fold change of gene } i|}{\sum_{\text{genes in } S} |\text{fold change}|}
\]

\[
P_{\text{not in set}} (\text{gene set } S, \text{rank } i) = \sum_{\text{genes not in } S \text{ with rank } \leq i} \frac{1}{\text{Total Number Genes} - \text{Genes in Gene Set}}
\]

After computing the enrichment score, the p value for the gene set is computed by randomly permuting the data labels (for example, switching experiment and control labels for some samples) 1,000 times and comparing the resulting enrichment scores to the score found for the actual data. A correction for multiple p value tests (as in section 1.2) is performed by comparing the actual enrichment score found with the scores found from running the statistical test on permutations of the data where the ranks of the genes have been randomly assigned. The score is then normalized by the average of these scores of the random permutations, treating those greater than zero and those less than zero separately.

**Gene Set Analysis**

Gene Set Analysis (GSA) (Efron and Tibshirani, 2007) follows the strategy of GSEA fairly closely. It makes two changes:

- Instead of using the Kolmogorov-Smirnov test, the authors devised an alternative statistic for the ranked list, the *maxmean* statistic, computed by dividing the genes in the gene set into two groups: those that have fold change greater than zero and those having fold change less than zero. For both of those two groups, they compute average fold change divided by size of the gene set. The number farther from zero is the score.
When making random data for statistical significance tests, GSEA only switches either the sample labels or the gene ranks but don’t change both at the same time. GSA randomly switches gene and experimental/control labels at the same time. They do this because they believe that gene labels are not truly independent of each other because many genes participate in functions together.

**Parametric Gene Set Analysis**

Parametric Gene Set Analysis (PGSEA) (Kim and Volsky 2005) takes an approach that does not involve ranking the genes. It invokes the Central Limit Theorem - that “the distribution of the average of randomly sampled n observations tends to follow normal distribution as the sampling size n becomes larger, even when the parent distribution from which the average is calculated is not normal”. PGSEA models the fold changes of the gene set and of the genome as two separate normal distributions. Then, the difference in means and standard deviations is used to compute statistical significance and direction of change.

**Limitation**

It is worth noting that all of these methods assume that a gene set is most expressed when all of the genes in the gene set change in the same direction from control, either higher or lower. In reality, many gene sets need some genes increasing from control and some genes decreasing from control in order to function differently from control. Some methods are beginning to model this (for example, Draghici et
al., 2007) but the larger problem is that there is no existing annotation database where the gene directions are given. Gene Ontology (The Gene Ontology Consortium, 2000) lists approximately 3000 biological pathway gene sets, which can contain over 100 genes per gene set. This makes doing a full annotation a daunting task.
1.4 Discovery of takeout as a Gene That Affects Aging

This single gene analysis of gene expression data for three treatments known to extend lifespan in *Drosophila melanogaster* (fruit flies) resulted in a publication in the journal *Aging* (Bauer, Antosh et al. 2010). The genes that changed significantly from control in each of the three treatments included the gene *takeout*, leading to its classification as a gene that affects aging. The three treatments are:

- **Dietary Restriction (DR)** (for example, Bross et al. 2005) – this affects the most biological processes of the 3 treatments. Flies were given food that was 3 times more watered down than the control flies.

- **Mutation of the gene Silent Information Regulator 2 (sir2)** (Rogina and Helfand 2004). sir2 is a transcription factor that suppresses transcription process of gene expression (through deacetylation). Adding an extra copy of sir2 extends lifespan.


sir2 is believed to be “downstream” of DR in the aging pathway, and p53 is believed to be downstream of sir2. In other words, it is believed that CR increases sir2 gene expression and that increased sir2 expression decreases p53 gene expression.

**Methods**

We chose a p value cutoff of 0.01, a fold change cutoff of 1.5-fold (50 percent larger or smaller than control) and used the present/absent method discussed in section
1.2 for use with microarray experiments. The p value and fold change cutoffs were chosen after an attempt by the Helfand lab to verify several of the microarray results using the experimental method quantitative polymerase chain reaction (qPCR) – in qPCR, the amount of complementary DNA (two strands of messenger RNA combined) is exponentially increased in a chain reaction and measured by a reagent specially designed to fluoresce when bonded to a chosen gene. The change in gene expression is quantified via comparison with a reference gene known to be constant across the experimental conditions. This method is more accurate than microarray experiments, but can only examine on the order of 100 genes at a time where microarrays can examine order 10,000.

As shown in figure 1.4.1, the chosen cutoffs led to the selection of genes that were most often selected as significant by both qPCR and microarrays. Each point on the graph represents a gene measured by both microarray and qPCR. The axes describe the fold change and p value of the microarray data. The red dots represent genes with a significant fold change (>1.2) in PCR, and the blue dots represent genes with a non-significant fold change in PCR. The dotted lines define a box of the region where the PCR data is most likely to be significant—fold change>1.5 (0.58 in log2) and p value<0.01. A fold change cutoff of 1.2 was used in the qPCR data because it is more accurate than microarray data.
Major Conclusions

The paper made 5 major conclusions from the gene expression analysis:

1. Genetic background affects DR mechanism
2. sir2 is similar to DR
3. p53 overlaps with sir2 and DR
4. 21 genes and 6 gene sets change significantly in all 3 treatments
5. One of these genes, takeout, is upregulated in several other experiments that extend lifespan

The rest of this section will discuss these conclusions using the analysis.

Genetic Background Affects DR Mechanism

Gene expression data was collected from DR experiments on flies of two different genetic backgrounds (called yw/w^{1118} and Canton-S) at age 10 days. As seen in figure 1.4.2, 1321 genes increased significantly with DR in yw/w^{1118} flies and 1140
genes decreased significantly. In Canton-S flies, 1286 genes increased significantly with DR and 1435 genes decreased significantly. Approximately 55-60% of significant genes were shared between the two backgrounds (765 up, 708 down). This shows DR as a wide-ranging effect (roughly 15% of all genes change significantly) that is strongly affected by genetic background (only just over half of the significant genes are the same).

![Comparison of genes upregulated and downregulated with DR in flies of genetic background yw/w1118 and Canton-S.](image)

**sir2 is similar to DR**

The sir2 genetic mutants are in the yw/w1118 genetic background, so the comparison is with the DR experiment from that background. As seen in figure 1.4.3A, 782 genes change significantly with sir2 addition. 525 of them are shared with DR, a 67% overlap. In the significantly upregulated genes, the overlap is 72%, greater than the percentage shared between two slightly different strains of fly on the same DR treatment (above). This suggests that DR and sir2 extend lifespan using a similar mechanism or perhaps the same mechanism. It further suggests that sir2 is part of the genetic pathway through which DR extends lifespan, because sir2 has roughly 1/3 of the number of significant genes as DR.
The high overlap continues under a GOStat analysis of the significant genes. Of the 148 gene sets found to be significant in both DR and sir2, there was a 72% overlap in upregulated gene sets and a 78% overlap in downregulated gene sets (overlap percentage is overlap divided by the number of gene sets in sir2).

Figure 1.4.3 Comparison of genes upregulated and down-regulated in DR, sir2 and p53 expressing long-lived flies at Day 10. (A) Venn diagrams comparing upregulated and downregulated genes in DR, sir2 and p53 in a yw/w1118 background at age 10 Days. Genes intersecting in all 3 sets are noted in box with arrow. (B) Heatmap comparing the average log2 fold changes for genes significantly altered in the yw/w1118 DR with the same genes in sir2 and p53 expressing flies.
p53 Mutation Overlaps with DR and sir2

As seen in figure 1.4.3A, 235 total genes changed significantly in the p53 data set. 37% of these genes overlap with DR, and 37% overlap with sir2. However, in upregulated genes the overlaps are 63% with DR and 65% with sir2. When comparing GOStat results, 6 of the 7 significantly upregulated gene sets overlap with both DR and sir2 and 0 of the 15 significantly downregulated gene sets overlap with DR or sir2.

The similarity between DR, sir2 and p53 is shown visually in figure 1.4.3B, which displays the direction of change for all of the genes that change significantly in DR across all three experiments.

21 Genes and 6 Gene Sets Change Significantly In All 3 Treatments

The genes of most interest will be the genes that change in all 3 aging treatments, because they will likely be tied to aging and not the other effects that these treatments cause, which differ between treatments. The 21 genes are named in figure 1.4.3A. 20 of the genes are upregulated, and 1 is downregulated. In the upregulated genes, 4 are associated with chromatin structure or maintenance (CG42249, CG5612, CG17325, CG4123), 3 are associated with circadian rhythm (CG10553, CG13928, takeout), 2 are involved in neural activity (Nplp3, synaptogyrin), 2 are involved in detoxification/chaperone activity (CG3091, CG6870), 2 are involved in muscle maintenance (Myo61F, CG14687). There were individual genes involved in immune function (IM3), growth factor activity (daw),
feeding behavior (takeout) and response to starvation (takeout). The downregulated gene is larval serum protein-2.

The 6 gene sets in common, all upregulated, are endopeptidases, peptidases, serine-type endopeptidases, serine-type peptidases, serine hydrolases and defense response. Peptidase means breaking down proteins and serine-type means that the protein is broken down at the amino acid serine. This set of peptidases and defense response may suggest an adaptive response that is turning over proteins at a faster rate to better control damage.

*takeout* Is Upregulated in Several Other Experiments That Extend Lifespan

From the 21 genes significantly changing in all 3 experiments, the lab focused on takeout as a potential aging gene because it was the only one of the group of 21 above to also increase in the long-lived flies with a mutation in the *Indy* gene (*Indy* stands for “I’m not dead yet”) (Neretti et al. 2009). The lab then used qPCR (described above) to find takeout significantly upregulated in several other genetic mutants that increase lifespan in flies: *Rpd3*, *chico* and *methuselah* (Lin, Seroude and Benzer 1998; Rogina, Helfand and Frankel 2002; Clancy et al. 2001). Then, the lab performed a lifespan experiment in flies with a mutation allowing for increased expression of takeout, and saw an increase in lifespan.

Discussion and Conclusion

The analysis done in this experiment demonstrated the relationship between DR, sir2 mutation and p53. All three experiments extend lifespan, but sir2 is
demonstrated to have fewer extra effects than DR, based on the number of genes that are significantly changing. p53 has even fewer extra effects. The gene sets changing in each condition indicate a turnover/repair and defense mechanism in common.

Single gene analysis is used to find a new aging gene. takeout is shown to be a common factor in the lifespan extending mechanisms of 7 different types of experiments, and then shown to extend lifespan on its own. takeout is primarily known as a part of circadian rhythm pathways (Sarov-Blat et al. 2000; So et al. 2000; Benito et al. 2010), but this is not the first link between biological rhythm and aging (Imai 2009). It has also been hypothesized to be involved with insect juvenile hormone (Meunier, Belgacem and Martin 2007; Noriega et al. 2006; Hamiaux et al. 2009), which has been linked to mammalian metabolism (Davey 2007; Marsh 1993) and grasshopper longevity extension (Tatar and Lin 2001).

The power of single gene analysis is demonstrated in an application where it leads to the discovery of a new aging gene from multiple experimental data sets.
1.5 Comparison of Biological Pathway Behavior Between Lifespan Extending Treatments

A second paper by our group (Antosh et al. 2011) compared the treatments from section 1.4 (DR, sir2, p53) using gene set analysis – specifically, GSEA. It adds gene expression data from a fourth treatment - resveratrol, a drug known to mimic the effects of DR (Pearson et al. 2008). This is an important addition because it is the only of the above treatments that can be reasonably done in humans (in practice, humans dislike undergoing dietary restriction). It also adds another DR gene expression data set that is gathered only from the head and thorax of flies (the original sets included the whole body, which is these section plus the abdomen, which contains eggs).

The paper makes four main conclusions, which will be discussed below:

- DR/sir2/p53 are very alike from a gene set perspective
- DR effect is tissue specific
- DR and resveratrol have similar mechanisms
- Genes Regulated by Juvenile Hormone are altered in DR, sir2, p53 and resveratrol

**DR/sir2/p53 Are Very Alike From a Gene Set Perspective**

As shown in section 1.4, these treatments are very alike in a single gene analysis approach. Figure 1.5.1 shows a heatmap of the results of a GSEA analysis using gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG)(Kanehisa 2000). With two small exceptions, every statistically significant gene set is going in the same direction across treatments. This suggests a common response. It is
noteworthy that the mechanisms of the two DR experiments from section 1.4 (two different genetic backgrounds) have a 100% overlap, since the single gene analysis showed only a 55-60% overlap between them. 72 gene sets were labeled as significant in the yw/w1118 DR, and 54 in the Canton-S DR. All 54 gene sets in Canton-S were in yw/w1118. Of the 54 gene sets statistically significant in both DR sets, 49 were also significant in sir2 (91% overlap) and 35 were significant in p53 (65% overlap).

Figure 1.5.1 Heat map of KEGG gene sets from sir2 and p53 long-lived flies show similarity with a DR signature gene set (significant in both genetic backgrounds). Red are gene sets that are statistically significantly upregulated, blue are gene sets that are statistically significantly downregulated and black are gene sets that are not changed by a statistically significant amount.
DR Effects Are Tissue Specific

Gene expression data for DR from the head and thorax tissue was taken in an attempt to eliminate the effect of eggs in the female abdomen on the experiment. As shown in figure 1.5.2, the GSEA results were very different. 54 gene sets were statistically significant in whole body (Canton-S was the genetic background for both tissue types) and 83 gene sets were statistically significant in head/thorax. Only 26 of these gene sets were significant in both datasets, less than 50% of the maximum possible overlap. Of particular note are a handful of mostly repair-related gene sets that changed in opposite directions: DNA polymerase, nucleotide excision repair, mismatch repair, base excision repair, proteasome, ubiquitin mediated proteolysis, GPI anchor biosynthesis, fatty acid elongation in mitochondria and oxidative phosphorylation.

DR and Resveratrol Have Similar Mechanisms

Gene expression data for resveratrol was taken from head and thorax tissues in female flies and compared with the DR data from the same tissue. As illustrated in figure 1.5.3, of the 83 categories statistically significant in DR, 81% were also statistically significant in resveratrol. This is similar to an 82% overlap shown in mouse liver (Pearson et al. 2008). A single gene analysis on the head/thorax data sets for DR and resveratrol (with the methods the same as section 1.4) shows only 28 statistically significant genes in resveratrol and 152 in DR. The overlap is 12 genes, 43% of the resveratrol set. If the p value cutoff is loosened to be 0.05 instead
of 0.01 and the fold change cutoff is loosened to be 1.2-fold instead of 1.5-fold, the result becomes 150 gene overlap out of 237 genes in resveratrol and 1708 in DR (63% overlap). The small number of significant genes despite a relatively large number of significant gene sets (as compared to whole body DR) demonstrates that GSEA may be a more efficient analysis for this data set. This difference may be caused by the gene expression changes being so small that a single gene approach would have a hard time detecting them; but a pattern of small coordinated changes at the gene set level may still be visible to a gene set approach.
Figure 1.5.2 Heat map of KEGG gene sets from DR of female whole body and female head and thorax appear very different. Red are gene sets that are statistically significantly upregulated, blue are gene sets that are statistically significantly downregulated, and black are gene sets that are not statistically significantly changed.

RNA is from Canton-S (C-S) DR and high calorie fed 10-day old female whole bodies or only heads and thoraces. KEGG categories and scores are in Supplemental Table 3.
Figure 1.5.3 Heat map of KEGG gene sets from DR and resveratrol female head and thorax appear very similar. Red are gene sets that are statistically significantly upregulated, blue are gene sets that are statistically significantly downregulated and black are gene sets that are not statistically significantly changed. RNA is from the same cohort of Canton-S (C-S) 10-day old female heads and thoraces on a DR diet, high calorie food or fed resveratrol.
Juvenile Hormone regulated genes are altered in DR, sir2, p53 and resveratrol

Since *takeout* is believed to play a role in both aging and juvenile hormone metabolism (one of two hormones in *Drosophila*, abbreviated JH), we created two gene sets of juvenile-hormone-metabolism-related genes in order to study if these genes are enriched in any of the experimental conditions we tested. One gene set contains the genes that are upregulated by juvenile hormone and one contains the genes downregulated by juvenile hormone, as determined by several studies that have appeared in literature (Flatt et al. 2008; Berger and Dubrovsky 2005; Dubrovsky, Dubrovskaya and Berger 2002; Dubrovsky et al. 2000). Using GSEA, we found that JH-upregulated genes are significantly upregulated in whole body DR and that JH-downregulated genes are significantly upregulated in whole body DR, sir2, p53 and head/thorax resveratrol (only one tissue type was used for sir2, p53 and resveratrol). JH-downregulated genes were also upregulated and just below statistical significance in head/thorax DR. The upregulation of genes that are downregulated by JH suggests that decreased JH signaling may be part of the mechanism of lifespan extension in flies. This would agree with the hypothesized result that JH is involved in life span determination (Sarov-Blat et al. 2000; Gaikova and Flatt 2010; Flatt, Tu and Tatar 2005; Herman and Tatar 2001; Tatar and Yin 2001). The upregulation of genes that are upregulated by JH contradicts this result, but may be explained by changes in the other *Drosophila* hormone, ecdysone.
Discussion and Conclusion

After trying both single gene analysis and gene set analysis on this data, it seems that trying both strategies is the optimal way to handle gene expression data. As a comparison of the two methods, we took a set of gene sets from the Gene Ontology database (The Gene Ontology Consortium 2000). The list of genes that change in both whole body DR data sets was run through this database using GOStat, giving 551 significantly changed gene sets. GSEA was run on each of the two data sets, and 713 gene sets were significant in both. Of the 551 and 713 gene sets, only 240 gene sets were the same – 44% of the GOStat sets and 34% of the GSEA sets. The methods seemed to give different answers, which makes sense because GSEA tests for changes being in the same direction while GOStat tests for many genes in the pathway to be changing significantly – two very different approaches.

Both single gene analysis and gene set analysis indicate that DR, sir2, p53 and resveratrol treatments have shared mechanisms in extending lifespan in Drosophila. This shared mechanism likely includes repair mechanisms, metabolism changes and juvenile hormone signaling decreases.
Chapter 1 References


Chapter 2: A New Algorithm for Comparing Gene Expression Data Sets

2.1 Introduction

As shown in chapter 1, one important strategy in gene expression analysis is the comparison between data sets in order to determine the similarity of the treatments and to find the genes that change significantly in both experiments. The most common method for comparing two gene expression experiments is to consider the two experiments separately and then find the genes that are determined to be significant in both experiments (for example, Bauer and Antosh 2010). This method depends on arbitrary cutoffs for test statistics, expression level (present/absent) and fold change (experimental average expression divided by control average expression). However, there is little difference between a gene with fold change 1.50 (labeled a significant change) and a gene with fold change 1.49 (labeled insignificant).

This chapter describes a new method for comparing two gene expression experiments, based on ranking the genes in each experiment and finding significant overlaps between different sections of the ranked lists. Labeling genes as significant depends on the statistical probability of these overlaps, not on arbitrary cutoffs. This method is a significant improvement on the methods used in the biologically focused paper by Antosh et al. (2011). The model for calculation of p value has been improved, and the method has been extended to allow for the more general case where the two experiments have a different number of differentially expressed genes.
2.2 Literature Survey

The use of ranked lists in gene expression analysis is a commonly used approach. For example, the algorithms behind GSEA (Subramanian et al., 2005) and GOrilla (Eden et al., 2009) rank genes by fold change in hopes of estimating the differential expression of gene sets. Eden et al. (2007) use ranked lists to compare experimental samples of immunoprecipitation data with background. Jurman et al. (2007) and Boulesteix and Slawski (2009) compare the algebraic stability between ranked lists of genes. However, none of these methods use ranked lists to compare two gene expression experiments.

There are two existing methods of ranked list comparison for gene expression data. Plaisier et al (2010) use a similar mathematical model to ours, but use a different strategy containing one important extra assumption. Yang et al. (2006) base their method around a similarity score. These methods will be compared with our method in detail in the discussion section.
2.3 Methods

The methods section will cover the following topics:

- Mathematical Model
- Calculation of Step P Values
- Maximization Approach to Determining Optimal Overlap Size
- Determining Optimal Step Size

**Mathematical Model**

The mathematical model is a “mixture model with correlation”. Genes in each experiment belong to one of two groups: significant or non-significant. It is assumed that most of the significant genes have higher fold changes than most of the non-significant genes. Based on this assumption, it is further assumed that the significant genes that are in common between two experiments will be correlated (highly ranked in both experiments) – this is because the genes that are significant in both experiments will be near the top of a list of genes ranked by fold change in each experiment. It is assumed that this correlation drops off at the point in the ranked lists where most of the genes are non-significant. The genes above this drop-off point in the ranked list are referred to as the significant set. Finally, it is assumed that the correlation holds over subsets of the significant set for a range of subset sizes.
**Calculation of Step P Values**

*Introduction to Steps*

For a given set of two ranked lists, the top \( m \) elements in list 1 and the top \( n \) elements in list 2 have an overlap of \( k \) elements. \( k \) increases by a value \( \Delta k \) when \( m \) increases by a step size \( \Delta m \) and \( n \) increases by a step size \( \Delta n \). Proceeding down the lists in steps is allowed because of the assumption that the correlation holds over subsets.

The statistical significance of each step is the probability that \( \Delta k \) found in the step is significantly more than the \( \Delta k \) that would be expected randomly.

![Diagram](Image)

*Step P Values Calculated Using 3x3 Contingency Tables*

The step p value can be represented using a 3x3 contingency table, as in table 2.3.1. Note that there are now 4 different types of overlap \( k \). This is because the different sections of lists produce distinct overlaps, as illustrated in figure 2.3.2. In the table, \( k_{AB} \) is the overlap between section A and section B in figure 2.3.2, \( k_{AD} \) is the overlap between sections A and D, etc. \( \Delta m \) is the step size in list 1, \( \Delta n \) is the step size in list 2.
### LIST 1 INDEX

<table>
<thead>
<tr>
<th>LIST 2 INDEX</th>
<th>1 to n₁</th>
<th>(m₁+1) to m₂</th>
<th>(m₂+1) to N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to n₁</td>
<td>k₁ = kₐB</td>
<td>k₉C</td>
<td>n₁-kₐB-k₉C</td>
<td>n₁</td>
</tr>
<tr>
<td>(n₁+1) to n₂</td>
<td>kₐD</td>
<td>k₉D</td>
<td>Δn-kₐD-k₉D</td>
<td>Δn</td>
</tr>
<tr>
<td>(n₂+1) to N</td>
<td>m₁-kₐB-kₐD</td>
<td>Δm-k₉C-k₉D</td>
<td>N-m₁-n₁-Δn-Δm+kₐB+kₐD+k₉C+k₉D</td>
<td>N-n₁-Δn</td>
</tr>
<tr>
<td>Total</td>
<td>m₁</td>
<td>Δm</td>
<td>N-m₁-Δm</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 2.3.1: 3x3 Contingency Table Applied to Ranked List Steps

---

Figure 2.3.2: Sections of Ranked Lists. Comparing this figure with figure 2.3.1, k₁ is the overlap of sections A and B and Δk is the sum of the overlaps between sections A and D (kₐD), sections B and C (k₉C) and sections C and D (k₉D).

Ghent (1972) derived an exact equation for the probability of a 3x3 contingency table. Applied to table 2.3.1, the probability is:

\[
p' = \frac{m₁!(Δm)!(N-m₁-Δm)!(n₁!(Δn))!(N-n₁-Δn)!}{N!kₐB!kₐD!k₉C!k₉D!(m₁-kₐB-k₉C)!(n₁-kₐB-kₐD)!(Δm-kₐD-k₉D)!(Δn-k₉C-k₉D)!(N-m₁-n₁-Δm-Δn+kₐB+kₐD+k₉C+k₉D)}
\]

(2.3.1)

The denominator is N factorial times the factorial of each element in the 3x3 table.
This p value can also be expressed in terms of binomial coefficients:

\[
p' = \left( \begin{array}{cccc} 
m_1 & m_1 - k_{AB} & \Delta m - k_{AD} & N - m_1 - \Delta m + k_{AB} + k_{AD} \\
k_{AB} & k_{BC} & k_{CD} & \Delta n - k_{BC} - k_{CD} \\
\end{array} \right) \\
\left( \begin{array}{c} 
N \\
n_1 \\
\end{array} \right) \\
\left( \begin{array}{c} 
N - \Delta m \\
n_1 - k_{AB} - k_{AD} \\
\Delta n \\
\end{array} \right)
\]

(2.3.2)

**Normalizing For \(k_{AB}\)**

The 3x3 table represents the probability that an exact set of k happens – meaning the exact combination (\(k_{AB}, k_{AD}, k_{BC}, k_{CD}\)). The probabilities will only add up to 1 over all possible combinations of the four k variables. However, only the case where \(k_{AB}\) equals the actual value of \(k_{AB}\) found is relevant to this problem. Normalization can be done to eliminate other possible values of \(k_{AB}\). This is done by dividing by the probability that \(k_{AB}\) is the actual value found. This probability is Fisher’s Exact Test, and the equation is

\[
p(k_{AB}) = \left( \begin{array}{c} 
m_1 \\
k_{AB} \\
\end{array} \right) \left( \begin{array}{c} 
N - m_1 \\
n_1 - k_{AB} \\
\end{array} \right) \\
\left( \begin{array}{c} 
N \\
n_1 \\
\end{array} \right)
\]

(2.3.3)

Another way to think about this normalization is the idea of conditional probability:

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}
\]

(2.3.4)
In words, “the probability of event A, given event that event B happened, is equal to the probability that A and B both happen divided by the probability that B happens.”

Event A is point 2 \((m_2, n_2, k_2)\) and event B is point 1 \((m_1, n_1, k_1)\). \(P(A|B)\) is the step probability (probability of A, given B), \(P(A \cap B)\) is from the 3x3 table (probability that A and B both happen) and \(P(B)\) is the Fisher’s Exact Test probability for point 1 (probability that B happens).

Note that two of the three terms in \(p(k_{AB})\) (equation 2.3.3) show up in the binomial coefficient expression of the 3x3 table probability (equation 2.3.2). Thus those two terms would disappear when normalized.

**Final Step P Value**

The final step \(p\) value is the probability that \(\Delta k\) is greater than or equal to the value of \(\Delta k\) found. This is a sum over equation 2.3.1 (or 2.3.2) divided by equation 2.3.3 for all values of \(\Delta k\) greater than or equal to the value found. This involves a sum over \(k_{AD}\), \(k_{BC}\) and \(k_{CD}\). However, the sum is shortened by some constraints:

\[
\begin{align*}
0 \leq k_{CD} &\leq \min(\Delta m, \Delta n) \\
0 \leq k_{AD} &\leq \Delta m - k_{CD} \\
0 \leq k_{AD} &\leq n_1 - k_{AB} \\
0 \leq k_{BC} &\leq \Delta n - k_{CD} \\
0 \leq k_{BC} &\leq m_1 - k_{AB} \\
k_{AD} + k_{BC} + k_{CD} &\geq k_2 - k_{AB} \text{ measured}
\end{align*}
\] (2.3.5)

For \(k_{CD}\), the first constraint is that the overlap (between sections C and D in figure 2.3.2) cannot be greater than the total number in the smaller of the step sizes – it is the overlap of segments the size of the step sizes. For \(k_{BC}\), the first constraint is that
only so many elements in section C are left to overlap with section B in figure 2.3.2 because some of the elements in section C could be in the overlap with section D, \( k_{CD} \). The second constraint on \( k_{BC} \) is the number of elements left to overlap in section B because some of the elements in section B could be in the overlap with section A, \( k_{AB} \). The constraints on \( k_{AD} \) are analogous to those on \( k_{BC} \). The final constraint is the actual p value constraint.

If \( \Delta n = \Delta m \), the number of terms in the sum before constraint is equal to \((\Delta m)^3\). Thus if \( \Delta m = 100 \), that number is 1 million. However, the second and fourth constraints alone remove roughly 60 percent of the terms. The speed of calculation is aided by the use of a table of factorial values instead of recalculating the factorial values in each term. Additionally, terms that stay the same throughout (for example, \( N! \)) are only calculated once.

**Maximization Approach to Determining Optimal Overlap Size**

We chose the statistic to be maximized, \( L \), as the measure of the total probability of getting to any point on the grid and then stopping. It is the product of the total probability \( T \) and the stability \( S \).

\[
L = T \cdot S \quad (2.3.6)
\]

In allowing the two significant sets to be of different size, the possible choices for the significant sets for list 1 and list2 can be represented by a 2-dimensional grid, as in figure 2.3.3, with a given point \((i,j)\) on the grid representing a significant set size choice of \((i \cdot \Delta m, j \cdot \Delta n)\). In practice, we keep \( \Delta m = \Delta n \). The total probability (\( T \)) for
a given point \((i,j)\) on the grid is equal to the sum over all possible paths from \((1,1)\) to \((i,j)\) of the probability of that path, \(R_{\text{path}}\).

\[
T_{i,j} = \sum_{\text{all possible paths from } (1,1) \text{ to } (i,j)} R_{\text{path}} \tag{2.3.7}
\]

The probability \(R_{\text{path}}\) of a given path depends on the length of the path and the probability \(p\) that the step will be taken.

\[
R_{\text{Path}} = \frac{1}{3^{\text{path length} - 1}} \prod_{\text{steps on path}} p_{\text{step}} \tag{2.3.8}
\]

The factors of 1/3 come from the 3 possible steps (we are allowing forward steps in only the \(x\) direction, only the \(y\) direction, or both) at each grid point. In order to choose the direction that each path takes, the other two directions must be discarded (for that particular path). The probability of choosing any of the 3 directions randomly is 1/3.

This is true with one exception. The first step has to be from \((0,0)\) to \((1,1)\), because a step to \((0,1)\) or \((1,0)\) will have a zero overlap by definition. Thus the first step has no factor of 1/3, producing the 1/3 to the power of path length minus 1.

The stability \(S\) is the probability of not taking a step in any of the three given directions, forwards or backwards.

\[
S_{i,j} = (1 - p_{(i,j) \to (i+1,j+1)})(1 - p_{(i,j) \to (i+1,j)})(1 - p_{(i,j) \to (i,j+1)}) p_{(i-1,j-1) \to (i,j)} p_{(i,j-1) \to (i,j)} p_{(i-1,j) \to (i,j)} \tag{2.3.9}
\]

The last three terms in equation 2.3.9 are a result of the probability of stepping backwards being 1 minus the probability of stepping forwards along the same step:

\[
p_{(i2,j2) \text{ back to } (i1,j1)} = 1 - p_{(i1,j1) \to (i2,j2)} \tag{2.3.10}
\]
Thus, the 1-p form of the first three terms becomes 1-(1-p) = p. Any backwards or forwards steps that would go off of the grid will not be considered. Combining equations 2.3.6 through 2.3.9 gives equation 2.3.11, the explicit formula for \( L \) at each grid point \( (L_{ij}) \):

\[
L_{ij} = \left( \sum_{\text{all possible paths from } (1,1) \text{ to } (i,j)} \frac{1}{3^{\text{path length} - 1}} \prod_{\text{steps on path}} (1 - p_{(i,j) \to (i+1,j)}) (1 - p_{(i,j) \to (i,j+1)}) p_{(i,j) \to (i-1,j)} p_{(i,j) \to (i,j+1)} p_{(i-1,j) \to (i,j)} \right)
\]

**Example**

Consider the overlap of two lists where the top 200 elements in list 1 are contained in the top 300 elements in list 2, with the overlap elements showing up in the same order in both list 1 and list 2. Calculate \( L \) for each point on the 3 by 3 grid of points (step size 100) representing the top of both lists (bottom left in figure 2.3.3 below). Figure 2.3.3 illustrates the significant steps for one simulation of data with these parameters. The blue dots mark the possible significant set choices for the given step size (100). The bottom left blue dot marks the possible significant set choice of 100 genes from set 1 and 100 genes from set 2. The blue dots are separated by step sizes of 100 and the grid shape allows for choices of significant set where different numbers of genes are chosen from the two ranked lists. The black lines represent a step that has a p value of 0.05 or less – for example, the step from \( (100,100) \) to \( (200,200) \) is significant and connected by a black line, but the step from \( (100,100) \) to \( (200,100) \) is not significant and not connected by a black line.

It happens that for our region of interest the probability of taking a step along the black lines is roughly 1 (it's about 1 – \( 10^{-11} \)) and the probability of taking a step that is not along a black line is exactly zero. Note the difference between probability and
p value – probability of a step is 1 minus the p value for that step. There is one exception – the step from (2,3) to (3,3) has probability 0.27.

In order to calculate L for the points on the 3 by 3 grid of points (step size 100) representing the top of both lists (bottom left 3x3 grid), we will first calculate its two components T and S.

\( L_{1,1} \): only one path, \((0,0)\)-(1,1)

\[ T_{1,1} = p_{(0,0) \rightarrow (1,1)} = 1 \]
\[ S_{1,1} = (1 - p_{(1,1) \rightarrow (2,2)}) (1 - p_{(1,1) \rightarrow (2,1)}) (1 - p_{(1,1) \rightarrow (1,2)}) p_{(0,0) \rightarrow (1,1)} = 0 \times 1 \times 1 = 0 \]
\[ L_{1,1} = 0 \]

\( L_{2,1} \): still only one path – \((0,0)-(1,1)-(2,1)\)

\[ T_{2,1} = p_{(0,0) \rightarrow (1,1)} * (1/3) p_{(1,1) \rightarrow (2,2)} = 1 \times 0 = 0 \]
\[ S_{2,1} = (1 - p_{(2,1) \rightarrow (3,2)}) (1 - p_{(2,1) \rightarrow (3,1)}) (1 - p_{(2,1) \rightarrow (2,2)}) p_{(1,1) \rightarrow (2,1)} = 0 \times 1 \times 0 = 0 \]
\[ L_{2,1} = 0 \]

\( L_{3,1} \): one path – \((0,0)-(1,1)-(2,1)-(3,1)\)

\[ T_{3,1} = p_{(0,0) \rightarrow (1,1)} * (1/3) p_{(1,1) \rightarrow (2,1)} * (1/3) p_{(2,1) \rightarrow (3,1)} = 1 \times 0 \times 0 = 0 \]
\[ S_{3,1} = (1 - p_{(3,1) \rightarrow (4,2)}) (1 - p_{(3,1) \rightarrow (4,1)}) (1 - p_{(3,1) \rightarrow (3,2)}) p_{(2,1) \rightarrow (3,1)} = 0 \times 1 \times 0 = 0 \]
\[ L_{3,1} = 0 \]
**L_{1,2}: one path – (0,0)-(1,1)-(1,2)**

\[ T_{1,2} = p_{(0,0)} \text{ to } (1,1) \times (1/3)p_{(1,1)} \text{ to } (1,2) = 1 \times 1/3 = 1/3 \]

\[ S_{1,2} = (1-p_{(1,2) \text{ to } (2,3)})(1-p_{(1,2) \text{ to } (2,2)})(1-p_{(1,2) \text{ to } (1,3)})p_{(1,1) \text{ to } (1,2)} = 0 \times 1 \times 1 = 0 \]

\[ L_{1,2} = 0 \]

**L_{1,3}: one path – (0,0)-(1,1)-(1,2)-(1,3)**

\[ T_{1,3} = p_{(0,0) \text{ to } (1,1)} \times (1/3)p_{(1,1) \text{ to } (1,2)} \times (1/3)p_{(1,2) \text{ to } (1,3)} = 1 \times 1 \times 0 = 0 \]

\[ S_{1,3} = (1-p_{(1,3) \text{ to } (2,4)})(1-p_{(1,3) \text{ to } (2,3)})(1-p_{(1,3) \text{ to } (1,4)})p_{(1,2) \text{ to } (1,3)} = 0 \times 1 \times 0 = 0 \]

\[ L_{1,3} = 0 \]

For the points with more than one path coming up, I will assume the ability to calculate R and S.

**L_{2,2}: three paths:**

<table>
<thead>
<tr>
<th>Path</th>
<th>R_{path}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)-(1,1)-(2,2)</td>
<td>((1 \times 1)/3^1 = 1/3)</td>
</tr>
<tr>
<td>(0,1)-(1,1)-(2,1)-(2,2)</td>
<td>((1 \times 0 \times 1)/3^2 = 0)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(1,2)-(2,2)</td>
<td>((1 \times 1 \times 1)/3^2 = 1/9)</td>
</tr>
</tbody>
</table>

\[ T = \text{sum of } R = 4/9 \]

\[ S = (1-1)(1-0)(1-1) \times 1 \times 1 = 0 \]

\[ L_{2,2} = 0 \]

**L_{3,2}: five paths**

<table>
<thead>
<tr>
<th>Path</th>
<th>R_{path}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)-(1,1)-(2,1)-(3,1)-(3,2)</td>
<td>((1 \times 0 \times 0 \times 1)/3^3 = 0)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(2,1)-(3,2)</td>
<td>((1 \times 0 \times 1)/3^2 = 0)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(2,1)-(2,2)-(3,2)</td>
<td>((1 \times 0 \times 1 \times 0)/3^3 = 0)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(2,2)-(3,2)</td>
<td>((1 \times 1 \times 0)/3^2 = 0)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(1,2)-(2,2)-(3,2)</td>
<td>((1 \times 1 \times 1 \times 0)/3^3 = 0)</td>
</tr>
</tbody>
</table>

\[ T = \text{sum of } R = 0 \]

\[ S = (1-1)(1-0)(1-1) \times 1 \times 0 = 0 \]

\[ L_{3,2} = 0 \]

**L_{2,3}: five paths**

<table>
<thead>
<tr>
<th>Path</th>
<th>R_{path}</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0)-(1,1)-(1,2)-(1,3)-(2,3)</td>
<td>((1 \times 1 \times 0 \times 1)/3^3 = 0)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(1,2)-(2,3)</td>
<td>((1 \times 1 \times 1)/3^2 = 1/9)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(1,2)-(2,2)-(2,3)</td>
<td>((1 \times 1 \times 1 \times 1)/3^3 = 1/27)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(2,2)-(2,3)</td>
<td>((1 \times 1 \times 1)/3^2 = 1/9)</td>
</tr>
<tr>
<td>(0,0)-(1,1)-(2,1)-(2,2)-(2,3)</td>
<td>((1 \times 0 \times 1 \times 1)/3^3 = 0)</td>
</tr>
</tbody>
</table>
\( T = \text{sum of } R = \frac{7}{27} \)
\( S = (1-0)(1-0.27)(1-0)*1*1*1 = 0.73 \)
\( L_{2,3} = \frac{21}{108} = 0.19 \)

\[ \begin{array}{|c|c|}
\hline
\text{Path} & R_{\text{path}} \\
\hline
(0,0)-(1,1)-(2,2)-(3,3) & \frac{1*1*1}{3^2} = \frac{1}{9} \\
(0,1)-(1,1)-(2,1)-(2,2)-(3,3) & \frac{1*0*1*1}{3^3} = 0 \\
(0,0)-(1,1)-(1,2)-(2,2)-(3,3) & \frac{1*1*1*1}{3^3} = \frac{1}{27} \\
(0,0)-(1,1)-(2,1)-(3,1)-(3,2)-(3,3) & \frac{1*0*0*1*1}{3^4} = 0 \\
(0,0)-(1,1)-(2,1)-(3,2)-(3,3) & \frac{1*0*1*0*1}{3^4} = 0 \\
(0,0)-(1,1)-(2,2)-(3,2)-(3,3) & \frac{1*1*0*1}{3^3} = 0 \\
(0,0)-(1,1)-(1,2)-(2,2)-(3,2)-(3,3) & \frac{1*1*0*1*1}{3^4} = 0 \\
(0,0)-(1,1)-(1,2)-(1,3)-(2,3)-(3,3) & \frac{1*1*0*1*0.25}{3^4} = 0 \\
(0,0)-(0,1)-(1,1)-(1,2)-(2,2)-(3,3) & \frac{1*1*1*0.25}{3^3} = \frac{1}{108} \\
(0,0)-(0,1)-(1,1)-(2,2)-(2,2)-(3,3) & \frac{1*1*1*0.25}{3^4} = \frac{1}{324} \\
(0,0)-(0,1)-(1,1)-(2,2)-(2,3)-(3,3) & \frac{1*1*1*0.25}{3^3} = \frac{1}{108} \\
(0,0)-(0,1)-(2,1)-(2,2)-(2,3)-(3,3) & \frac{1*0*1*1*0.25}{3^4} = 0 \\
\hline
\end{array} \]

\( \text{T = sum of } R = \frac{55}{324} \)
\( S = (1-0)(1-0)*1*0.27*1 = 0.27 \)
\( L_{3,3} = (55/324)*0.27 = 0.05 \)

Based on the method, (2,3) is selected as the best point, which is the correct answer based on the inputs to the simulations. The one term that is not equal to 1 in \( S_{3,3} \) represents the fact that the step from (2,3) to (3,3) has only a 25% probability – the consideration of the backwards steps in \( S \) caught the fact that the step from (2,3) to (3,3) was improbable, which should make (2,3) a more logical solution.

Notice that the 13 paths in \( L_{3,3} \) can be divided up as the paths through the 3 points that are allowed to step to (3,3):

- 3 steps through (2,2) - yellow
- 5 steps through (3,2) - blue
- 5 steps through (2,3) - green
Computationally, this represents the ability to compute \( L \) recursively for every point, given all of the step probabilities. \( S \) is calculable for every point, and \( T \) can be expressed as:

\[
T_{ij} = T_{i-1,j} \frac{1}{3} p_{(i-1,j) \to (i,j)} + T_{i,j-1} \frac{1}{3} p_{(i,j-1) \to (i,j)} + T_{i+1,j} \frac{1}{3} p_{(i,j) \to (i,j)}
\]  

(2.3.11)

When \( i \) or \( j \) is equal to 1, the formula becomes a simplified version of (2.3.11).

**Determining Step Size**

The only adjustable parameter in this method is the step size (\( \Delta m \) or \( \Delta n \)), which sets the size of the grid describing the data (as in figure 2.3.3). \( \Delta m \) and \( \Delta n \) are allowed to be two possible values: 0 or an overall step size \( \Delta \) (the allowance of 0 is to create the steps in only one list or the other list). Figure 2.3.4 illustrates the effect of step sizes on a set of real data. The data is RNA sequencing data from the Helfand Lab (Wood et al. 2012) for female flies aging (day 10 to day 40) on a normal diet and aging on a dietary restriction (DR) diet (DR is analyzed in chapter 1). As shown in the figure, \( \Delta = 50 \) and \( \Delta = 100 \) choose smaller significant set sizes than those chosen by larger values of \( \Delta \). \( \Delta = 50 \) chooses 2400 genes from list 1 and 2850 from list 2, \( \Delta = 100 \) chooses 4000 genes from list 1 and 3900 genes from list 2 and \( \Delta = 200, 300 \) and 400 choose 4800 genes from each list. The black lines on figure 2.3.4 represent steps with False Discovery Rate (Benjamini and Hochberg 1995) less than 0.05 for \( \Delta = 100 \). It appears that \( \Delta = 50 \) and \( \Delta = 100 \) stop because of one step that is insignificant in the data – there is a visible gap right after the stopping point for \( \Delta = 100 \), followed by
significant steps out to (4800,4800). The smaller the step size, the more likely it is that random fluctuations will happen. The larger step sizes will jump past small random fluctuations.

Thus, we propose to handle fluctuations by testing multiple step sizes and find the point where increasing step size no longer increases the resulting set sizes. For a set of step sizes such as (50, 100, 200, 400, 800) where the step sizes increases by a common factor (2), the best step size is the first step size to have the step size after it select set sizes that are within one step of it. In other words, a set of results like table 2.3.2 would suggest using Δ=200, because the result for Δ=400 is within 400 of the result for Δ=200.

<table>
<thead>
<tr>
<th>Δ</th>
<th>Set Size Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>(950,900)</td>
</tr>
<tr>
<td>100</td>
<td>(1500,1500)</td>
</tr>
<tr>
<td>200</td>
<td>(2200,2200)</td>
</tr>
<tr>
<td>400</td>
<td>(2400,2400)</td>
</tr>
<tr>
<td>800</td>
<td>(2400,2400)</td>
</tr>
</tbody>
</table>

Table 2.3.2 Example of Set Size Selection. Step size Δ=400 is the first step size to select the same result as the step before it, to within the step size.

An alternative that does not use as much computational time is to use a plot similar to figure 2.3.4 (with only the position for Δ=100 shown) in order to determine whether or not the answer given has been affected by a random fluctuation. For example, figure 2.3.4 shows that a step size of 100 still has several significant steps past the chosen significant sets (blue point). This would suggest trying a larger step size, such as Δ=200. Trying Δ=200 would yield a point with no significant steps past.
the significant set choice (green point). This method could choose a step size after only two trial step sizes, whereas the method described above may try many more step sizes and take longer computationally.

Figure 2.3.4 Illustration of different step size choices. Red is $\Delta=50$, which chooses significant set (list1,list2)=$\Delta=50$, which chooses significant set (list1,list2)=(2400,2850), blue is $\Delta=100$, green is $\Delta=200$, 300 and 400 (4800,4800), magenta is $\Delta=800$, which barely finished above (4800,4800), $\Delta=250$ chose (4750,5000) and $\Delta=150$ chose (3300,2850). Black lines connect steps with $\Delta=100$ and corrected p value $\leq 0.05$. Several black lines occur after the blue point, but few occur after the green point.
2.4 Simulations

This section will cover three topics:

1. Simulation Methods
2. Basic Simulation Results
3. Simulations Where Correlation Decreases Gradually

Simulation Methods

The data simulation is meant to be basic in order to provide a case with a known answer. The simulation skips expression values and goes immediately to setting ranks – each list will consist of the values 1 through N, where N is the list length. For inputs of set size 1 (m), set size 2 (n), overlap (k) and list length (N),

- Set list 1 equal to 1 through N
- Randomly select k indices between 1 and m for list 1 and between 1 and n for list 2
- Put both sets of indices in order – this produces a more uniform overlap.
- Set the k selected indices in list 2 equal to the k selected indices in list 1, creating the overlap.
- Randomly fill out the set 1 through n in list 2 such that the overlap size of the top n elements in list 2 and the top m elements in list 1 (which are 1 through m) is exactly k. This amounts to choosing from the set of values between m+1 and N to fill out rows 1 through n in list 2.
- Randomly fill the rest of list 2 with the remaining indices.
In the gradually decreasing correlation simulation, 200 rows that are exactly the same are added on at the top of both lists – this amounts to adding 200 to the existing lists and putting 1 through 200 at the start of each list.

**Basic Simulation Results**

Figure 2.4.1 shows the simulation results for the case \( m=200, n=400, k=200, N=1200 \) with a step size \( \Delta=50 \). In the majority of 100 simulations, the method chooses the correct set size. This is also true for step sizes \( \Delta=100 \) and \( \Delta=200 \).

![3D scatter plot showing simulation results](image)

**Figure 2.4.1 Simulation Results** for \( m=200, n=400, k=200, N=1200, \Delta=50 \). 100 data sets were simulated, and the probability of selecting each set size is displayed.
Figure 2.4.2 shows the simulation results for m=400, n=400, k=200, N=1200 with a step size Δ=50. The algorithm chooses (250,250) and (300,300) at roughly an equal rate, with some bias towards (300,300). Although this is not the set size specified, it illustrates the stopping point of the algorithm. Assuming the 50% overlap specified in the inputs holds on average, the step from (250,250) to (300,300) has a p value of roughly 0.006 (before FDR correction). On average, the step from (300,300) to (350,350), rarely made, has a p value of 0.07 before Benjamini-Hochberg False Discovery Rate correction (Benjamini and Hochberg 1995). In chapter 1, a p value cutoff of 0.01 was used instead of a False Discovery Rate correction.

![Figure 2.4.2 Simulation Results](image)

Figure 2.4.2 Simulation Results for m=400, n=400, k=200, N=1200, Δ=50. 1000 data sets were simulated, and the probability of selecting each set size is displayed.

As a negative control, simulation results for random data (m=n=k=N=1200) with a step size Δ=50 chose the smallest possible size on the grid: (1,1).
Simulations Where Correlation Decreases Gradually

As a more realistic simulation, we chose the simulation parameters used in figure 2.4.2 (which already has a gradually decreasing correlation) and added 200 rows of perfect overlap to the top of the lists (as discussed in the simulation methods above). The results (figure 2.4.3) depend on step size, but center around (500,500) or (600,600). Assuming 50% overlap, the step from (500,500) to (550,550) has a p value of 0.07 and the step from (500,500) to (600,600) has a p value of 0.09. Notice that for step size 400, (400,400) is chosen, but for step size 350, (700,700) is chosen. This illustrates the fact that the results from a given step size are only accurate to within one step.

Figure 2.4.3 Simulated Data With Decreasing Correlation. Green (400,400) is chosen by $\Delta=400$, red (500,500) is chosen by $\Delta=50, 100$ and 250; blue (600,600) is chosen by $\Delta=150, 200$ and 300; magenta (700,700) is chosen by $\Delta=350$. Black lines connect steps with $\Delta=100$ and corrected p value ≤ 0.05.
2.5 Application to Aging and Dietary Restriction Data

The method was run on the data set described in section 2.3, and additionally on the male flies from the same experiment. It was also run on the data from chapter 1.4. The analysis yielded the following biological conclusions:

- Overlap between *Drosophila* Females Aging on High Calorie and Low Calorie Diets is 4320 genes up and 2612 genes down
- Overlap Between *Drosophila* Males Under Dietary Restriction at Young and Old Age is 1738 genes up and 6640 genes down
- Comparison of *Drosophila* Females with sir2 Mutation with Females Under Dietary Restriction Shows Significant Set Sizes of 2600 Genes Up in sir2 and 3800 Genes Up in Dietary Restriction

Overlap between *Drosophila* Females Aging on High Calorie and Low Calorie Diets is 4320 genes up and 2612 genes down

Figure 2.3.4 used genes going up in *Drosophila* aging on either a comparatively high calorie (control) or low calorie (DR) diet to illustrate the use of different step sizes in order to choose an optimal set size for each list. We chose to use $\Delta=200$ in each list, because the results do not increase significantly with larger step sizes. Using the results from $\Delta=200$, the final result would be a selection of 4800 genes up from high calorie food and 4800 genes up from low calorie food (“up” means upregulated in old compared to young, etc; a gene is included in the up comparison if it is upregulated in one of the two conditions being compared). Using the same method, we find 3200 genes down from high calorie food and 3200 genes down from low
calorie food. These set sizes produce an overlap of 4320 genes up and 2612 genes down. These overlaps are consistent with *Drosophila* females aging similarly on low calorie and high calorie diets. A GOStat analysis (Beissbarth and Speed 2004) of these genes indicates a biological response including:

Genes Up
- Cell Communication
- Neurogenesis
- Immune System
- Detection of Stimuli
- Circadian Rhythm
- Tissue Death/Autophagy
- Aging
- Organ Development
- Regulation of Gene Expression

Genes Down
- Mitochondria Functions (Electron Transport, Lumen, Respiratory Chain, NADH)
- Oxidoreductase
- Transport
- Chromatin Modification
- Metabolism

All of these changes seem to resemble the changes seen in p53/sir2/DR in chapter 1 and low dose radiation in chapter 3.
Overlap Between *Drosophila* Males Under Dietary Restriction at Young and Old Age is 1738 Genes Up and 6640 Genes Down

From the same data set, we now compare males on dietary restriction (DR) at ages 10 days (young) and 40 days (old). After trying multiple step sizes and using the results from $\Delta = 300$ for upregulated genes and $\Delta = 200$ for downregulated genes, the final result would be a selection of 2400 genes up from day 10 DR and 3300 genes up from day 40 DR. Using the same method, we find 6800 genes down from day 10 DR and 7000 genes down from day 40 DR. These set sizes produce an overlap of 1738 genes up and 6640 genes down. A GOStat analysis (Beissbarth and Speed 2004) of these genes indicates a biological response including:

- **Genes Up**
  - Cell Communication
  - Neurogenesis
  - Neuropeptide Hormone Activity
  - Circadian Rhythm
  - Upregulation of Metabolism
  - Heterochromatin
  - Aging
  - DNA Repair
  - Response to Stimuli

- **Genes Down**
- Ribosome
- Mitochondria (lumen, envelope, electron transport, respiratory chain)
- Metabolism (of protein, carboxylic acid, mRNA, phosphorus)
- Oxidoreductase
- Detection of light stimulus
- Nuclear, RNA Transport
- Neurogenesis
- Apoptosis
- Response to Stress
- Circadian Rhythm
- Neurotransmitter Secretion
- Tissue Death
- Immune System
- Negative Regulation of Signal Transduction
- Aging
- Chromatin Binding

As noted previously, DR has wide-ranging effects. Not all of these effects are necessarily related to aging, but this analysis shows that these effects carry through old age in restricted diet flies. Several categories are shared between the list of functions from genes up and genes down. In particular, the genes upregulated in DR at both ages have cell communication and the genes downregulated at both ages have negative regulation of cell communication. This suggests more strongly that
cell communication is increasing. These results are consistent with dietary restriction affecting males in a similar manner at young and old age.

Comparison of *Drosophila* Females with sir2 Mutation with Females Under Dietary Restriction Shows Significant Set Sizes of 2600 Genes Up in sir2 and 3800 Genes Up in Dietary Restriction

In chapter 1, we labeled genes as either significantly changing or not depending on cutoffs of p value, fold change and present/absent. We return to this data set to see if this algorithm shows sir2 mutants and flies under dietary restriction (DR) to have significantly different set sizes. The expectation would be for DR to have more genes because it is known to be a more wide-ranging effect than a sir2 mutation. Trying different set sizes and settling on $\Delta=200$, we find that 7400 total genes are selected in sir2 mutants (2600 up, 4800 down) and 8600 total genes are selected in DR (3800 up, 4800 down). The result does select more genes from DR than sir2, with the effect coming from genes upregulated from control. While this is a smaller percentage difference in set size between sir2 and DR than the difference seen in section 1.4, this result is consistent with sir2 being “downstream” of DR in the *Drosophila* aging pathway.
2.6 Discussion

Comparison with Plaisier et al.

The methods of Plaisier are fairly simple and elegant. For every possible combination of significant set sizes in each list (m genes from experiment 1, n genes from experiment 2, overlap k and total number of genes N), they compute the hypergeometric probability based on Fisher’s Exact Test:

\[
p \text{ value} = \sum_{\text{measured } k} \left( \frac{m}{k} \right) \left( \frac{N - m}{n - k} \right) \left( \frac{N}{n} \right) \left( \frac{m}{m + k - n} \right)
\]  

(2.6.1)

They select the set (m,n) that gives the minimum p value as the location of the significant sets. They do allow for two selections – one representing the overlap between genes up and one representing the overlap between genes down. They correct these p values for multiple hypothesis testing (although this shouldn’t affect which values are minimum, only their significance level). Finally, they produce a heatmap showing the logarithm of the p value for all combinations of (m,n).

This method makes an implicit assumption that the point with the highest significance is the end of the significant correlation between the two lists. This assumption is not correct.

Consider the following scenario. In two ranked lists of length 6000,

- Elements 1-100 in each list have a complete overlap (m=n=k=100)
- Elements 101-200 in each list have overlap 50 (m=n=200, k=150)
- Elements 201-300 in each list have overlap 25 (m=n=300, k=175)
• The rest of the list is randomly distributed

The hypergeometric p value of \((m,n,k)=(100,100,100)\) is of order \(10^{-220}\), the hypergeometric p value of \((200,200,150)\) is of order \(10^{-209}\) and the hypergeometric p value of \((300,300,175)\) is of order \(10^{-169}\). The p values will get worse once the random overlap begins. Thus, the method of Plaisier et al. method would select somewhere around \(m=n=100\).

Our step p value method, described in section 2.3, measures the significance of small steps in the data. Taking a step down the list of size 100 gives:

- 0 to 100: \(10^{-209}\)
- 101 to 200: \(10^{-67}\)
- 201 to 300: \(10^{-15}\)

Stepping down the list to elements 101-200 and 201-300 still adds a significant number of genes, even if it isn't the maximum overall significance. Phrased another way, selecting \((m,n,k) = (300,300,175)\) would select 75 more genes in the overlap.

Another argument for stepping through the list involves the step from 301 to 400 in the example above. If the overlap added is 6 genes, the p value for the step using our method is 0.58, not significant. However, the overall overlap (in equation 2.6.1, \(m=n=400, k=181, N=6000\)) is of order \(10^{-119}\). A method based solely on the total overlap and not stepping may miss the transition from significant correlation between lists to random correlation between lists.
Comparison with Yang et al.

As mentioned previously, Yang’s method centers on a newly defined similarity score. For a given ranked list index $m$, the score is the overlap of the top and bottom $m$ genes in each list multiplied by an exponential decay term:

$$S_\alpha = \sum_{m=1}^{N} e^{-\alpha m} (k_{m,\text{top}} + k_{m,\text{bottom}})$$

Here, $k$ represents an overlap, and $\alpha$ is a tuning parameter. The choice of $\alpha$ is analogous to choosing the set size in our method, because the score is set to zero when the exponential factor reaches a certain magnitude. They determine $\alpha$ by calculating the similarity score for permutations of the data and fashioning the results into a Receiver Operating Curve for several trial values of $\alpha$. In their bioconductor software package *OrderedList*, their default trial values of $\alpha$ amount to trying set sizes of 100, 200, ..., 2500. This is analogous to setting a step size of 100.

This method has several important differences from our method:

- Yang’s method assumes that the top and bottom significant sets in each list (4 total significant sets) are all of the same size. Our method allows the 4 sets to all be of different sizes. This is a more general solution.

- The exponential factor in equation 2.6.2 places extra importance on the overlap between genes at the very top of the ranked lists. However, the overlap of genes near the end of the significant set may still have very significant fold changes and be very important.

- Both methods essentially set a step size, but only our method looks at the probabilities of the individual steps.
• It seems that multiple uses of data permutation to create a random data set may create some problems - often, comparisons involve two things that are somewhat alike. So permutation of treatment “labels” (for example, cancer type 1 samples with cancer type 2 samples or controls), the result may not be completely random.

• Yang’s method depends on a process that removes individual expression data replicates (getting the “true positives” for the Receiver Operator Curves). Yang’s example data set has roughly 12 replicates per condition (experiment or control), but many or most data sets have about 3 replicates. With this number of replicates, it becomes detrimental to remove a replicate. Our method always keeps each replicate.
2.7 Conclusion

This chapter describes a new algorithm for comparing gene expression data based on ranking the data by fold change and finding the statistical significance of overlap between different sections of the ranked lists. There are two innovations in the method – finding the p value for a given step in the list and the use of these p values to determine the point of significance. There is only one parameter, the step size, and multiple step sizes can be used to find an optimal step size. The method performs as expected on simulated data, and identifies interesting biological similarities and differences when applied to *Drosophila* datasets from both sequencing and microarray experimental data. The method is general enough that it could be applicable to data from fields outside of gene expression and biology.

With the growing number of gene expression datasets, comparisons between datasets will become increasingly important. This method is an improvement on existing ranked list methods and should prove a useful tool in the field.
Chapter 2 References


Chapter 3: Response to Low-Dose Radiation in *Drosophila melanogaster*

3.1 Introduction

It is important for science and for safety to determine the correct biological response to low doses of radiation. There are many situations where a small amount of extra radiation is imparted to human beings, such as the use of cell phones, full body scanners at airports, medical imaging procedures and the addition of nuclear power plants. Depending on the correct model of radiation damage, this dose could be beneficial, harmful or neutral. As one example of a low dose radiation addition, approximately 0.01 percent of background ionizing radiation currently comes from nuclear power plants (United Nations Scientific Committee on the Effects of Atomic Radiation, 2000). Would the addition of more nuclear power plants pose a cancer risk? A second example is a study of over 53,000 patients by the National Cancer Institute (National Lung Screening Trial Research Team, 2011) which found that lung cancer patients imaged using computed tomography (CT) scans had 20% fewer deaths than those imaged using medical x-rays despite the increased radiation from CT scans. It seems in this case that the additional quality of image in the CT scans outweighed any potential damage from radiation, but is there still a risk from the low doses of radiation?

There are currently three models of biological response to low dose radiation. The Linear No Threshold Theory (LNT) is that all doses of radiation are harmful, regardless of dose, and that the damage scales linearly with dose. For example, if
dose X causes 1000 deaths from cancer, dose 10*X will cause 10,000 deaths and
dose 0.01*X will cause 10 deaths. The threshold model is that doses below a certain
level are not harmful, likely because a repair mechanism can handle the radiation-
induced damage at that level. The hormesis model is that low doses of radiation are
beneficial.

A theoretical argument for threshold or hormetric effects is as follows. Suppose that
radiation damage increases linearly, as in figure 3.1.1A. Then suppose that there is a
cell damage repair mechanism. Figure 3.1.1B illustrates two possible responses –
one that quickly increases to a maximum response at a low dose, and one that
reaches the maximum response level more slowly. Figure 3.1.1C shows the net
response, which is damage minus repair response. The faster response results in a
hormetric effect and the slower response results in a threshold effect. A slower or
nonexistent response could result in a linear or linear-like effect.
Although the idea of a repair response seems intuitive, the National Academy of Sciences supports the LNT. Figure 3.1.2 is from a book by the Committee to Assess...
Health Risks from Exposure to Low Levels of Ionizing Radiation (Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation 2006). Their experimental evidence is the excess rates of solid cancers from Japanese atomic bomb survivors. They find that a quadratic (nonlinear) fit cannot be separated from a linear fit using this data. All of the low dose radiation papers regarding lifespan and *Drosophila* (section 3.2) support a threshold or hormesis model.

![Graph showing Excess Solid Cancer Risk vs. Radiation Dose](image)

Figure 3.1.2: Excess Solid Cancer Risk vs. Radiation Dose. Note 1 Sv = 1 Gy = 100 rem. (Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation 2006)

The first two chapters have covered the analysis of gene expression data, with all of the example data coming from aging data sets. This chapter describes my own experimental work testing the hypothesis that lifespan effects in flies follow the threshold model and perhaps the hormesis model. As with the work in the first two
chapters, this is a low dose stressor; so the analytical methods used in the first two chapters will also be applicable to this data set.
3.2 Literature Search

One of the ideal low dose radiation experiments is a longevity experiment on *Drosophila melanogaster*. As a model organism, *Drosophila* has been studied for over a century. Longevity studies on flies last approximately three months, whereas studies on mice last approximately three years. Flies also produce many more offspring than mice, allowing for much larger studies. One caveat of using *Drosophila* is that most of the cells in an adult fly are postmitotic (no longer dividing). However, stem cells are known to be present in the gut (for example, Ohlstein and Spradling 2005).

This survey of existing literature focuses on studies of low dose radiation and longevity on *Drosophila Melanogaster*. A few studies from other insects are introduced as useful, as are aspects of studies on effects on effects besides longevity. A good starting point for low dose radiation literature in mammals is the book “Radiation Hormesis” by Luckey (1990). Many of the experimental methods are discussed in detail, for comparison with my experimental methods later on in the chapter.

The literature survey is divided into four sections:

- Early Work: Sacher, Lamb, Baxter and Blair, Planel and Geiss
- Interaction of Lifespan, Diet and Radiation-Induced Sterility – Carey
- Physical Properties of Low-Dose Radiation Flies
- Towards Genetic and Molecular Mechanisms
Early Work: Sacher, Lamb, Baxter and Blair, Planel and Geiss

George Sacher

George Sacher (1963) wrote one of the earliest detailed papers on low-dose radiation in *Drosophila melanogaster*. He also summarized the few existing papers that came before his:

- Strehler (1962) gave flies single x-ray doses of 5,000 rem. He saw a 40% increase in median lifespan. However, when he took care to remove pathogens in the fly vials this effect disappeared. In addition, he did not change the food at all for the duration of the experiment – when he did, the median lifespan doubled.

- Gowen and Stadler (1962) found that single x-ray doses of 12,500 rem did not change the lifespan of the flies and that a dose of 62,500 rem reduced lifespan by 60%.

- Henderson (along with Baxter and Blair, whose other work is detailed below) demonstrated that radiating the flies’ food did not affect the overall lifespan, up to doses of 1,000,000 rem.

- Davey (1917, 1919) and Cork (1957) experimented on the Flour Beetle, *Tribolium confusum*. Davey used x-rays, Cork used Cobalt-60 gamma rays. Both found “small” increases in lifespan; Cork’s increase occurred at a dose of 3,000 rem.

- Terzian (1953) gave mosquitoes doses of 5,000-30,000 rem and found that their resistance to avian malaria increased.
Lifespan improvements in mice are due to a decrease in various infectious diseases (Sacher 1961 and 1962).

In an attempt to relate the results from different insect species, Willard and Cherry (1975) investigated radiosensitivity in 12 different types of insects, using a regression analysis to find that 46.3% of differences between species can be attributed to control lifespan and 32.6% of differences between species can be attributed to control body weight.

Sacher’s own work started with a fly food containing pieces of raisins (for breeding flies only), a change from the sugar and yeast mixes employed in contemporary experiments. Sacher gave different groups of flies doses of x-irradiation (200 k.v., 15 milliamp) at different intervals: every day, every second day, every third day and just once. The daily doses were given at a dose rate of 500 rem/minute and the one-time doses were given at a rate of 2,000 rem/minute.

Sacher finds that the flies live longer than non-irradiated flies in the 1,500-3,000 rem/day range, that in that range dose per day is the most important determining factor of lifespan and that the lifespan decreases “in linear fashion” after the peak (figure 3.2.1). In the figure, note the spread in points at zero radiation – there was significant variability in the control sets.
Figure 3.2.1: Sacher's Radiation Lifespan Data (1 Roentgen approximately equals 1 rem).

Sacher includes the survivorship plots for flies given the same dose every day (figure 3.2.2), and finds that there is no increase in the maximum lifespan with low-dose radiation, only in the average lifespan.
Sacher notes that there is much more variability in the control sets than in the radiation sets, and attempts to explain these two phenomena by hypothesizing that the improvement in average lifespan at low doses of radiation is caused by the elimination of harmful “random variables”. The random variables strike the control sets at varying degrees, but are more controlled in the radiation sets, resulting in less variance.

There are potential problems with this hypothesis. In the control sets, the two that lived the longest had the largest variance. This could mean that those two were simply error prone. A statistical analysis in Sacher’s paper showed the correlation between mean lifespan in each vial and standard deviation in each vial to be statistically insignificant. Marion Lamb (1964, mentioned in more detail below) argues directly against Sacher’s model – her data has flies that live longer in mean and maximum, with variances that match their controls.
Sacher hypothesized a damage model with two components: a damage term that increases with radiation and a term that causes the benefit. The benefit term increases quickly at first, and then levels off at a maximum value. This view is summarized by figure 3.2.3 below. This model is very similar to the hormesis model used today.

![Figure 3.2.3: Sacher's Model of Repair.](image)

Figure 3.2.3: Sacher’s Model of Repair. The solid line with negative slope is the same line as the single dose curve from figure 3.2.1 above, representing the damage. The dashed line is fitting that line to the multiple dose data. The solid line at the bottom of the figure represents the difference between the damage line and the actual data, illustrating the “advantage factor”.

Marion Lamb

Marion Lamb’s first paper (1964) is about work done with *Drosophila subobscura*, not *melanogaster*. However, it is widely quoted as one of the basic early works for low dose radiation. She discovers a large lifespan increase (from 70 days control to
around 110 days), and speculates that this is due to sterilization. To investigate this, she made sterile mutants, using the gene *granddaughterless*. The radiation does not extend their lifespan, leading to the conclusion that sterility is causing the lifespan extension. However, Lamb cautions that she did the two experiments on different foods. She then does an experiment where she changes food levels and freshness, finding that only the “nutritionally suboptimal conditions” show lifespan extension. This work suggests the possibility that nutritional stress and radiation lifespan extension work through the same pathways.

In the book *Radiation and Ageing* (1969), Lamb argues that higher doses of radiation at later ages linearly reduce the remaining life expectancy. However, she claims lifespan increases in male *Drosophila Melanogaster* on low doses of radiation. She adds that her study of diploid and triploid cells (1965) demonstrates that somatic cell lethal mutations are not the cause of lifespan shortening. She shows that the linear dependence at high dose still occurs at different temperatures. There is almost a threshold effect at doses below 20,000 rem in her experiments, with irradiations at 3, 9, 15, 21, 27 and 35 days.

In another paper from 1969, Lamb takes the data of Baxter and Blair (mentioned in detail below) and fits it to her model where the remaining life expectancy (from age at irradiation) is reduced linearly. The fit works well enough to convince Baxter and Blair, who write a paper agreeing with her in 1970. She argues for a model where the number of surviving important target “organs” (probably cells) is

\[ S = Ne^{-kR} \]  

(3.2.1)
Where $N$ is the original number of cells, $k$ is a constant and $R$ is the dose. The implications of this model are that aging is a process of wearing the organs out, and one hit will destroy the cell. The effect in Lamb’s model actually ends up being linear in log scale:

$$\ln S = C - kR$$

(3.2.2)

Where $C$ is another constant.

(A side note: one possible biological mechanism of Lamb’s aging effects is oxidation.

Parashar et al. (2008) investigated age dependent death due to high dose radiation and paraquat, both of which induce oxygen radicals. The dose necessary to kill half the flies in 2 days is found using a Cs-137 source with dose rate of 227.9 rem/minute. The lethal dose for half of the flies in 2 days ($LD_{50/2}$) goes down from about 124,000 rem when irradiated at day 1 to 48,000 rem when irradiated at day 50. In a separate experiment, radiation is given early in life and followed by paraquat at day 10 or 20. The result is the same age-increasing lack of resistance.)

**Baxter and Blair**

Even though Sacher and Lamb are most often cited as founding work on low dose radiation in fruit flies, Baxter and Blair were contemporaries of Sacher and Lamb who developed their own models of radiation and aging.

A 1967 paper (“Kinetics of Aging”), predicts a threshold dose of 18,000 - 25,000 rem. Baxter and Blair do a small set of experiments on 100 male flies each, and find no significant change in lifespan below 15,000 rem.
Like Marion Lamb’s 1964 paper, they argue that these might only happen in suboptimal conditions. They point to Noethel’s 1965 paper, where lifespan extension was seen at doses of 6,000 and 11,000 rem and only a small lifespan decrease was seen at 67,000 rem. However, Noethel’s lifespans were dramatically shorter than usual (only 19.4 days for control females), likely because the fly cages were not changed often enough (every third day; most laboratories today change cages every day or every second day).

(Another side note: Baxter and Blair ran 28 different lifespan experiments for control males and females done at different times of the year. This is a good data set to use in investigating the “seasonal effect” in fly lifespan. I ran an analysis of variance (ANOVA) on this data, and found two factors to be significant: time of year (p value 0.006) and gender (p value 1.1x10^{-5}). The interaction between time of year and number of flies was also significant (p value 0.012), as was the interaction between time of year and gender (p value 0.035). The time of year is given as month of eclosion. I represented this in the analysis by the minimum number of months separating that month from January (so that for example December would be closest to January and not farthest).)

A second paper in 1967 (Age of Death in Drosophila Following Sublethal Exposure to Gamma Radiation) provides a set of lifespan data for several levels of radiation.

<table>
<thead>
<tr>
<th>Dose (rem)</th>
<th>Survival (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>1,000</td>
<td>38</td>
</tr>
<tr>
<td>2,500</td>
<td>42.8</td>
</tr>
<tr>
<td>5,000</td>
<td>43.4</td>
</tr>
<tr>
<td>15,000</td>
<td>42.4</td>
</tr>
<tr>
<td>25,000</td>
<td>40</td>
</tr>
</tbody>
</table>

Table 3.2.1: Baxter and Blair’s Low-Dose Radiation Lifespans
Cobalt-60 gamma (dose rate approx. 1070 rem/min) irradiation was given to *Drosophila* of background Swedish-R. 200-500 males and females were used in each lifespan experiment. Light ether was used on the flies when switching vials in order to prevent escape, something that current labs do not do. A small experiment with 50 males and 50 females showed that no effect on lifespan due to ether. Etherized flies were checked for death using signs of movement and abdomen shape. Flies were switched between vials every day. Food was present in the irradiation chamber, and temperature was not controlled.

Baxter and Blair again test for the existence of a threshold dose, trying different ages and levels of radiation below 25,000 rem in both x-rays and Cobalt-60 gamma rays (table 3.2.2).

<table>
<thead>
<tr>
<th>Number of flies</th>
<th>Age of exposure</th>
<th>Dose (kR)</th>
<th>Age of death (Male)</th>
<th>Age of death (Female)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>35</td>
<td>0</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>50</td>
<td>35</td>
<td>5</td>
<td>45</td>
<td>47.3</td>
</tr>
<tr>
<td>50</td>
<td>35</td>
<td>10</td>
<td>44.7</td>
<td>48</td>
</tr>
<tr>
<td>50</td>
<td>35</td>
<td>20</td>
<td>44.3</td>
<td>47</td>
</tr>
<tr>
<td>354</td>
<td>1</td>
<td>6.8</td>
<td>40.5</td>
<td>48.9</td>
</tr>
<tr>
<td>375</td>
<td>25</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>350</td>
<td>1</td>
<td>13.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>1</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>2.5</td>
<td>42.8</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>5</td>
<td>43.4</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>15</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>25</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

*The upper seven groups were exposed to γ-radiation, and the lower six groups to X-irradiation. Only 25 kR of X-rays would be expected to shorten life (t).*

Table 3.2.2: Formal Low-Dose Radiation Experiments by Baxter and Blair

Lifespan increases 11 and 7 percent in females for doses of 6,800 and 13,700 rem given at day 1. Most of the gamma ray experiments (top half of the table) suggest a
threshold response, even the ones given at a later age (35 days). Baxter and Blair conclude that the effect is small and hard to determine exactly.

**Planel and Geiss**

Planel and Geiss showed a reverse hormetic effect (1973A, 1973B). Flies were kept in two groups: control, and shielded by 10cm of lead in order to have the flies protected from background radiation. The flies were kept at constant temperature and were ventilated, but both sets of flies were kept in the dark and their food was only changed every 2 weeks (standard procedure in labs today is changing food every 1-2 days). In addition, the flies were only counted once per week. Still, the results are significant. The survival plot shown in figure 3.2.4 shows a mean lifespan difference of about 2 weeks, with the control flies living longer.

![Figure 3.2.4: Survival for Drosophila females in background radiation (solid curve) and shielded (dashed)]
Interaction of Lifespan, Diet and Radiation-Induced Sterility – Carey

James Carey’s lab works with the Mediterranean fruit fly, not *Drosophila melanogaster*. However, their work on the interaction of longevity, sterility and diet advances the work of Lamb (1964) and is notable for the enormous number of flies in their lifespan experiments. In a 2001 paper, he used 536,000 total flies over 8 experimental conditions, many more than the 100-500 flies per condition that most other fly aging labs are able to use.

The 8 experimental conditions come from 3 variables with 2 values each: gender, diet (either sugar-only or a full diet of sugar and yeast), and irradiation (zero or a “sterilizing dose” of 14,000 rem from a Cobalt-60 source 2 days before emergence).

As seen in figure 3.2.5, the males had very similar performance across all conditions, although the full-diet irradiated flies did a small bit better and the sugar-diet irradiated flies did a small bit worse.

The female responses are very different – only the two sugar diets have similar shapes. Carey argues that the shapes of the female curves are very suggestive of an interaction between sterility and diet. He says that the nutritional (protein) needs of fertile female flies require the full diet to avoid using reserves, and that sterilization lessens or eliminates those needs. A comparison of the two full diet female lifespan data curves reveals that the full diet intact females die at a lesser rate than the full diet sterile females until day 13. He says that this is due to the protective effect of eggs. Carey runs the data through a 2-way Multivariate Analysis of Variance, concluding that sterilization and diet affect the lifespan of females, as well as the interaction of the two treatments.
In a study from 1995, Carey and Liedo get a different result when sterilizing flies. The males see a significant decrease in lifespan (from 16.6±1.21 days to 10.9±0.71 days), while the females see a significant increase in lifespan (from 13.1±0.68 days to 15.4±0.40 days).

**Physical Properties of Low-Dose Radiation Flies**

The earliest generation of fly researchers could generally only test lifespan as an endpoint. As technology and methods improved, the next step was to look at biological mechanisms of the low dose response.
Alexander Vaiserman


In one of his first papers (2003), he measures egg viability (what percent eclose (“hatch”)) after irradiating fly eggs – only 6.1% are not viable in control, roughly 20% at 25-75 rem and increasing after that, up to 70% at 400 rem. They claim lifespan extension at 50 and 75 rem in males, but not females – he argues that reproductive effects often mask effects in females. Since fewer flies eclosed (emerge) when irradiated, this difference could be due to selection.

He also studies the number of S1 nuclease-sensitive sites in the DNA of flies two weeks after irradiation. S1 is used to quantify the number of breaks in DNA, and also can be an indicator of the amount of transcription. He finds that S1 nuclease-sensitive sites drop by up to a factor of 3 in the longer-lived male flies, suggesting that the males have less damage and probably less overall transcription. The sir2 mutants discussed in chapter 1 also have less overall transcription.

In a follow-up study, Vaiserman (2004A) irradiates flies as larvae, which is a later stage in development than eggs. This time, viability is the same as control for low doses. Median lifespan stays about the same as control for males up to 420 rem while going down by roughly 15% for females in the same range. Mortality is greater than control for the first two weeks after irradiation, then becomes less. Egg laying goes down, but only for the first two weeks after irradiation, and only for the higher doses. This study seems to have a change point around 2 weeks for mortality
and egg laying, perhaps suggesting a time of repair. It also indicates a threshold response in lifespan.

Vaiserman then looks (2004B) at responses to physical stresses in irradiated flies as well as their offspring, in an attempt to see how much the radiation-induced changes are passed on genetically. He runs 8 different tests – climbing to the light from dark, climbing after being tapped down, eggs laid, weight, survival from heat shock, survival from starvation, lifespan and metabolic rate. He finds that body weight is less in offspring of irradiated flies and also in the lower doses the irradiated flies – this is attributed to selection. In the highest dose, the offspring are heavier. Irradiated flies and their offspring are more active (the two climbing tests) – this is shown to correlate with body weight. The smaller flies were all less resistant to heat shock and starvation, and the larger flies more resistant, except for male flies on one particular dose (50 rem). Several of the male offspring were more stress resistant than control after irradiation. Egg laying was non-significantly decreased with irradiation in both generations. Very small increases in median lifespan were discovered for 50 rem in males and 25 and 50 rem in females. Metabolic rate increased in all irradiated groups except irradiated males.

In an earlier study (in Russian), Vaiserman (2000) claims to see increases in longevity, activity, fertility, resistance to heat shock and resistance to starvation at doses of 25, 50 and 75 rem. The increase in fertility is a new and surprising result; however, earlier researchers rarely if ever looked as doses less than 1000 rem.


**Adaptive Response**

Another mechanism is the phenomenon that flies given a small dose of radiation before a large (harmful) dose show an increased resistance to the large dose. Fritz-Niggli and Schaeppi-Buechi (1991) took females from three strains of flies (control and two mutants with repair deficiencies) and subjected them to a dose of 2 rem and then a dose of 200 rem, with the second irradiation coming at the age of 4 days. The 4-day-old females were mated with males, and their eggs were studied for dominant lethality (eggs that did not hatch). In all three strains of fly, the eggs show fewer eggs not hatching than would be expected for the additive effect of the two doses.

**Towards Genetic and Molecular Mechanisms**

With technological advances in the past decade, gene expression experiments and similar technologies have made it possible to target specific genes as part of the response to low (and high) doses of radiation. Ogura et al. (2009) provide evidence for hormesis. Eye color mutations were measured in flies 2 generations removed from irradiation, with initial doses of 0.05, 10, 50 or 1000 rem. The results (figure 3.2.6) show a nonlinear relationship and the first two points have fewer average mutations than control. The first point is statistically significant. In addition, this group took gene expression (microarray) data for fly eggs (4-8 hours after egg laying) irradiated with a total dose of 0.05 rem. Samples were taken at 5, 30, 60 and 90 minutes post-irradiation.
Figure 3.2.6: Eye Color Mutation Frequency with Radiation Dose in Ogura et al. Note that 1 Gy (Gray) is equal to 100 rem. Also note that the x-scale is log scale except for control on the left.

The microarray data, along with a more accurate gene-by-gene test (qPCR) showed up-regulation of genes associated with heat and chemical stress responses as well as grim, a gene that increases apoptosis.

Seong et al. (2010) performed another microarray gene expression study. They irradiated male drosophila eggs with doses of 20 and 400 rem at dose rate 80 rem/minute (a low rate) with a Cs-137 gamma ray source. There is a very small (2%) lifespan increase with the 20 rem, and a roughly 25% lifespan decrease with the 400 rem. They use microarray analysis to look at the gene expression basis for this lifespan extension, and find roughly 13% of the genome to change. 39 of the changing genes are aging-related.

Moskalev et al. (2011) look at individual gene mutants as part of the pathway of low-dose radiation response. Canton-S background control flies were given gamma radiation from a Radium-226 source. They used doses of either 40 rem over a 10-day period from embryo to eclosion from a Radium-226 source, a 3,000 rem dose.
shortly after eclosion from a Cobalt-60 source or the combination of both. The Canton-S background flies showed a small but statistically significant increase of lifespan in the smaller dose (median 50 to 51 days in males, 54 to 57 days in females) as well as a significant adaptive response (Table 3.2.3).

<table>
<thead>
<tr>
<th>Dose (rem)</th>
<th>Male Median Lifespan (d)</th>
<th>Female Median Lifespan (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50</td>
<td>54</td>
</tr>
<tr>
<td>40</td>
<td>51</td>
<td>57</td>
</tr>
<tr>
<td>3000</td>
<td>34</td>
<td>50</td>
</tr>
<tr>
<td>40, then 3000</td>
<td>47</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 3.2.3: Low and High-Dose Response, Moskalev’s Control Flies

Moskalev et al. attempted the same experiment with mutant flies missing a copy of a certain gene that they think could be involved with the response. The experiments included stress signal sensors (ATM, ATR), stress signal transducers (Sirt1, p53, JNK, FOXO), the transcription factors Sir2 and Sirt2 and a set of autophagy genes (ATG7, ATG8A). The autophagy genes are a guess at a specific mechanism – protein turnover (inspired by the work of Seong (2010)). The flies missing copies of any one of the signal sensors or signal transducers lost the ability to have longer lifespan with radiation, and also lost the adaptive response. This generally occurred in both genders, although JNK-missing females still had the “adaptive response” and the Sirt2-missing males still had lifespan extension at low doses. In addition, the flies missing Sirt2 did not have show any change in adaptive response and flies missing either of the two autophagy genes showed no significant change in radiation response.

One particularly interesting result is the Sir2-missing males: with zero radiation, the flies lived only 24 days median; with the 40 rem dose they lived 50 days median, equal to the wild type controls. In females, the Sir2-missing irradiated flies all lived
around 10% longer than control at each dose. The male result suggests that low-dose irradiation may induce a higher expression of Sir2.

In a similar experiment by Moskalev et al. (2009), heat shock proteins are shown to be involved in the radiation response (Vaiserman (2000) also found an increase in heat shock resistance). Two low doses (40 rem, 4 rem) were used with the high dose of radiation replaced by a dose of the poison paraquat, which produces oxidative damage. The control flies demonstrate an adaptive response to paraquat, but flies missing both copies of the heat shock response genes Hsf and Hsp70Ba (one gene missing for each experiment) do not. In addition, the adaptive response is found to be missing for females (but not males) in the control Canton-S background as well as in flies missing both copies of Hsp22, Hsp67Bb or one copy of Hsp83.

In an earlier paper by the same group (Shaposhnikov, Turysheva and Moskalev 2009), radiating wild-type flies and DNA repair ability mutants (with the mutants having less repair ability) with 3,000 rem prolonged the larval stage and increased pupal lethality (eggs that do not hatch). Radiation with a total dose of 20 rem on wild-type flies and one type of mutant induces a shortening of the larval stage – the opposite of high radiation. In mutants having a deficiency in DNA repair, larval stage prolongation and increase in death rate start at doses of 20-40 rem, which are similar levels to the hormetic dose on the wild-type flies. Radiation of wild-type flies with constant doses of 6 and 60 rem doses over 10 days resulted in adult whole body weight gain, which is another type of hormetic effect (although it conflicts with the result that Vaiserman’s result that low-dose flies were smaller).
Summary

All of the studies above that deal with lifespan and low-dose radiation in *Drosophila melanogaster* suggest a threshold response and sometimes a hormetic response. All of these studies (with the exception of Carey’s) only have a sufficient number of flies to see mortality rates greater than or equal to 1/100. This is why studies regarding mechanism are necessary.

These studies suggest possible response mechanisms related to stress response, diet response, fertility and possibly immune response. The phenomenon where radiation only increases lifespan in suboptimal conditions raises the possibility that many responses to stresses (caused by suboptimal conditions) are the same or similar.
3.3 Experimental Methods

The experiment is comprised of two parts: a lifespan experiment and a gene expression experiment. There is a pilot experiment and a full experiment.

**Lifespan Experiment**

A cohort of male flies was maintained on a regular diet (called “1.5 Normal Ethanol Food” in the Helfand laboratory) and passed between vials onto new food every second day. At the time of passing, dead flies were counted. 100 flies per condition were used in the pilot experiment, 300 per condition were used in the full experiment. Roughly 10 percent of the flies escaped over the course of the experiment during passing. Doses used in the pilot experiment were (in kiloRoentgens, abbreviated kR) 0.01, 0.1, 1, 10, 20, 40, 60, 80 and 100. Doses used in the full experiment were (in kR) 0, 0.01, 0.1, 1, 5, 10, 20 and 40. All doses are in addition to background radiation. Roentgens are approximately the same as rem (they are measured relative to absorption in water (rem) and air (Roentgens), a roughly 10% difference in an experiment where doses are always much more than 10% different).

**Gene Expression Experiment**

A second cohort of flies, grown and irradiated with the lifespan cohort, were frozen in liquid nitrogen and processed into mRNA for a gene expression measurement on an Illumina RNA sequencing machine at Brown University. The resulting sequence fragment reads were matched (aligned) to the *Drosophila melanogaster* genome
using the Cufflinks software (Trapnell et al. 2010). In the pilot experiment, one sample was taken at days 2 and 10 post-irradiation for doses 0, 0.1 and 10 kR. In the full experiment, three samples of roughly 30 flies each were taken at days 2, 10 and 20 post-irradiation for doses 0, 0.01, 0.1, 1, 5, 10, 20 and 40 kR. Three samples per condition and per time point are standard procedure, including for the work described in chapter 1.

Irradiation Protocol
Flies were irradiated at the Brown University Irradiator, a Cesium-137 gamma ray source. The dose rate was 86.22 kR/hour for all doses greater than or equal to 1 kR. For doses less than 1 kR, the dose rate was 35.4 kR/hour (in the full experiment and likely in the pilot experiment, although dose rate was not notated in the pilot experiment for the lowest doses). The reason for the difference was the lower limit of exposure time accuracy in the irradiator. However, all doses took under an hour (and most took under a minute), so the opportunity for biological response was not significantly longer in any samples, particularly in comparison with experiments like Moskalev’s where some doses were given over 10 days.
Food was not irradiated. The flies were provided with drops of water in the vial stopper and were not off of food for greater than two hours.

Gene Expression Analysis
Single gene analysis with GOStat (Beissbarth and Speed 2004) was performed on the pilot and available full experiment data. PGSEA (Kim and Volsky 2005) using gene sets from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa 2000) and
the ranked list comparison method described in chapter 2 were performed on the available full experiment data. A fold change cutoff of 1.5-fold was used. For the full experiment, a present/absent expression cutoff was determined using the “NOTEST” option on the Cufflinks software (Trapnell et al. 2010). P values were ignored, since there is currently only zero or one sample available for each combination of dose and time post-irradiation.
3.4 Results

At the time of writing, the progress of this experiment was limited to the complete pilot experiment with lifespan experiments and gene expression, the full lifespan experiment and a subset of the gene expression data from the full experiment. The results thus far can be summarized as:

- Full and pilot experiment lifespan data show a threshold effect
- Lifespan data compared with existing data supports measuring radiation damage in terms of total incident photons
- Pilot and available full experiment gene expression results suggest stress, metabolism, reproduction and mitochondria function as mechanisms behind the threshold effect

Full and Pilot Experiment Lifespan Data Show a Threshold Effect

Figures 3.4.1 and 3.4.2 show the survivorship curves (fraction remaining alive with time) for the pilot and full experiments. Both figures show all doses below 10 kR as being the same within error. After that, the median lifespan decreases with dose (note: median lifespan is the day where the survivorship crosses 50%) and the curves began to take on the square shape described by Sacher. In the larger full experiment, two of the three lowest doses (0.01 and 1 kR) actually outperform control, but a larger study with more flies would be needed to see if it is truly different. The best measure of a difference in lifespan is a change in the slope or the intercept of the mortality, which is the percentage of flies left alive that die on a
given day (essentially the slope of survivorship). The mortality curves for the full experiment are shown in figure 3.4.3.

The third of the lowest doses (0.1 kR) has a different survivorship curve shape than the others, suggesting the occurrence of a random harmful event (for example, flies occasionally excrete too much slime and drown in it – one vial in another dose was excluded from the full experiment for this reason).

Figure 3.4.1 Pilot Experiment Survivorship.
Figure 3.4.2 Full Experiment Survivorship. Note that it has the same features of the pilot experiment.

Figure 3.4.3 Full Experiment Mortality Rate. The 40 kR dose has greatest slope and intercept, with 20 kR also separated from the other doses. 10kR rises above the lower doses at approximately day 50 and the lower doses are nearly indistinguishable.
Lifespan Data Compared with Existing Data Supports Measuring Radiation Damage in Terms of Total Incident Photons

A good existing dataset for comparison is that of Baxter and Blair (Kinetics of Aging, 1967). Figure 3.4.4 shows the median lifespan data from Baxter and Blair with our pilot experiment data. The two experiments look to be behaving very differently – the very acute deaths happen at 100,000 R for our experiment and at 180,000 R for Baxter and Blair.

Figure 3.4.4 Median Lifespan Versus Dose for Baxter and Blair Data and Antosh Pilot Experiment. Note that we treat 1 rem equal to 1 Roentgen (assuming absorption is similar enough between air and water; really is a factor roughly of 0.9).

These two experiments were done using different gamma ray sources. Baxter and Blair used Cobalt-60, which has twin energy peaks at 1.17 and 1.33 million electron volts (MeV), and we used Cesium-137, which has its main energy peak at 0.66 MeV. This means that for the same dose, a fly would be receiving roughly twice as many photons from the Cesium-137 source as from the Cobalt-60 source.
In an attempt to reconcile this, we decide to model radiation damage in terms of incident photons instead of energy deposited. This model is based on two assumptions:

- The breaking energy of a DNA bond is of order 10 electron volts, whereas each photon from a radiation source is orders of magnitude greater – thus, the DNA will very likely be broken if it interacts with a photon regardless of the photon energy.

- In addition to this, the photons will usually only interact once. An example calculation of Compton Scattering in flies (section 3.7) gives a probability of order 0.05 for a photon to interact once and a probability of order $(0.05)^2$ that it will interact twice.

These assumptions suggest an effect where the majority of the interactions are photons that break one bond and then leave the fly regardless of photon energy. Using the fact that the fly weights approximately 1 milligram (Kearsey and Kojima 1967), we can calculate the number of incident photons for each dose for a given source. We can also scale the median age of death by the median age of death in control flies in an attempt to reconcile the work of two different laboratories using slightly different strains of fly.

Converting to photons and scaling the y-axis, the two experiments look much more similar (figure 3.4.5). The full experiment (figure 3.4.6) does not fit the data quite as well, but the pattern is still visible. The largest change is in the 40 kR point (approximately $4 \times 10^9$ photons). Roughly 6-10 days before the full experiment 40 kR flies died, paralysis was observed in some flies but the flies were still alive.
Experimentally, the death rates of paralyzed flies will vary with the stickiness in each vial and in the prepared food – the flies die more quickly if they become stuck to the food while paralyzed. The food used in the full experiment was drier than that used in the pilot experiment. Paralysis was seen in all doses greater than or equal to 40 kR in the pilot experiment, and the measurement of death at the 100 kR dose was affected most. The overlap is put into better perspective when the number of photons is fit to the log scale (figure 3.4.7). This plot resembles Lamb’s log-linear model of radiation damage (equation 3.2.2) combined with a threshold.

Figure 3.4.5 Comparison with Control Lifespan versus Number of Incident Photons. In the pilot experiment, the lowest radiation dose (0.01 kR) is treated as control.
Figure 3.4.6 Data Comparison with Full Experiment Added.

Figure 3.4.7 Data Comparison in Log Scale. The added dose zero (regular background) points were not included.
**Pilot and Available Full Experiment Gene Expression Results Suggest Stress, Metabolism, Reproduction and Mitochondria Function as Mechanisms Behind the Threshold Effect**

**Pilot Experiment**

After running a GOStat analysis (Beissbarth and Power 2004) of the pilot data and discussing the output with the Helfand lab, the significantly changing biological functions are as summarized in table 3.4.1.

<table>
<thead>
<tr>
<th></th>
<th>Day 2</th>
<th>Day 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 kR</td>
<td>mitochondria production, stress response, metabolism</td>
<td>male gonad function</td>
</tr>
<tr>
<td>10kR</td>
<td>stress response, senses, male gonad function</td>
<td>stress response, mitochondria maybe metabolism</td>
</tr>
</tbody>
</table>

Table 3.4.1 Summary of Pilot Gene Expression Results

The results suggest that the lower dose (0.1 kR) is stressed at day 2 post-irradiation, but not at day 10, suggesting a recovery. In particular, the change in direction of male gonad function may represent a repair or an accommodation of more gene expression to deal with less available functioning data. The results suggest that the higher dose (10 kR) is stressed at both days 2 and 10 post-irradiation. Many of the stress responses at 0.1 kR at day 2 are similar to GOStat analyses on lifespan extending treatments (such as those in chapter 1), for example increases in metabolism and protein turnover. Many of the stress responses in 10 kR are more indicative of damage, such as apoptosis. The results also indicate mitochondria as an
affected function. The temporary increase in mitochondria production found at day 2 in the 0.1 kR dose flies may be due to an increased demand for energy to respond to biological stress. The decrease in mitochondria function found at day 10 in the 10 kR dose flies may indicate a permanent loss of energy.

*Full Experiment GOStat Analysis*

At the time of writing, one sample of gene expression data was available from the doses 0 (control), 0.01 kR, 1 kR and 5kR at days 2, 10 and 20 post-irradiation. Three samples are the usual standard for a full experiment, so the analysis is preliminary. The results of a GOStat analysis are summarized in table 3.4.2.

<table>
<thead>
<tr>
<th>Day 2</th>
<th>0.01 kR</th>
<th>1 kR</th>
<th>5 kR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Synthesis</td>
<td>Defense/Immune Response</td>
<td>Up</td>
<td>Oxidoreductase</td>
</tr>
<tr>
<td>Spermatogenesis</td>
<td>Defense/Immune Response</td>
<td>Peptidase Inhibitor</td>
<td>Oxidoreductase</td>
</tr>
<tr>
<td>Protein CK2</td>
<td>Neurotransmitters</td>
<td>Development</td>
<td>Oxidoreductase</td>
</tr>
<tr>
<td>Pheromone Binding</td>
<td>Protein CK2</td>
<td>Down</td>
<td>Peptidases</td>
</tr>
<tr>
<td>Receptors</td>
<td>DNA Synthesis</td>
<td>Downregulation of DNA replication</td>
<td></td>
</tr>
<tr>
<td>Endopeptidase Inhibition</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4.2 GOStat Analysis Results for Available Full Experiment Data

<table>
<thead>
<tr>
<th>Day 20</th>
<th>0.01 kR</th>
<th>1 kR</th>
<th>5 kR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein CK2</td>
<td>Up</td>
<td>Oxidoreductase</td>
<td></td>
</tr>
<tr>
<td>Spermatogenesis</td>
<td>Defense/Immune Response</td>
<td>Mitochondria</td>
<td></td>
</tr>
<tr>
<td>Reproduction</td>
<td>Peptidases</td>
<td>Down</td>
<td>Mitosis</td>
</tr>
<tr>
<td>Endopeptidase Inhibition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein CK2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DNA Synthesis
Spermatogenesis
Protein CK2
Downregulation of DNA replication
Neurotransmitters
Defense/Immune Response
Oxidoreductase
Receptors
Endopeptidase Inhibition
Protein CK2
Causes
Spermatogenesis
Receptors
Peptidases
Hormone Activity
Protein CK2
Spermatogenesis
Receptors
Endopeptidase Inhibition
Protein CK2
Reproduction
Causes
Spermatogenesis
Protein CK2
DNA Synthesis
Spermatogenesis
Protein CK2
Spermatogenesis
Protein CK2
Spermatogenesis
Protein CK2
Spermatogenesis
The results suggest that very little is different from control by day 20 - these are the flies that lived the same length as control, so it appears the flies may be going back to being like control. This is particularly evident in the 5 kR flies, where the lifespan curves in figure 3.4.2 are the most similar to control. Based on the lifespan results, it is possible that the 0.01 kR and 1 kR changes are beneficial.

The pilot experiment (table 3.4.1) suggests the method of radiation response as changes in defense/immune response, mitochondria, sensing and metabolism. The full experiment result is very similar, with neuropeptides likely related to sensing and oxidoreductase and electron carrier activity related to mitochondria. Defense/Immune Response seems to be changing often, switching between up and down. This difference in direction may be attributable to the limitations of gene set methods discussed in chapter 1 - gene sets don't actually change only if all the genes go in one direction (up or down). This means that it isn't clear whether the defense/immune system is doing more or doing less, but it is likely affected. The protein CK2 pathway changes at least once for each dose, always up. CK2 affects mitosis, growth and signaling.

The two parts of the fly that will be affected most strongly by radiation are the gut and testis, which are the two parts that still undergo mitosis post-emergence. The spermatogenesis and reproduction pathways may be indicative of repair to the testis, or it may be that the portion of it that is still working are working harder in order to make up for the portion that is damaged. The 5 kR day 2 pathways up may indicate something happening in the gut - for example, damage to the gut may affect
the bacteria that resides in the gut, leading to the defense response pathways changing.

**Full Experiment PGSEA**

A PGSEA analysis on the Kyoto Encyclopedia of Genes and Genomes gene sets showed very few changes. Most notably:

- At dose 0.01 kR, a change in metabolism at days 2 and 10 and a change in DNA polymerase and repair at day 10.
- At dose 1 kR, a change in the breakdown of amino acids at days 2 and 20 (possibly defense response) and a change in repair and DNA polymerase at day 20.
- At 5 kR, a change in repair at day 2, protein breakdown at day 10 and energy metabolism at day 20.

The few changes are again similar to the pilot experiment.

**Full Experiment Ranked List Comparison**

A ranked list analysis was performed in order to determine the degree of similarity between the different radiation doses and different time points. The overlaps were calculated (using Δ=200) between every available combination of dose and day. The results were used to cluster the samples hierarchically (average linkage method) and the results are shown in a dendrogram (figure 3.4.8). The dendrogram suggests that the three doses cluster together by day and not by dose level. In each time cluster, the 1 kR and 5 kR doses are more similar to each other than to 0.01 kR.
Figure 3.4.8. Dendrogram of Low Dose Radiation Treatments. Similarity measure used was number of overlap genes as determined by ranked list comparison. Distance measure between samples was the distance from the maximum overlap found.

Full Experiment Conclusion

Overall, it seems that there are fewer changes in the low dose samples compared to control at day 20. And the pathways found to change are similar between the pilot and full experiments. Investigation into single genes will be more relevant when repeat samples are processed, as this will enable the calculation of p values and improve the overall quality of the data. With more replicates, gene set analyses will better capture the trends of the data.
3.5 Discussion

The results of this experiment (while still ongoing) indicate a threshold response in *Drosophila* lifespan with radiation. A pilot experiment followed by a larger full experiment successfully repeated lifespan results. Comparison with existing data indicates that radiation damage in lifespan may be better measured as number of incident photons instead of energy imparted. Our data predicts a Cesium-137 threshold of 5-10 kR, which is the same number of photons as 9.5-19 kR in Baxter and Blair’s Cobalt-60 source. In that paper (Kinetics of Aging), Baxter and Blair predict a threshold around 15 kR, which falls in the middle of this range.

A change in the measurement units of radiation damage (if necessary) could represent a shift in the practice of radiation safety. The model used in this chapter suggests a difference of approximately 100 percent between the damage from a Cesium source compared to the damage from a Cobalt source for the same dose in energy/mass. More experimental and analytical work is necessary before making a final decision, but the evidence here is suggests that this is the case. With *Drosophila*, it would be helpful to do an experiment testing the lifespan response with different sources and also with varying dose rates to measure repair.

The mechanisms of radiation response seem to include stress responses, apoptosis, reproductive function, metabolism and mitochondria. These are very similar to those seen in treatments such as DR, sir2 and p53 (in chapter 1). Further experiments should investigate the link between these treatments. Once the full data set is in for this data (the current data available is one sample per condition;
three is considered an acceptable data set), the analysis described in chapter 2 may be used to compare between data sets. In addition, an experiment should be done testing the radiation response of flies that are also on DR. If the flies at low dose are unaffected by DR (or less affected), DR and low dose radiation are likely working through the same gene expression pathways.

A further experiment should be done to see if the lower doses of radiation actually extend lifespan. More flies per dose would be needed in order to more accurately test this, since it appears that the lifespan effect is small (approximately 10% of control lifespan).

It is worth noting that the doses used on fruit flies are orders of magnitude higher than the doses used on mammals. This is because most of the fly's cells are post-mitotic, meaning that they are no longer dividing and thus cannot become cancerous. However, gut stem cells in the fly are believed to have the capability to still divide and become cancerous. Studies targeting these cells are beginning in the Tatar lab at Brown, with promising initial results.

It may also be that the acute damage dose of radiation across flies and mammals scales to some extent as weight or effective length or vital organs. A damaging dose in humans is approximately 10 rem (Committee to Assess Health Risks from Exposure to Low Levels of Ionizing Radiation 2006), a damaging dose in mice is approximately 1000 rem (for example, Travis 1979) and a damaging dose in flies can be considered as approximately 10,000-100,000 rem based on the results of this experiment.
3.6 Conclusion

The overall goals of this work are determine if a threshold response to radiation exists with *Drosophila melanogaster* lifespan, to indicate the gene expression mechanisms of radiation damage and response, particularly near what appears to be a threshold dose, and whether or not energy per unit mass is the best way to measure damage. The lifespan data suggests a threshold response as well as the possible existence of a hormetic (hormesis model) response. The initial work identifies a set of genes and biological pathways (gene sets) that change with radiation, and the lifespan experiments suggest that radiation damage is better measured in terms of number of incident photons. Gene expression data is still being prepared, and further experimentation will advance the work.
3.7 Calculation of Compton Scattering Interaction Probability for A Photon in Water

Note: all equations in this section were found in Leo (1994).

Compton Scattering is the one of the dominant photon/matter interaction effects at the range of photon energies considered in this chapter (roughly 1 MeV) (Leo 1994). The Compton Scattering differential cross section can be calculated from quantum electrodynamics and is known as the Klein-Nishina formula (equation 3.7.1). It depends on the classical electron radius \( r_e = 2.82 \cdot 10^{-15} \) m, the ratio of photon energy to electron mass energy (electron mass energy is 511 keV, the ratio is \( \gamma \)) and the angle through which the photon is scattered (\( \theta \)).

\[
\frac{d\sigma}{d\Omega} = r_e^2 \frac{1}{2[1 + \gamma(1 - \cos \theta)]^2} \left( 1 + \cos^2 \theta + \frac{\gamma^2(1 - \cos \theta)^2}{1 + \gamma(1 - \cos \theta)} \right)
\]

Integrating over the solid angle \( d\Omega = \sin \theta \cdot d\theta \cdot d\phi \) gives an equation for the cross section (equation 3.7.2). The cross section is the total probability per electron for a Compton Scattering to occur.

\[
\sigma = 2\pi r_e^2 \left( \frac{1 + \gamma}{\gamma^2} \left[ \frac{2(1 + \gamma)}{1 + 2\gamma} - \frac{1}{\gamma} \ln(1 + 2\gamma) \right] + \frac{1}{2\gamma} \ln(1 + 2\gamma) - \frac{1 + 3\gamma}{(1 + 2\gamma)^2} \right)
\]
For $\gamma = 1$ (photon energy 511 keV), this gives $3 \times 10^{-29}$ m$^2$; for $\gamma = 2$ (photon energy 1.022 MeV) this gives $2 \times 10^{-29}$ m$^2$.

Multiplying the cross section by the number density of electrons gives the linear absorption coefficient ($\mu$). Approximating a fly as water, the number density of electrons (equation 3.7.3) depends on the number of electrons ($Z$), density ($\rho$) and molecular weight of water ($A$), multiplied by Avogadro's number ($N_A$).

$$\frac{Z \cdot N_A \cdot \rho}{A} = \frac{(18)(6.02 \cdot 10^{23} \text{ mol}^{-1})(10^6 \text{ g/m}^3)}{18.02 \text{ g/mol}} = 6 \cdot 10^{29} \text{ 1/m}^3 \quad (3.7.3)$$

This gives the linear absorption coefficient as $17$ m$^{-1}$ for $\gamma = 1$ and $12$ m$^{-1}$ for $\gamma = 2$.

The linear absorption coefficient can be used to calculate the fraction of photons that interact (equation 3.7.4).

$$\text{fraction not interacting} = 1 - e^{-\mu \text{ (length of material)}} \quad (3.7.4)$$

Estimating a fly as roughly 3mm long (1mm each for head, thorax and abdomen), equation 3.7.4 gives a probability of 5 percent for $\gamma = 1$ and 4 percent for $\gamma = 2$.

What about the probability of interacting twice? The energy of the scattered electron is related to the original energy with the Compton Scattering energy relation (equation 5), which depends on the same $\gamma$ and $\theta$ as the Klein-Nishina equation.
\[ E' = \frac{E}{1 + \gamma(1 - \cos \theta)} \]  

(3.7.5)

The Klein-Nishina equation indicates that forward scattering (higher resulting photon energies and thus less likely to interact a second time), is more likely than backward scattering (lower energies and more likely to interact); but an easy and conservative estimate will be to assume all scattering angles are equally likely. A numerical integration with this assumption gives \( E'/E \) equal to 0.45 for \( \gamma = 1 \) and 0.58 for \( \gamma = 2 \). Plugging these new energies into equations 3.7.2 through 3.7.4 (and approximating that the average first interaction happens halfway through the fly, cutting down the effective length by a factor of 2) gives a probability of second interaction as 3 percent for \( \gamma = 1 \) and \( \gamma = 2 \). These probabilities are the probabilities of having a second interaction after having the first interaction – so roughly 97 percent of interacting photons at \( \gamma = 1 \) and \( \gamma = 2 \) do not interact a second time. This justifies our approximation in section 3.4 that most photons only interact once.
Chapter 3 References:


