QUANTITATIVE STUDIES OF AGING USING STATISTICAL MECHANICS AND PROBABILISTIC APPROACHES

BY DIANA DAVID-RUS

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Biophysics and Computational Biology

Written under the direction of

Professor Joel L. Lebowitz

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2009

ABSTRACT OF THE DISSERTATION

Quantitative studies of aging using Statistical Mechanics and Probabilistic approaches

by Diana David-Rus

Dissertation Director: Professor Joel L. Lebowitz

The goal of understanding aging is not just about fulfilling the age-old quest of immortality, but, rather in trying to answer the question of "what is aging?" we expect to generate insight that can be used to improve the health span of an aging organism. Recently, the biology community has come to play a role into this quest with identifications of gene pathways that can double, even triple the life spans for certain organism such as C .elegans and Drosophila melanogaster. Aging biology finds itself in a post-genomic era. Hopes of bringing methods developed in mathematics, physics or statistics into the biology realm are widespread. The goal and unifying theme of my thesis is to get a better understanding of this new and exiting field (and at the same time ancient subject) of aging as a complex process, using quantitative methods. By combining molecular and biophysical modeling with statistical and mathematical tools, my goal was to provide a multi-scale view of the complex biological process that is aging. The approach I am taking involves consideration of the problem on several levels–from

transcriptional regulation of gene expression, modeling of biological pathways and interaction networks, to the development of mathematical and statistical methods; from trying to understand the aging process at a transcriptional level, and analyzing and understanding how stochastic factors might come to play a role in aging in understanding aging as an epigenetic process.

Acknowledgements

I would like to thank my committee members for help and understanding during the process of writing this thesis.

I would like to add special thanks to my advisers Prof. Joel L. Lebowitz and Prof. Monica Driscoll. I couldnt have got in this stage of my life without their constant presence and help.

Also to my colleges for useful discussions!

Nevertheless I would like to thank my family: my husband, father and my little sister for their support.

I am grateful to have such great people around me! I wouldnt have accomplish anything without their tremendous patience love and care!!!

Dedication

To my father,

You always have been an example for my life

Thank you!

Table of Contents

Chapter 1

Introduction

My interest lies at the interface between biological, mathematical and physical sciences. By combining molecular and biophysical modeling with statistical and bioinformatics tools, my goal was to provide a multi-scale view of complex biological processes such as aging and its connection with stochastic and epigenetic processes. The approach I am taking involves consideration of the problem on several levels–from modeling of biological pathways and interaction networks, to the development of computational and statistical methods to understanding transcriptional regulation of gene expression.

Since stochasticity plays a major role into the aging process, in the first chapter I am developing methods for solving/describing two dimensional stochastic processes that involve interaction. The model I am working on involves the "simple" but biologically important problem of protein interaction and stochastic interactions therein. I am building an artificial-toy model that describes a two dimensional generation/degradation process of two protein types that interact with each other in the following manner: the rate of generation of one type of protein changes once the quantity of the second protein is above a certain threshold. In this work, we study a protein synthesis degradation process by defining a general mathematical model and showing a route to gain some analytical insight to the problem. We discuss the model in the steady state situation in two study cases for a particular choice of states and rules of state transitions and

find exact solutions using generating function technique. Overall, this relatively simple model can be used to evaluate the impact of stochastic factors in protein folding on biological fitness. Such analysis would constitute a core unit for considering the complexities of multiple stochastic processes that are relevant to aging.

Epigenetic regulation of multiple heritable cell fates involves transcriptional repression or activation of the expression levels of genes, over possibly many cell cycles, without altering the underlying genetic sequence. At the heart of one important mechanism of epigenetic control is the accessibility of DNA packaged into higher order structures known as chromatin. More than two decades ago, it was proposed that 'aging of proliferating cells' results from genome reorganization occurring during the division cycle. However, progress in the areas of epigenetics and cellular and organismal aging has been slow and sporadic.

Maintenance of alternative chromatin states through cell divisions pose some fundamental constraints on the dynamics of histone modifications. In the second chapter, we study the systems biology of epigenetic inheritance by defining and analyzing general classes of mathematical models. We discuss how the number of modification states involved plays an essential role in the stability of epigenetic states. In addition, DNA duplication and the consequent dilution of marked histones act as a large perturbation for a stable state of histone modifications. The requirement that this large perturbation falls into the basin of attraction of the original state sometimes leads to additional constraints on effective models. Two such models, inspired by two different biological systems, are compared in their fulfilling the requirements of multistability and of recovery after DNA duplication. We conclude that in the presence of multiple histone modifications that characterize alternative epigenetic stable states, these requirements

are more easily fulfilled.

Microarrays are chips on which genes are attached for hybridization to mRNA samples–hybridization signals indicate which genes are expressed as messages and can speak to relative abundance and changes in gene expression over time. In the third chapter of my thesis I'm using methods developed for data mining microarray experiments, adapted for aging research. Methods bridge knowledge of statistical mechanics with data mining methods developed in statistical mathematics. Such analyses can reveal how the transcriptional regulation of genes might coincide, thereby implicating proteins as parts of networks acting together towards a common biological function.Such experiments are most useful for complex biological traits that result from the presumed functioning of several molecular pathways. Aging is one such biological phenomenon that incorporates numerous molecular mechanisms underlying environmental stimulus sensing, metabolic regulation, stress responses, reproductive signaling, hibernation, and transcriptional regulation. Current models of aging emphasize different mechanisms as driving forces behind aging and lifespan determination. However, an integrated understanding of exactly how these mechanisms drive aging has not yet been formulated. Using an unsupervised approach based on concepts from statistical mechanics, I identified an interesting gene expression pattern that suggests that a gene expression switch at midlife. This switch coincides with the onset of biomarkers of aging including age pigment accumulation

Age-related muscle decline, a condition referred to as sarcopenia and defined as loss in muscle mass and muscle strength over time, is one of the most pervasive problems of the elderly, such that significant declines in strength and mobility affects essentially every old person. Although the rate of decline is relatively slow (estimated to be only

1 procent loss annually), ultimate losses are substantial, such that nearly a 50 procent loss of muscle mass can occur by age 90. Decreased physical strength is a central contributor to loss of independence. We have found that aging C. elegans body wall muscle undergoes a process remarkably reminiscent of human sarcopenia. Both have mid-life onset and are characterized by progressive loss of sarcomeres and cytoplasmic volume; both are associated with locomotory decline.

To extend understanding of this fundamental problem, in the forth chapter I have focused microarray analyses on C. elegans muscle aging. C. elegans genes expressed in muscle have been experimentally defined. I surveyed expression of all known muscle related genes to describe a profile of transcriptional changes in muscle that transpires during adult life and aging. Importantly, the intersection of this dataset with that from aging flies and some human studies can suggest conserved genes that might impact the process most strongly. Hypotheses I formulate will be used to drive experiments at the bench and perhaps to focus attention for human therapies. This research will advance understanding of conserved aging mechanisms; data should influence novel strategies to extend healthspan.

Chapter 2

PART I- Protein synthesis-degradation a stochastic approach

2.1 Introduction

In this work, we study a protein synthesis degradation process by defining a general mathematical model and showing a route to gain some analytical insight to the problem. We discuss the model in the steady state situation in two study cases for a particular choice of states and rules of state transitions and find exact solutions using generating function technique.

Proteins are essential macromolecules that serve both as structural components of the cell and as its enzymatic machinery. The turnover of these proteins (synthesis and degradation) is a dynamic process that plays a critical role in the maintenance of cellular homeostasis. Most of the reported studies have focused on the protein synthesis aspect of protein turnover, as opposed to protein degradation, and there is a general consensus that protein synthesis does decline with age (Van Remmen et al. 1995; Rattan 1996; Ward and Richardson 2000). Although protein synthesis is an obvious important process, it is worth noting that protein degradation is of equal importance, as evidenced by the number of critical physiological functions it serves. These functions include the maintenance of plasma amino acid concentrations, the removal of abnormal

and post-translationally modified proteins and many more. Different pathways of protein degradation are affected by aging. Continuous turnover of intracellular proteins is essential for the maintenance of cellular homeostasis and for the regulation of multiple cellular functions. The first reports showing a decrease in total rates of protein degradation with age are dated more than 50 years ago, when the major players in protein degradation where still to be discovered.

Protein synthesis is a tightly regulated cellular process that affects growth, reproduction, and survival in response to both intrinsic and extrinsic cues, such as nutrient availability and energy levels. A pronounced, age-related increase or decline of the total protein synthesis rate has been observed in many organisms, including humans. The molecular mechanisms underlying this increase-decline and their role in the aging process remain unclear. A series of recent studies in the nematode, have revealed a novel link between protein synthesis and aging. Remarkably, these research findings, in their totality, converge to indicate that reduction of mRNA translation prolongs life in worms (Syntichacky et.al, Molecular Mechanisms and Models of Aging, Nektarios Tavernarakis et al. 2006) Signal transduction cascades implicated in aging, such as the insulin/insulin growth factor-1 pathway, interface with mechanisms regulating protein synthesis via a battery of key mRNA translation factors. One possibility is that the effects of these pathways on aging are mediated, in part, by alterations in protein synthesis.

Given the implications of the protein synthesis/degradation on all major biologically phenomenon including aging the importance of studying such a process is clear. Here I study a simple model of protein synthesis and degradation process as described by a protein -protein interaction.

Intracellular randomness has long been predicted from basic physical principles (1) and observations of phenotypic heterogeneity (2,3).Such 'noise' affects all life processes and has recently been measured using green fluorescent protein (GFP) (see 4-8). Random fluctuations in genetic networks are inevitable as chemical reactions are probabilistic and many genes, RNAs and proteins are present in low numbers per cell.

Biochemical processes frequently involve small numbers of molecules (e.g. a few molecules of a transcriptional regulator binding to one 'molecule' of a DNA regulatory region). Such reactions are subject to significant stochastic fluctuations. Traditionally Monte Carlo methods with the Gillespie algorithm (see also BKL) are employed to study the functional consequences of the fluctuations and simulate processes that cannot be modeled by continuous fluxes of matter.

The aim of this work is to explore minimal models of protein synthesis- degradation in order to gain some analytical insight of the process described by a general stochastic model presented in the next section.

2.2 A general stochastic model of protein synthesis-degradation

We envision a protein synthesis degradation process as a continuous in time birth death Markov process with a discrete (very large) state space. When two different species or types of proteins with a large number of possible states are involved the stochastic model that describes the process can be considered a two dimensional birth death Markov process. The master equation gives the flow for $\pi(j,k;t) = \pi_{jk}$, the probability of there being j copies of the 1st species, k copies of the second, at time t. For a simple case where each reaction either creates or annhilates one and only one component, and for the simple case where birth/decay rates are constant, we have the following master

equation describing the time evolution of the probability distribution

$$\frac{d\pi_{jk}}{dt} = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}(\pi_{j-1k}\beta + \pi_{jk-1}b_{jk} - \pi_{jk}j\delta - \pi_{jk}kd) \tag{2.1}$$

where:

π_{jk} is the joint probability distribution of type I and II to have j respective k quantities.

δ and d are the rates of annihilation of protein type 1 respectively 2 which are some constants proportional to the existing number of protein copies. The rates of creation for protein type 1 with j copies is β and for protein type 2 with k copies is b_{jk},. The two protein types interact in the following way: when protein type 1 gets to a certain threshold θ, protein type 2 is changing the birth rate as following:

$$b_{jk} = \begin{cases} b_0, & \text{when } j < \theta \\ b_1 & \text{when } j \geq \theta \end{cases}$$

We will study analytically the stochastic model formulated above in the steady state case for a particular choice of states and rules of state transitions. To gain some analytical insight, we will use an analytical approach described in next section.

2.3 Generating function technique

To solve the master equation analytically for the long time behavior of π_{jk} is generally an impossible task when the state space is very large. One, therefore, has to resort on various techniques. One such technique often used successfully in stochastic processes literature is the "generating function technique" (Bailey,N.T.J.1990, Karlin,S. and H.Taylor.1975,1981,1998, Linda S. Allen. 2003).

In this section we remind the reader of some well-known aspects of the generating function technique commonly used in stochastic processes literature, in order to make the present discussion self-contained.

Assume X is a discrete random variable and assume, for convenience, the state space is $\{0, 1, 2...\}$. Let f denote the probability mass function of X and suppose the probabilities are given by:

$f(j) = Prob(X = j) = p_j, j = 0, 1, 2...$,where $\sum_{j=0}^{\infty} p_j = 1$

The mean of X satisfy: $\mu_X = E(X) = \sum_{j=0}^{\infty} jp_j$

The probability generating function (p.g.f.) of the discrete random variable X is defined by

$$\mathcal{P}_X(t) = E(t^X) = \sum_{j=0}^{\infty} p_j t^j$$

for some $t\epsilon\mathcal{R}$

Because $\sum_{j=0}^{\infty} p_j = 1$, the above sum converges absolutely for $|t| \leq 1$

As the name implies, the p.g.f. generates the probabilities associated with the distribution $\mathcal{P}_X(0) = p_0$, $\mathcal{P}'_X(0) = p_1$, $\mathcal{P}"_X(0) = 2!p_2$, and in general $\mathcal{P}^k_X(0) = k!p_k$.

The p.g.f. gives entire information associated with the distribution.

2.4 Two state model

2.4.1 Describing the model

There are 2 types of proteins: type I and type II undergoing a birth- death process with interaction. Type I protein can have any number of copies/states. Type II protein can have 2 possible copies/states: 0 or 1, meaning we have no protein or just one protein. For simplicity, the rates of annihilation for both protein types are: some constants

δ respectively d proportional to the existing number of protein copies. The rate of creation for the first protein type is a constant β. The rate of reaction of the second protein, b_{jk}, can have two possible values, b_0 respectively b_1, depending on whether the quantity of the protein type I is bellow or above a certain threshold taken for simplicity to be 1.

$$b_{jk} = \begin{cases} b_0, & \text{when } j < 1 \\ b_1 & \text{when } j \geq 1 \\ 0 & \text{when } k > 1 \forall j \end{cases}$$

An example of such situation in real life would be a genetic switch on/off. The two protein types with j respective k copies interact in the following way: when is no protein type I present than the protein type II will have a constant birth rate b_0. When is at least one copy of protein type I in the system, protein type II is changing the birth rate from b_0 to b_1.

Note that for the case when we have just one protein type undergoing a birth death process with a constant decay/birth rate is a well known fact that in steady state, its stationary probability distribution is a Poisson distribution (Karlin,S. and H.Taylor.1975).

2.4.2 Equations which describe the birth death process of the two protein types in steady state

In steady state the left side of the master equation (1) describing the time evolution of the probability distribution is 0.

For this model I have in steady state 4 possible situations. Bellow are the equations

that defines each situation.

when $j = 0, k = 0$:

$$\pi_{00}(\beta + b_0) = \pi_{10}\delta + \pi_{01}d \tag{2.2}$$

when $j \geq 1, k = 0$:

$$\pi_{j0}(j\delta + \beta + b_1) = \pi_{j+1,0}(j+1)\delta + \pi_{j1}d + \pi_{j-1,0}\beta \tag{2.3}$$

when $j = 0, k = 1$:

$$\pi_{01}(\beta + d) = \pi_{11}\delta + \pi_{00}b_0 \tag{2.4}$$

when $j \geq 1, k = 1$:

$$\pi_{j1}(j\delta + \beta + d) = \pi_{j+1,1}(j+1)\delta + \pi_{j0}b_1 + \pi_{j-1,1}\beta \tag{2.5}$$

Using generating function technique we simplify our problem by transforming the above equations into ODE's satisfied by the generating function.

Let $f_k(x) = \sum_{j=0}^{\infty} \pi_{jk}x^j$ be the p.g.f.

The marginal W_0 giving the probability that species 2 has k elements, k=0,1 is given by

$W_k = \sum_{j=0}^{\infty} \pi_{jk} = \sum_{n=0}^{\infty} f_k^n(0)/n!$ where k=0,1

Since protein 2 doesn't influence protein 1 the marginal distribution of protein 1 is given by:

$$p_j = \sum_{k=0}^{1} \pi_{jk} = \frac{1}{j!}\left(\frac{\beta}{\delta}\right)^j e^{-\beta/\delta}$$

where $p_j = \pi_{j0} + \pi_{j1}$

it follows that:

$$f_0(x)+f_1(x) = \sum_{j=0}^{\infty} x^j(\pi_{j0}+\pi_{j1}) = \sum_{j=0}^{\infty} x^j e^{-\beta/\delta}\frac{1}{j!}\left(\frac{\beta}{\delta}\right)^j = e^{-\beta/\delta}\sum_{j=0}^{\infty}\frac{1}{j!}\left(x\frac{\beta}{\delta}\right)^j = e^{-\beta/\delta}e^{x\beta/\delta} = \tag{2.6}$$

$$= e^{(x-1)\beta/\delta}$$

and

$$f_0^{'}(x) + f_1^{'}(x) = \frac{\beta}{\delta}e^{(x-1)\beta/\delta} \tag{2.7}$$

Using generating function technique on equations 2,3 an ODE equation (eq.8) satisfied by a generating function is derived (see Appendix A for a detailed derivation)

$$x\delta\frac{d}{dx}f_0(x) + (\beta + b_1)f_0(x) + (b_0 - b_1)f_0(0) = \delta\frac{d}{dx}f_0(x) + f_1(x)d + f_0(x)\beta x \tag{2.8}$$

same procedure applied on eq 4,5 and obtain the following equation:

$$x\delta\frac{d}{dx}f_1(x) + (\beta + d)f_1(x) = (b_0 - b_1)f_0(0) + \delta\frac{d}{dx}f_1(x) + f_0(x)b_1 + xf_1(x)\beta \tag{2.9}$$

Steps toward obtaining $f_0(x)$:

Using condition (6) in eq.(8) I obtain a new equation (eq.10)in the $f_0(x)$ as unknown which once solved gives me the expression for $f_0(x)$ generating function.

$$(x-1)\delta\frac{d}{dx}f_0(x) + (-\beta(x-1) + b_1 + d)f_0(x) = de^{(x-1)\beta/\delta} + (b_1 - b_0)f_0(0) \tag{2.10}$$

This is a first order ODE;

Using integrand factor method one gets after some calculations (see Appendix B for a detailed derivation in solving eq.10)

$$f_0(x)(1-x)^{(b_1+d)/\delta}e^{-\beta/\delta x} - f_0(0) = \tag{2.11}$$

$$-de^{-\beta/\delta}\frac{1}{b_1+d}(1-(1-x)^{(b_1+d)/\delta}) - \frac{(b_1-b_0)f_0(0)}{\delta}\int_0^x e^{-\frac{\beta}{\delta}y}(1-y)^{\left(\frac{b_1+d}{\delta}-1\right)}dy$$

Setting $x = 1$ in equation 11 one obtains $f_0(0)$ and than going back at eq.11 one gets an expression for $f_0(x)$

2.4.3 RESULTS

from eq. 11 for x=1,

$$f_0(0) = \frac{de^{-\beta/\delta}}{b_1+d}\left[\frac{1}{1-\frac{b_1-b_0}{\delta}\int_0^1 e^{-(\beta/\delta)y}(1-y)^{\frac{b_1+d}{\delta}-1}dy}\right]$$

Given that the derivatives of generating fct at zero gives the probabilities associated with the distribution, we have:

$$\pi_{00} = f_0(0)$$

$$\pi_{10} = f_0'(0)$$

$$\pi_{20} = f_0''(0)$$

.....

$$\pi_{n0} = f_0^n(0)$$

From (6) we have $f_1(x) = e^{(x-1)\beta/\delta} - f_0(x)$ and then for $x = 0$ one obtains: $f_1(0) = e^{-\beta/\delta} - f_0(0)$

therefore,in same way I can easily get:

$$\pi_{01} = f_1(0)$$

$$\pi_{11} = f_1'(0)$$

$$\pi_{21} = f_1''(0)$$

.....

$$\pi_{n1} = f_1^n(0)$$

Using the result above one can determine the probability in the stationary state of having a given number of proteins type I and II in the system given that the protein type II can have just 2 possible states.

Next I will expand this result for the case when protein type II can have more than two states involved in the process.

2.5 Three state model

2.5.1 Describing the model

Having explored a two-state model in the previous section, we now study a simple three-state model of protein synthesis - degradation.

The system is the same as before except that

the type II protein can now have 3 possible copies/states: 0 or 1, and 2. As before the rates of annihilation for both protein types are: δ respectively d some constants proportional with the existing number of protein copies. The rate of creation for the first protein type is a constant β. The rate of creation of the second protein b'_{jk}, can have two possible values, b'_0 respectively b'_1 depending on wheater the quantity of protein type I is bellow or above a certain threshold taken for simplicity in this case to be 2.

$$b'_{jk} = \begin{cases} b'_0, & \text{when } j < 2 \\ b'_1 & \text{when } j \geq 2 \\ 0 & \text{when } k > 2 \forall j \end{cases}$$

The two protein types with j respective k copies interact in the following way: when is no more then one protein type I present than the protein type II will have a constant birth rate b_0. When are 2 or more protein type I in the system, protein type II is changing the birth rate from b'_0 to b'_1.

Again I will use that for the case when we have just one protein type undergoing a birth death process with a constant decay/birth rate is a well known fact that in steady state, its stationary probability distribution is a Poisson distribution (Karlin,S. and H.Taylor.1975).

2.5.2 Equations which describe the birth death process of the 2 types I and II in steady state

The same as in previous model, in steady state, the left side of the master equation (1) describing the time evolution of the probability distribution is 0. By difference with the two state model, one gets in steady state 9 possible situations in which quantities of protein type I and II, j and k, can be. We can write an equation for each such situation.

π_{jk} is probability of type I and II to have j respective k quantities.

when $j = 0, k = 0$:

$$\pi_{00}(\beta + b'_0) = \pi_{10}\delta + \pi_{01}d \tag{2.12}$$

when $j = 1, k = 0$:

$$\pi_{10}(\delta + \beta + b_0^{'}) = \pi_{20}2\delta + \pi_{01}\beta + \pi_{11}d \tag{2.13}$$

when $j \geq 2, k = 0$:

$$\pi_{j0}(j\delta + \beta + b_1^{'}) = \pi_{j+1,0}(j+1)\delta + \pi_{j1}d + \pi_{j-1,0}\beta \tag{2.14}$$

when $j = 0, k = 1$:

$$\pi_{01}(d + \beta + b_0^{'}) = \pi_{11}\delta + \pi_{00}b_0^{'} + \pi_{02}2d \tag{2.15}$$

when $j = 1, k = 1$:

$$\pi_{11}(\delta + d + \beta + b_0^{'}) = \pi_{21}2\delta + \pi_{12}2d + \pi_{01}\beta + \pi_{10}b_0^{'} \tag{2.16}$$

$j \geq 2, k = 1$:

$$\pi_{j1}(j\delta + \beta + b_1^{'} + d) = \pi_{j+1,1}(j+1)\delta + \pi_{j2}2d + \pi_{j-1,1}\beta + \pi_{j0}b_1^{'} \tag{2.17}$$

when $j = 0, k = 2$:

$$\pi_{02}(\beta + 2d) = \pi_{12}\delta + \pi_{01}b_0^{'} \tag{2.18}$$

when $j = 1, k = 2$:

$$\pi_{12}(2d + \beta + \delta) = \pi_{22}2\delta + \pi_{11}b_0^{'} + \pi_{02}\beta \tag{2.19}$$

when $j \geq 2, k = 2$:

$$\pi_{j2}(j\delta + \beta + 2d) = \pi_{j+1,2}(j+1)\delta + \pi_{j-1,2}\beta + \pi_{j,1}b_1^{'} \tag{2.20}$$

As before, using generating function technique we simplify our problem by transforming the above equations into ODE's satisfied by the generating function.

Let $f_k(x) = \sum_{j=0}^{\infty} \pi_{jk} x^j$ be the p.g.f.

The marginal W_k giving the probability that species 2 has k elements, k=0,1,2 is given by

$$W_k = \sum_{j=0}^{\infty} \pi_{jk} = \sum_{n=0}^{\infty} f_k^n(0)/n!$$

Since protein 2 doesn't influence protein 1 the marginal distribution of protein 1 is given by:

$$p_j = \sum_{k=0}^{2} \pi_{jk} = \frac{1}{j!} \left(\frac{\beta}{\delta}\right)^j e^{-\beta/\delta}$$

where $p_j = \pi_{j0} + \pi_{j1} + \pi_{j2}$

it follows that:

$$f_0(x) + f_1(x) + f_2(x) = \sum_{j=0}^{\infty} x^j (\pi_{j0} + \pi_{j1} + \pi_{j2}) = \sum_{j=0}^{\infty} x^j e^{-\beta/\delta} \frac{1}{j!} \left(\frac{\beta}{\delta}\right)^j = \tag{2.21}$$

$$e^{-\beta/\delta} \sum_{j=0}^{\infty} \frac{1}{j!} \left(x\frac{\beta}{\delta}\right)^j = e^{-\beta/\delta} e^{x\beta/\delta} = e^{(x-1)\beta/\delta}$$

and

$$f_0^{'}(x) + f_1^{'}(x) + f_2^{'}(x) = \frac{\beta}{\delta} e^{(x-1)\beta/\delta} \tag{2.22}$$

Using generating function technique on the equations (12-20) together with the conditions (21-22), one can derive three ODE's satisfied by the generating function (see Appendix C for detailed calculations):

$$(x-1)\delta f_0(x)^{'} + (\beta + b_1^{'} - x\beta) f_0(x) - d f_1(x) + (b_0^{'} - b_1^{'})\pi_{10} x + (b_0^{'} - b_1^{'}) f_0(0) = 0 \tag{2.23}$$

$$(x-1)\delta f_1^{'}(x)+f_1(x)(b_1^{'}+d+\beta-\beta x)+f_1(0)(b_0^{'}-b_1^{'})+x\pi_{11}(b_0^{'}-b_1^{'})-2df_2(x) \quad (2.24)$$

$$-b_1^{'}f_0(x)-f_0(0)(b_0^{'}-b_1^{'})-\pi_{10}x(b_0^{'}-b_1^{'})=0$$

$$(x-1)\delta f_2^{'}(x)+f_2(x)(\beta+2d-\beta x)-f_1(0)(b_0^{'}-b_1^{'})-f_1(x)b_1^{'}-\pi_{11}x(b_0^{'}-b_1^{'})=0 \quad (2.25)$$

and then reduce everything,(see Appendix D), at one eq.

$$(x-1)\frac{\delta^2}{d}f_0^{"}(x)+(x-1)\frac{\delta}{d}f_0^{'}(x)(C_1-2x\beta)+f_0(x)\frac{1}{d}[x^2\beta^2-(x-1)\delta\beta-x\beta C_2+C_3]-C_6x^2+C_4x+C_5=0 \quad (2.26)$$

where C_1, C_2, C_3 are known constants fully determined by the birth/death rates:

$$C_1=\delta+2\beta+2b_1^{'}+3d$$

$$C_2=2\beta+2b_1^{'}-d$$

$$C_3=(\beta+b_1^{'})(\beta+b_1^{'}+d)+2d^2+2\beta d+b_1^{'}d$$

and

$$C_4=(b_0^{'}-b_1^{'})[-f_0(0)\beta/d+\frac{f_0^{'}(0)}{d}(\delta-\beta-b_1^{'}-2d)+f_1^{'}(0)]$$

$$C_5=(b_0^{'}-b_1^{'})[f_1(0)+\frac{f_0(0)}{d}(\beta+b_1^{'}+2d)-\frac{f_0^{'}(0)}{d}\delta]$$

$$C_6=(b_0^{'}-b_1^{'})\frac{f_0^{'}(0)}{d}\beta$$

depend of $f_0(0), f_1(0), f_0^{'}(0), f_1^{'}(0)$

where we know that $f_0(0) = \pi_{00}$, $f_1(0) = \pi_{01}$, $f_0'(0) = \pi_{10}, f_1'(0) = \pi_{11}$

Equation (26) is a second order ODE with non-constant coefficients;

One can solve such equation by using "power series solutions method" (see Appendix E for detailed derivations of eq.(26)) ; Basically assume eq. has a solution of the form:

$$y = \sum_{n=0}^{\infty} a_n x^n = f_0(x)$$

$$y' = \sum_{n=1}^{\infty} n a_n x^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n = f_0'(x)$$

$$y'' = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} = \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n = f_0"(x)$$

Substitute this solution back in eq(26)

Doing this, in the end I'm left with the following recursion relation (see Appendix D):

$$a_{n+2} = \frac{a_n - L}{(n+1)(n+2)K} = \frac{1}{(n+1)(n+2)K} a_n - \frac{L}{(n+1)(n+2)K} \tag{2.27}$$

and therefore obtain for even coefficients:

$$a_{2n} = \frac{1}{2n! K^n} a_0 - \frac{L}{2n!} \sum_{l=0}^{n-1} \frac{2l!}{K^{n-l}}$$

and for odd coefficients:

$$a_{2n+1} = \frac{1}{(2n+1)! K^n} a_1 - \frac{L}{(2n+1)!} \sum_{l=0}^{n-1} \frac{(2l+1)!}{K^{n-l}}$$

going back to the assumption I've made on the solution of the 2nd order diff. eq. and using the expression for the coefficients obtained above, the solution it can be written as:

$$y = \sum_{n=0}^{\infty} a_n x^n = f_0(x) =$$

$$a_0 + a_1x + a_2x^2 + a_3x^3 + .. + a_nx^n ... = \sum_{n=0}^{\infty} a_{2n}x^{2n} + \sum_{n=0}^{\infty} a_{2n+1}x^{2n+1}$$

Therefore:

$$f_0(x) = \sum_{n=0}^{\infty}[\frac{1}{2n!K^n}a_0 - \frac{L}{2n!}\sum_{l=0}^{n-1}\frac{2l!}{K^{n-l}}]x^{2n} + \sum_{n=0}^{\infty}[\frac{1}{(2n+1)!K^n}a_1 - \frac{L}{(2n+1)!}\sum_{l=0}^{n-1}\frac{(2l+1)!}{K^{n-l}}]x^{2n+1}$$

where

$$K = \frac{G}{H}\frac{R}{S} + F$$

$$L = \frac{G}{H}\frac{T}{S};$$

F,G,H,R,S,T,C_1....C_6 are written explicitly bellow:

$$F = \left[\frac{1}{C_4} + \frac{1}{C_5}\right]\frac{\delta^2}{d}$$

$$G = \left[(C_1 + 2\beta)\frac{1}{C_4} + \frac{C_1}{C_5}\right]\frac{\delta}{d}$$

$$H = \left[\frac{1}{C_4}\beta(\delta + C_2) + \frac{C_3}{C_5}\right]$$

$$R = -\frac{1}{C_5}\frac{\delta^2}{d}$$

$$S = \frac{1}{C_6}2\frac{\delta}{d} - \frac{1}{C_5}\frac{\delta}{d}C_1$$

$$T = \frac{1}{C_6\beta}\frac{1}{2}\beta^2 + \frac{1}{C_5}C_3$$

and C_1, C_2, C_3 are known constants fully determined by the birth/death rates:

$$C_1 = \delta + 2\beta + 2b_1^{'} + 3d$$

$$C_2 = 2\beta + 2b_1^{'} - d$$

$$C_3 = (\beta + b_1^{'})(\beta + b_1^{'} + d) + 2d^2 + 2\beta d + b_1^{'}d$$

meanwhile

$$C_4 = (b_0^{'} - b_1^{'})[-f_0(0)\beta/d + \frac{f_0^{'}(0)}{d}(\delta - \beta - b_1^{'} - 2d) + f_1^{'}(0)]$$

$$C_5 = (b_0^{'} - b_1^{'})[f_1(0) + \frac{f_0(0)}{d}(\beta + b_1^{'} + 2d) - \frac{f_0^{'}(0)}{d}\delta]$$

$$C_6 = (b_0^{'} - b_1^{'})\frac{f_0^{'}(0)}{d}\beta$$

still depend by $f_0(0), f_1(0), f_0^{'}(0), f_1^{'}(0)$

The next step is to obtain $f_0(0), f_1(0), f_0^{'}(0), f_1^{'}(0)$; Once I will get this I will have an explicit formula for $f_0(x)$. To obtain $f_0(0), f_1(0), f_0^{'}(0), f_1^{'}(0)$; one can go back in the equations (23-25) and using the conditions (21),(22) for $x = 0, x = 1$ will obtain the following equations:

for x=1 and 0 in eq. 23 one obtains:

$$(b_0^{'} - b_1^{'})f_0(0) + (b_0^{'} - b_1^{'})f_0^{'}(0) + b_1^{'}f_0(1) - df_1(1) = 0 \tag{2.28}$$

and respectively:

$$(\beta + b_0^{'})f_0(0) - \delta f_0^{'} - df_1(0) = 0 \tag{2.29}$$

for x=1 and 0 in eq. 24:

$$-(b_0^{'}-b_1^{'})f_0(0)-f_0^{'}(0)(b_0^{'}-b_1^{'})-f_1(0)(b_0^{'}-b_1^{'})-b_1^{'}f_0(1)+f_1^{'}(0)(b_0^{'}-b_1^{'})+f_1(1)(b_1^{'}+d)-2df_2(1) = 0 \tag{2.30}$$

and respectively:

$$-b_0{}'f_0(0) + f_1(0)(d + \beta + b_0^{'}) - \delta f_1^{'}(0) - 2df_2(0) = 0 \tag{2.31}$$

for x=1 and 0 in (25):

$$-f_1(0)(b_0^{'}-b_1^{'})-(b_0^{'}-b_1^{'})f_1^{'}(0)-b_1^{'}f_1(1)+2df_2(1)=0 \tag{2.32}$$

and respectively

$$-f_1(0)b_0^{'}+(\beta+2d)f_2(0)-\delta f_2^{'}(0)=0 \tag{2.33}$$

also conditions (21) and (22) become

$$f_0(0)+f_1(0)+f_2(0)=e^{-\beta/\delta} \tag{2.34}$$

$$f_0^{'}(0)+f_1^{'}(0)+f_2^{'}(0)=\beta/\delta e^{-\beta/\delta} \tag{2.35}$$

where we have:

$$f_0(1)+f_1(1)+f_2(1)=1 \tag{2.36}$$

In total the equations (28-36) is a system with 9 linear equations in 9 unknowns: $f_0(0), f_0^{'}(0), f_1(0), f_1^{'}(0), f_2(0), f_2^{'}(0) f_0(1) f_1(1) f_2(1)$

2.5.3 RESULTS:

solving the above system of equations I'm obtaining the following expression for $f_0(0)$, $f_1(0)$ and $f_2(0)$:

$$f_0(0)=\frac{E_3}{E_1}-\frac{\frac{E_2}{E_1}E_3-E_2\frac{D_3}{D_1}}{E_2-E_1\frac{D_2}{D_1}}$$

$$f_1(0)=\frac{E_3-E_1\frac{D_3}{D_1}}{E_2-E_1\frac{D_2}{D_1}}$$

$$f_2(0)=-f_0(0)-f_1(0)+e^{-\frac{\beta}{\delta}}$$

where:

$$E_1 = [-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}+1/d-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}(b_1^{'}/d)+b_1^{'}/2d^2-$$

$$-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}((b_1^{'})^2/2d^2)][1+(\frac{b_0^{'}(\beta+2d)}{2d\delta})-([1-\frac{\beta+2d}{2d}])\frac{(1-\frac{b_0^{'}}{2d})}{\frac{\delta}{2d}}]+$$

$$[-\frac{1}{2d^2+(b_1^{'})^2+2db_1^{'}}((b_1^{'})^2/2d)-\frac{d}{2d^2+(b_1^{'})^2+2db_1^{'}}-$$

$$-\frac{b_1^{'}}{2d^2+(b_1^{'})^2+2db_1^{'}}+(1/2d)]\frac{(1-\frac{b_0^{'}}{2d})}{\frac{\delta}{2d}}$$

$$E_2 = -[-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}+1/d-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}(b_1^{'}/d)+b_1^{'}/2d^2-$$

$$-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}((b_1^{'})^2/2d^2)]\{[-\frac{b_0^{'}}{\delta}+\frac{(\beta+2d)(d+\beta+b_0^{'}}{2d\delta}]-[1-\frac{\beta+2d}{2d}]\frac{(1+\frac{d+\beta+b_0^{'}}{2d})}{\frac{\delta}{2d}}\}+$$

$$[-\frac{1}{2d^2+(b_1^{'})^2+2db_1^{'}}((b_1^{'})^2/2d)-\frac{d}{2d^2+(b_1^{'})^2+2db_1^{'}}-$$

$$-\frac{b_1^{'}}{2d^2+(b_1^{'})^2+2db_1^{'}}+(1/2d)][1+\frac{(1+\frac{d+\beta+b_0^{'}}{2d})}{\frac{\delta}{2d}}]$$

$$E_3 = [-\frac{1}{2d^2+(b_1^{'})^2+2db_1^{'}}((b_1^{'})^2/2d)-\frac{d}{2d^2+(b_1^{'})^2+2db_1^{'}}-$$

$$-\frac{b_1^{'}}{2d^2+(b_1^{'})^2+2db_1^{'}}+(1/2d)]e^{-\frac{\beta}{\delta}}-[-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}+1/d$$

$$-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}(b_1^{'}/d)+b_1^{'}/2d^2-$$

$$-\frac{(b_1^{'}+2d)}{2d^2+(b_1^{'})^2+2db_1^{'}}((b_1^{'})^2/2d^2)]\{[1-\frac{\beta+2d}{2d}]e^{-\frac{\beta}{\delta}}+\frac{\beta}{\delta}e^{-\frac{\beta}{\delta}}\}$$

$$D_1 = \beta + b_0^{'} - \delta(\frac{b_0^{'}(\beta + 2d)}{2d\delta}) + ([1 - \frac{\beta + 2d}{2d}])\frac{\delta(1 - \frac{b_0^{'}}{2d})}{\frac{\delta}{2d}}$$

$$D_2 = \delta[-\frac{b_0^{'}}{\delta} + \frac{(\beta + 2d)(d + \beta + b_0^{'}}{2d\delta}] + \frac{[1 - \frac{\beta+2d}{2d}]\delta(1 + \frac{d+\beta+b_0^{'}}{2d})}{\frac{\delta}{2d}} - d$$

$$D_3 = \beta e^{-\frac{\beta}{\delta}} + \delta(\beta e^{-\frac{\beta}{\delta}} + \delta)e^{-\frac{\beta}{\delta}}$$

are now constants fully determined by the birth/death rates.

Now the expression for

$$f_0(x) = \sum_{n=0}^{\infty}[\frac{1}{2n!K^n}a_0 - \frac{L}{2n!}\sum_{l=0}^{n-1}\frac{2l!}{K^{n-l}}]x^{2n} + \sum_{n=0}^{\infty}[\frac{1}{(2n+1)!K^n}a_1 - \frac{L}{(2n+1)!}\sum_{l=0}^{n-1}\frac{(2l+1)!}{K^{n-l}}]x^{2n+1}$$

is fully determined, therefore the joint probability distribution of having proteins type I/II as well. Using the result above one can determine the probability in the stationary state of having certain number of proteins type I/II given that the protein type II is present in one or two copies or totally absent. This is a generalization of the case when I have just a presence or absence of the protein I/II involved in the process.

2.6 Discussions and conclusions

The above models are very general and have the potential of being applied in many various concrete biological scenarious. An example of an alternative way to think about this models is if we imagine the following scenario.

We have 2 different types of proteins, let's call these protein types as in the above model, protein type I and II. The protein type I is benefic for the organism, the protein type II, each time it gets created even just one protein is immediately detected and

than destroyed by enzymes with an protective role against this bad for the organism protein type II. Apparently the two protein types have no connection, is just that when the good protein type I gets to a certain threshold the bad protein, type II, is increasing it's rate of being created. This has as consequence the fact that the enzyme in charge with detecting the bad protein type II created at the previous rate can not cope anymore with destroying the bad protein at the new rate of being created. This gives the potential for the bad protein to get accumulated into the organism and in time to harm the organism.

These simple but very general models provide an mathematical framework which might help for a better understanding of protein synthesis/degradation when biochemical processes frequently involve small numbers of molecules (e.g. a few molecules of a transcriptional regulator binding to one 'molecule' of a DNA regulatory region). As mentioned such reactions are subject to significant stochastic fluctuations, and therefore the stochastc behavior in the process can not be ignored anymore. As mentioned at the beginning the process of synthesis/degradation of proteins is a process known to affect many phenomenon including aging, therefore having a general mathematical framework in which to include this models would be an important step toward a better understanding of protein synthesis/degradation when stochasticity is involved.

Chapter 3

Inheritance of Epigenetic Chromatin Silencing

3.1 Introduction

3.1.1 Chromatin modifications and the impact on aging

More than two decades ago, it was proposed that 'aging of proliferating cells' results from genome reorganization occurring during the division cycle (Macieira-Coelho,1980). However, progress in the areas of epigenetics and cellular and organismal aging has been slow and sporadic. Reasons for this include the complexity of aging, as well as a lack of specific defined targets that could be probed with specific reagents. The recent discovery that over-expression of Sir2, a NAD+-dependent histone deacetylase extends yeast and worm lifespan (Hekimi and Guarente,2003) has triggered a renewed interest in the possible role of chromatin remodeling in aging and replicative senescence. It has been proposed that reassembly of repressive chromatin domains (heterochromatin) may contribute to senescence and aging process (Howard, 1996). Chromatin modifications has been linked to the biology of aging in several ways. Aging produces phenotypic chromatin defects such as telomere shortening and general heterochromatinization, which correlates with a decrease in the repair of chromatin aberrations (T. Lezhava, Chromosome and aging (2001), W. E. Wright, 2002). The first clear link between chromatin-modifying activities and the aging process was described in yeast

and pointed to Sir2p (see Kaeberlein Perspective, 2001).

3.1.2 Chromatin modifications and epigenetic processes

Epigenetic regulation of multiple heritable cell fates involves transcriptional repression or activation of the expression levels of genes, over possibly many cell cycles, without altering the underlying genetic sequence EpigeneticsAllis. Such regulation is crucial in eukaryotic development where specialized cells with identical genetic information differentiate early on to serve distinct functions. At the heart of one important mechanism of epigenetic control is the accessibility of DNA packaged into higher order structures known as chromatin. The basic unit of such packaging is the nucleosome comprising 146 base pairs of DNA wrapped around a core histone octamer (two each of H2A, H2B, H3 and H4) in $1\frac{3}{4}$ superhelical turns [1] MCB. These histones are some of the most evolutionarily conserved proteins known. Covalent post-translational modifications of these histones have been identified to be a critical player in cellular memory. At least seven such modifications (or 'marks') are documented and have been studied extensively in recent years; methylation, acetylation, phosphorylation, ubiquitination, sumoylation and ribosylation. These 'marks' create a favorable binding site for specific regulatory proteins, and thereby play a pivotal role in controlling transcriptional activation and repression, as well as other cellular processes like mitosis/meiosis and DNA repair; for a recent overview see Peterson. Another important epigenetic mark is CpG methylation of DNA. In this work we will be mostly concerned with histone modification, rather than DNA modification, although some of the issues raised may apply to DNA methylation as well.

[1] Nucleosome may also contain linker histones, e.g. H1 and variants in higher-order structure like the 30 nm chromatin fiber.

One of the defining properties of epigenetic phenomena is its stability— the ability of the cell to maintain its epigenetic state through many cell divisions. The marks responsible for the epigenetic effects, be they on DNA itself or on the histones, are bound to get diluted during DNA replication by introducing newly synthesized DNA and histone proteins, indicating that these heritable states must be robust against significant perturbations in the concentration of marks. The aim of this work is to explore minimal models of epigenetic silencing in order to identify the necessary conditions for stability of chromatin states that correspond to distinct epigenetic phenotypes.

In order to provide a concrete example, let us focus on the tails of histones H3 and H4 which exhibit a number of modifications. Methylation/acetylation of Lysines(K) and Argininies (R), phosphorylation of Serines(S) and Threonines(T) on multiple positions on these tails are some examples. Moreover, Lysine residues can accept from one to three methylations groups and Arginines can be mono- or di-methylated. The majority of these post-translational marks occur on amino-terminal (also called 'N-terminal tail') and carboxy-terminal (also called 'C-terminal tail') domains, though examples of modifications within the central domains are beginning to be unraveled. As an example of an N-terminal tail modifications, consider the case of H3K9. This Lysine can be acetylated or methylated and, as already mentioned, there are three methylated states. There is no detectable H3K9 methylation in *S. Cerevisiae*, however in *S. Pombe*, *Drosophila* and mammals, methylation of H3K9 have been associated with trascriptional silencing and acetylation has been associated with transcriptional activation Peterson, Turner, Strahl, Lachner. A combination of such marks defines an epigenetic state, and some of these states are possibly stabilized by histone modifications influencing the presence of one another.

Various enzymes coordinate histone modifications and others bind to modified tails, like chromatin modifying proteins and trascriptional regulatory proteins. From the Silenced Information Regulator (SIR) proteins in budding yeast, regulating repression of gene expression from hidden mating loci and from telomeres MCB, to silencing of developmentally important Hox genes in metazoans by the Polycomb group of proteins Gilbert, mechanisms of chromatin silencing involve enzymes that can act on more than one nucleosome in its neighborhood GrewalMoazed. This non-locality of action opens the possibility of interesting collective aspects of stability of epigenetic states.

3.2 A general stochastic model of epigenetic inheritance

3.2.1 The model

We consider a lattice of size L whose sites correspond to nucleosomes ordered along the length of the chromatin. The nucleosome corresponding to site i, has multiple states, corresponding to particular combinations of modifications of a set of side chains that we are interested in. These states are labeled by $s = 1, ..., N$. The rates of transition at site i from state s' to state s, namely, $R_{iss'}[s_1, \ldots, s_{i-1}, s', s_{i+1}, \ldots, s_L]$, depends not only on the local state but also on the states of all the neighbors within a range l. In practice, this dependence arises because particular modifications of a site leads to recruitment of particular histone modifying enzymes that could affect modification rates of the neighboring nucleosomes. Fig.6 provides a schematic representation for the model and its dynamics.

The master equation describing the time evolution of the probability distribution

$P[s_1, \ldots, s_L; t]$ is given by

$$\begin{aligned} & \frac{d}{dt} P[s_1, \ldots, s_L; t] \\ = & \sum_{i=1}^{L} \sum_{s'} \big(R_{i s_i s'}[s_1, \ldots, s_{i-1}, s', s_{i+1}, \ldots, s_L] P[s_1, \ldots, s_{i-1}, s', s_{i+1}, \ldots, s_L; t] \\ & - R_{i s' s_i}[s_1, \ldots, s_{i-1}, s_i, s_{i+1}, \ldots, s_L] P[s_1, \ldots, s_{i-1}, s_i, s_{i+1}, \ldots, s_L; t] \big) \end{aligned} \tag{3.1}$$

for times between DNA replication. At the point of DNA duplication, existing histones components like H3-H4 octamers and H2A-H2B dimers get distributed with equal probability to the resulting pair of DNA molecules Sogo, Krude, GasserSogo. This process retains some memory of the original state. In addition, newly synthesized histones also get deposited. Thus the process of DNA duplication and subsequent reassembly of nucleosomes retain, as well as dilute, the information carried by epigenetic marks.

While considering the result of duplication, we would always track one of the two resulting cells. In this work, we ignore the variability of histone marks over the cell cycle. We assume that, independently at each site i, there is one half probability of having the parental histones with epigenetic marks and one half probability of it being replaced by a newly synthesized histones where the state of histone modification s comes with probability p_s. The process of de novo assembly of histones can be thought to be independent of existing histone modifications. Therefore, we represent the evolution of the probability distribution from the parental cell to one of its progeny, due to replication and reassembly, as follows:

$$P[s_1, \ldots, s_L; nT+] = \sum_{s'_1, \ldots, s'_L} \prod_{i=1}^{L} \left(\frac{1}{2} \delta_{s_i, s'_i} + \frac{1}{2} p_{s'_i} \right) P[s'_1, \ldots, s'_L; nT-]. \tag{3.2}$$

where $nT+$ and $nT-$ refer to the times just after and just before the n-th round of DNA duplication happening with a time period of T. We assume that DNA duplication happens instantaneously (in reality, fast compared to the time between two duplication

events), namely it occurs at times $t = nT$, n being an integer.

We will study, computationally, the stochastic model of epigenetic inheritance formulated above for a particular choice of states and rules of state transitions. However, to gain some insight, it will be useful to carry out a parallel analytical approach which will be described in the next section.

3.3 Mean field theory

To solve the master equation analytically for the long time behavior of $P[s_1, \ldots, s_L; t]$ is generally an impossible task. One, therefore, has to resort to some sort of approximation. One such approximation often used successfully in statistical mechanics is the "mean field" approximation (Reichl,1997). In this approach one approximates $P[s_1, \ldots, s_L; t]$ by a factorized form $\prod_i p_i[s_i; t]$. Using this approximation one derives that the evolution equation for $p_i[s_i; t]$ is going to be

$$\frac{d}{dt} p_i[s_i; t] = \sum_{s'} (\bar{R}_{is_is'} p_i[s'; t] - \bar{R}_{is's_i} p_i[s_i; t]) \tag{3.3}$$

where the definition of the average rates $\bar{R}_{iss'}$ is

$$\bar{R}_{iss'} = \sum_{s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_L} R_{iss'}[s_1, \ldots, s_L] p_1[s_1; t] \ldots p_{i-1}[s_{i-1}; t] p_{i+1}[s_{i+1}, t] \ldots p_L[s_L; t]. \tag{3.4}$$

Notice that these averaged rates $\bar{R}_{iss'}$ are polynomials in $p_i[s; t]$ making Eq. 3.3 a nonlinear equation.

We also need the equivalent of Eq. 3.2, capturing the effect of DNA duplication.

$$p_i[s'_i; nT+] = \sum_{s'_i} \left(\frac{1}{2} \delta_{s_i,s'_i} + \frac{1}{2} p_{s'_i} \right) p_i[s'_i; nT-] = \frac{p_i[s_i; nT-] + p_{s_i}}{2}. \tag{3.5}$$

In the mean field analysis of all the models discussed in this work, we will ignore the spatial variation of 'marks' and replace them by average concentrations corresponding

to an entire region of chromatin, namely $p_i[s_i;t] = p[s_i;t]$. We thereby focus on regions of chromatin with one epigenetic fate and in the spirit of exploring minimal dynamical models, we claim that the study of just few histone modification states can already lead to nontrivial insight about the dynamical system. For example, recent studies see Sengupta has addressed one such model of silencing that included spatial structure, leading to predictions about the propagation of silencing, . In this work we will be concerned with inheritance of 'uniform' states. The equations for the variables $p[s;t]$

$$\frac{d}{dt}p[s;t] = \sum_{s'}(\bar{R}_{ss'}p[s';t] - \bar{R}_{s's}p[s;t]) \tag{3.6}$$

where $\bar{R}_{ss'} = \bar{R}_{iss'}$, is given by Eq. 3.3. They are independent of i because the rules of transitions are translation invariant and we ignore boundary effects. The equivalent of Eq. 3.2, indicating the effect of DNA duplication, in the mean-field context is

$$p[s;nT+] = \frac{p[s;nT-] + p_s}{2}. \tag{3.7}$$

We remind the reader of some well-known aspects of the mean-field approximation commonly used in statistcal physics, in order to make the present discussion self-contained. On incorporating recruitment and cooperative behavior multiple neighboring sites of a site influence the probability of the state at that site, therefore, the transition rates are dependent on what happens on neighboring sites. In what sense can these rules of transition be thought as depending solely on the state of histone modification on the site? To answer this, we suppose that the rates $R_{is_is'}[s_1,\ldots,s_{i-1},s',s_{i+1},\ldots,s_L]$ depend only on the fraction of sites in a given state in the neighborhood of i within separation l, where $1 << l$ (we could still have $l << L$ to be physically meaningful). That mean field theory is applicable, and very often an excellent approximation, can be understood by defining mean-field averaged quantities, i.e., coarse-graining the system.

We can group L sites into L/l clusters of l sites each. We redefine the probabilities $p_i[s_i, t]$ of state s_i at site $i \in [1, L]$ by the averaged probability $\bar{p}_j[s, t]$ of state S at any cluster $j \in [1, L/l]$, where formally

$$\bar{p}_j[S, t] \equiv \frac{1}{l} \sum_{i=jl-l+1}^{jl} p_i[s_i, t] \tag{3.8}$$

Now we can assume that the averaged probabilities are approximately site independent. The approximation turns out, a posteriori, to be justified when the chemical noise in the concentrations of the states is relatively small, and the system is not near a dynamical critical point. The new states S are not binary corresponding to the presence or absence of marks but a discrete spectrum of states that can be approximated by the concentration of marks in a cluster. This mean-field equivalence of the local probability of a binary state at a site to the probability density (or normalized concentration) of states in a 'coarse-grained cluster' is going to be exploited in the rest of the work implicitly in writing down mean-field differential equations for the dynamics of the system. We will not introduce in the rest of the work the formal redefinitions of probabilities done above.

3.4 Two state model

Abiding by our goal of identifying a minimal model of epigenetic silencing, we outline in this section a two-state model of stable epigenetic marks and observe that without cooperativity one cannot obtain bistability in such models. This is instructive in appreciating the role of multiple heritable histone modifications in stable epigenetic states.

Here the epigenetic state s could be just the presence (A) or absence of a mark (U), and therefore the probabilities are, with notational simplification, $p_j[A, t] = a_j(t)$

and $p_j[U,t] = u_j(t) = 1 - a_j(t)$, where, for example, $a_j(t)$ could be the probability of finding the acetylation mark A on H4K16 on a nucleosome of the chromatin of budding yeast *S. Cerevisiae* and $u_j(t)$ of finding that lysine unmodified (deacetylated). The rate constant for an acetylated mark to be deacetylated owing to Histone deacetylase (HDAC) activity and natural degradation is given by γ_A, i.e., $R_{jUA} = \gamma_A$. To include the effect of recruitment of acetylases by acetylated marks we define a rate constant of recruitment α_A. We obtain the mean-field expressions for this rate as follows,

$$\bar{R}_{jAU} = \frac{\alpha_A}{2}\{\, a_{j+1}(t) + a_{j-1}(t)\,\} \tag{3.9}$$

$$\bar{R}_{AU} \approx \alpha_A a(t) \tag{3.10}$$

Similarly, we also include the effect of recruitment of deacetylases by unmodified sites, for example, SIR2 protein complex is known to have deacetylation activity and is recruited by deacetylated sites, and the rate constant for this process is denoted by η_A. The constant rate of acetylation of an unacetylated mark is denoted by χ_A. With these definitions, we obtain the equation for the rate of acetylation. In fig. 7 is depicted a schematic representation of the two state model dynamics.

$$\frac{da(t)}{dt} = (1 - a(t))(\chi_A + \alpha_A a(t)) - (\gamma_A + \eta_A(1 - a(t)))a(t) \tag{3.11}$$

In the spirit of this work, this is the simplest model one can examine. This model has only one stable state given by

$$a^* = \frac{1}{2\bar{\alpha}_A}\left(\bar{\alpha}_A - \gamma_A - \chi_A + \sqrt{4\bar{\alpha}_A\chi_A + (\bar{\alpha}_A - \gamma_A - \chi_A)^2}\right) \tag{3.12}$$

where $\bar{\alpha}_A \equiv \alpha_A - \eta_A$. This solution goes to one for vanishing rate of degradation γ_A. This behavior is insufficient as far as epigenetics is concerned— the model fails to produce bistability even in the absence of a cell cycle. Including DNA duplication in

the model will not produce multiple dynamical attractors. This very simple analysis leads us to conclude that cooperativity (of histone modifications) is necessary in a two state model to attain bistability, as we shall soon present. In the context of the specific example of silencing in *S. Cerevisiae* [Kurdistani and Grunstein(2003)], SIR complex of proteins bind cooperatively at a deacetylated site; see [Sedighi and Sengupta(2003)] for modeling of this system.

Thus, if we allow the deacetylated and acetylated sites in the above model to recruit enzymes cooperatively to deacetylate and acetylate neighboring sites respectively, then the above model is modified to,

$$\frac{da(t)}{dt} = (1 - a(t))\{\chi_A + \alpha_A a^n(t)\} - \{\gamma_A + \eta_A(1 - a(t))^m\}\, a(t) \tag{3.13}$$

where the degree of cooperative acetylation is n and the degree of cooperative deacetylation is m. Assume that the basal rates are very small— χ_A and γ_A can be ignored to the lowest order approximation. For the simplest case of cooperative behavior ($n = m = 2$), the fixed points of the model are

$$\left\{a = 1, a = 0, a = \frac{\eta_A}{\alpha_A + \eta_A}\right\} \tag{3.14}$$

where the first two are stable fixed points, showing explicitly that both a high mark and a low mark state is stabilized by cooperative effects. More generally, call $f(a)$ the RHS of Eq. 3.13 with $n = 2$, $f(a)$ will have three zeros, $a_1 < a_2 < a_3$ in the interval $[0, 1]$. The scenario relevant to us is when a_1 and a_3 is stable and is separated by unstable a_2.

Any initial states with $a(0) < a_2$ will eventually be attracted to a_1 and any initial state with $a(0) > a_2$ will eventually be attracted to a_3. Now suppose that the cell undergoes mitosis with a typical cell-cycle period of T. For simplicity, assume that mitosis exactly halves the concentration of marks on chromatin. If $a_2 \geq \frac{a_3}{2}$ then, for

cell-cycle time T considerably larger than the timescale of histone modification rates, only one fixed point will be stable to cell-cycle perturbations over many cell-cycles, and this fixed point will be approximately a_1. This can be understood as follows. Even when the system starts close to a_3 (corresponding to high concentration of marks), the concentration of marks after mitosis will be less than a_2 and, therefore will be in the basin of attraction of the stable fixed point a_1 (low concentration of marks). However, for $a_2 < \frac{a_3}{2}$ and T fulfilling the same conditioned stated earlier, two fixed point will be stable to such cell-cycle perturbations. This condition implies that $\eta_A < \alpha_A$ for stability when χ_A and γ_A are negligible. For fairly explicit expressions for T in terms of $f(a)$ and restrictions on the parameters entering $f(a)$ and T obtained from requiring stability.

Going beyond mean field theory, we use simulations to explore the tolerance of the system to changes in the rate parameters and its stability against cell-cycle perturbation and chemical noise.

Comparison of the simulation of this model against mean-field theory is shown in Fig. 3.1. The most important conclusions from this study are the following. We have already observed that even at the mean-field level, the requirement of stability against cell-cycle perturbations impose constraints on the rate parameters. In particular, the constraint $\eta_A < \alpha_A$ implies that the cooperative conversion of U's into A's is stronger than the cooperative conversion of A's into U's. Therefore, even when the rates of γ_A and χ_A (i.e., the rates for spontaneous creation and decay of A) are small, which it should be in order for the epigenetic marks to be stable within a cell-cycle period, the fluctuations in U turning into A are magnified compared to the fluctuation in A turning into U. As an example of this "instability" of the system for a reasonable choice of

values for the rate parameter, see Fig. 3.2. The concentration $a(t)$ is plotted against time for two initial states, $a(t_0) = 0$ and $a(t_0) = 1$. In all these studies, we always consider cell-cycle period to be much larger than the typical relaxation times to reach a stable state. Nevertheless, spontaneous fluctuations may flip a low A state to a high A state eventually, often within a few cell-cycles. This phenomena is quite striking when compared to the behavior of the three-state model we introduce in the next section. To anticipate our results, we observe that a three-state model is more stable in the above sense, and we thereby postulate that presence of multiple epigenetic marks is a design criterion for epigenetic stability.

An alternative way to think about this phenomenon is as follows. Let us ask ourselves how can we go beyond mean field theory. Even if the uniform solution with a nearly zero is stable in mean field theory, there is always a non zero probability of nucleating a cluster of few A sites among all the U's. This configuration has two boundaries between the all A phase and the all U phase. The condition $\eta_A < \alpha_A$, a consequence of the constraint imposed by the states surviving through cell cycle, implies that, on the average the boundary would propagate into the all U region. This is the phenomenon of front propagation between two stable states AronsonWeinberger, CrossHohenberg. The linear growth of acetylation shown in Fig. 3.1 is the consequence of such a constant front velocity.

The only way we could make the unacetylated state survive for many rounds of cell cycle is by having the probability of the initial nucleation lowered. This indeed happens in models where the range of interaction l is large, as we have seen from our simulation of related models (data not shown). The nucleation probability is also low for the the three state model as we will argue, later.

3.5 Three state model

Having explored a two-state mean-field model and its limitations in the previous section, we now study a simple three-state mean-field model of histone modification, originally proposed in the context of silencing in fission yeast *S. pombe* Dodd, where the states are unmodified (U), methylated (M) and acetylated (A). This model is a simple example from a class of models where we will prove that bistability is a result of the presence of recruitment of multiple marks. For the sake of clarity, a concrete example of a three-state model could be the acetylation and methylation marks on H3K9. We belabor the spirit of this analysis— we are not pretending that these modifications on the histone are independent of other modifications, or that a high acetylation or high methylation on any histone tail protein leads to identical functional outcomes, we are, instead, interested in clarifying the distinctions in stability of epigenetic inheritance obtained in the presence of multiple marks. The stable fixed points we analyze could as well be combination of various histone modifications.

Coming back to the example, a methylated site recruits further methylation of neighboring nucleosomes and an acetylated site similarly recruits further acetylation. The epigenetic states s are high methylation, high acetylation and unmodified site. Therefore, we denote the mean-field probabilities as $p[M;t] = m(t)$, $p[U;t] = u(t)$ and $p[A;t] = a(t)$. These probabilities obey the conservation law $m(t) + u(t) + a(t) = 1$. Let α_M be the net (recruited) enzymatic activity of histone methyltransferase (HMT) which converts U to M and of histone demethylase (HDM) which converts M to U. Similarly, let α_A be the net (recruited) enzymatic activity of histone acetyltransferase (HAT) which converts U to A and histone deacetylase (HDAC) which converts A to U. We also include recruited conversion of A to U in the presence of M parametrized by

the enzymatic activity β_M, and M to U in the presence of A parametrized by β_A. The kinetic equations for the concentrations are given by

$$\frac{dm(t)}{dt} = \alpha_M\, u(t)m(t) - \beta_M\, m(t)a(t) \tag{3.15}$$

$$\frac{da(t)}{dt} = \alpha_A\, u(t)a(t) - \beta_A\, a(t)m(t) \tag{3.16}$$

One should include basal rates of conversion of U to M and U to A given by rate constants χ_M and χ_A, natural degradation and conversion rates of M to U and A to U given by rate constants γ_M and γ_A, and we will do so shortly.

We can further embellish this minimal model to suit other observed features like protein regulations, intermediate states like di- or mono-methylation etc., but the key aspect of bistability is already captured at this level of sophistication, and we think it is instructive to present that without complicating the model. The fixed points of the above equations are determined by the simultaneous roots of the quadratic polynomial, obtained by setting the LHS of Eqs. 3.15 and 3.16 to zero. They are given by

$$\{a^* = 1, m^* = 0\}, \{a^* = 0, m^* = 1\}, \{a^* = 0, m^* = 0\},$$
$$\left\{a^* = \frac{\alpha_M\beta_A}{\alpha_A\beta_M + (\alpha_M + \beta_M)\beta_A}, m^* = \frac{\alpha_A\beta_M}{\alpha_A\beta_M + (\alpha_M + \beta_M)\beta_A}\right\}$$

It can be easily checked that the first two fixed points are stable, the third fixed point is an unstable saddle point and the fourth point is unstable. It is not hard to convince oneself that if one includes small basal rates the stability of the model remains unaffected, and we come back to this later.

This simple level of modeling may already be quite relevant. We observe that in the absence of active chromatin remodeling processes which may dictate basal rates for conversion and degradation of marks, recruitment alone ensures that methylated and acetylated states are quite robust against mitotic perturbations. During mitosis,

the parental nucleosomes with marks are distributed randomly to daughter chromatins, however, newly synthesized nucleosomes are modified by recruitment from neighbors, restoring the original state. Cooperativity is not necessary. One can argue that the prevalence of multiple modifications of histones, instead of just unmodified and uniquely modified histones (a two-state scenario), is owing to this efficient robustness achieved through multiple states. The reason for this increased stability lies in the higher dimensionality of the space of configurations and the fact that multiple transitions (say, $M \rightarrow U \rightarrow M$, at more than one neighboring sites) need to take place before one nucleates the other stable phase.

For the sake of completeness, we now analyze the model by including basal rates for conversion and degradation. In fig. 8 is depicted a schematic representation of the three state model dynamics. The new equations are

$$\frac{dm(t)}{dt} = \alpha_M\, u(t)m(t) - \beta_M\, m(t)a(t) + \chi_M u(t) - \gamma_M m(t) \tag{3.17}$$

$$\frac{da(t)}{dt} = \alpha_A\, u(t)a(t) - \beta_A\, a(t)m(t) + \chi_A u(t) - \gamma_A a(t) \tag{3.18}$$

A plot of the flow lines when high A and high M states are stable is shown in Fig. 3.3. Points are evenly distributed on a grid and allowed to evolve for a fixed time in generating the flow lines numerically. The hue of the plotted lines is changed linearly in time. A similar plot for the scenario when the high A and high M states are unstable as shown in Fig. 3.4. This is the case when the degradation rates are too high. The lattice-averaged concentration of mark $a(t)$ as a function of time is plotted in Fig. 3.5.

3.6 Conclusion

We have formulated a mathematical model of inheritance of epigenetic silencing and showed how we have two routes to producing stable epigenetic states: one via co-operativity of silencing factor recruitment and the other via the presence of multiple marks, where there are barrier states between an active and a repressed states. We also found that multiple marks allow the cell higher stability to cell-cycle perturbations, in comparison to a single mark system.We believe that the robustness of these models to cell-cycle perturbation may be a reason why multiple histone modifications are observed frequently in epigenetic design.We note however, that at a fundamental level these two are not entirely distinct routes.The presence of intermediate states naturally lead to cooperative effects when each of the intermediate states recruit enzymes for further modification. Moreover, protein complexes that induce further enzymatic activity often possess domains that simultaneously recognize histone modifications at adjacent sites. This is thought to be the case with SIR protein complex and also for the polycomb silencing mechanism. Effective cooperativity can emerge on eliminating transient intermediate states in models with first order rates.

We have phrased the mean field theory in terms of coarse-grained quan- tities like the fraction of sites with a particular mark in a cluster. For those readers familiar with statistical physics, a natural question is how does effec- tive model change if we continue the coarse-graining to larger length scales. In other words: how does the model renormalize" under the blocking trans- formation (Reichl, 1997)? In practice, setting up a reasonable scheme for doing such block transformation may be dificult. However we could make some educated guesses about what would happen. In absence of any conser- vation law, there is no obvious reason why this system should not have

a finite (although long) correlation length in space and, similarly, a finite corre- lation time. The system would not have genuinely multiple phases. All these effects, which are missed by mean field theory, would, in principle, show up in renormalization group transformations. We had already mentioned how the system could get out of one of the phases, by nucleation of the other phase, and showed the numerical evidence. Such nucleation gives rise to do- main boundaries, which are responsible for finite correlation length in the system. Technically, therefore, the system becomes very weakly coupled if we coarse-grain to blocks with size larger than the correlation length. Having said that, in the biological context, the domains usually incorpo- rate few hundred nucleosomes and epigenetic states are stable for somewhere between 10-100 cell cycles. It is enough for the model to produce correlation lengths and correlation times in those ranges. Mean field theory gives us a hint when such correlated states appear. However, in this approximation, 13 long, but finite, lifetimes become infinite.

As we saw, for both states to be long-lived, we need suppression of the probability of spontaneous nucleation of the more stable state (as measured by average front velocity helping to spread the state). This can be achieved either by having a more complex model which requires multiple marks to occur before nucleation happens, or by having a long range model where many sites have to have unlikely changes before the nucleation is complete.

In practice, for the systems biology of silencing, the possibility of more complex models is worth serious consideration, especially when there is no obvious mechanism of cooperativity and there appears to be a plethora of histone marks that are involved in the process. In addition, these models have different degree of robustness to variation of conditions from cell to cell. Many of the parameters in the model are not just chemical

reaction rates but also depend upon abundances of certain proteins in the cell. For example, the effect of the neighbors is often through recruitment of histone modifying enzymes not explicitly modeled. Variation in the abundance of those enzymes would change the effective parameter from cell to cell. On the other hand, if the biochemistry dictates that the basal modification rates are very small, say compared to modification due to recruited enzymes, the basal is unlikely to become significant player in any of the cells. If one neglects the basal rates, the two state model has an additional constraint on the nonzero parameters, in addition to constraints of multistability, whereas the three state model does not have such an additional condition. As a result, we expect the functionality of the second model to be more immune to cellular variability.

The interaction between cell cycle and epigenetic silencing is a rich subject in biology. We have only focused on one aspect of it in these models, namely, the recovery of the epigenetic information after the dilution caused by DNA duplication, and ignored other phenomena like cell cycle dependent histone modifications. However, even within our simplest setup, different classes of models give rise to interesting differences in performance. Exploring such models in combination with experiment designed to test qualitative predictions valid for a broad class of models is the way to gain insight into the nature of epigenetic inheritance.

References

[Allis et al.(2007)Allis, Jenuwein, and Reinberg] Allis, C. D., Jenuwein, T., Reinberg, D. (Eds.), 2007. Epigenetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

[Aronson and Weinberger(1975)] Aronson, D., Weinberger, H., 1975. Non-linear diffusion in population genetics and nerve pulse propagation. In: Goldstein, J. (Ed.), Partial Differential Equations and Related Topics. Vol. 446 of Springer Lecture Notes in Mathematics. Springer, NY, pp. 5–49.

[Cross and Hohenberg(1993)] Cross, M. C., Hohenberg, P. C., 1993. Pattern formation outside of equilibrium. Reviews of Modern Physics 65 (3, part II), 851–1112.

[Dodd et al.(2007)Dodd, Micheelsen, Sneppen, and Thon] Dodd, I. B., Micheelsen, M. A., Sneppen, K., Thon, G., 2007. Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129, 813–822.

[Gasser et al.(1996)Gasser, Koller, and Sogo] Gasser, R., Koller, T., Sogo, J., 1996. The stability of nucleosomes at the replication fork. J Mol Biol. 258 (2), 224–239.

[Gilbert(2003)] Gilbert, S., 2003. Developmental Biology. Sinauer, Sunderland, MA.

[Grewal and Moazed(2003)] Grewal, S. I. S., Moazed, D., 2003. Heterochromatin and epigenetic control of gene expression. Science 301, 798–802.

[Krude and Knippers(1991)] Krude, T., Knippers, R., 1991. Transfer of nucleosomes from parental to replicated chromatin. Mol Cell Biol. 11 (12), 6257–67.

[Kurdistani and Grunstein(2003)] Kurdistani, S. K., Grunstein, M., 2003. Histone acetylation and deacetylation in yeast. Nature Reviews. Molecular Cell Biology 4, 276–284 .

[Lachner et al.(2003)Lachner, O'Sullivan, and Jenuwein] Lachner, M., O'Sullivan, R. J., Jenuwein, T., 2003. An epigenetic road map for histone lysine methylation. Journal of Cell Science 116, 2117–2124 .

[Lodish et al.(2004)] Lodish, H., et al., 2004. Molecular Cell Biology. WH Freeman, New York, NY.

[Peterson and Laniel(2004)] Peterson, C., Laniel, M., 2004. Histones and histone modifications. Curent Biology 14 (14), 546–551.

[Sedighi and Sengupta(2003)] Sedighi, M., Sengupta, A. M., 2003. Epigenetic chromatin silencing: bistability and front propagation. Phys Biol 4, 246–255 .

[Sogo et al.(1986)Sogo, Stahl, Koller, and Knippers] Sogo, J., Stahl, H., Koller, T., Knippers, R., 1986. Structure of replicating simian virus 40 minichromosomes. the replication fork, core histone segregation and terminal structures. J Mol Biol. 189, 189–204.

[Strahl and Allis(2000)] Strahl, B. D., Allis, D., 2000. The language of covalent histone modifications. Nature 403, 41–45.

[Turner(2002)] Turner, B., 2002. Cellular memory and the histone code. Cell 111 (3), 285 – 291.

[G.M. Schtz In: C. Domb and J. Lebowitz Ed., Phase Transitions and Critical Phenomena vol. 19(2001)]

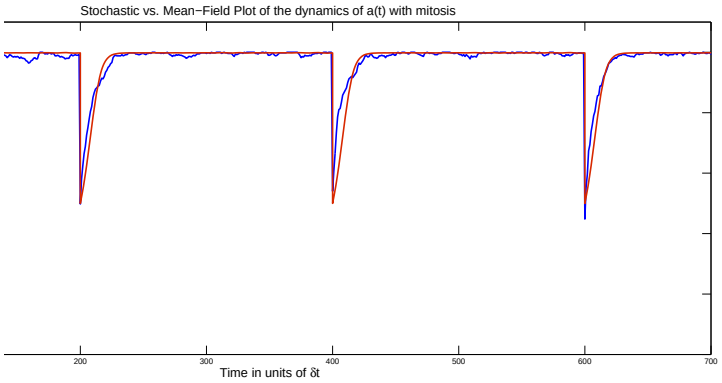


Figure 3.1: For chapter 3: Two state model's stochastic simulation, averaged concentration $a(t)$ and mean-field ODE solution fit. Values of parameters: $\alpha_A = 5; \eta_A = 2.5; \gamma_A = 0.1; \chi_A = 0.01$. For the ODE fit, the fitting time-scale is $\delta t = 56/3$.

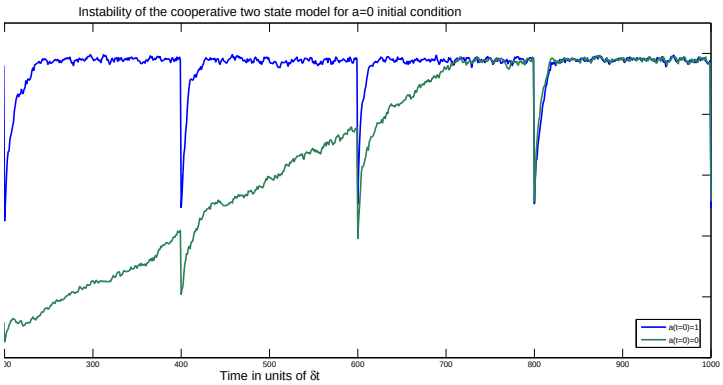


Figure 3.2: For chapter 3:Two state model's stochastic simulation, averaged concentration $a(t)$ starting from initial states $a(t = 0) = 1$ and $a(t = 0) = 0$. Values of parameters: $\alpha_A = 5; \eta_A = 2; \gamma_A = 0.1; \chi_A = 0.01$. Though mean field theory would predict stability, fluctuations compromise it.

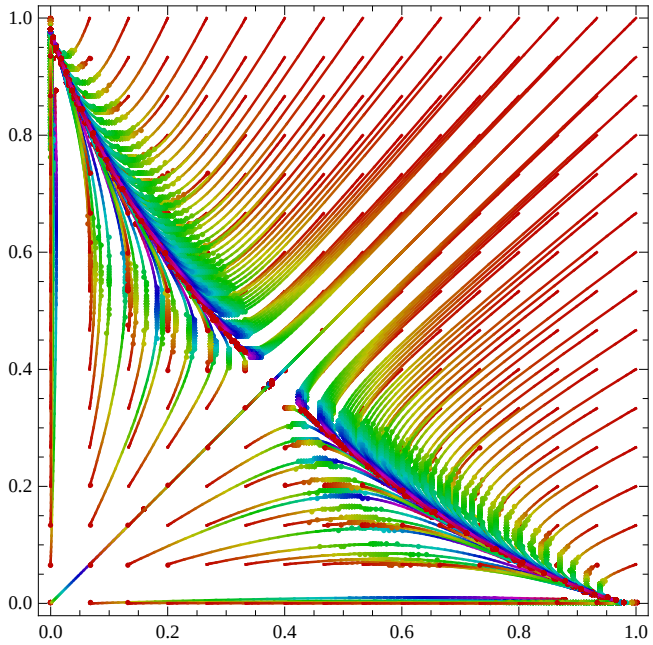


Figure 3.3: For chapter 3:Three state model's phase flow with high A and high M stable, x-axis is $m(t)$ and y-axis is $a(t)$. Values of parameters used: $\alpha_A = \alpha_M = 5; \beta_A = \beta_M = 3; \gamma_A = \gamma_M = 0.1; \chi_A = \chi_M = 0.01$

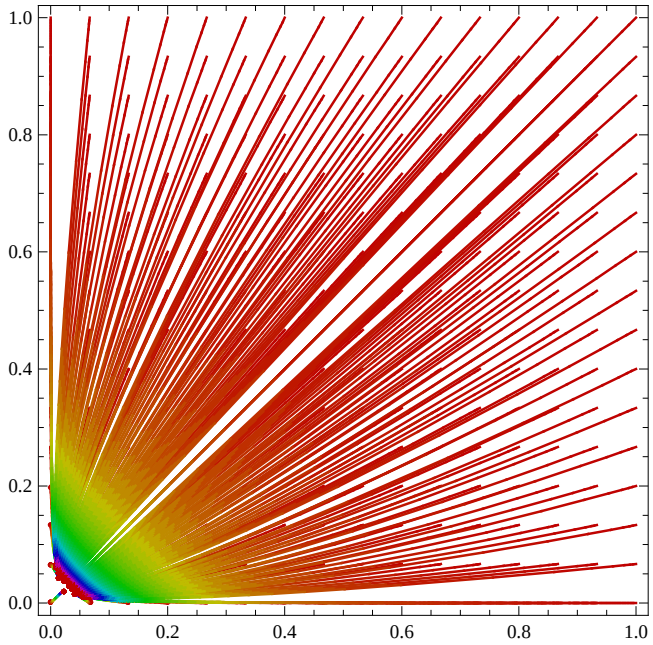


Figure 3.4: Three state model's phase flow with U stable, x-axis is $m(t0$ and y-axis is $a(t)$.Values of parameters used: $\alpha_A = \alpha_M = 5; \beta_A = \beta_M = 3; \gamma_A = \gamma_M = 5; \chi_A = \chi_M = 0.01$

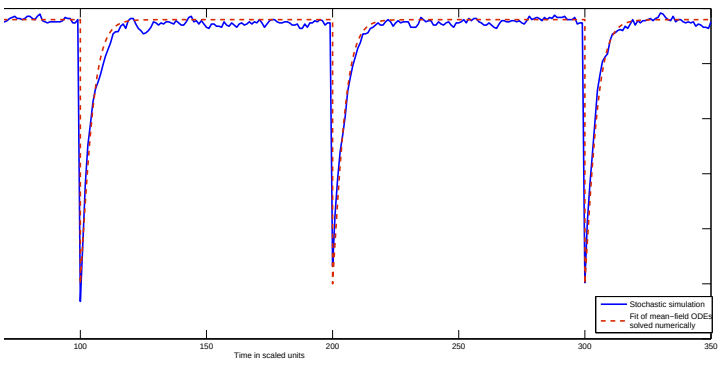


Figure 3.5: Three state model's stochastic simulation, averaged concentration $a(t)$ and mean-field ODE solution fit. Values of parameters: $\alpha_A = \alpha_M = 5; \beta_A = \beta_M = 3; \gamma_A = \gamma_M = 0.1; \chi_A = \chi_M = 0.01$. For the ODE fit, the fitting time-scale is $\delta t = 15.5$.

PART II

Chapter 4

Quantitative transcriptional analysis of aging *C. elegans*

4.1 Introduction

Given the existence of several mechanisms of aging (those conserved across species), simple models have become important tools for elaborating the basic biology of aging. Our lab uses the nematode *C. elegans* to dissect conserved mechanisms of aging.

Caenorhabditis elegans was chosen by Sydney Brenner in 1965, as a model organism to study animal development and behavior. This soil nematode has proven to have a great potential for genetic analysis, partly because of its rapid (3-day) life cycle, small size (1.5-mm-long adult), and ease of laboratory cultivation. *C. elegans* natural way of breeding is as a self-fertilizing hermaphrodite. (see Fig. 1- from *C.elegans* II)

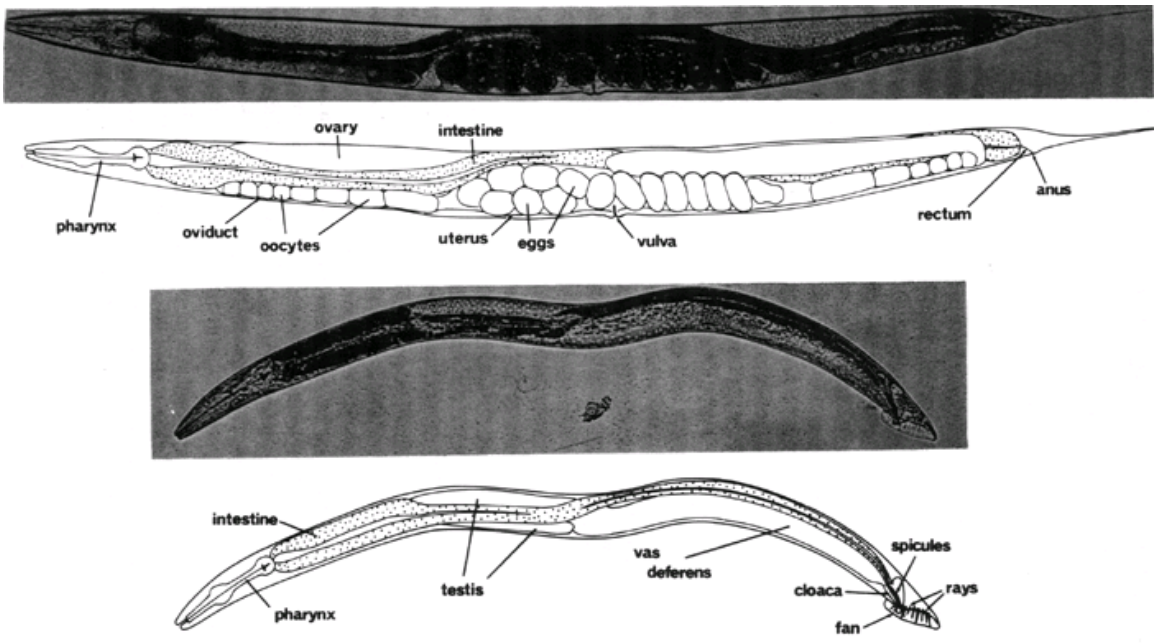


Fig. 1 *C. elegans* anatomy (from from *C. elegans* II book)

Another advantageous feature of this nematode is its size, just 20 times that of *E. coli* and its simple anatomy of 959 cells, including the 302-cell nervous system. With such a small nervous system, *C. elegans* was the first animal model in which its circuitry was completely reconstructed by serial-section electron microscopy (White et al. 1986, 1988). The wild-type reconstructions showed all the connections of all the neurons in the hermaphrodite nervous system. Other unique advantages offered by this organism are the transparency of the body, the constancy of cell number, and the constancy of cell position from individual to individual. Due to such advantages, the complete wild-type cell lineage from fertilized egg to adult was determined by observation of cell divisions and cell migrations in living animals (Sulston et al. 1988).

An essentially complete *C. elegans* sequence was published in *Science* (Vol. 282 11 December 1998) and the last remaining gap in the sequence was finished in October 2002. The completed *C. elegans* genome sequence is represented by over 3,000 individual clone sequences which can be accessed through WormBase. The worm informatics group at the Sanger Institute play an important role in assembling the whole database. WormBase is the repository of mapping, sequencing and phenotypic information for *C. elegans* (and some other nematodes). WormBase is based on the Acedb database system. Acedb was originally developed by Jean Thierry-Mieg (CNRS, Montpellier) and Richard Durbin (Sanger Centre) for the C. *elegans* genome project, from which its name was derived (A C. *elegans* DataBase). However, the tools in it have been generalized to be much more flexible and the same software is now used for many different genomic databases from bacteria to fungi to plants to man. It is also increasingly used for databases with non-biological content.

The entire genome of *C. elegans* encodes approximate 19000 genes. Note that ~50% of *C. elegans* genes have clear human homologs and many molecular processes are strikingly conserved from nematodes to higher organisms. Over 400 mutations or RNAi interventions that extend the ~ 3 week *C. elegans* lifespan have been identified. Although the longevity phenotype is a focal point for much of the work in the field, much less is understood about whether longevity genes actually act to extend *healthspan* (the period of mid-life “vigor” that precedes decline), enabling the animal to live a high quality or "youthful-like" life for longer.

One simple question is what happens to gene expression as animals age--can analysis of transcriptional profiles inform about the biology of aging and suggest ways we might extend healthspan? In the first part of my thesis, I have therefore been analyzing transcriptional profiles of aging animals with a focus on two areas suggested by ongoing work in the Driscoll lab: one addresses the question of whether there is a major 'crisis' or transcriptional transition during midlife, and a second focuses on age-associated muscle deterioration.

4.2 Transciptional profiling to characterize aging and identify genes that might impact healthspan

Microarrays are chips on which genes are attached for hybridization to mRNA samples--hybridization signals indicate which genes are expressed as messages and can speak to relative abundance and changes in gene expression over time. We prepared replicate RNA samples over the course of *C. elegans* adult life and hybridized to near complete genome arrays to ask how transcription changes during adulthood and to correlate some of these with aging phenotypes. My analysis uses methods developed for data mining microarray experiments, adapted for aging research. The method I'm using bridges knowledge of statistical mechanics with data mining methods developed in statistical mathematics. Such analyses can reveal how the transcriptional regulation of genes might coincide, thereby implicating proteins as parts of networks acting together towards a common biological function. Such experiments are most useful for complex biological traits that result from the

presumed functioning of several molecular pathways. Aging is one such biological phenomenon that incorporates numerous molecular mechanisms underlying environmental stimulus sensing, metabolic regulation, stress responses, reproductive signaling, and transcriptional regulation. Current models of aging emphasize different mechanisms as driving forces behind aging and lifespan determination. However, an integrated understanding of exactly how these mechanisms drive aging has not yet been formulated.

I used supervised and unsupervised methods for gaining a better understanding of the gene expression changes that might impact the aging process. When interpreting the data using a supervised approach, I've tried to address the major biological theories currently known that describe aging. To address the oxidative damage theory of aging, for instance, I highlight stress response genes that exhibit statistically significant changes, and then ask whether the expression patterns of these genes share a common pattern. Overall, my work includes surveys of insulins, longevity-implicated genes, dauer-related genes, autophagy-related genes, muscle, neuronal and germline genes as groups of interest relative to aging and healthspan that I analyze in detail.

Using an unsupervised approach based on concepts from statistical mechanics, I identified an interesting gene expression pattern that suggests that a gene expression switch at midlife. This switch coincides with the onset of biomarkers of aging including age pigment accumulation. I conclude my work with

the description of the gene expression sets that underlie new hypotheses about impact on later aging.

Several experiments using microarray technologies to address the different models of aging have been published. A second phase of my analysis was to look at how my data intersects with similar studies performed. I will focus on comparison of just two other experiments due to their similarity with our experiment. I've searched for the overlap between experiments and the common trends in expression pattern using different statistics methods for normalizing and filtering the data. This part of my work will not be included in the thesis however will be included in the paper in preparation (see **David-Rus, Driscoll et. al 'A search for mid-life gene expression changes that might influence aging', in preparation)**.

A first foray into *C. elegans* microarray analysis was performed using Affymetrix oligonucleotide-spotted chips.(Hill et.al. 2000). This study compared gene expression profiles from 18,791 predicted open reading frames, mainly over developmental time points. One mid-life time point (post-fertilization day 14) is included in the study for comparison. One-way ANOVA analysis was performed on the data, which then was normalized to have a mean value of zero and a variance of one. 4221 ORF's with statistically significant variations in frequency ($p < 0.001$) were identified. Of these ORF's, subclusters (clustered by self-organizing maps (SOM)) of expression patterns that exhibit *declining* expression at the 2-week time point further were studied (Hill et al., 2000).

The first focused microarray study devoted to studying aging in *C. elegans* utilized a probe DNA-spotted microarray of 17, 871 open reading frames to study aging nematodes over a series of time points spanning pre-reproductive adulthood to old age (Lund, et al., 2002). In this study a combination of mutation and strains has been used, see table 1 below:

Table 1. Time Course of *C. elegans* from 3 to 19 Days of Age

Day	3	4	6–7	9–11	12–14	16–19
fer-15	2	0	1	1	1	0
spe-9;fer-15	3	3	2	5	2	3
spe-9;emb-27	1	0	1	0	1	
Total = 26	6	3	4	6	4	
	Young adult		Oocyte production ends			<25% survival

The number of arrays and the strain of worm included in each time point are shown. Notable characteristics of the population are indicated.

Table 1 from Lund et. al. 2002

An ANOVA analysis was undertaken to identify statistically significant variations among the time points, then the data was normalized to the earliest time point (non-aging-related gene expression). Open reading frames showing variations in expression over time were clustered together in groups showing common expression changes. Those genes that changed only from the pre-reproductive to first reproductive time point were labeled as maturity genes. Those genes that had any changes in expression over the successive time points were designated as aging genes. After statistical filtering, 201 genes exhibited changes over time; 34 maturity genes and 167 aging genes. Three genes were subsequently discarded due to strong correlation with a particular strain, and 72 of the remaining genes were found to encode proteins conserved across species.

A second study (McElwee et al., 2003) compared the gene expression from *daf-16(-/-)* and *daf-16(+/+)* worms on a *daf-2 (-/-)* reduction of function mutant background, only on the first day of adulthood. This microarray utilized DNA probes corresponding to 17,871 *C. elegans* genes. 1646 genes were isolated that showed differential expression of greater than 1.5-fold. 602 genes were up-regulated in *daf-16 (+/+),* animals, while 1044 genes were down regulated.

The third study (Murphy et al., 2003) of gene expression changes in aging *C. elegans* utilized DNA probes corresponding to 18455 open reading frames. This study also compared samples from *daf-2*-deficient and *daf-2;daf-16*-deficient animals, as well as from wild-type animals. However, this study went further than the McElwee study by comparing the results across multiple time points. The time points begin at

a pre-reproductive age and continue until mid-adulthood (later than the time point collected in the McElwee study, but earlier than the Lund et al. study.

A fourth study compares results from microarray studies of aging across species, including *C. elegans* (from the Murphy et al. data), *D. melanogaster*, *S. cerevisiae*, and *H. sapiens*. (McCarroll et al., 2004). In this experiment were performed specific comparisons between *C. elegans* and *D. melanogaster* at two points early adulthood and mid-life adulthood) included several hundred ortholog gene pairs that are conserved in expression across the two species.
A final study is a time-course study of an aging wild-type (N2) and non-aging *daf-2* (-/-). What distinguishes this study is that target samples were prepared from individual nematodes rather than populations, thereby bypassing the variation in aging inherent to worm populations (Golden and Melov, 2004).

Beyond *C. elegans*, other studies have looked at genome-wide transcriptional profiles of aging in specific tissues of other organisms, as well as in the whole organisms of flies and yeast. Such studies have surveyed aging in mouse liver, mouse heart, mouse brain, mouse muscle, rat hippocampus, rat kidney and pituitary, rat muscle, human blood, and human muscle.

When interpreting their data, several of the studies took a similar, supervised approach in the context of current theories of aging. To address the oxidative damage theory of aging, for instance, the studies identify stress response genes that exhibit statistically significant changes, then ask whether the expression patterns of these genes share a common pattern. Conversely, several of the studies have taken an unsupervised approach and examine expression cluster groups for patterns that indicate possible relevance for the biology of aging, looking for commonalities of function among the listed genes, in addition to the relevance of individual genes for aging.

Hill et al. study focuses on cluster groups where the expression patterns decrease with age. One such cluster group was found to be enriched for genes for metabolic activity, including oxidoreductases, amino acid metabolism genes, carbohydrate metabolism genes, and protein synthesis genes. Other down-regulated genes were found to belong to common functional groups, such as muscle-related genes and genes coding for extra-cellular matrix proteins.

In the Lund et al. study, a more supervised approach was taken. Insulins, aging-related genes, dauer-related genes, heat shock genes, transposons, muscle, neuronal and germline genes all were singled out and their expression profiles was examined. Key findings here include: 1) that both aging and dauer-related genes cluster to mount 15 in the Kim expression map when multiple gene expression experiments are combined, 2) while specific insulin genes change in expression over time, the insulin signaling pathway genes do not change over time, 3) both muscle and neuronal genes show an increase in expression in later life, indicative of an up-regulation of the expression of these genes, or indicative of a general down-regulation of the majority of other genes 4) heat shock genes are up-regulated initially, then down-regulated at the latest time points, suggestive of the lack of a stress-response as being possibly causative for the ultimate demise of the organism. Lund results are comparable with our results where we identified heat shock genes as up-regulated as well in cluster G11. By difference with Lund, most of our heat shock genes stay up-regulated over the entire life of the nematode (see more at cluster analysis-G11 cluster) 5) mitochondrial genes and genes involved in oxidative stress resistance do not change over time, 6) transcription of transposable elements is increased, perhaps indicative of a less stable genome with age, and 7) germline genes only are down-regulated at the latest time-point, indicative that the cessation of oocyte production is not dependent on gene expression regulation. Also indicated is that persisting germline tissues into old age may retain functional abilities for reproduction if exposed to a favorable environment. Additionally, an unsupervised approach was taken and 167 genes that show any type of significant change over time are identified. This relatively low

number of aging-related genes supports a model where environmental damage contributes more to aging than genetic influences. However, another plausible argument is that the modulation of the expression of a few genes may directly influence life span determination.

In the McElwee et al. study, (Aging cell 2003) results look at comparing an aging (*daf-2 -/-; daf-16 -/-*) and delayed-aging (*daf-2 -/-; daf-16 +/+)* population of worms. In a delayed-aging population, heat shock and oxidative stress-response genes are observed to increase. This increase correlates with the decrease in the heat stress-response genes in the aging Lund et al. population. Furthermore, in the delayed-aging population, metabolic genes are observed to decrease. This finding might speak to the theory that a higher metabolism leads to greater tissue damage and enhanced aging due to buildup of damaging metabolic byproducts. Finally, again corroborating the Lund et al. data, *ins-7* is observed to increase in the delayed-aging population (Lund et al.'s aging population shows a decrease in *ins-7*). Interestingly, no gene expression changes were observed in protein synthesis or protein degradation genes (both proteosomal, and more specific, non-proteosomal genes). This finding does not support the theory that reduced protein turnover in a cell might lead to a buildup of protein damage and an enhancement of aging phenotypes.

The Murphy et al. study, is a similar study comparing *daf-16* (+/+)(delayed-aging) and *daf-16*(-/-)(aging), includes many results comparable to McElwee et al.. For example, both oxidative and heat shock stress response genes increase in expression in the delayed-aging population. Furthermore, metabolic genes were decreased in the delayed-aging population. Conversely, however, this study finds *ins-7* decreasing in the delayed-aging (*daf-16 +/+*) population, while increasing in the aging (*daf-16 -/-*) population. I will present later how this study compares with our study. The authors point out further that the gene identified in the McElwee study as *ins-7* really is *ins-30* based on the cosmid name. *ins-18* is up-regulated in the delayed-aging population, in contrast to the up-regulation of *ins-18* in the aging

population in the Lund et al. study. Further findings include: 1) that anti-microbial genes are up-regulated in the delayed-aging population. This finding supports a theory that the ultimate demise of the animal is due to bacterial infections overcoming the weakened organism in old age, 2) vitellogenin is down-regulated in the delayed-aging population, consistent with the theory that excessive expression of non-essential genes also may contribute to the aging and demise of the organism, 3) in contrast to the McElwee study, several proteases were repressed in the delayed-aging population, 4) lysosomal genes were up-regulated in the delayed-aging population, and finally 5) genes from the glyoxylate cycle, which are up-regulated during dauer and hibernation, also are up-regulated in the delayed-aging population. This finding is consistent with the data from Lund et al. that shows dauer and aging genes co-segregating on the same gene expression mountain (Mount 15).

The cross-species comparison between *C. elegans* and *D. melanogaster* (Steven A McCarroll et. al, Nature genetics 2004) reveals trends common to both species or unique to each species. Trends in common include: a downregulation of many mitochondrial and oxidative metabolism genes (including mitochondrial membrane genes, genes for components of the electron transport chain, ATP synthase genes, and genes in the citric acid cycle), a downregulation of peptidases, proteins for DNA repair, and genes coding for ATP-dependent transporters. Gene expression changes unique to aging *C. elegans* include: a downregulation of collagens, histones, trasnposases, and DNA helicases. Gene expression changes unique to *D. melanogaster* include upregulations of cytochrome p450 genes, glycosylase genes, and peptidoglycan receptors.

A more recent (Golden and Melov, 2004) *C. elegans* microarray study utilized individual nematodes compared wild- type (N2) to *daf-2*(-/-) nematodes. Interestingly, greater gene expression changes were observed between the two strains rather than between different ages of a single strain. This is consistent with the Lund et al. data that found few changes in gene expression with age. Many

individual genes that relate to current models of aging were identified as different among N2 and daf-2(-/-). In N2 nematodes, an increase in antimicrobial peptides, mitochondrial electron transport chain proteins, proteasomal components, and actin all can be explained by the earlier onset of aging in N2 worms, thereby increasing the needs for defenses and anabolism to compensate for the age-related deterioration.

While these studies all focus on aging, the varying time points present a problem for the specific study of the mid-life aging, or healthspan, of the organism. Therefore, we chose to perform a time-course study of aging wild-type nematodes, from reproductive to old aged, with an emphasis on covering the mid-life time points, represented by the post-embryonic days 9-12 when grown at 25 degrees Celsius. Previous studies from our lab reveal that the mid-life changes in an animal may be critical in determining the ultimate lifespan of that animal. All of our samples utilize the same sterile mutant strain and replicates were harvested at the same time point (give or take an hour). Affymetrix oligonucleotide arrays were chosen based upon the good coverage of open reading frames on the array, and based on the optimization of the Affymetrix system. We chose to use a clustering system based upon the statistical mechanics of disordered granular ferromagnets and developed in the Domany lab (M. Blatt, S. Wiseman, and E. Domany, (1996)). This clustering system has proven superior to other clustering methods for a variety of biological problems (E, Domany et.al, (1997, 1998).

4.3 A search for mid-life gene expression changes that might influence healthspan- Experimental design

Previously we showed that neuronal cells do not physically deteriorate whereas muscle cells deteriorate morphologically with age, starting in mid-life (see Herndon L. et all Nature 2002).Interestingly my microarray analysis capture such tendencies. I found that neuronal gene expression shows little change for a certain group of

neuronal related genes and I showed an overall decline in transcription of muscle expressed genes. Of the muscle related genes, a group of muscle genes is expressed at lower levels during midlife.

Further more, using unsupervised approaches I've identified an interesting gene expression pattern that suggests that a previous unknown genetic switch might occur during midlife day 10 from the time of hatching, consistent with patterning of changes in age pigment accumulation rates. I noted a similar gene expression pattern for day 11 in Kenyon data when I clustered this data (see results at comparison section)

I've also performed a comparative analysis with data from other microarray studies. Here I've looked for the overlap between experiments and the common trend in pattern expression, deploying different statistics methods for normalizing and filtering the data. I've also clustered using Domany algorithm each of the data set I used for comparison.

With an interest in tracking gene expression changes over the *C. elegans* adult lifespan and in identifying genes that are similarly regulated in aging microarray experiments in independent laboratories, we performed microarray analysis using RNA isolated from adults of increasing age.

We identified genes for which expression changes over adult life, using the oligonucleotide type of chip. These chips were specially designed for *C. elegans* by the Hoffmann-LaRoche company, from Basel, Switzerland, as Affymetrics format which effectively covered 87% of the actual predicted genes. The raw data identified by Affymetrix ID annotations is found in the supplementary materials (see Table X).

To grow the worms in a synchronous way, and at the same time limit use of multiple mutations, we used *spe-9*(hc88) which is a temperature sensitive sterile mutation. We cultured *spe-9* mutants at restrictive temperature of 25.5^{o}C. Independent samples containing ~ 20000 worms were taken on different days: day3, day6, day9, day10, day 11, day12, day15. Day 3 is the first day of adulthood, with day 0 the day of hatching. For each day, the mRNA of 3 independent samples was extracted. Each sample was labeled and hybridized to the *C. elegans* genome

chip so for each day, we have 3 samples hybridized to 3 independent chips.

Because of our focus was on potential relevant changes at the midlife transition, suggested by changes in rate of age pigment accumulation (Gertsbrein et. al.) we also prepared another triplicate experiment in which we harvested nematodes at days 9, 10 and 11. Data from these middle time-points were combined with those in the more extended trials to increase the significance of findings at days 9, 10 and 11 (we therefore analyzed six total independent repeats for the middle life time points).

4.4 Identifying the 2000 genes that show greatest variance over time points.

The next steps, called data preprocessing, are important to address several issues related to removal of the effects of systematic sources of variation; to identify if there are still any discrepant observations and to transform the data into a scale suitable for analysis. Preprocessing can greatly enhance the quality of any analysis, therefore is critical to choose the right methods appropriate to the particular type of data and the questions that will be analyzed.

The microarray data can be represented in a matrix form. The rows are the genes covered on our chips, and the 7 columns are the seven time points in which we were interested.

In order to detect the outliers we used the Nalimov outlier test, an outlier exclusion test. For each gene per *condition* a modified Nalimov outlier test (Kaiser R, Gottschalk G (1972)) is performed for data points representing replicate experiments. In contrast to the original test, we used a modified version called "Nalimov1". A normal distribution model is calculated for data points to be tested, and outliers are removed at a 95% confidence level. This means that only in 5 of 100 cases a data point is removed erroneously. Since the test is rather conservative, Nalimov outlier removal normally improves the quality of results, since (chip) artifacts are quite reliably removed. Note that the test requires at least 3 data points (*replicate* experiments) for an experimental condition, otherwise no outliers can be

detected. We used, the standard Nalimov 95% confidence level.

Once we identified and discarded the outliers, we scaled the data on each chip and between chips. In order to achieve this, for each chip, we calculated the median signal intensity over all probe sets. The median of this median signal intensity from all chips was calculated. Than, every chip, was scaled to this median value.

The next step was to transform the data using a logarithmic transformation in base 2. The reason we do this is that is preferable to work with logged intensities rather than absolute intensities since the variation of logged intensities tends to be less dependent on the magnitude of the values; taking log reduces the skewness of the distributions, comparing with a Gaussian distribution, and improves variance estimation. Sometimes, "thresholding" is used as part of the preprocessing- any data that have an expression level below the chosen threshold is discarded. We were interested in all the data, since we consider that a low level of expression at a certain time point can be significant for what we were looking for. We therefore didn't consider using any threshold level on our data. We've 'estimated' the data based on the values of the K nearest neighbor genes estimator (see Troyanskaya, T., Tibshirani, R., Botstein, D. & Altman, R. B. (2001)). In this sense I've chosen KNN=12, meaning the range of neighbors for the estimation process is 12. I average these 12 values and replace the smallest value I would like to 'estimate' with that average. This method is considered a better estimator than discarding data through the 'threshold' method.

We also filtered the data. We consider that genes that vary in expression the most over the adult life identify the most regulated genes and therefore tell us more about aging expression of this organisms. Genes were filtered on the basis of their variation across the time samples. We choose a set of 2000 genes that exhibit the greatest variation. In presenting the results of our unsupervised as well as supervised method, we will refer to the list of 2000 genes determined based on highest variation filtering.

The next step for preprocessing the data was normalization. For normalize the data, we performed two steps: first, we centered to the median. We subtracted

from each component of the initial vector the median value between the components of that vector, to obtain a new vector. We than normalized the newly obtained vector by dividing each component of the newly obtained vector to its norm i.e.square root of the sum of the squares of the components. By normalizing all genes (or raw in the matrix) we get to the stage of being able to compare the genes with 7 samples between each other and as consequence apply a classification method. Given that we don't know exactly what we expect to find in the data, an unsupervised method is the right method to choose. We choose to cluster the data.

Before clustering the data, we wanted to identify the main direction of variations of the genes, and to get a better understanding of the structure of the data we wanted to cluster. In order to do this, we performed a Principal component analysis (PCA). Using this model one can identify the most important gradient of variation in the data points, identifies the first and second eigenvalues, than rotates the data points such that the maximum variability becomes visible, i.e. by plotting the data on the corresponding first and second eigenvectors.

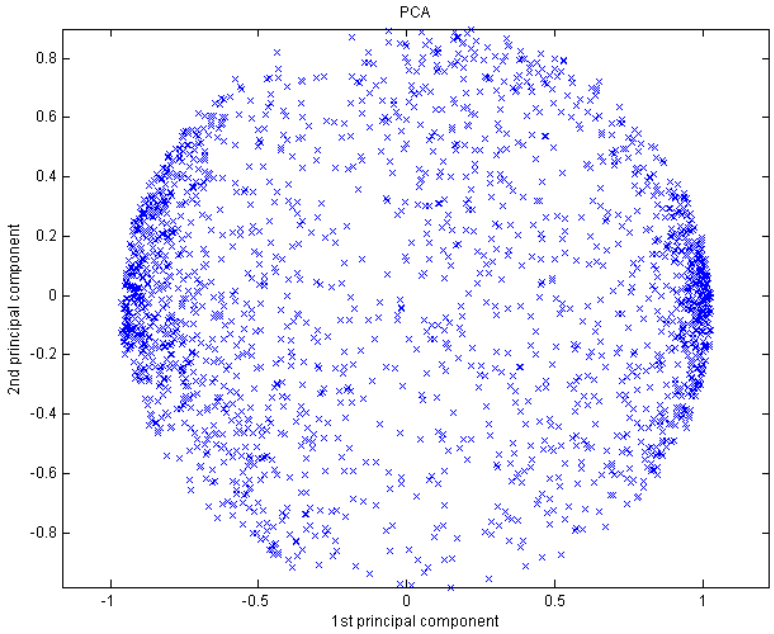


Fig. 2 Principal component analysis performed on the list of 2000 genes.

The Principal component analysis provides a first clue of the potential structure existing in the data. For a better understanding of the structure and possible classes existing in the data we will use a clustering algorithm as an unsupervised method. If we would have had already a first classification on the data traditionally we would have used a supervised method. Given that we had no prior classifications on the data, an unsupervised method for classifying the data is required. A supervised method is used when you already have some kind of knowledge on your data, as for example a data classification and you are using that knowledge to learn more from the data, by contrast with unsupervised methods which you are using when you know nothing about the data and you learn first hand from the data. Given that this is our case, we cluster our data.

The clustering method we chose is based on the physical properties of a magnetic system and enables identification of clusters that have not been obtained by other unsupervised clustering methods as Tree- View which are based on Pearson coefficient. This method has a number of unique advantages:

1 Number of the "macroscopic" clusters is an output of the algorithm

2 Hierarchical organization of the data is reflected in the way the clusters split or merge when a control parameter is varied.

3 Being a Monte Carlo based method, the results are insensitive to the initial conditions.

Comparing this SPC algorithm with other clustering algorithms, the draw back of any other methods (such as Tree View) is the high sensitivity to initialization, poor performance when the data contains overlapping clusters; and the most serious problem: lack of cluster validity criteria. None of these methods provide an index that could be used to identify the most significant partitions among those obtained in entire hierarchy. At the same time, we did not want to use a clustering method based on K means algorithm, since this method is known for highly overlapping cluster results that do not necessarily correspond to the biological process. The fact that K means method can place the same gene in two different clusters does not necessarily indicate that one gene can be part of several biological process, but instead reflects incapacity of this algorithm to deal with

simply overlapping data (see references for K means). SPC eliminates several of this concerns (see Domany et.al, (1998)).

Using the SPC algorithm- a Monte Carlo based method, stability, is an attribute of the clusters. The Swendsen-Wang Monte Carlo method has been used, due to its known ability to speed up the algorithm and make it faster. As we look "more deeply" into the data (by increasing a control parameter), and unveil the hierarchical structure of the data, we performed 2500 cycles, with cycle corresponding to a one step increase in control parameter. The number of cycles in which a cluster remains intact, before it is split in other clusters, is called stability. We consider that clusters with higher stability to be more meaningful for biological interpretation of the data (see, M. Blatt, S. Wiseman and E, Domany, Neural Computation 9, 1805-1842 (1997) for more on algorithm).

4.5 Clustering results and interpretation

We clustered the list of 2000 genes using SPC approach to identify 34 clusters, classified, based on size (number of genes in each cluster) and stability.

The hierarchical organization of the data has a graphical representation as a tree, called a dendogram. See fig1 (dendogram_main patterns) with main patterns observed in the data highlighted. The hierarchical organization of the data is reflected in the way clusters split or merge. First, the entire data set of 2000 genes is considered to be part of one giant cluster. As we vary the control parameter the giant cluster will split into multiple small clusters. The clusters or nodes we obtained were annotated as G1-G34, each with a distinctive pattern. The constraints we choose for the clustering algorithm were: minimal cluster 10 (any cluster with less than 10 genes will not be accepted) and stability 3 (meaning, any cluster with stability less than 3 will not be accepted, or, in other words, any cluster which breaks down sooner than 3 cycles).

We performed 2500 cycles on the list of 2000 genes, and we used KNN = 12 (KNN-are nearest neighbors – see M. Blatt, S. Wiseman and E, Domany, Neural Computation (1997), for more on the algorithm, also G. Getz, E. Levine & E.

Domany, Department of Physics of Complex Systems, Weizmann Inst., Rehovot, Israel, 2001).

Besides classifications of clusters based on the size and stability criterion mentioned above (found in size/stability table), we attempted a classification based on patterns of gene expression identified in each such cluster. The results of this clustering analysis were compiled for an easy access in a web- based design that facilitates their analysis.

Table 2 below is a sample of the main web page. The rest of the Tables and figures which are web based design can be found in the Supplemental data. Main clusters are displayed each with their respective size, stability and pattern in gene expression. A color coded dendogram based on cluster patterns found is presented in fig1.

Green are all clusters with an up-regulated pattern, yellow, with a down-regulated and pink with a day 10 up-ward peak. Smaller sub-patterns are the orange coded with a down-peak day 10, a grey color for pattern of high peak at day 6 and a lila color senescence pattern,

Red	high stability;
Yellow	down regulated pattern, cluster break/split from G18
Green	up regulated pattern, cluster break/split from G22
pink	up regulated day 10 pattern

Table 2 Color codes corresponding to cluster patterns depicted in the dendogram from Fig. 3

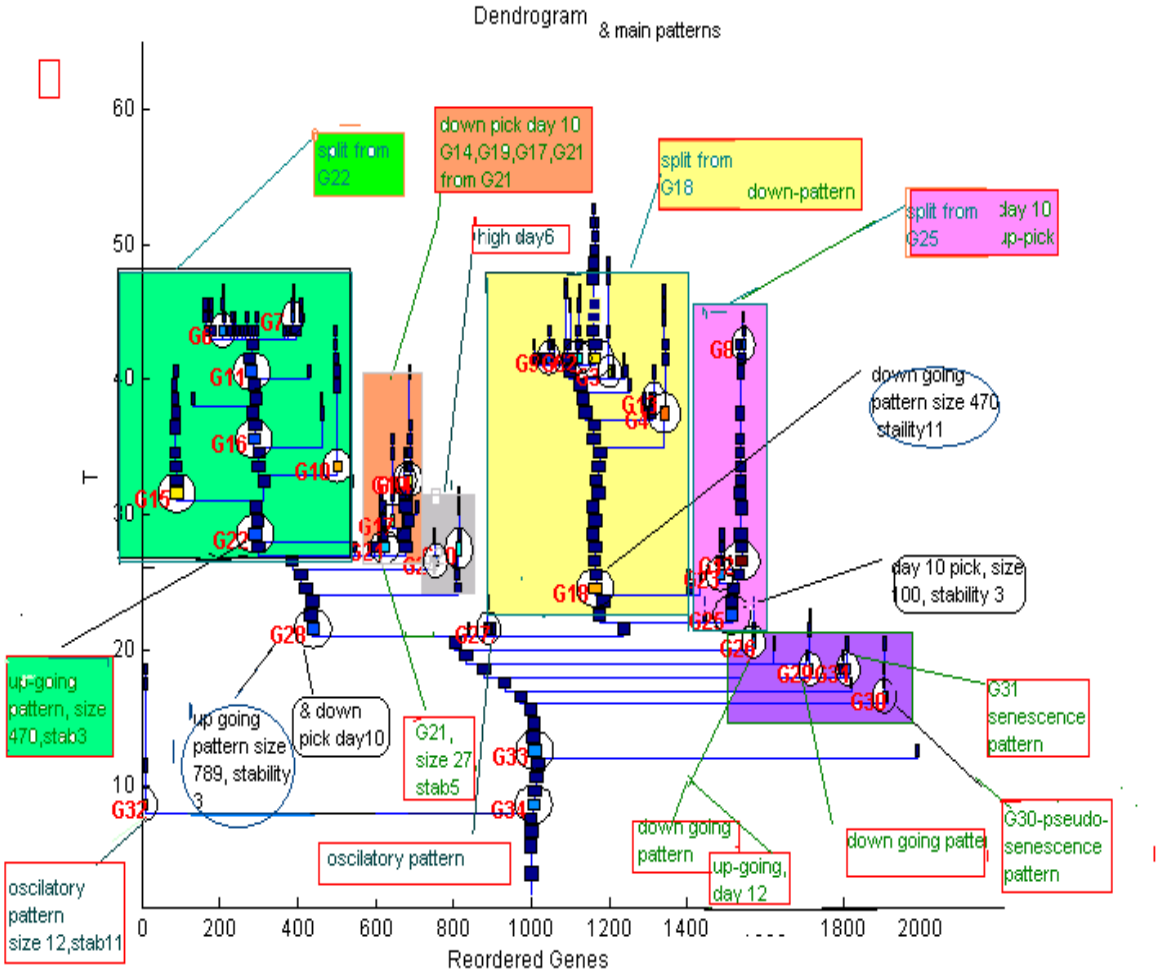


Fig. 3 dendogram with cluster patterns

-TABLE 2-exerpt from the main web page with cluster results from using SPC algorithm:

Clusters of Genes: Clusters are annotated from G1-G34; G2 Stability=10 Size=63 down regulated pattern (see G18 on the dendogram), split from G18
G3 Stability=9 Size=13 oscillatory down regulated pattern (see G18 on the dendogram); split from G18; 5 out of 13 collagen
G4 Stability=12 Size=24 7 out of 24 collagen; low peak day 6, stay low; split from G18; (see G18 on the dendogram)
G5 Stability=6 Size=28 oscillatory down regulated pattern (see G18 on the dendogram); it split from G18,)
G6 Stability=4 Size=20 down peak day 10 in upward overall pattern (see G6 on dendogram); it split from G22
G7 Stability=3 Size=11 down regulated peak day 10, in upward overall pattern (see G7 on dendo); it split from G22
G8 Stability=3 Size=11 upward regulated peak day10; split from G25
G9 Stability=3 Size=12 downward pattern; (split from G18 see dendogram))
G10 Stability=11 Size=26 high peak day6, left over the rest of *C. elegans* development split from G22
G11 Stability=3 Size=284 upward regulated pattern; major size node ; split from G22
G12 Stability=16 Size=51 upward regulated peak day10; splits from G25

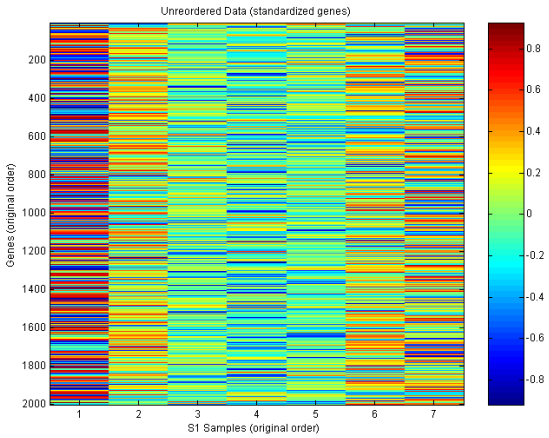
G13 Stability=3 Size=11 down regulated peak day10 sub-pattern in down regulated general pattern (split from G18 see dendogram)
G14 Stability=9 Size=15 down regulated peak day 10 split from G28
G15 Stability=10 Size=40 up regulated pattern; split from G22
G16 Stability=3 Size=340 up regulated pattern; major size node split from G22
G17 Stability=7 Size=12 down regulated peak day 10 split from G28
G18 Stability=11 Size=470 down pattern; major size node from which merge:G2,G3,G5,G9; G4 (dendogram:G18) collagen cluster 67 members are collagen related.
G19 Stability=3 Size=11 down peak day 10 split from G28
G20 Stability=6 Size=12 high- peak day 6, down-going main pattern, down-peak day12
G21 Stability=5 Size=27 down peak day10; from G21 splits G14,G17,G19. G21,splits from G28
G22 Stability=3 Size=470 up-regulated pattern & low peak day 10; high peak day6 major size node from which merge: G6, G7,G10 , G11, G15,G16
G23 Stability=4 Size=27 up-regulated peak day10; splits from G25
G24 Stability=3 Size=14 high day 6, decreasing pattern rest of life
G25 Stability=3 Size=100 up-regulated peak day10; splits from G25
G26 Stability=4 Size=11 oscillatory down pattern, senescence pattern-increase day 12-day15

G27 Stability=3 Size=12	up-peak pattern day 10
G28 Stability=3 Size=789	upward pattern major size node
G29Stability=5 Size=11	down going pattern -senescence as pattern; high expression day 12-day15
G30 Stability=5 Size=10	down peak day 9, pseudo- senescence pattern
G31 Stability=3 Size=12	senescence pattern
G32 Stability=11 Size=12	oscillatory pattern
G33 Stability=4 Size=1941	
G34 Stability=4 Size=1978	

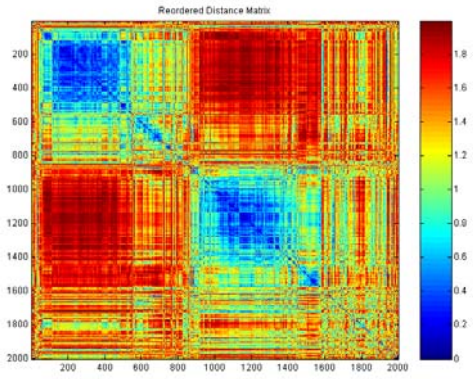
TABLE 2 extract from main web page, SPC results.

The entire informational content of the web based clustering design is displayed graphically or in tables. Links from the main web page can be found for:

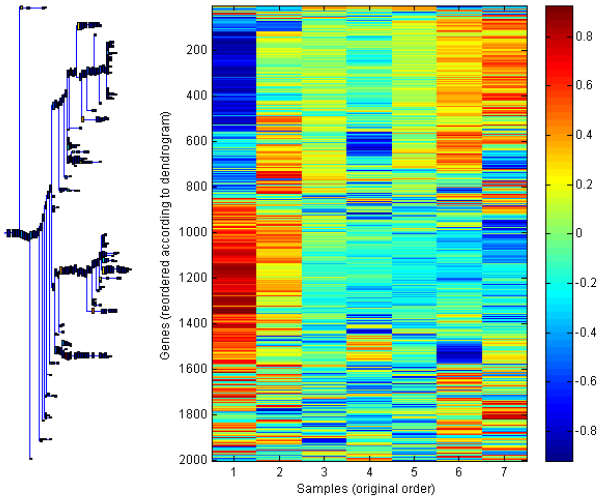
- Unreordered Data (standardized genes) : a hit-map graph with all the genes normalized before being clustered.



- PCA –a graph displaying the principal component analysis
- KNM –graph with K mutual nearest neighbors; helped in building 'distance matrix'.
- Reordered distance matrix graph-based on which clusters have been identified.



- Dendrogram with Stable Clusters –web based accessible dendogram
- Reordered Data : the entire data list of 2000, reordered after clustering, hit-map graph
- Dendrogram next to Reordered Data : hit-map graph & dendogram



- Reordered Genes : table with all 2000 genes and the clusters where they fit.
- Samples : time points
- Parameters for SPC

The access to the above is through a link from the main web page (from G1(S1)). Each cluster can be accessed from the main web page and is represented graphically in two plot formats: as a heat map and as gene expression level changes over time. In addition, a short description of the biological content, of each cluster, correspondence of the cluster with any other clusters, and the list of gene members found in the respective cluster is included. Two tables with clusters sorted based on the stability and size are also presented (see Tables: clusters of genes sorted according to stability and respectively clusters of genes sorted according to size).

Besides size and stability criterion, the genes in a cluster can be hypothesized to have a functional relationship. An examination of the identity of genes in a cluster can allow the potential nature of this relationship to be addressed. We identified sets of genes that could be grouped by some functional criterion. For example, the genes in some groups shared a protein motif or enzymatic function. In others, the genes were shown to have similar expression patterns or regulation.

We assume in this experiment we are analyzing the gene expression of the wild type *C. elegans* and that the *spe-9* mutation we used for age synchronicity, has little or no influence on the aging phenotype. The *spe-9* gene is required for fertility in *Caenorhabditis elegans* and encodes a sperm transmembrane protein with an extracellular domain (ECD) that contains 10 epidermal growth factor (EGF) repeats. Evidence suggests that (see Singson A. et.al, Dev. Biol.2004) EGF repeats can be mutated to create animals with temperature-sensitive (ts) fertility phenotypes.

4.5.1 Clustering interpretation

1) The general up-regulated pattern: Cluster G28:
heat-shock genes, insulin like ligands, male- specific genes, genes involved in life expectancy, linkages genes (genes involved in cross-talk).

G28 features genes that increase in expression levels in adult life; has the size of 789 genes and stability 3. This cluster breaks quickly (after 3 cycles) the pattern of up-regulation is maintained in the G22 cluster of size 478 and again,

stability 3, and all clusters that merge from it : G15,G16,G11,G6,G7. In the Table 2, as well as in the dendogram (fig1 above & web based dendogram), these clusters are highlighted in green.

We analyzed the main G28 cluster, the large cluster that groups genes that can be broadly considered to be increasing in expression level as the animals age. An Table with cluster G28 members can be found at Supplemental data. The G28 cluster includes heat shock proteins, insulin-like ligands, and male-specific genes.

Also in this cluster are genes that were found to be expressed at a higher level in long-lived *C. elegans* mutants as compared with wild-type and short-lived mutants. To determine this, we used the list of short/extended life genes from Murphy et. al. 2003, Nature 424 of ~ 200 genes. The fact that we see an increase in the expression of the group of genes that are suggested to have a role in the shortening life expectancy of this organism suggests an increase in the relative activity of the products of these genes with obvious consequence of shortening the life of the nematode. On the other hand the increase in the expression of the group of genes involved in long-lived mutants might suggest that beneficial stress resistance genes turn on to protect against aging. Modulating longevity or shortivity genes can impact lifespan

In the case of heat shock genes, most are found in a single sub-cluster, G11, of size 284 and stability 3 whereas, the other functional groups, are broadly distributed in several different sub-clusters. A Table with cluster G11 members can be found at Supplemental data.

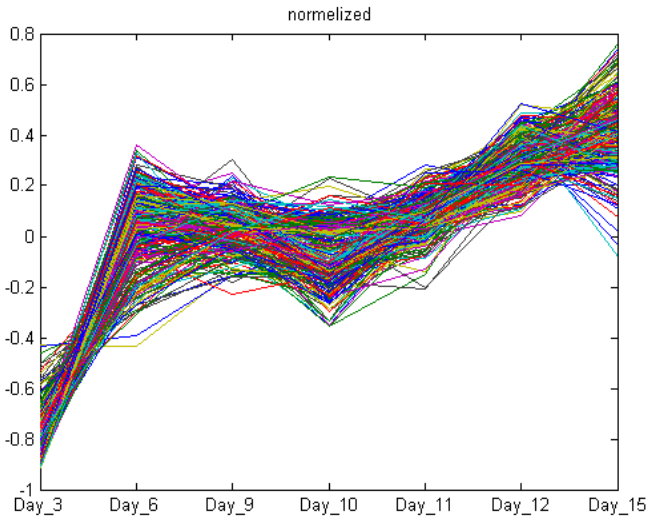


fig4:G11 cluster mostly heat shock genes

An increase in the expression of the heat-shock genes might be considered to support the cumulative damage aging theory. The heat shock response gene is a highly conserved biological response, occurring in all organisms. In response to elevated temperature, proteins misfold. As the organism ages, the damaged proteins that misfold accumulates in the organism. Indeed, this might be reflected by the increase in the expression of heat shock genes we are noticing.

Error-repair mechanism in *C. elegans* doesn't need to be induced with age:

We should mention here that heat shock genes are not responsible for example for error repair from DNA replication or transcription. Given that we don't notice genes responsible for DNA replication/ transcription repair in the up-regulated cluster G11 we might infer that for the wild type *C.elegans,* this mechanism is not need to be induced in aging animals. We further infer that damage accumulation that might induce aging in the nematode is not due to internal error accumulation like DNA error and transcription accumulation but rather is induced by external stimulus which creates damage that accumulates in organism. The error repair mechanism might work properly and therefore wouldn't require any surplus in

gene expression activity. Alternatively, a decrease in expression over time might be required for a proper function of the wild type nematode.

An increase in expression for male-specific genes might be explained by the research done in Andy Singson lab. Interestingly the Singson lab has shown that male mating desire changes with age. One possibility is that the increase in expression of male specific genes might be a signature of age.

An increase in the activity of the genes involved in "cross-talk" is noticed in the G15 cluster-this might tell us that with age the communication between cells and in the cells is diminished. Also, from a stability criterion point of view, cluster G15, is interesting since has a high stability of value 10. This cluster has 40 gene members, see table 3 bellow:

1	B0228-5_at	q09433 caenorhabditis elegans. probable thioredoxin. 11/1995
2	B0284-2_at	best hit: p25386 saccharomyces cerevisiae (baker`s yeast). intracellular protein transport protein uso1. 7/1998 7.0e-10 22%
3	C08E3-4_at	o17194 c08e3.4 protein. 5/2000
4	C17H1-5_at	best hit: o40947 orf 73. 6/2000 2.0e-09 25%
5	C17H1-6_at	o45257 c17h1.6 protein. 1/1999
6	C31B8-4_at	"best hit: q13439 trans-golgi p230 (256 kda golgin) (golgin-245) (72.1 protein) (golgi autoantigen, golgin subfamily a, 4). 5/2000 4.0e-12 25%"
7	C34D1-3_at	ce08571 locus:odr-3 guanine nucleotide-binding protein (cambridge) tr:q18434 protein_id:cab01489.1. 0/0
8	C34E11-3_at	"best hit: p10587 gallus gallus (chicken). myosin heavy chain, gizzard smooth muscle. 12/1998 2.0e-42 22%"
9	C53A5-9_at	ce08958 ring canal protein like (cambridge) tr:o17700 protein_id:cab03989.1. 0/0
10	C53D6-6_at	best hit: q17894 similar to hobo element transposase hfl1. 11/1998 3.0e-27 23%
11	EGAP9-2_at	p91200 cosmid egap9. 5/2000
12	F09F9-3_at	q19283 cosmid f09f9. 11/1998
13	F11A1-1_at	q19331 f11a1.1 protein. 1/1999
14	F13H8-8_at	q19432 cosmid f13h8. 11/1998
15	F15A4-9_at	best hit: o01749 similar to human dihydroxyvitamin d3-induced

		protein. 11/1998 4.0e-17 24%
16	F26D11-6_at	o61963 f26d11.6 protein. 11/1998
17	F26F2-3_at	q9xv56 f26f2.3 protein. 11/1999
18	F26F2-4_f_at	q9xv55 f26f2.4 protein. 11/1999
19	F26F2-5_i_at	q9xv54 f26f2.5 protein. 11/1999
20	F33D11-8_at	o44779 f33d11.8 protein. 11/1998
21	F36H5-3_at	p91298 cosmid f36h5. 5/2000
22	F43B10-2_at	best hit: p21997 volvox carteri. sulfated surface glycoprotein 185 (ssg 185). 10/1996 3.0e-15 51%
23	F44G3-8_at	ce16039 f-box domain. (cambridge) tr:o62239 protein_id:cab05520.1. 0/0
24	F44G4-6_at	q20416 f44g4.6 protein. 5/2000
25	F53B6-4_at	"best hit: p40631 tetrahymena thermophila. micronuclear linker histone polyprotein (mic lh) [contains: linker histone proteins alpha, beta, delta and gamma]. 12/1998 1.0e-14 38%"
26	F56H6-2_at	o45580 f56h6.2 protein. 5/2000
27	F59C6-2_at	best hit: cab92119 dj50o24.4 (novel protein with dhhc zinc finger domain). 7/2000 5.0e-13 37%
28	F59E11-10_at	ce11512 zinc finger protein (st.louis) tr:o16752 protein_id:aab66229.1. 0/0
29	H27M09-B_at	"best hit: p40631 tetrahymena thermophila. micronuclear linker histone polyprotein (mic lh) [contains: linker histone proteins alpha, beta, delta and gamma]. 12/1998 4.0e-16 31%"
30	K07H8-4_at	ce18024 (st.louis) tr:o45179 protein_id:aac04425.1. 0/0
31	K09D9-12_at	aaf39930 hypothetical protein k09d9.12. 7/2000
32	M162-6_at	ce18896 (cambridge) protein_id:cab05252.1. 0/0
33	R07B1-2_at	q09605 caenorhabditis elegans. probable galaptin lec-7. 12/1998
34	T07D3-1_at	o16727 t07d3.1 protein. 5/2000
35	T22C8-3_at	"ce02351 zinc finger, c2h2 type (cambridge) tr:q22676 protein_id:caa88875.1. 0/0"
36	W01B6-9_at	best hit: o44929 microtubule binding protein d-clip-190. 6/2000 3.0e-12 22%
37	Y102A5C-19_at	best hit: o17578 c06h5.2. 5/2000 8.0e-19 29%

38	Y102A5C-8_g_at	q9xx81 y102a5c.8 protein. 6/2000
39	Y53F4A-2_f_at	cab54462 y53f4a.2 protein. 8/2000
40	ZK632-11_at	p34656 caenorhabditis elegans. hypothetical 51.8 kda protein zk632.11 in chromosome iii. 11/1997

Table 3 G15 cluster members

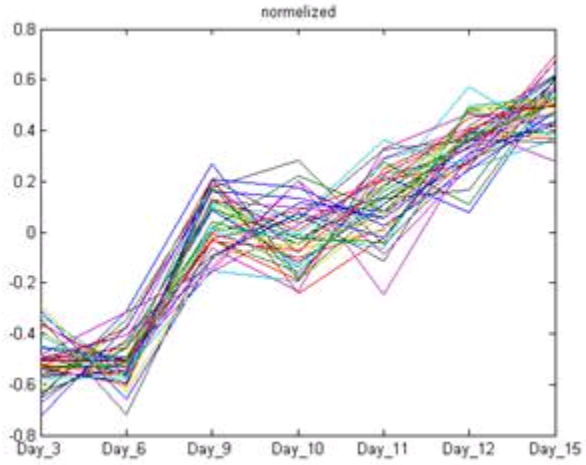


Fig. 5 **G15 cluster** pattern

This G15 cluster has mostly linker and binding type of proteins, intracellular protein transport, canal proteins type, micronuclear linker proteins, linker histones proteins, zink finger proteins, microtubule binder proteins. One reason for this might be that with age the connections and linkages at cellular and intracellular level and the cross-talk in and between cells is weakening therefore an increase in the activity of the products of the genes involved in such processes might be required.

2) A down-regulated pattern is noticed in **G18** (**collagen cluster**), **has mostly collagen and muscle related genes.** Size of this cluster G18 is of 470, stability 11. The down regulated pattern is maintained in all clusters that merge from

it:G2,G3,G4,G5,G9,G13. In TABLE 2, as well on the dendogram these genes are highlighted on yellow. It is worth mentioning that the G18 cluster has a high stability value 11 in spite of it's size.
The cluster G18 groups genes that can broadly be considered to be decreasing in expression as the animals age. Cuticle collagens are very strongly overrepresented in this cluster—almost one third of the 200+ collagen genes in the worm are found in this cluster.

Kim, et.al (2002 Nature 418: 975-979), identified a large group of genes that are expressed at a relatively high level in muscle tissue; these genes are also in the G18 cluster. Many of the muscle-enriched genes are collagens, but if we exclude collagens from the analysis of muscle genes, we still see most of muscle genes in cluster G18. We also examined 60 genes that are known to have function in muscle (e.g. muscle myosin and other muscle motor proteins), and most of them are also in this G18 cluster. Many of the genes that have human homologs in the total muscle enriched dataset of David Miller(Genome Biology 2007, vol.8, issue 9) are found in this cluster as well. We've used this preliminary results genes for a more careful analysis we've performed later when we've analyzed sarcopenia process in *C. elegans* (see Chapter 2)

We find two distinct gene expression patterns among genes involved in cellular damage protection: heat-shock genes vs. oxidative stress genes, indicating 2 distinctive gene classes among genes with roles in damage protection.

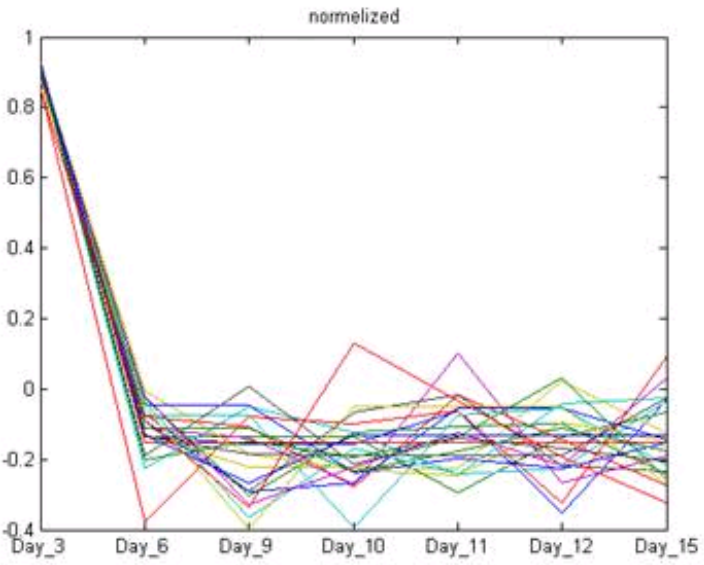
We find that peroxidases, cytochrome p450s and glutathione S-transferases are more predominant in cluster G18 with a decrease in gene expression pattern; these types of proteins have a variety of biological functions, but all are commonly involved in detoxification and protection from oxidative stress.

Cellular damage can be induced by heat shock, oxidative stress and various toxic substances. Several pathways are involved in cellular damage. The cumulative

theory of aging implies that over time such damage accumulates in organism. Such accumulation might be reflected at the transcription level in an increase in gene expression due to an increase in the activity required from the genes with a protective role such as heat shock related genes, and observed in the pattern of G11 cluster.

The finding that genes involved in oxidative stress and detoxification show a decrease in expression pattern opposed to the pattern noticed for heat-shock genes from the cluster G11, might suggest existence of 2 distinct classes among genes involved in cellular damage protection. Misfolding proteins might be associated with an increase in gene expression, as the heat shock genes in cluster G11 have, whereas the oxidative stress theory of aging might have surprisingly an opposite signature of decreasing in gene expression level as can be depicted in the cluster G18. Possible hypothesis might be that oxidative stress gene activity in the nematode doesn't have to increase with age given that the oxidative damage from some reason doesn't accumulate with age in *C. elegans*. It would be interestingly to experimentally see the correspondence between, the type of damage these genes are involved i.e type of toxic agent and the pattern and class category enters.

The high stability value of G18 cluster is maintained in G2,G3 and G4. The G3 cluster contains 5 -collagen related genes out of 13 gene (it's size value), and has an oscillatory down -going pattern. The G4 cluster has an interesting pattern of high expression day3, and low expression rest of the times. It has 24 members genes of which 7 are collagen- related genes (see Fig. 6)



x-axis: time points; y-axis: normalized gene expressions

Fig.6: G4 cluster pattern

The G2 cluster has 36 collagen- related genes out of 61 it's total size and has a similar expression pattern as G3.

G2 and G3 clusters show the strongest decline.(see Fig. 7)

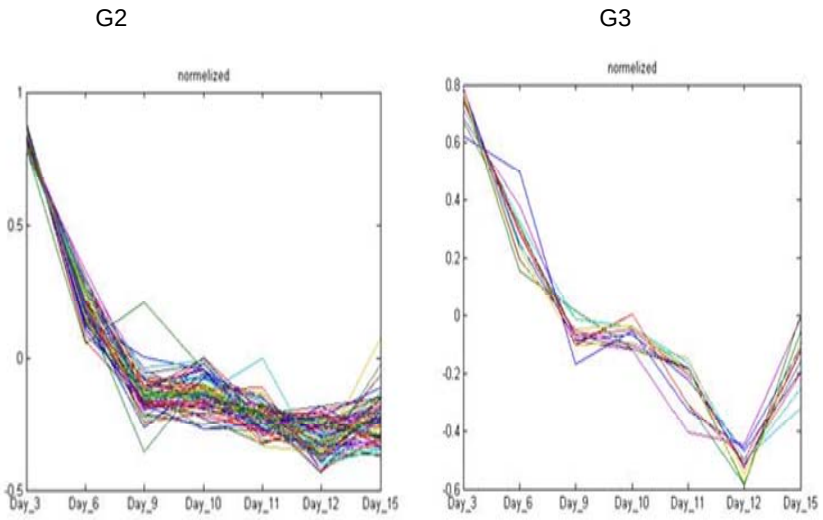


Fig.7: G2 and G3 cluster patterns; x-axis: time points; y-axis: normalized gene expressions

3) The day 10 pattern can be seen as an oscillatory pattern over all time points with the particularity that for day 10, the pattern of the expression for down or up peak is more pronounced.

3.a) An oscillatory pattern with an up-peak pattern at day 10 is observed in G25: signaling and transcription factors genes might have a common regulatory loop with germ line genes.

G25 cluster has size 100, and stability 3, and all the subsequent clusters merge from G25 as: G23, G12, G8. The G25 cluster includes many signaling and transcription factors genes. This group is defined by an inferred role in regulation, e.g. kinases, receptors, G proteins, and the like.

We examined the 23 signaling and transcription factor genes in G25, and found that 7 of them had well-characterized functions in the germline or in early embryonic development. The G23 cluster has the most prominent day 10 change pattern.

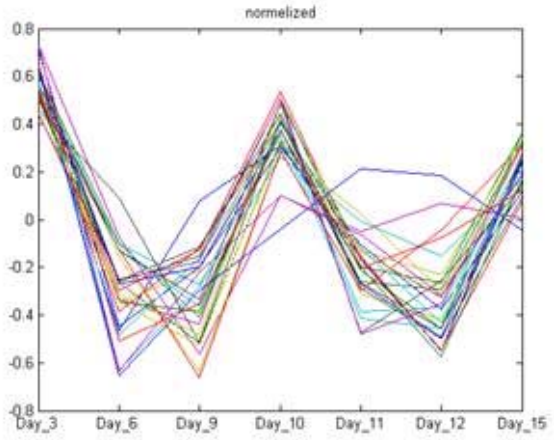


Fig. 8 **G23 cluster oscillatory pattern with a day 10 peak** pattern; **x-axis: time points; y-axis: normalized gene expressions**

The fact that an oscillatory pattern describes signaling and transcription factors genes is expected. What was a surprise for us was that this oscillatory pattern has high expression peak for day 10. When we later examined a group of germ line enriched genes we find that they have same gene expression pattern. In case that this oscillatory day10 up regulated pattern might indeed be a germ line enriched gene signature than all the rest of genes in this G25 cluster as signaling and transcription factors genes might have a common regulatory loop with the germ line genes. See also the discussion on the germ line genes .

3.b) an oscillatory pattern with a pronounced down regulation pattern for day 10.

This pattern can be seen first in G21, and then the sub-clusters G14,G17,G19. This down peak pattern at day 10 can also be seen in the clusters with a general, up-going trend as pattern, as in clusters G6, and G7, which merged from G22 (the main

up-going pattern); as well as in the clusters with down-going general pattern as can be seen in the G13 cluster, which merge from G18.

All the clusters with the down peak pattern at day 10 can be seen in Table2 as well as in dendograms fig 3 (highlighted in light pink). G14 is the cluster with highest stability among cluster with down going day 10 pattern (size 15 and stability 9). G14 contains the expression of sri-29 chemoreceptor, gcy-21 which is a protein kinase, fmo-1 which has disulfide oxidoreductase activity and monooxygenase activity, also genes with proteolysis function. See Table 4 bellow for cluster members also fig. 9 for cluster pattern.

1	B0454-4_at sri-29	o17170 b0454.4 protein. 6/2000 chemoreceptor, sri family - (*Serpentine Receptor, class I*)
2	F08E10-2_r_at	q9xxp2 f08e10.2 protein. 5/2000
3	F22A3-4_at	ce04440 contains similarity with human homeotic protein pbx2 (st.louis) tr:q19696 protein_id:aaa83195.1. 0/0
4	F22E5-3_g_at or gcy-21	ce09555 locus:gcy-21 protein kinase (st.louis) tr:o16715 protein_id:aab66169.1. 0/0
5	F36H12-16_at	o76718 f36h12.16 protein. 11/1998;contains similarity to Lactobacillus delbrueckii Abc transporter ATP-binding protein
6	F53G2-1_at	best hit: q23181 similarity to c.elegans early embryogenesis zyg-11 protein. 5/1999 4.0e-66 29%
7	K08C7-2_at or fmo-1	ce21038 dimethylaniline monooxygenase (cambridge) tr:q21311 protein_id:caa94291.1. 0/0; dimethylaniline monooxygenase (N-oxide-forming) activity disulfide oxidoreductase activity; monooxygenase activity

8	T06A1-5_at	best hit: q21003 similarity to a putative single-stranded nucleic acid binding protein. 11/1998 5.0e-47 29% contains similarity to Pfam domain PF01697 (Domain of unknown function) molecular_function unknown
9	T23B12-5_at	similarities with: p70561 fgf receptor activating protein frag1. 8/1998 3.0e-09 27%, is a gene from Ratus norvegicus; FRAG1, a gene that activates fibroblast growth factor receptor by C-terminal fusion through chromosomal rearrangement.';
10	W07B8-1_at	ce14674 thiol protease (st.louis) tr:o16289 protein_id:aab65343.1. 0/0 proteolysis and peptidolysis
11	Y47D7A-F_at	aaf60634 hypothetical protein y47d7a.f. 7/2000
12	Y48E1B-6_at	o18200 y48e1b.6 protein. 1/1999
13	Y54G9A-1_at	q9xwh1 y54g9a.1 protein. 11/1999
14	Y54G9A-2_at	q9xwh2 y54g9a.2 protein. 11/1999 contains similarity to Giardia lamblia Median body protein.;
15	ZK250-5_at	o17299 zk250.5 protein. 5/2000

Table 4- G14 cluster member. -Some of the genes have direct links to worm base. The yellow highlighted gene has human similarities

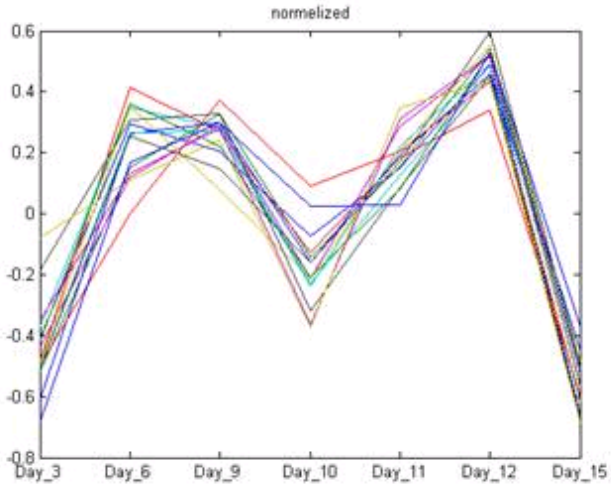


Fig. 9 pattern of cluster G14:
x-axis: time points; y-axis: normalized gene expressions

As mentioned these genes are involved in signaling or are transcription factors.

Our data reveal a dramatic change in gene expression around day 10 in at least 13 clusters.

The consistency of the day10 pattern suggest a significant physiological transition in the *C. elegans* organism during the middle life span time window day 9-day12 of the nematode. Given that some clusters have an oscillatory up-regulated day 10 pattern and others have a down-regulated day 10 pattern might suggest existence of an complementory process among signaling and transcription factors genes involved in the up-regulated day 10 pattern versus the genes involved in the down-regulated pattern.

The reason we were able to narrow down this middle life time window changes focused arround day 10 is due to the statistical analysis performed, as well as the design of the experiment. This day 10 pattern persists even when we repeated the experiment We note that day 10 is measured from day0, the moment of hatching, that the worms were grown at 25C in this study.

Sub-clusters patterns:

- **Senescence pattern, neuronal related genes:**

Besides this main patterns defined by large clusters we will mention two other patterns described by smaller sized clusters and found in the list of 2000 adult regulated genes obtained after filtering.

One such pattern is a "senescence" pattern. The cluster with such a pattern is G31. This pattern can be characterized as a relative low, constant level of expression that spans the life of the *C. elegans* from day 3, toward the end of life of the nematode at day 12 with a drastic increase in expression level between day 12 and day 15. This last day is when most of the nematodes grown at the temperature of 25C have already died and the viable animals we assigned are all decrepit. Day 12 is the time when the decay of *C. elegans* as an organism is easily noticeable.

The cluster G31 includes heat shock genes, a homolog of human fetal brain protein, olfactory receptor, stress-inducible protein and sodium neurotransmitter see Table 5 below with cluster G31 members:

1	C01G6-9_at	q17575 c01g6.9 protein. 1/1999
2	C12D8-4_at	ce05268 transthyretin-like family (cambridge) tr:q17937 protein_id:caa98234.1. 0/0
3	C49A9-5_at	o44151 c49a9.5 protein. 11/1998
4	EGAP1-1_at	q19073 cosmid egap1. 11/1998
5	F08H9-4_at	ce09268 heat shock protein (cambridge) tr:q19228 protein_id:cab01147.1. 0/0
6	F25H5-7_g_at	ce15903 protein-tyrosine phosphatase (cambridge) tr:o17840 protein_id:cab02988.1. 0/0

7	F59B2-9_at	ce00236 f-box domain. (cambridge) sw:p34484 protein_id:caa77586.1. 0/0
8	T07D4-2_at	best hit: q15777 homo sapiens (human). fetal brain protein 239 (239fb). 5/2000 4.0e-50 41%
9	Y37A1C-1B_r_at	q9xtc4 y37a1c.1b protein. 11/1999
10	Y61B8A-1_at	best hit: p91118 similarity in c. elegans olfactory receptor odr-10. 6/2000 4.0e-21 31%
11	ZK1010-9_at	ce23490 sodium:neurotransmitter symporter (cambridge) tr:o18288 protein_id:cab04975.1. 0/0
12	ZK328-7_at	ce05072 stress-inducible protein sti1 (st.louis) tr:q23468 protein_id:aaa91253.1. 0/0

Table 5 G31 cluster members

This senescence pattern (See fig. 10 cluster G31 pattern) was also observed when we made the analysis of the previous data with just 3 replicates per each time point.

Note that G31 cluster has a majority of neuronal genes in the form of neurotransmitters, human homologies as fetal brain or olfactory genes. Also comparing with the sub-pattern of 'young adult-day 6' note that the neurotransmitter involved in the senescence pattern is Na-related by comparison with the day 6 young-adult cluster which has K-related neurotransmitter. See the gene members in G10 cluster.

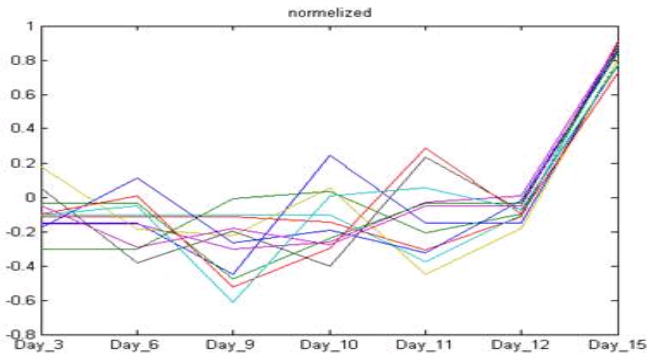


fig. 10 cluster G31

x-axis: time points; y-axis: normalized gene expressions

Both “day 10 pattern” as well as the “senescence pattern” are patterns that were never noticed in previous experiments performed by other labs. We consider that this is due in part of the design of our experiment, in particular, the way we chose the time point sampling, and in part, of the statistical analysis performed, in particular the clustering algorithm used.

- **‘Young adult pattern, - day6 pattern’-has genes specific for larval development and adult morphogenesis and again neuronal genes.**

The K-related channel is member of the young adult cluster by difference with Na-channel related which can be found in ‘senescence pattern’ cluster.

We found in G10 an opposite pattern from the senescence pattern as day 6 pattern

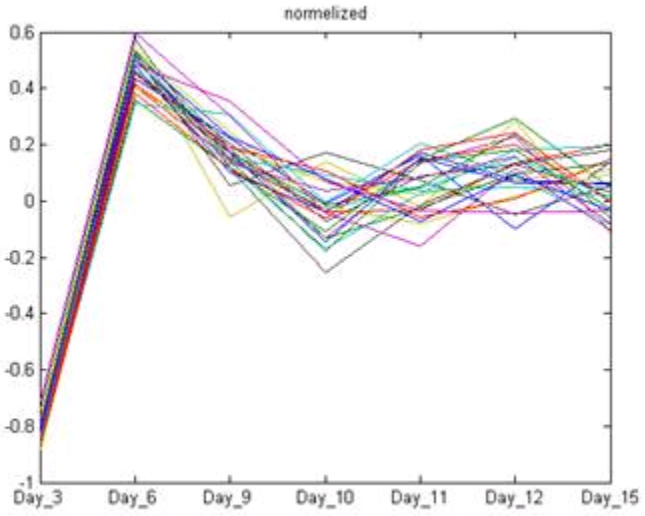


fig. 11 cluster G10; x-axis: time points; y-axis: normalized gene expressions

The G10 cluster contains kqt-1 a human homolog, which encodes one of three *C. elegans* KCNQ-like potassium channel subunits that, with respect to humans, is most similar to the KCNQ2-5 subfamily of channel proteins; genes similar to human

necrosis factor-alpha-induced protein with voltage gated potassium channel activity.

It might be important to note that 'young adult day-6' cluster contains K-channel related genes by difference with 'senescence' cluster which has Na-channel related gene. Also both clusters G31 and G10 have a gene expression similar with protein tyrosine phosphatase.

Otherwise, G10 also includes genes specific for larval development and morphogenesis as insulin like families that affect dauer formation and eating behavior like *ptr-3* gene which is in same family with *daf-6*; *acn-1*, required for larval development and adult morphogenesis; the hypodermal expression of *acn-1* appears to be controlled by *nhr-23* and *nhr-25*. Another dauer related gene is *crb-1* which affects dauer formation and eating behavior. Below is Table 6 with members of cluster G10.

1	C16C8-9_at		p91047 cosmid c16c8. 11/1998
2	C18A3-8_at		ce01800 helix-loop-helix transcription factor (st.louis) tr:q09961 protein_id:aaa68375.1. 0/0
3	C25B8-1_at or kqt-1		ce08386 locus:klq-1 voltage-gated potassium channel (st.louis). 0/0 The kqt-1 gene encodes one of three C. elegans KCNQ-like potassium channel subunits that, with respect to humans, is most similar to the KCNQ2-5 subfamily of channel proteins;
4	C33E10-1_at		best hit: q18385 similar to protein tyrosine phosphatase. 6/2000 1.0e-11 25% encodes an protein containing an F box (motif considered to mediate protein/protein interaction)
5	C40A11-3_at		best hit: p91563 similar to human necrosis factor-alpha-induced protein b12. 6/2000 1.0e-34 45% voltage-gated potassium channel activity
6	C41D7-2_at		best hit: p91184 similar to c. elegans protein f44f4.4. 6/2000 3.0e-93 29%

	or ptr-3		from PaTched Related family; in this family is ptr-7 as well, known as daf-6
7	C41H7-5_at		best hit: q21396 similarity to c. elegans proteins c18h2.1 and t28d9.9. 11/1998 1.0e-20 25% also,(as second hit)similarity with gene CBG19613 from C. briggsae which has similarity with SGD:YKL129C from S.cerevisiae which is an class I myosin; One of two class-I myosins; localizes to actin cortical patches; deletion of MYO3 has little affect on growth, but myo3 myo5 double deletion causes severe defects in growth and actin cytoskeletion organization; myosin I
8	C42D8-5_at or acn-1		ce06951 peptidase (st.louis) tr:q18581 protein_id:aaa98719.1. 0/0 acn-1 encodes an ACE-like protein required for larval development and adult morphogenesis, is expressed in hypodermal cells, vulval precursor cells, and ray papillae in the male tail; the hypodermal expression of acn-1 appears to be controlled by nhr-23 and nhr-25. acn-1(RNAi) animals have arrested larval development
9	C56E6-6_at		ce04278 leucine-rich repeats (st.louis) tr:q18902 protein_id:aaa81094.1. 0/0 similarity with H. sapiens Insulin-like growth factor binding protein complex acid labile chain precursor and with S.cerevisiae protein required for START A of cell cycle
10	EEED8-6_at		"best hit: baa91749 cdna flj10682 fis, clone nt2rp3000072. 7/2000 2.0e-57 40%" biological fct:proteolysis and peptidolysis molecular fct: carboxypeptidase A activity
11	F02C12-3_at		ce23626 (cambridge) tr:q19109 protein_id:caa91020.2. 0/0

12	F11C7-4_at or crb-1		ce07053 egf-like repeats (st.louis) tr:q19350 protein_id:aac69012.1. 0/0 proteins with same EGF-like domain are: UNC-52 plays essential roles in muscle structure development and regulation of growth factor-like signaling pathways; other genes which encode for proteins with EGF like domain:eat-20,spe-9,mec-9 and so on; crb-1 encodes a homolog of Drosophila CRUMBS that affects dauer formation and feeding behavior
13	F11D11-3_at		o62153 f11d11.3 protein. 1/1999
14	F11D11-3_g_at		o62153 f11d11.3 protein. 1/1999
15	F23D12-5_at		best hit: o14607 homo sapiens (human). ubiquitously transcribed y chromosome tetratricopeptide repeat protein (ubiquitously transcribed tpr protein on the y chromosome). 7/1999 5.0e-50 28%
16	F26D10-12_at		ce19812 lectin c-type domain (cambridge) protein_id:cab02321.1. 0/0
17	F27E11-1_at		ce09732 nucleoside transporter (st.louis) tr:o16192 protein_id:aab65255.1. 0/0
18	F32H2-6_at		ce09881 fatty acid synthase (n-terminus) (cambridge) tr:p91866 protein_id:cab04239.1. 0/0
19	F55C9-3_at		q9xuy9 f55c9.3 protein. 11/1999
20	F55C9-5_at		q9xuz0 f55c9.5 protein. 5/2000
21	H16D19-4_at		q9xx93 h16d19.4 protein. 11/1999
22	K08B4-2_at		caeelgn; k08b4-2; -. 7/100; similarity with SW:O35598 from M. Musculus, contributes to the normal cleavage of the cellular prion protein.
23	R13H4-7_at		p90946 r13h4.7 protein. 1/1999
24	T02E9-5_at		q9u382 t02e9.5 protein. 5/2000
25	T14B4-6_at dyp-2 or rol-2		p35799 caenorhabditis elegans. cuticle collagen dpy-2 precursor. 11/1997

26	W04A8-4_at		q9xul8 w04a8.4 protein. 11/1999

Table 6: G10 cluster members; the highlighted genes might be representative for the biological theme of the G10 cluster.

Higher gene expression after day 6 for the gene members of 'day-6 young adult' cluster pattern might induce shortening in life span of the nematode.

A high peak at day 6 and than decreasing pattern for the rest of time points, is also maintained in the G24 cluster of size 14 and stability 3. G24 contains *unc-44*, a collagen related gene, and *ces-2*, which is required to activate programmed cell death in the sister cells of the serotoninergic neurosecretory motor (NSM) neurons, and is transcriptionally inhibited by activated LET-60. We can hypothesize that the genes in cluster G10 and G24 are genes important for adult morphogenesis; cell growth and in general genes involved in cell homeostasis.

Considering the cluster pattern of high peak at day 6 and than a general steady gene expression for the rest of the life of this nematode we might consider that gene members of this clusters might act as 'left-on' genes.

Day 6- young adult pattern + senescence pattern

G20 cluster is interesting because it has the high peak pattern at day 6, and the "senescence" pattern, of increase at end stages as well. It contains transcription factors and heat shock protein. See the gene pattern and gene members in fig. 10 and respectively Table 7 below

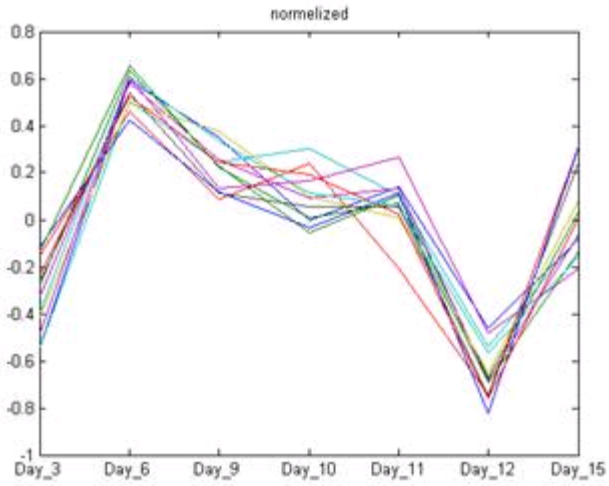


fig.12 G20 cluster; x-axis: time points; y-axis: normalized gene expressions

1	C09G12-5_f_at	o44457 c09g12.5 protein. 11/1998
2	C09G12-5_i_at	o44457 c09g12.5 protein. 11/1998
3	C50F4-3_at	ce05468 thiol protease (cambridge) tr:q18740 protein_id:caa94738.1. 0/0
4	C54F6-8_at **C54F6.8**	ce17603 nuclear hormone receptor (st.louis) tr:o16443 protein_id:aab65942.1. 0/0; transcription factor:; associated with: • **asp-4--**biological fct :induction of non-apoptotic programmed cell death! • **csn-3** -biological fct(b.f.):control development; • **csp-3** –b.f.:destruction of protein or peptides by hydrolysis (proteolysis, peptidolysis) • **clp-3** –involved in neurodegeneration caused by necrotic cell death • **unc-71** - proteolysis and peptidolysis;

5	F02E11-5_at	best hit: o44932 vespid allergen antigen homolog. 6/2000 2.0e-23 35%
6	F35D11-8_at	q20037 cosmid f35d11. 11/1998
7	F35H8-2_at	best hit: q22481 similarity to c. elegans hyupothetical protein. 11/1998 4.0e-12 32%
8	F40G9-7_at	q9tz78 f40g9.7 protein. 5/2000
9	F59D6-1_at	o16344 f59d6.1 protein. 11/1998
10	K03D3-5_at **K03D3.5**	• o45643 k03d3.5 protein. 1/1999 • **K03D3.5-** by blast- best match with a heat shock protein from of B. aphidicola organism
11	T02D1-7_at	o45726 t02d1.7 protein. 1/1999
12	ZK1290-1_at	q23439 cosmid zk1290. 11/1998

Table 7 G20 cluster members

G4-Young adult-day3-day6 pattern:

G4, splits from G18 with a distinct pattern. G4 has a size of 24 expression genes, and stability 12. It starts with a high pattern at day 3, and than a relative constant low expression. G4 has 7 collagen related genes. Because of high expression pattern at day 3 the genes in this cluster might play an important role in the young adult life of this nematode-see Table 8 with gene members of G4 cluster as well as Fig. 11 for cluster pattern.

1	B0024-2_at	ce05147 collagen (cambridge) tr:q17418 protein_id:caa94875.1. 0/0
2	C09G5-6_at	q09457 caenorhabditis elegans. putative cuticle collagen c09g5.6. 11/1997
3	C29E4-1_at	p34340 caenorhabditis elegans. putative cuticle collagen c29e4.1. 11/1997
4	E03H12-2_at	o02128 cosmid e03h12. 11/1998
5	F11E6-2_at	q9u3j9 f11e6.2 protein. 5/2000
6	F12E12-	best hit: q17724 similar to the insect-type alcohol

	C_at	dehydrogenase/ribitol dehydrogenase family. 5/2000 2.0e-73 55%
7	F36A4-11_at	q20088 similarity to collagen. 11/1998
8	F36A4-1_at	q20096 cosmid f36a4. 11/1998
9	F37B1-6_at	ce09993 glutathione s-transferase (cambridge) tr:q93699 protein_id:cab02292.1. 0/0
10	F40A3-6_at	o16266 f40a3.6 protein. 11/1998
11	F41G4-1_at	best hit: q26630 axonemal dynein light chain p33. 11/1998 6.0e-41 44%
12	F42A10-7_at	q20312 cosmid f42a10. 11/1998
13	M18-1_at	ce06193 collagen (cambridge) tr:q21556 protein_id:caa92826.1. 0/0
14	R09A8-4_at	ce03540 cuticle collagen (cambridge) tr:q21855 protein_id:caa92006.1. 0/0
15	T01B10-2_at	o02153 cosmid t01b10. 11/1998
16	T10E10-G_i_at	q22326 similar tocollagen. 6/2000
17	T11F9-8_at	ce06420 zinc metalloprotease (cambridge) tr:q22400 protein_id:caa98532.1. 0/0
18	T20B3-2_at	ce20087 troponin (cambridge) protein_id:cab04737.1. 0/0
19	W09B7-B_at	aaf60391 hypothetical protein w09b7.b. 7/2000
20	Y57A10B-6_at	best hit: p41991 caenorhabditis elegans. pes-10 protein. 11/1995 2.0e-38 28%
21	ZC101-2E_at	q06561 caenorhabditis elegans. basement membrane proteoglycan precursor (perlecan homolog). 7/1999
22	ZC373-6_at	q23262 zc373.6 protein. 1/1999
23	ZK1193-3_at	q23410 similarity over a short region to tenascin precursors. 5/2000
24	ZK1193-4_at	q23412 cosmid zk1193. 11/1998

Table 8 G4 cluster members

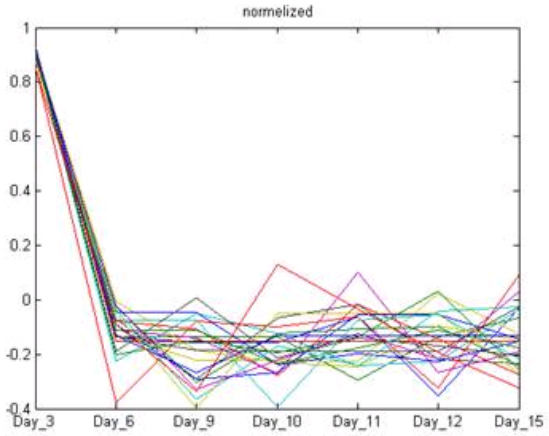


fig.13 G4 cluster; x-axis: time points; y-axis: normalized gene expressions

Summary clusters result:

We've analyzed and present some clusters that include the 5 major expression patterns. We put more emphasis on clusters with high stability. Any other clusters that are not discussed here can be found in Table clusters in the Appendix for Chapter 1.

In Table 10 is a summary of the clusters with a short description and color based representation of the cluster pattern.

Red	**high stability;**
Yellow	**down regulated pattern, cluster break/split from G18**
Green	**up regulated pattern, cluster break/split from G22**
pink	**up regulated day 10 pattern**

Table 9 legend for the color based representation in the Table below with cluster summary .

G2 Stability=10 Size=63 down pattern (see G18 on the dendogram) split from G18
G3 Stability=9 Size=13 oscillatory down pattern (see G18 on the dendogram); split from G18;
5 out of 13 collagen
G4 Stability=12 Size=24 7 out of 24 collagen; low pick day 6,stay low; split from G18; (see G18 on the dendogram)
G5 Stability=6 Size=28 oscillatory down pattern (see G18 on the dendogram); it split from G18,)
G6 Stability=4 Size=20 down pick day 10 in upward overall pattern(see G6 on dendogram); it split from G22
G7 Stability=3 Size=11 down pick day 10,in upward overall pattern(see G7 on dendo); it split from G22
G8 Stability=3 Size=11 upward pick day10; split's from G25

G9 Stability=3 Size=12 downward pattern;(it split fromG18see dendogram))

G10 Stability=11 Size=26 high pick day6, left over the rest of C.elegans development split from G22

G11 Stability=3 Size=284 upward pattern; major size node ; split from G22

G12 Stability=16 Size=51 upward pick day10; split's from G25

G13 Stability=3 Size=11 down pick day10 sub-pattern in downward general pattern (it split from G18 see dendogram)

G14 Stability=9 Size=15 down pick day 10 split from G28

G15 Stability=10 Size=40 upward pattern; split from G22

G16 Stability=3 Size=340 upward pattern; major size node split from G22

G17 Stability=7 Size=12 down pick day 10 split from G28

G18 Stability=11 Size=470 down pattern; major size node from which merge:G2,G3,G5,G9; G4(dendogram:G18) collagen cluster
67 members are collagen related.

G19 Stability=3 Size=11 down pick day 10 split from G28

G20 Stability=6 Size=12 high- pick day 6,down-going main pattern, down-pick day12

G21 Stability=5 Size=27 down pick day 10 ; from G21, splits G14,G17,G19. G21,splits from G28

G22 Stability=3 Size=470 upward pattern& low pick day 10;high pick day6 major size node from which merge: G6, G7,G10 , G11, G15,G16

G23 Stability=4 Size=27 upward pick day10; split's from G25

G24 Stability=3 Size=14 high day 6, decreasing pattern rest of life

G25 Stability=3 Size=100 upward pick day10; split's from G25

G26 Stability=4 Size=11 oscillatory down pattern, senescence pattern-increase day 12-day15

G27 Stability=3 Size=12 up-pick pattern day 10

G28 Stability=3 Size=789 upward pattern major size node

G29 Stability=5 Size=11 down going pattern -senescence as pattern; high expression day 12-day15
G30 Stability=5 Size=10 down pick day 9, pseudo- senescence pattern
G31 Stability=3 Size=12 senescence pattern
G32 Stability=11 Size=12 oscillatory pattern
G33 Stability=4 Size=1941
G34 Stability=4 Size=1978

Table 10: Clustering results: list, pattern description.

4.6 Specific group of genes analysis-Supervised analysis

Besides performing an un-supervised analysis I was interested in specific groups of genes. To analyze these genes I used a supervised type of analysis. For a better understanding of gene pattern data was normalized using same method described in Section1.

Germ line enriched genes- pattern analysis

We analyzed a list of ~ 500 genes known to be germ line enriched (Lund et.al. 2002). We found 472 genes out of the 500 on our arrays. We checked the pattern

behavior of these genes in entire data set as well as in various lists filtered based on the 'variance threshold-method 'described for my assembly of list of 2000 greatest variance genes.[1] Using such a filtering method I obtained 2 more gene lists of 4500 genes and respectively 1100 genes. We analyzed the gene expression pattern of all 472 germ line enriched gene list find in our data of 18615 genes as well as in the 4500 gene list and in the 1100 gene list. In the list of 4500 genes I identified 220 out of ~ 500 germ line enriched genes and in the list of 1100 genes I've find 51 genes out of the 500 germ line enriched genes. For each list we normalized the gene expression in order to facilitate comparison of the gene patterns. Below is the graph of germ line gene expression pattern corresponding to 472 out of 500 found in 18617 gene list, 220 out of 500 found in 4500 list and 51 gene germ line enriched out of 500 found in 1100 gene list.

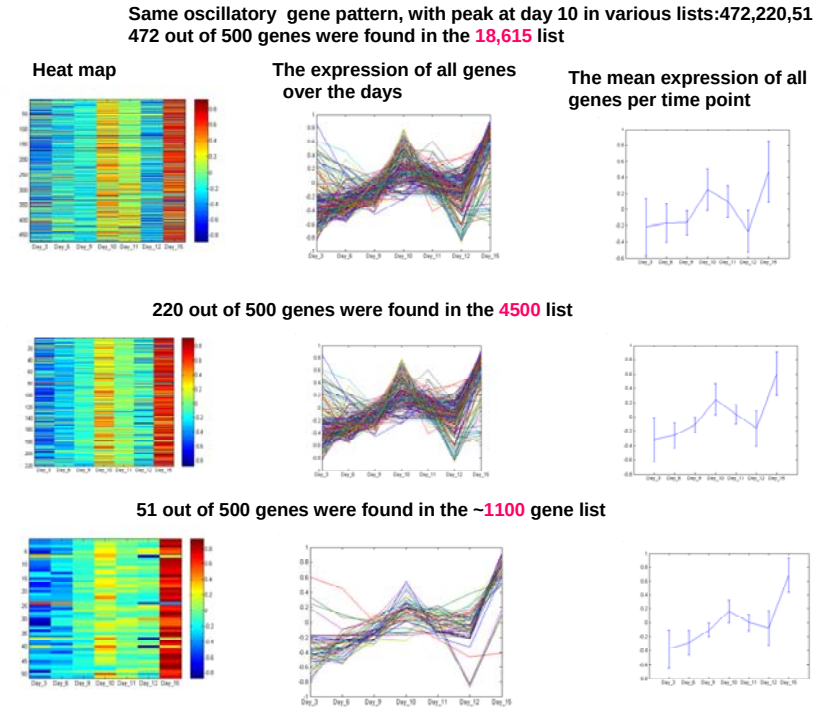


Fig. 14 Germ line enriched genes.

Legend fig. 14 :

[1] We computed the variance for each gene per time point. A ranking between all variances has been performed and we choose the first 4500 genes with highest variance. In the same way we identified the list of 1100 genes.

- First row are the 472 germ line genes found in our raw data in 3 graph representation: heat map, gene expression over time, mean expression of all gene per time point.
- Second raw: 220 germ line genes found in 4500 gene list with highest variation genes in 3 graph representation: heat map, gene expression over time, mean expression of all gene per time point.
- Third raw: 51 germ line genes found in 1100 gene list with highest variation genes. in 3 graph representation: heat map, gene expression over time, mean expression of all gene per time point.

A predominant oscillatory pattern with a 'day 10 peak' can be clearly depicted in all three lists we analyzed. The same pattern we depicted when we clustered the data and find cluster G25. The main biological theme of this cluster was also of germ line genes. Since the genes find in the cluster G25 have a common pattern regulation with germ line genes we might consider that all genes in G25 cluster might have a common loop regulation with germ line enriched genes (see also the comments for cluster G25). Given that this day 10 up-regulated peak was detected when we analyzed directly genes known to be involved with germ line and found also in one of the clusters with a predominantly germ line genes as a biological theme, we might conclude that such pattern is the signature of germ line related genes and of that genes which might have a common loop regulation with germ line enriched genes. Further we might consider that germ line genes have a major role in the transition identified at day 10.

Tissue specific gene expression analysis

Insulin related genes

The gene expression pattern of insulin genes suggests involvement in all major processes.

Given the importance of insulin pathway in aging biology I wanted to get an understanding of the expression of insulin related genes I have on the microarray chips. I identified 23 insulin related genes, see bellow Table 11:

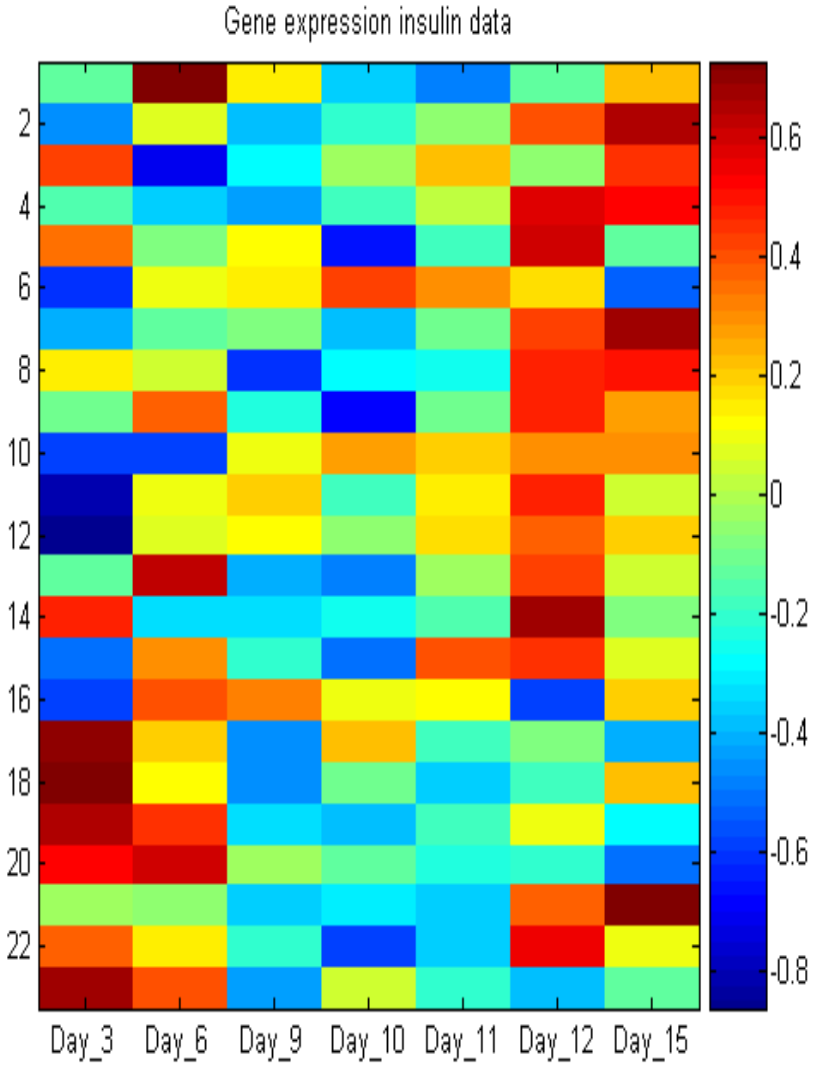
AFFY_ID	names	rdesc							
ZK75-2_at	ins-2	q09627 caenorhabditis elegans. probable insulin-like peptide beta-type 2 precursor. 7/1999							
ZK75-3_at	ins-3	q09628 caenorhabditis elegans. probable insulin-like peptide beta-type 3 precursor. 7/1999							
ZK75-1_at	ins-4	q09626 caenorhabditis elegans. probable insulin-like peptide beta-type 1 precursor. 7/1999							
ZK84-3_at	ins-5	best hit: p56173 caenorhabditis elegans. putative insulin-like peptide beta-type 6. 7/1998 1.0e-45 89%							
ZK84-6_at	ins-6	p56174 caenorhabditis elegans. probable insulin-like peptide beta-type 5 precursor. 7/1998							
ZK1251-2_at	ins-7	q23430 caenorhabditis elegans. probable insulin-like peptide beta-type 4 precursor. 7/1999							
C17C3-4_at	ins-11	q18060 caenorhabditis elegans. probable insulin-like peptide gamma-type 1 precursor. 7/1998							
F56F3-6_at	ins-17	q20896 f56f3.6 protein. 5/2000							

T28B8-2_at	ins-18	ce16518 insulin-like growth factor i like (cambridge) tr:o18149 protein_id:cab03444.1. 0/0
M04D8-1_at	ins-21	q21507 caenorhabditis elegans. probable insulin-like peptide alpha-type 1 precursor. 7/1998
M04D8-2_at	ins-22	q21508 caenorhabditis elegans. probable insulin-like peptide alpha-type 2 precursor. 7/1998
M04D8-3_at	ins-23	q21506 caenorhabditis elegans. probable insulin-like peptide alpha-type 3 precursor. 7/1998
ZC334-3_at	ins-24	q9u1p6 zc334.3 protein. 5/2000
ZC334-1_at	ins-26	q9xui9 zc334.1 protein. 11/1999
ZC334-2_at	ins-30	q9xui8 zc334.2 protein. 11/1999
T10D4-4_at	ins-31	q9tzf3 t10d4.4 protein. 5/2000
Y8A9A-6_at	ins-32	q9tyk2 y8a9a.6 protein. 5/2000
W09C5-4_at	ins-33	q9u333 w09c5.4 protein. 5/2000
F52B11-6_at	ins-34	ce18726 locus:ins-34 (cambridge) protein_id:cab05196.1. 0/0
K02E2-4_at	ins-35	ce18839 locus:ins-35 (cambridge) protein_id:cab04546.1. 0/0
F08G2-6_at	ins-37	ce19778 locus:ins-37 (cambridge) protein_id:cab04062.1. 0/0
T13C5-	daf-9	ce04942 cytochrome p450 (st.louis) tr:q27523 protein_id:aaa80380.1. 0/0

1_at		
R13H8-1_at	daf-16	best hit: o16850 fork head-related transcription factor daf-16b. 5/2000 0.0e+00 96%

TABLE 11: 23 insulin related genes that are on our array.

Below is the hit map for the 23 insulin genes-. Besides genes where represented also in a graph with expression of each gene over time in fig.15



Legend: each row is a gene expression, each column is a time point sample.

Red is high level gene expression, blue is low level gene expression. Gene expressions are normalized.

Fig.15 heat map of the 23 insulin genes

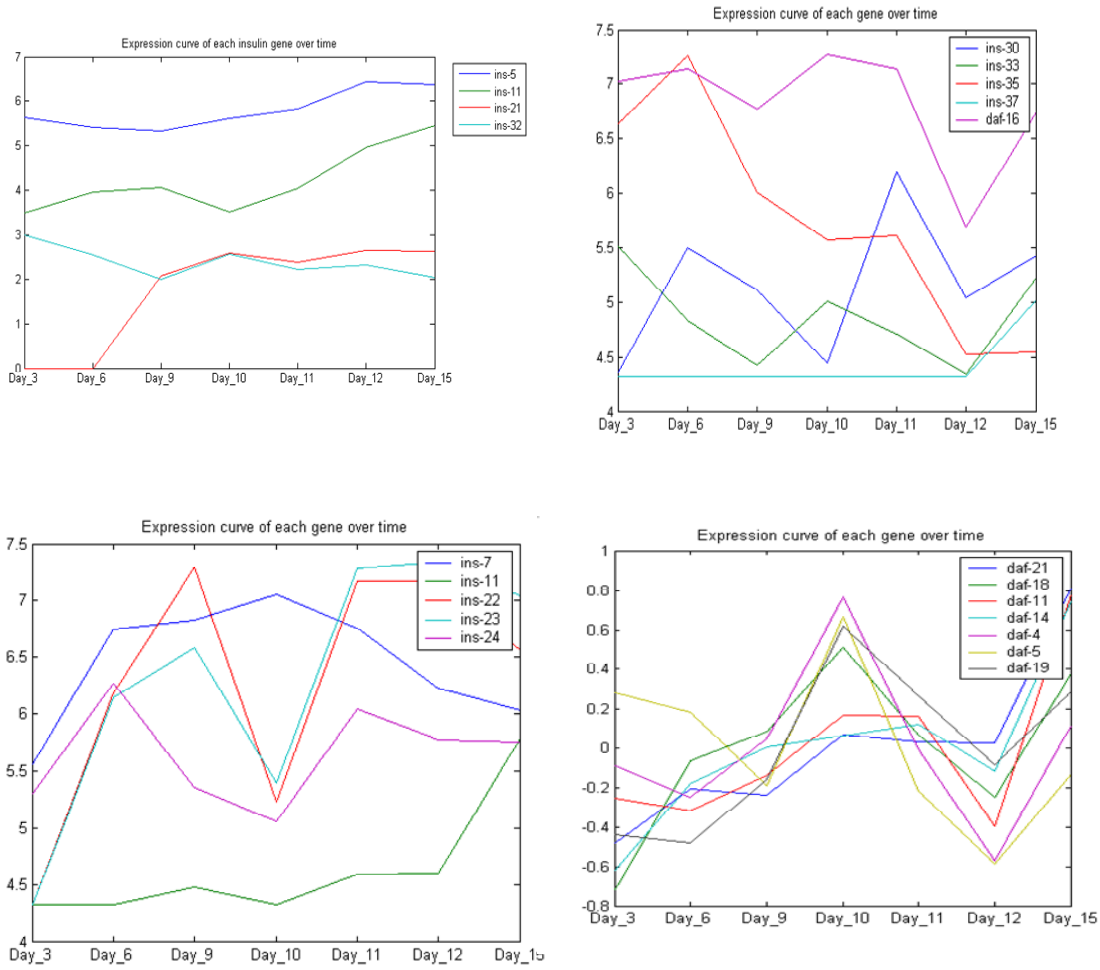


Fig. 16 expression of insulin related gene out of 23 genes. The 4th fig right down are dauer related gene expressions normalized.

The first 3 plots from fig 16 are expression profile of insulin related genes, approximate 5 genes per plot. The 4th plot contains dauer related genes found on our array.

Various patterns can be identified in this plot. For example 'day 10 up-peak pattern' found in the cluster G25 can be seen in the dauer related genes at the bottom of fig 10, right side, the genes: *daf-21 daf-18;11;14;4;5;19* and in expression of *ins-7* gene. A down-peak at day 10 can be seen for *ins-22, ins-23, ins-24.*

The insulin genes *ins-5;11;21;32* show an up-regulated pattern.

ins-35 and *daf-16* show an oscillatory down-regulated pattern and the *ins-30* and *ins-33* show an oscillatory pattern. The *ins-37* has an interesting 'senescence pattern' of steady state over all time points and than an abrupt increase in the expression between day12-15.

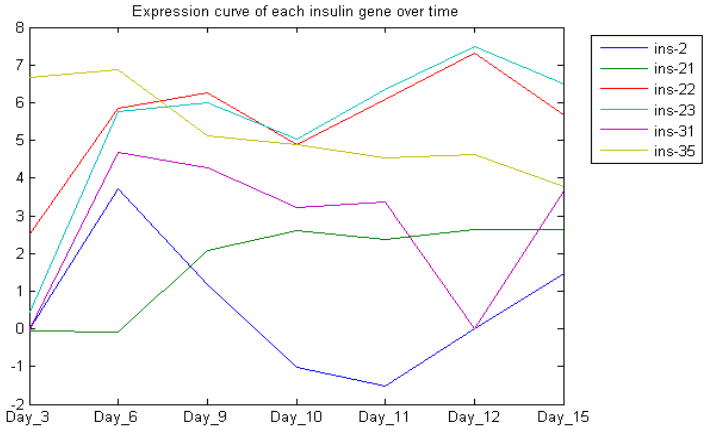
Basically, we can identify all 4 patterns found when the general data were clustered for insulin and dauer related genes . This finding can be interpreted as following: if we consider the 4 patterns found in the clustered data as patterns that are describing the main biological processes in this organism, than by observing that insulin genes are expressed in all 4 patterns suggests that insulin genes are involved in all main processes which this nematode undergoes: aging, development, homeostasis. I may further infer that insulin and dauer genes might affect the longevity of *C. elegans* just in an indirect way. The fact that when mutations occur on insulin pathway this in turn affects longevity might be just a signature that insulin pathway actually affects some other vital biological processes which in turn will have an affect on the length of the life time of the *C. elegans*.

- ***daf-21, daf-18, daf-11, daf-14, daf-4, daf-5, daf-19* and *ins-7* might share same regulatory loop as *daf-2, daf-16* and *ins-7***

Another interesting aspect is that all daf genes mentioned together with *ins-7* have the same oscillatory pattern with a day 10 up-peak. We know from Kenyon results (Nature 424,2003) that when DAF-2 is active, DAF-16 activity is inhibited and *ins-7* is expressed, allowing further DAF-2 activation. When DAF-2 activity is reduced, DAF-16 is activated and *ins-7* expression is inhibited. Our finding that other daf

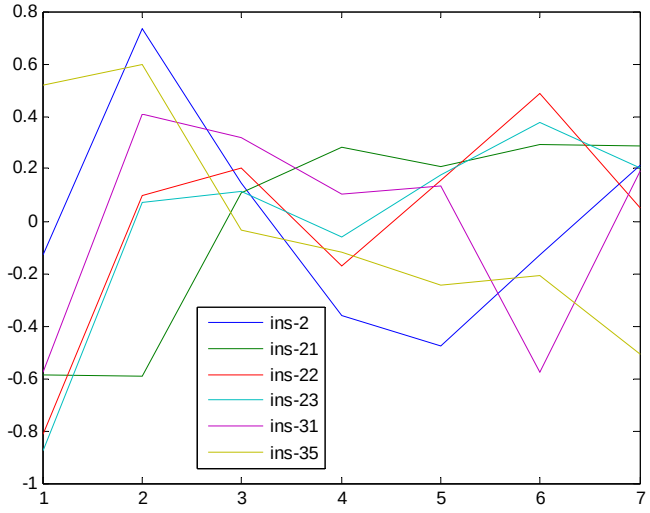
genes: *daf-21, daf-18, 11,14,4,5,19*- share the same pattern with *ins-7* suggests other insulins may participate in the same or similar regulatory loops.

From this survey of 23 insulin related genes, 6 of genes were in our list of 2000 genes, expressed with highest variance. Below is the plot for expression of this genes-see Fig. 17a. The second plot represents the normalized data-see Fig. 17b



x-axis: time points; y-axis: gene expressions-not normalized, log2 applied.

Fig. 17a-6 insulin genes in our 2000 gene list with highest variance



x-axis: time points; y-axis: gene expressions- normalized, log2 applied.

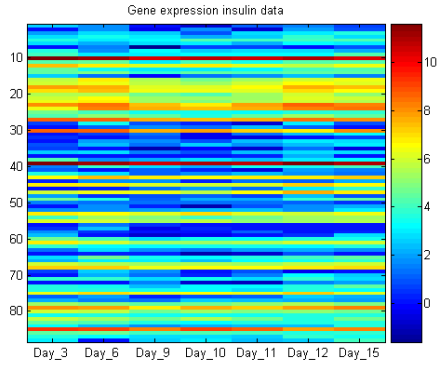
Fig. 17b-6 insulin genes in our 2000 gene list with highest variance-normalized data.

Differences in insulin gene patterns might be a direct reflection of the stochastic behavior of gene expression.

Comparing our data on insulin gene expression with Kenyon results (Nature 424,2003), I note that in the Kenyon data *ins-2* is increased and *ins-21* decreases slightly, whereas in our data, *ins-21* definitely increases and *ins-2* has an oscillatory down-going pattern. Various explanations might be offered for the difference in the patterns, including the differences in the chips and technology used and the difference in the biological strains used. Nevertheless, the pattern difference might also be considered a direct reflection of the stochastic behavior of the gene expressions in *C.elegans*. I will return to these issues of interpreting the comparison between data sets in different labs when I make a careful comparison of our data with other two microarray experiment data.

Neuronally expressed genes:

For the neuron expressed genes we focused on a list of approximate 90 genes. The heat map for the expression of these ~ 90 genes, 88 genes to be more precise, can be seen in fig.18.



Each row is a gene expression, each column is time point sample.

Red is high level gene expression, blue is low level gene expression. Gene expressions are normalized.

fig. 18 Heat map of the 88 neurons

From the graph we can already depict two gene categories: a large number of genes with a relative steady expression value and another category of genes where each of the gene has a distinct pattern of expression.

- **Neuronal related genes might be regulated by environmental cues**

We were interested to find out what genes from the list of ~ 90 genes expressed in neurons are in our list of 2000 genes that exhibits highest variation.

Out of the ~ 90 genes I examined, I found 10 genes in the 2000 filtered list based on highest variation. See neurons_table. The gene expression pattern of each neuron related gene appears to be a separate pattern.

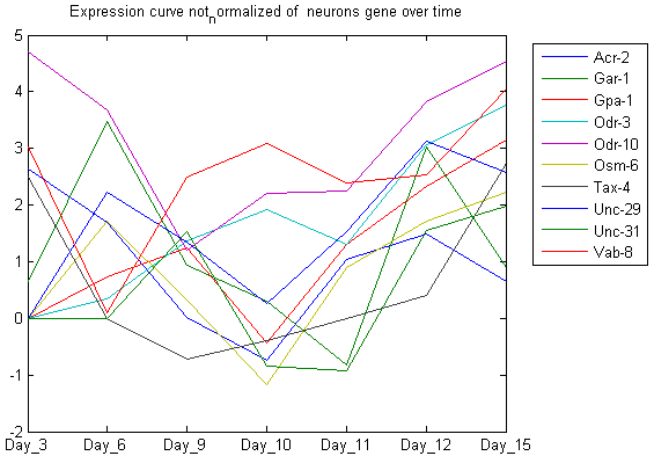


fig. 19 10 neuronal related genes (log2) out of 88 in the 2000 list; **x-axis: time points; y-axis: gene expressions- normalized, log2 applied.**

This might be a signature for the fact that some neuronal related genes might be regulated by various stimulus and environmental cues over entire life of this organism and that these cues are perceived and integrated in a complex and

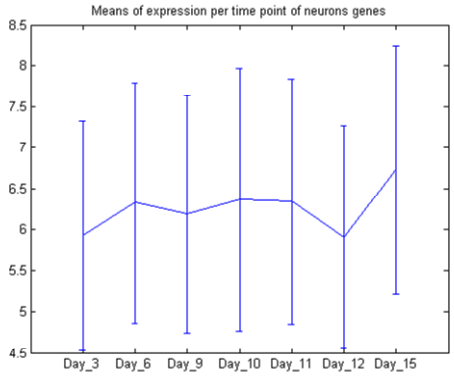
sophisticated fashion by specific neurons. These 10 genes might be a signature of the neuronal genes that do change over time and consequently impact various processes at various time points.

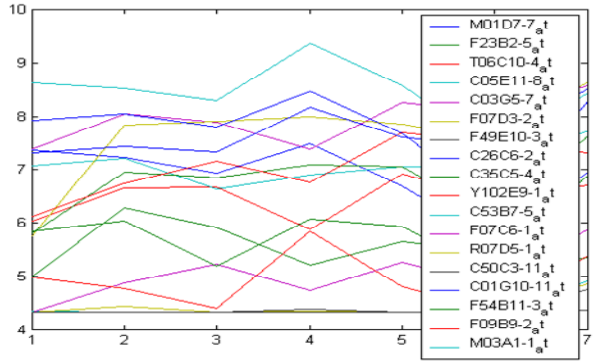
'Steady state' expression pattern-a signature consistent with no morphological changes at neuron level.

We wanted to see if neuronal related genes might have any other pattern. In this sense I enlarged the list of genes with highest variation at 4000. We found 18 more genes out of the 90 genes. They present a distinct clear 'steady' gene expression pattern over the life span of *C. elegans* See fig 20.

Neurons related genes

18 genes out of 88 genes in the list with highest variation





Neurons show a relative steady gene expression over the life span of the nematode.

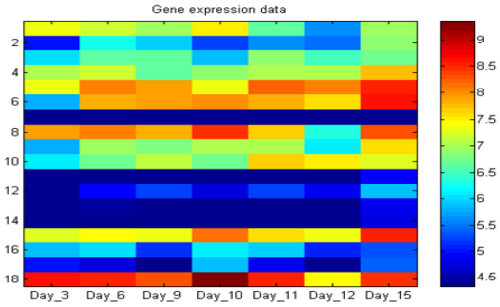


Fig. upper right: gene expression over time **x-axis: time points; y-axis: gene expressions- normalized, log2 applied.**

Fig down: Each row is a gene expression, each column is time point sample.

Red is high level gene expression, blue is low level gene expression. Gene expressions are normalized.

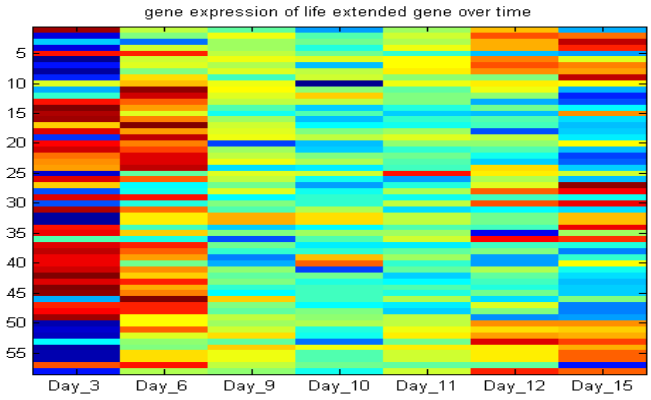
Fig. 20 neuronal 18 genes out of ~90 genes

- **Two distinct categories of neuron related genes:**

We concluded that neuronal related genes can be grouped into two distinct categories: one which shows steady expression over the life span of the nematode and another with variable gene expression over time. The steady state pattern of gene expression might be a signature of Driscoll lab hypothesis that neurons don't show significant morphological changes during the life of the nematode. The unchanged group of genes is represented by 18 neuronal expression genes.
On the other hand, the group of 10 genes implies that gene expression in aging animal might be regulated by environmental cues and that these cues are perceived and integrated in a complex and sophisticated fashion by specific neurons. These 10 genes might be a signature of the neuronal genes that do change over time and consequently impact various processes at various time points.

Genes predicted to impact longevity

Next we've checked a list of genes considered to be involved in life extension of *C. elegans*. The list of 260 genes was taken from Murphy et. al. 2003, (Nature 424). Approximately one in four genes from the list of life extended genes of 260 genes is included in our list of 2000 genes with highest variation. The heat map for the 58 genes in this graph can be seen in fig 21.



Each row is a gene expression, each column is time point sample.

Red is high level gene expression, blue is low level gene expression. Gene expressions are normalized.

Fig 21. Heat map of **58** genes out of 260 **life extended genes** from Murphy et.al. 2003, (Nature 424)

For a complete description of the 58 life-extended genes see Appendix B

- **Life extended genes- regulate key developmental switch, metabolic rate and core processes in general, in accord with evolutionary theory of aging.-two class genes:**

Some of these genes regulate a key developmental switch, while the others control core processes, such as the overall rate of metabolism. These are exactly the kinds of processes predicted to be important to longevity by the evolutionary theory of aging. This theory suggests that competition for metabolic resources between processes such as growth, reproduction and cellular maintenance lies at the heart of the ageing process.

Based on the clustering analysis these genes can be broadly classified into two classes: one of increasing and the other of decreasing in expression pattern.

- **General gene expression pattern suggests that the events are taking place at the beginning of the adulthood day 3-day6 might influence the final days of this organism**.

In our data the 58 life extended genes show a high expression pattern at day 3, day 6, a decrease in expression up to day 12 and than an increase again between day12-15

If we consider the possibility that an increase in expression pattern reflects an increase in the gene activity, than the genes responsible for prolonging the life span of the nematodes have an increase in activities at the beginning of the adulthood and than again at the end of the life of this organism suggesting that the events might take place at the beginning of the adulthood day 3-day6 might determine the final days of this organism.

Autophagy related genes and their implications in aging studies.

A recent paper provides evidence that macroautophagy is an essential downstream pathway for one of the mutations known to extend life span (A. Melendez, B. Levine, *Science* (2003)) Autophagy, or the degradation of intracellular components by the lysosomal system, was thought for a long time to be a catabolic process responsible for cellular cleanup. However, in recent years, we have learned that autophagy comes in different sizes and shapes, macroautophagy being one of them, and that this cellular maid plays many more roles than previously anticipated. Activation of autophagy is essential in physiological processes as diverse as morphogenesis, cellular differentiation, tissue remodeling, and cellular defense, among others. Furthermore, macrophautophagy participation in different pathological conditions, including cancer and neurodegeneration, is presently a subject of intense investigation. A role in aging has now been added to this growing list of autophagy functions. The activity of different forms of autophagy decreases with age, and this reduced function has been blamed for the accumulation of damaged proteins in old organisms. Research shows that there is much more than trash to worry about when autophagy is not functioning properly. We wanted to investigate the expression gene pattern of authophagy genes in our data set. We identified 7 authophagy related genes that are significant regulated during adult life. Their expression pattern is presented in fig.22:

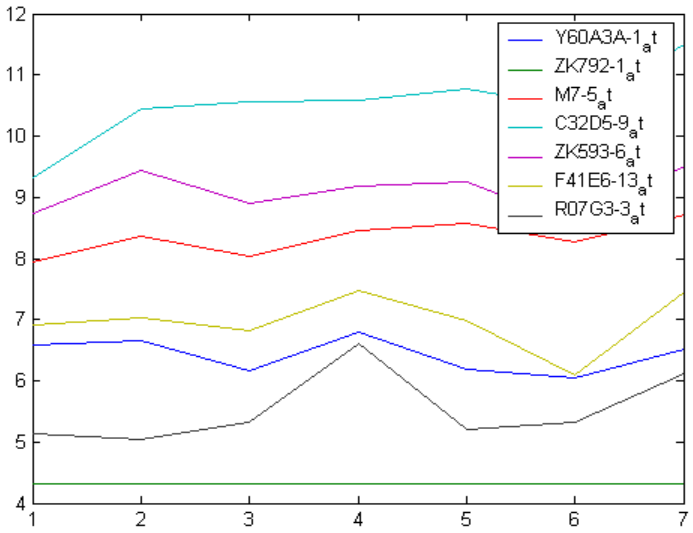


fig.22 authophagy genes: gene expression over time; **x-axis: time points; y-axis: gene expressions.**

- **Gene expression pattern suggests a common regulatory loop between certain daf genes and autophagy genes.**

As can be seen in the table 12 bellow and fig. 22 the autophagy related genes have similar pattern and have a role in dauer larval development as well.

R07G3.3	npp-21	authophagy, dauer larval development
F41E6.13	atg-18	authophagy, dauer larval development
ZK593.6	lgg-2	not required for dauer larva development or extended life
C32D5.9	lgg-1	authophagy, dauer larval development
M7.5	atg-7	authophagy, dauer larval development
ZK792.1		nematode larval development, autophagy

Y60A3A.1	unc-51	authophagy, dauer larval development

Table 12-autophagy related genes

Both dauer formation (a stage of developmental arrest) and adult life-span in *Caenorhabditis elegans* are negatively regulated by insulin-like signaling, but little is known about cellular pathways that mediate these processes. Dauer formation is associated with increased autophagy (A. Melendez, B. Levine et al., Essential role of autophagy genes in dauer development and lifespan extension in *C. elegans. Science* (2003)).

Interestingly, indeed in our data we find that same oscillatory pattern with 'day 10' peak is shared by both dauer genes (see fig 23) and *npp-21*, *atg-18* and *unc-51* autophagy genes.

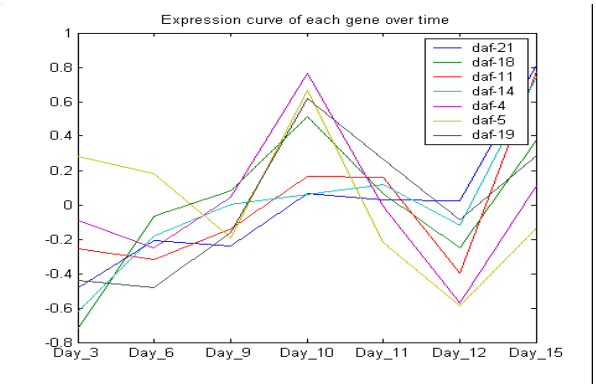


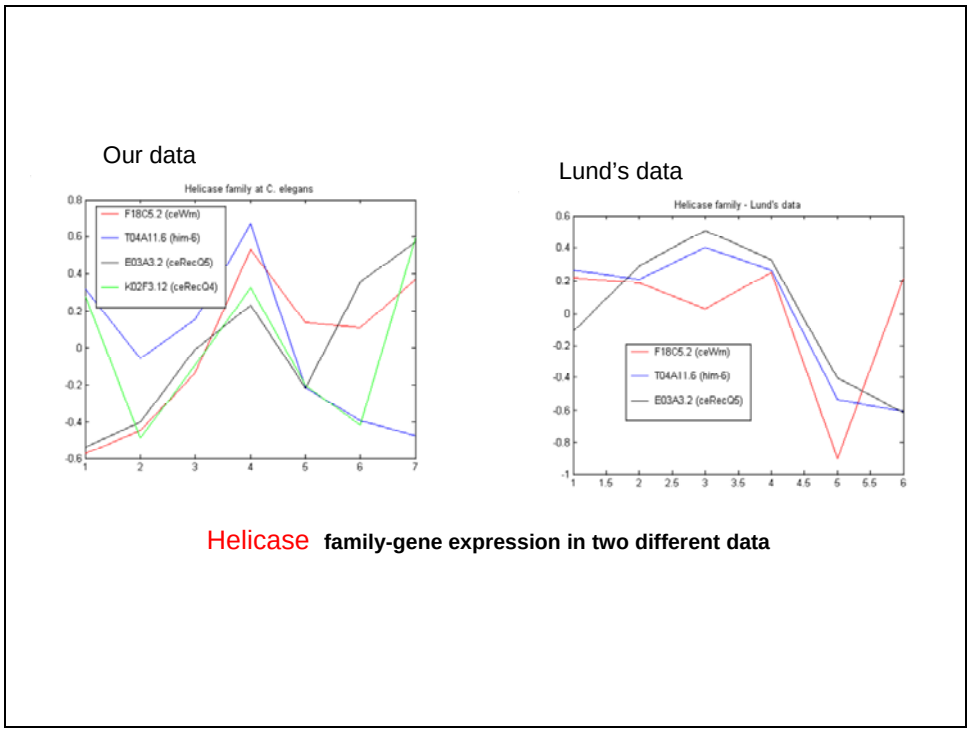
Fig.23 dauer genes gene expression over time **x-axis: time points; y-axis: gene expressions- normalized, log2 applied.**

This finding might suggests that *daf-21,18,11,14,4,5,19* and *npp-21*, *atg-18*, *unc-51* may participate in the same or similar regulatory loops.

Helicases related genes and aging:

Deficiency in a helicase of the RecQ family is found in at least three human genetic disorders associated with cancer predisposition and/or premature aging. The RecQ helicases encoded by the *BLM*, *WRN* and *RECQ4* genes are defective in Bloom's, Werner's and Rothmund–Thomson syndromes, respectively. Cells derived from individuals with these disorders in each case show inherent genomic instability.
We identified in our data four genes known to be helicase related:
ceWrn him-6 ceRecQ5 ceRecQ4. see fig 24 left. The data presented in fig 24 is normalized for comparison purpose.

Given the importance of the helicase genes in aging related diseases we wanted to see how similar is their expression in one more data set. Lund et. al. performed an experiment similar to our experiment. In spite of some major differences as the type of chip used, time points and strains used, differences which will be stressed later, fundamentally, the design of the experiment has some similarities in the sense that the time points in both data sets cover the entire life span of *C. elegans* and both experiments have replicates, therefore we've checked the helicase gene expression in Lund et. al. data as well, see fig. 24, right.



Helicase **family-gene expression in two different data**

left: our data; ox-axis: 7 time points: day3,day6,day9,day10,day11,day11,day15. oy-axis: normalized gene expression. Log 2 was applied.

right: Lunda data; ox-axis: 6 time points: 1/day3 ; 2/day4 ;3/day6-7 ; 4/day9-11 ; 5/day12-14; 6/day16-19; oy-axis: normalized gene expression. Log 2 was applied.

fig.24 Helicases genes

The same oscillatory pattern of day 10 can be notice for *him-6* in both data sets. The *RecQ-5* gene as well as the *ceWrn* gene related with Werner syndrome in humans show a different pattern in each data set. This might be interpreted as a lack of robustness of this gene for various stochastic factors that influences *ceWrn* and *RecQ-5* gene expression. The fact that the gene expression pattern is not repetitive in the two experiments could be considered a feature of the genes involved with genomic instability.

DNA helicases are molecular motors that catalyse the unwinding of energetically unstable structures into single strands and have therefore an essential role in nearly all metabolism transactions. Defects in helicase function can result in human syndromes in which predisposition to cancer and genomic instability are common features. RecQ helicases are a family of conserved enzymes required for maintaining the genome integrity that function as suppressors of inappropriate recombination. Mutations in RecQ4, BLM and WRN give rise to various disorders characterized by genomic instability and increased cancer susceptibility. One of the signatures of such genes involved in genomic stability might be exactly this inconsistency in gene expression between experiments due to the stochastic factors modulating the expression level of such genes.

Muscle related genes and aging

The behavioral study of ageing nematodes showed a significant decrease in mobility. Age-associated locomotory defects increase progressively in severity over time. Progressive locomotory impairment during *C. elegans* ageing could be the consequence of a decline in muscle function.

For muscle related genes we had several lists we wanted to check in our data. First we checked a list of 60 muscle genes , obtained from Lund et al., 2001, Curr Biol. 12(18):1566-73.

In entire raw data of 18 668 the expression of this group of 60 genes after normalization is as follows: high expression for day 3 to day 6, a relatively steady expression between day 9 to day 12 and again an increase in expression between day 12 to day 15. A graphic representation of this genes can be seen in fig. 25 see bellow.

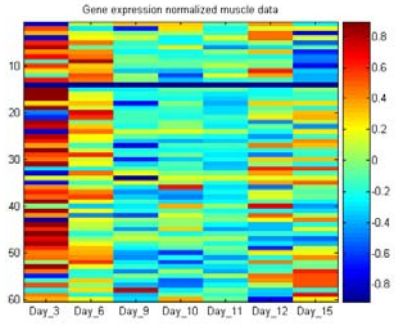


fig. 25 Muscle data_60 genes normalized

We've classified the 60 genes in structural, development/assembly, contraction and anchoring.

When we've checked this 60 genes in our list of 2000 genes with highest variation we found 14 genes. See below their expression pattern in fig 26.

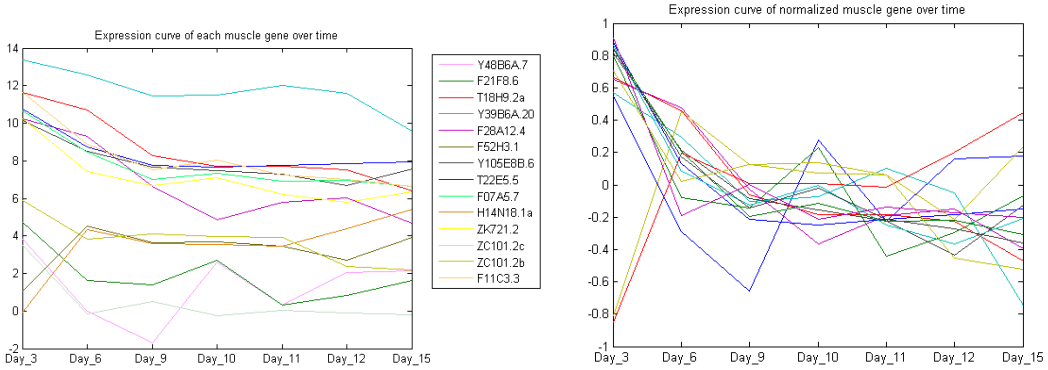


Fig26. a) 23 muscle out of 60 identified in the list with 2000 genes b) same 14 genes normalized.

A relatively steady expression patterns between day 9 to day 12 can be seen in most of the expression patterns of the 14 genes.

When I examined the list of 1283 muscle related genes proposed by Kim (Roy,Kim et.al, 2002, Nature 418) we found 1187 genes, on our chips among which 111 were within the list of 2000 filtered genes.

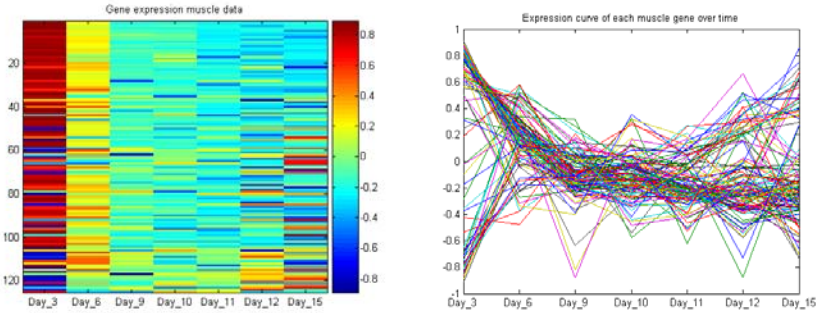


fig. 27 125 genes normalized in common between 2000 list & 1187

As for the previous muscle lists we analyzed, most of the genes are highly expressed at day 3, than a relative steady expression from day 6 to day 12 and than again the expression pattern of this genes become higher. For complete references on the genes name list and a short description of 1187 muscle related genes, 111 muscle related genes.

In conclusion, all muscle expressed genes checked in our data and involved in structural, development/assembly, contraction and anchoring share a similar pattern: high expression day 3, day 6, steady expression at the middle life span of *C. elegans* (day9-12), and again high expression toward the end of life of the organism (day12-15). It is interesting that this pattern is similar to that of the long life gene pattern (from Kenyon et. al) previously analyzed. The apparent changes, beginning in mid-life, muscle structure (see Driscoll et.al, Nature 2003) might have as signature at transcriptional level the low expression level between day6-day12. Besides, as later research I've done will show, another clearly pattern muscle related is of decrease in expression level starting with day 3, all the way up to day 15.

We performed a much thorough analysis later on the data muscle gene when we wanted to understand sarcopenia process. Later results will conclude that are actually two distinctive patterns at the muscle level: one of low expression between day 6-12, and another of decrease in expression for all time points between day3-day15. One use of this finding might be if the muscle related group of genes which start to show an increase of the expression level between day 12-15 might be considered as genes with the potential of reverting the sarcopenia.

4.7 Concluding Remarks

Microarray experiments seek to obtain readouts of gene expression levels over the whole transcriptome . This information can be useful for determining how the transcriptional regulation of genes might coincide, thereby implicating proteins as parts of networks acting together towards a common biological function. Such experiments are particularly useful for complex biological traits that result from the presumed functioning of several molecular pathways. Aging is one such biological trait that apparently incorporates numerous molecular mechanisms underlying environmental stimulus sensing, metabolic regulation, stress responses, reproductive signaling, and transcriptional regulation. Current models of aging emphasize different mechanisms as driving forces behind aging and lifespan determination. However, an integrated understanding of exactly how these mechanisms drive aging has not yet been formulated.

The methods I used for gaining a better understanding of the mechanisms which might underline the aging process where supervised and unsupervised. When interpreting the data, using a supervised approach, I tried to follow the major biological theories currently known that describe aging. In this sense to address the oxidative damage theory of aging, for instance, I've identified stress response genes that exhibit statistically significant changes, and then ask whether the expression patterns of these genes share a common pattern. Also I've looked into insulin and dauer pathway, all being considered important leads in aging studies. Insulins, aging-related gene, dauer-related genes, , autophagy related genes, muscle, neuronal and germline genes all are singled out and their expression profiles examined.

I've investigated and addressed two major aging hypothesis, both being developed in Driscoll's lab; one is pointing out to a major 'crisis' which is going on in the midlife period of time of the organism especially at the muscle level, which might be critical for determining the ultimate lifespan of that animal, the other is underlining the idea that aging must be understood as an stochastic process due to stochastic cues acting on the organism over the entire lifespan of that organism. The implications of aging as a stochastic process can be seen at the transcription level in

various specific group of genes as I've pointed out when I've analyzed insulin or neuronal related genes.

Chapter 5

In the search of molecular signature of sarcopenia in *C. elegans*

5.1 Introduction

Sarcopenia is an "age-related" loss of muscle mass leading to muscle weakness, limited mobility, and increased susceptibility to injury. Overall changes with age that contribute to sarcopenia include declines in androgenic and growth hormone concentrations, declines in spontaneous physical activity, and changes in dietary intake of protein and/or energy. Specifically, in skeletal muscle there is a selective loss of muscle fibers, decline in total muscle area and reduced muscle capillarization, shortening velocity, and maximal force .

To begin to identify the molecular basis for the loss of muscle mass with age, investigators have measured in mammals changes in gene expression on a global scale during aging in skeletal muscle using serial analysis of gene expression, cDNA arrays, and oligonucleotide-based microarrays. These studies have reported changes in gene expression consistent with decreased protein synthesis, impaired oxidative defense, and decreased activity of mitochondrial proteins. They have also reported differential expression of genes involved in energy metabolism, DNA damage repair, stress response, immune/inflammatory response, RNA binding and splicing, and proteasome degradation. Although these studies have provided insight into the age-related changes in gene expression and therefore the aging process, the

human studies in particular have limitations with regard to sample size, number of genes surveyed, overall smaller differences in gene expression, and pooling of samples. Importantly, investigating the molecular mechanisms underlying sarcopenia in humans with the use of microarrays is also complicated by the inherent variability in human gene expression profiles. This variability is likely due to differences in genetics, diet, environment, and habitual patterns of activity, making it more difficult to identify true age-specific alterations. In fact, investigators using the human Affymetrix microarrays to study young vs. older males found that the intragroup (n = 8) variability was so high that a special ratio method was needed to be developed to reduce the within-group variance .

Using *C. elegans* as animal model in an effort to better understand the biology of aging we put an emphasis on mid-life changes that, we consider might influence aging, and give us an insight into sarcopenia as a process. Studies in our lab and others have suggested that critical events during the mid-life of the nematode can influence the aging of the organism. The small nematodes have less variability in gene expression profiles, and a lot of muscle-related genes (50% have human homologie), which presents clear advantages for aging and sarcopenia microarray studies.

In the present study, Affymetrix GeneChips special designed for *C. elegans* (by Hoffmann-LaRoche company, from Basel, Switzerland) were used to interrogate the expression of 18,612 genes (open reading frames).

5.2 Experimental design

In this study we used same experimental design as described in first chapter. Our experiment includes time points covering the reproductive and post-reproductive periods, with a series of consecutive mid-life time points.

In order to be able to grow the worms in a synchronous way, we used only spe-9 (hc88), which is a temperature sensitive mutation. This strain does not produce

progeny at 25 C;

We cultured *spe-9*(hc88) mutants at 25.5^{o}C in order to avoid contamination of a synchronous culture with young animals. *spe-9*(hc88) is defective in spermatogenesis and produce unfertilized oocytes when reared at high temperature. Any "escapers" were visually identified and eliminated manually, so cultures were highly synchronous. At days 3, 6, 9, 10, 11, 12 and 15 of culture (as measured from egg deposition) animals just reach adulthood at day 3. We harvested ~ 20,000 worms per time point in three trials and used RNA from each of these for three independent hybridizations. Because of our focus on potential relevant changes at the midlife transition, we also prepared another triplicate experiment in which we harvested nematodes at days 9, 10 and 11.

Data from these middle time-points were combined with those in the more extended trials to increase the significance of findings at days 9, 10 and 11.

Therefore, in our data we have six total independent repeats for the middle life time points day 9, 10, 11 and 3 repeats for day3, 6, 12, 15.

5.3. Data Analysis-Methods

5.3.1 Outlier detection:

In order to detect the outliers we used (as outlier exclusion test) the Nalimov outlier test. For each gene per *Condition* a modified Nalimov outlier test is performed for data points representing Replicate experiments. (see, Kaiser R, Gottschalk G (1972)).

5.3.2 Scaling

In order to achieve, scaling of the data on the chip and between chips, for each chip, we calculated the median signal intensity over all probe sets. The median of this median signal intensity from all chips was calculated. Then, every chip, is scaled to

this median value.

5.3.3 KNN estimation method

The data was estimated based on the values of the K nearest neighbor genes estimator, (Tibshirani, R., Botstein, D. & Altman, R. B. (2001).

We also transformed the data using a logarithmic transformation in base 2, X= $\log_2(X)$. The reason we do this is that is preferable to work with logged intensities rather than absolute intensities since the variation of logged intensities tends to be less dependent on the magnitude of the values; taking logs, reduces the skewness of the distributions and improves variance estimation.

5.3.4 Filtering

A special method for filtering genes based on highest variance was designed. Genes were filtered on the basis of their variation across the samples. A set of 2000 genes were chosen, on the basis of their standard deviations. I analysed also other lists choosing as, for example, 2500 genes, 3000, and 5000 genes; or filtering data using ANOVA as way of filtering and then used FDR test for checking on false positive and obtained a list of 1241 genes, however after careful unsupervised analysis of all the lists mentioned, we conclude that the lists of 2000 genes obtained from the filtering based on highest variation is more suitable for answering question in regard of aging, sarcopenia and that the ANOVA method is too conservative for our purpose. Also the lists of 3000 and 5000 genes were unnecessary large from the point of biology novelty. Therefore in presenting the results of our unsupervised as well as supervised method, we will refer to the list of 2000 genes obtained based on the highest variation filtering.

5.3.5 Normalization

In order to normalize the data, I went through two steps: first, the step of what is known in statistics literature as “center mean”, and obtain this way a new vector.

Then the new vector is divided to its “standard deviation”, meaning normalize the newly obtained vector. (See Methods Chapter 1).

5.3.6 Unsupervised method for data mining- clustering

I’ve chosen to use for our data a new approach to clustering based on the physical properties of a magnetic system. The algorithm is a Monte Carlo-based method (in particular Swedsen-Wang Monte Carlo method) and uses KNN for defining the neighbors. We were able to find clusters that have not been obtained by other unsupervised clustering methods as Tree- View or K means clustering. The reason is that this method has a number of unique advantages:

4 The number of the “macroscopic” clusters is an output of the algorithm.

5 The hierarchical organization of the data is reflected in the way the clusters split or merge when a control parameter is varied.

6 The results are insensitive to the initial conditions.

Comparing this algorithm with other clustering algorithms, the drawback of methods like “Tree View” or “K means algorithms” is that they have high sensitivity to initialization and they have poor performance when the data contains overlapping clusters; The most serious problem is lack of cluster validity criteria; none of these methods provide an index that could be used to identify the most significant partitions among those obtained in entire hierarchy (Methods Chapter 1 and Domany et. al. Physical Review ‘96 for more on SPC algorithm).

5.3.7 Supervised methods

For compiling various lists I used AQL language which is a new query language for the Acedb database system. It borrows syntax and ideas from OQL, the ODMG's proposed query language for object-oriented databases (which is supported by O2),

Lorel, a language for querying semi-structured data in the Lore database system developed at Stanford, and Boulder.

I also used the Gene Ontology database, (GO_term). (http://www.geneontology.org).

5.4 Results

5.4.1 Sarcopenia signature

I have expanded on previous work in *C. elegans* (see Kim et al. 2002, Kenyon et. al 2003) studies by increasing the number of samples over the midlife time window, and, most importantly, focusing our efforts on defining a molecular signature of sarcopenia rather than a general survey of gene changes with age. We also extended our previous analysis on muscle related genes from Chapter 1.

In our sarcopenia studies I used a combination of supervised and un-supervised methods. For this purpose I compiled several lists of genes muscle- related genes. The definition of sarcopenia at the phenotypic level is that is an "age-related" loss of muscle mass leading to muscle weakness and reduce mobility. At the genotypic level we might suspect that any changes in muscle- related genes might be the reason of an sarcopenia phenotype. In the rest of the work I will call a sarcopenia signature any major changes in muscle-related gene expression. Given the phenotypic aspect of the sarcopenia one might expect to identify as sarcopenia signature a down-regulated gene expression pattern. As I will show in this work, these will not be always the case. I will call as 'positive connection with the sarcopenia phenotype' any down-regulated gene expression pattern.

At first I compiled a list of genes considered to be expressed in muscle cells. In order to do this I used a combination of bioinformatics tools special designed for searches in Worm Database. I used AQL language (see Methods section) which is a new query language for the Acedb database system to compile a list of 829 genes from the worm database expressed in muscle cells. Out of the 829 gene list in our raw data we identified 721 genes. We wanted to understand how many of these 721

genes considered to be expressed in cell muscle are in our list of 2000 genes with highest variation. To obtain the list of 2000 genes I used a selection method based on the highest variation of the gene expression in conjunction with a k-nearest neighbor estimator (see Methods section). We identified 42 genes as being at the intersection between the 721 genes with the list of 2000 genes.

The set of 42 muscle expressed genes are included in the list of 2000 genes, that show greatest variation during adult life, and might serve as a signature of *C. elegans* sarcopenia. These 42 genes might be representative as genes involved in muscle if we consider as underlying hypothesis that genes that vary the most should be more involved in the biological process than genes that show a relative steady state of gene expression.

We analyzed the expression of each of the 42 genes which might comprise the signature of sarcopenia over time. The list of the 42 genes with ORF annotations, CDG annotation (or 3 letter names) as well as a concise description for each gene as was found in worm base using AQL query language can be seen in Table 1.

'T14A8.1'	'ric-3'	Biological process: embryonic development ending in birth or egg hatching (IMP) growth (IMP) nematode larval development (IMP) protein targeting to membrane
'H30A04.1'	'eat-20'	'"eat-20 encodes a paralog of the C. elegans and Drosophila genes crb-1 and crumbs, expressed in pharynx, head neurons, hypodermis. Biological process: hermaphrodite genitalia development
'C12C8.1'	'hsp-70'	'hsp-70 encodes a member of the hsp70 family.' Biological process determination of adult life span
'F11C3.3'	'unc-54'	'"unc-54 encodes a muscle myosin class II heavy chain (MHC B); UNC-54 is the major myosin heavy chain expressed in C. elegans ; Biological process: body morphogenesis (IMP) inositol lipid-mediated signaling (IPI) locomotion (IMP) muscle contraction (IMP) muscle thick filament assembly (IMP) oviposition (IMP) pharyngeal pumping (IPI) positive regulation of locomotion
'C13B9.4'	'...'	"C13B9.4 is orthologous to the human gene CALCITONIN RECEPTOR Biological process: G-protein signaling, adenylate cyclase activating pathway (ISS) cellular calcium ion

		homeostasis (ISS) locomotion (IMP) positive regulation of adenylate cyclase activity
'ZK721.1'	'tag-130'	The precise role is not known; detected in such tissues as body wall muscle, hypodermis, intestine, pharynx, and the gonad
'W02C12.3'	'hlh-30.'	'"W02C12.3 is orthologous to the human gene TRANSCRIPTION FACTOR BINDING TO IGHM ENHANCER 3 (TFE3; OMIM:314310), Helix loop helix transcription factor EB
'ZC101.2'	'unc-52'	biological processes: cell adhesion (IEA) cell migration (IGI) determination of adult life span (IMP) embryonic development ending in birth or egg hatching (IMP) epidermal growth factor receptor signaling pathway (IGI) locomotion (IMP) molting cycle, collagen and cuticulin-based cuticle (IMP) muscle development (IMP) muscle morphogenesis (IGI) nematode larval development (IMP) positive regulation of growth rate. UNC-52 is synthesized by the hypodermis and localizes to the extracellular matrix between hypodermis and muscle And following Molecular function: calcium ion binding, structural molecule activity
'C16D9.2'	'rol-3'	biological process: collagen and cuticulin-based cuticle development, embryonic development ending in birth or egg hatching, locomotion positive regulation of growth rate (IMP) protein amino acid phosphorylation (IEA)Cellular component integral to membrane (IEA)Molecular function ATP binding (IEA) protein kinase activity (IEA) protein serine/threonine kinase activity (IEA) protein tyrosine kinase activity (IEA)
'F56D12.1'	'alh-6'	'"alh-6 is orthologous to the human gene ALDEHYDE DEHYDROGENASE 4 FAMILY, MEMBER A1 (ALDH4A1; OMIM:606811), biological process: locomotion (IMP) metabolic process, positive regulation of growth rate (IMP) positive regulation of locomotion (IMP) proline biosynthetic process (IEA) reproduction (IMP) Cellular component mitochondrial matrix (IEA)Molecular function 1-pyrroline-5-carboxylate dehydrogenase activity (IEA) oxidoreductase activity
'F40F9.10'	'...'	Has larval expression: pharynx; anal depressor muscle; body wall muscle;Adult Expression: pharynx; anal depressor muscle; body wall muscle;

'H28G03.6'	'mtm-5'	MTM-5 is expressed in adult pharynx, intestine, and body wall muscle, but has no obvious function in RNAi assays
'R07B7.11'	'gana-1'	'"R07B7.11 is orthologous to the human gene ALPHA-GALACTOSIDASE B (GALB; OMIM:104170), which when mutated leads to Schindler disease. Description: gana-1 encodes a protein with homology to both human alpha-galactosidase (alpha-GAL) and alpha-N-acetylgalactosaminidase (alpha-NAGA) enzymes; GANA-1 is expressed in body wall muscle and intestinal cells and in coelomocytes; Biological process: carbohydrate metabolic process, glycoside catabolic process , metabolic process
'T05C12.10'	'qua-1'	Biological process: cell communication , embryonic development ending in birth or egg hatching, locomotion (IMP) molting cycle, collagen and cuticulin-based cuticle (IMP) multicellular organismal development (IEA) nematode larval development (IMP) positive regulation of multicellular organism growth (IMP) proteolysis (IEA)Molecular function: peptidase activity (IEA)
'F48F7.1'	'alg-1'	'A homolog of rde-1 that is involved in RNA interference and affects developmental timing. Biological process: determination of adult life span (IMP) embryonic development (IGI) embryonic development ending in birth or egg hatching (IMP) hermaphrodite genitalia development (IMP) locomotion (IMP) molting cycle, collagen and cuticulin-based cuticle (IMP) nematode larval development (IMP) positive regulation of growth rate (IMP) positive regulation of locomotion (IMP) positive regulation of multicellular organism growth (IMP) vulval development
'C26C6.5'	'dcp-66'	Biological process: embryonic development ending in birth or egg hatching (IMP) growth (IMP) hermaphrodite genitalia development (IMP) locomotion (IMP) morphogenesis of an epithelium (IMP) negative regulation of vulval development (IMP) nematode larval development (IMP) positive regulation of growth rate (IMP) positive regulation of vulval development (IMP) reproduction
'F11E6.2'	'grl-24'	is expressed in body wall muscle and intestine; No gene ontology terms have been assigned to grl-24
'T01B10.2'	'grd-14'	Biological process: locomotion (IMP) positive regulation of multicellular organism growth (IMP) vulval development
'F37H8.5'	'...'	Gamma-interferon inducible lysosomal thiol reductase
'F53A9.10'	'tnt-2'	Biological process: locomotion (IMP) positive regulation of growth rate (IMP) positive regulation of locomotion (IMP) reproduction
'F42G4.3'	'zyx-1'	'"zyx-1 encodes a zyxin homolog that physically interacts with

		P granule components (GLH proteins); Biological process: reproduction
'C44H4.4'	'...'	Uncharacterized conserved protein
'Y38F1A.6'	'...'	Biological process: metabolic process (IEA) Molecular function: metal ion binding (IEA) oxidoreductase activity
'D2045.9'	'...'	Biological process: hermaphrodite genitalia development (IMP) lipopolysaccharide biosynthetic process (IEA) locomotion (IMP) morphogenesis of an epithelium (IMP) positive regulation of growth rate (IMP) reproduction
'Y75B8A.7'	'...'	Biological process: growth (IMP) hermaphrodite genitalia development (IMP) nematode larval development (IMP) positive regulation of growth rate (IMP) rRNA processing (IEA) reproduction
'T01C8.5'	'...'	Biological process amino acid metabolic process (IEA) biosynthetic process (IEA) positive regulation of growth rate
'F25H2.1'	'tli-1'	Description: none available
'ZK112.2'	'ncl-1'	ncl-1 encodes a B-box zinc finger protein that may be a repressor of RNA polymerase I and III transcription; has much larger neuronal nucleoli than normal
'F42A10.3'	'...'	Molecular function: methyltransferase activity
'F54C9.11'	'...'	Guanine nucleotide exchange factor
'F07A5.7'	'unc-15'	The unc-15 gene encodes a paramyosin ortholog; Biological function: body morphogenesis (IMP) carbohydrate metabolic process (IEA) growth (IMP) locomotion (IMP) muscle thick filament assembly (IMP) nematode larval development (IMP) oviposition (IMP) regulation of cytoskeleton organization and biogenesis
'F11A1.3'	'daf-12'	daf-12 encodes a member of the steroid hormone receptor superfamily that affects dauer formation downstream of the TGF- and insulin signaling pathways, and affects gonad-dependent adult longevity together with DAF-16. Is homologous to human VITAMIN D RECEPTOR. Biological process: negative regulation of multicellular organism growth (IMP) positive regulation of growth rate (IMP) regulation of development, heterochronic (IMP) regulation of transcription, DNA-dependent
'F38E11.2'	'hsp-12.6'	HSP-12.6 is required in vivo for normal lifespan; hsp-12.6 encodes a small heat-shock protein; HSP-12.6 is predominantly and ubiquitously expressed in L1 larvae without any obvious induction by stressors; but, in adult hermaphrodites, at least one HSP-12 is also expressed in spermatids (and perhaps spermatocytes), as well as in some vulval cells; hsp-12.6(RNAi) animals are shorter-lived than

		normal;
'F56H9.4'	'gpa-9'	gpa-9 encodes a member of the G protein alpha subunit family of heterotrimeric GTPases; it is expressed in ASJ, PHB, PVQ, pharyngeal muscle, and the spermatheca
'C40C9.1'	'twk-20'	Concise Description: none available; Biological process potassium ion transport
'T27E4.3'	'hsp-16.48'	hsp-16.48 encodes a 16-kD heat shock protein (HSP) that is a member of the hsp16/hsp20/alphaB-crystallin (HSP16) family of heat shock proteins Biological process determination of adult life span.
'E03D2.2'	'nlp-9'	Concise Description: none available
'C02F4.2'	'tax-6'	tax-6 encodes an ortholog of calcineurin A Biological process: chemosensory behavior (IMP) chemotaxis (IMP) dauer larval development (IGI) hyperosmotic response (IGI) locomotion (IMP) olfactory behavior (IGI) olfactory learning (IMP) positive regulation of growth rate (IMP) positive regulation of multicellular organism growth (IMP) thermosensory behavior (IMP) thermotaxis
'C36B7.7'	'hen-1'	'"hen-1 encodes a secretory protein that contains a low-density lipoprotein receptor class A domain. GFP reporter is expressed in pharyngeal muscles, the vulva, and weakly in a subset of neurons;Biological process: associative learning
'F42A10.2'	'nfm-1'	nfm-1 encodes a homolog of human merlin/schwannomin (NF2), which when mutated leads to neurofibromatosis.
'F11E6.5'	'elo-2'	'"The elo-2 gene encodes a palmitic acid elongase, homologous to polyunsaturated fatty acid (PUFA) elongases such as ELO-1 Biological process: positive regulation of growth rate
'T20B3.2'	'tni-3'	Biological process: muscle contraction (IMP) nematode larval development (IMP) oviposition (IMP) post-embryonic body morphogenesis (IMP) Cellular component: sarcomere

Legend: The blue colored names are down-regulated genes

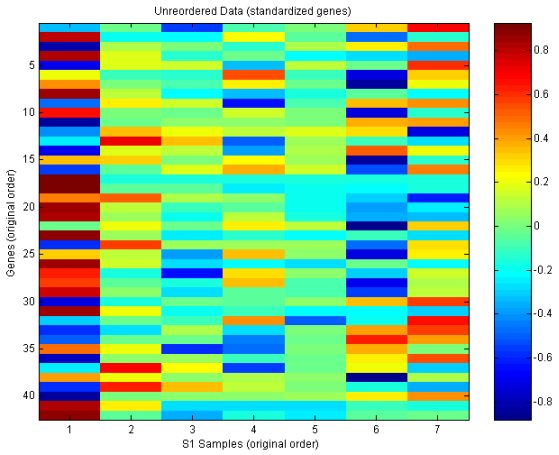
Table 1 List of 42 genes that might be representative for sarcopenia signature.

The next step should be to identify the gene patterns in this list. In this sense I clustered the list of 42 genes *C. elegans* using the SPC approach to identify 8 clusters, classified (see Chapter 1) based on size (number of genes in each cluster) and stability. The hierarchical organization of the data reflected in the way clusters split or merge has a graphical representation as a tree, called a dendogram (see fig 1). For details on the clustering algorithm see Chapter 1 as well as in E. Domany et. al,

Neural Computation (1997). The clusters or nodes I obtained were annotated as G1-G8, each with a distinctive pattern. Besides classifications of clusters based on the size and stability criterion mentioned above (found in size/stability table), I attempted a classification based on patterns of gene expression identified in each such cluster. The results of this clustering analysis were compiled for an easy access in a web- based design that facilitates their analysis.

The entire informational content of the web-based clustering design is displayed graphically or in tables. I will mention below some of the links which can be found in the main web page:

- heat-map graph with all the genes normalized before being clustered.



- PCA –a graph displaying the principal component analysis (see Chapter 1 for details and references on PCA method)
- Dendrogram with Stable Clusters –web based accessible dendogram
- Dendrogram next to Reordered Data (after clustering): hit-map graph & dendogram
- Reordered Genes : table with all 42 genes and the clusters where they fit.
- Samples : time points
- Parameters for SPC
- Access to each cluster for pattern visualization and gene members.

Each cluster can be accessed from the main web page and is represented graphically

in two plot formats: as a heat map and as gene expression level changes over time. In addition, a short description of the biological content, of each cluster, correspondence of the cluster with any other clusters, and the list of gene members found in the respective cluster is included. Two tables with clusters sorted based on the stability and size are also presented.

Among the 42 genes I depicted 8 clusters, using SPC clustering algorithm (see more about SPC in Chapter 1).

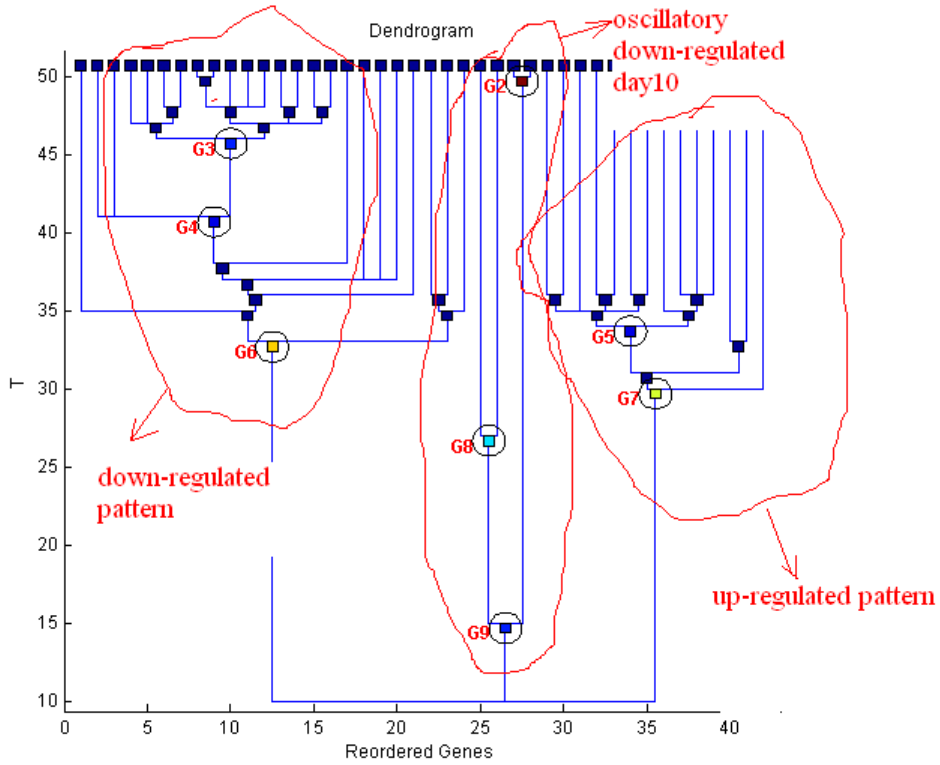
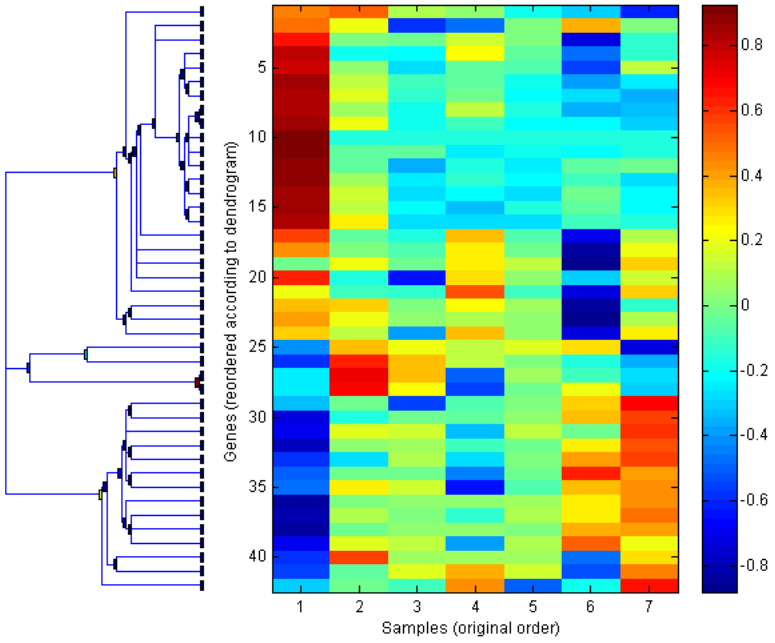


Fig. (1). Dendogram showing 8 clusters annotated G2-G9 and the pattern categories they enter

Fig.(1) shows 8 clusters annotated from G2 - G9; the dendogram shows the hierarchical organization found in the data based on SPC algorithm.

Below is the dendogram and the heat map graph of the 42 genes;



Legend: x-axis-7 time points: 1/day3; 2/day6; 3/day9; 4/day10; 5/day11; 6/day12; 7/day15; **y-axis**-42 gene expressions; color bar: from red-high gene expression too dark blue-low gene expressions.

Fig. 2 Dendogram and heat map of the 42 genes after clustering.

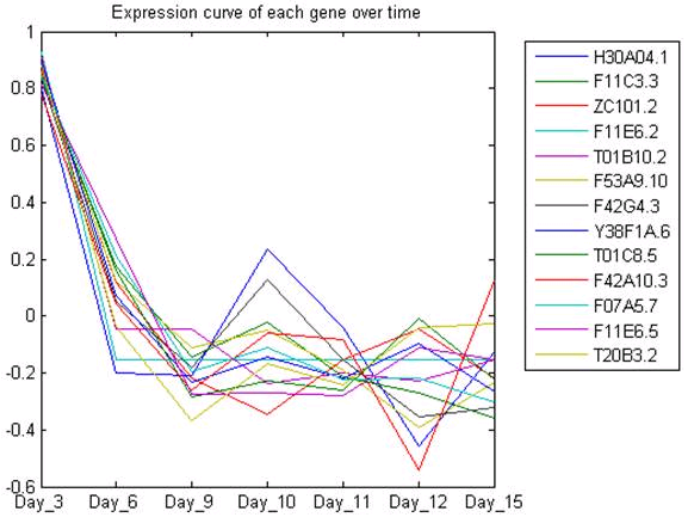
I depicted 3 main patterns in the clustered data: a down-regulated pattern, oscillatory pattern and up-regulated pattern. The dendogram in fig.1 shows the 8 clusters and the pattern category each cluster undergoes.

The down-regulated pattern –sarcopenia signature: direct and positive connection between sarcopenia signs and 24 genes that exhibits a down-regulated pattern. Genes are involved in developmental and locomotion.

The down-regulated pattern can be depicted in the cluster G6 and the 2 clusters that merge/split from G6: clusters G4 and G3. The main cluster G6 with down-regulated pattern has 24 genes, which is more than 50% of the genes in the list of 42 genes. A down regulation of expression of these 24 genes with age might be a direct reflection of the "age-related" loss of muscle mass leading to muscle

weakness. In this sense, the 24 genes might be the representative genes for sarcopenia signature. The genes can be found in Table 1 highlighted in blue.

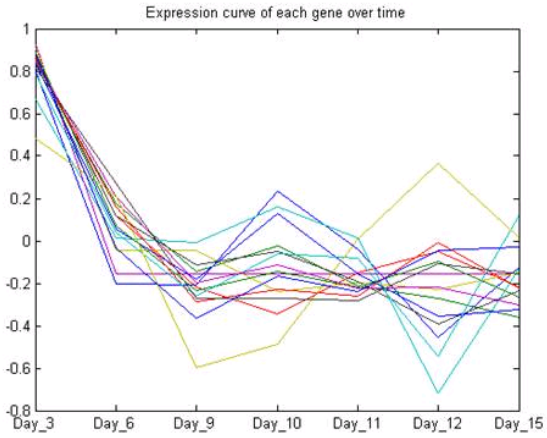
The common biological theme of the 24 genes is involvement in determination of adult life, locomotion, positive regulation of growth rate, larval developmental. As mentioned the down-regulated pattern of cluster G6 is maintained in clusters G3 and G4 as well as the common biological theme. Below are the down-regulated cluster patterns for the clusters G3 with the highest stability of size 13 and stability 5. The intermediate cluster G4 has size 15 and stability 3. The same common biological theme as in cluster G6 is preserved in G3 cluster also



Legend: x –axis: time points, y-axis: gene expressions., log2 was applied.

Fig. 3 **G3 cluster pattern: stability 5, size 13**

Also, below is cluster G4 pattern.



Legend: x –axis: time points, y-axis: gene expressions., log2 was applied.

Fig. 4 G4 cluster pattern: stability 3, size 15.

The emergent biological theme for the down-regulated pattern genes is the involvement in developmental and locomotion.

Given our finding of the 24 genes, with a decrease in gene expression, that have a role in locomotion and that the sarcopenia phenotype is defined by signs of decreasing in locomotion functions might be that this 24 genes could play an important role in sarcopenia. The fact that the phenotypic outcome over time of this nematode shows a clearly slow movement with age and that we notice a decrease in gene expression with age of the 24 genes involved in locomotion might be an expression of a direct and positive connection between sarcopenia signs and the 24 genes.

Oscillatory down-regulated day 10 pattern: is found in 3 clusters G8,G9,G2 of very small size of 4 and 2 genes but of very high stability: 35, 12 and 5. The cluster G2 below is the cluster with highest stability among the 3.

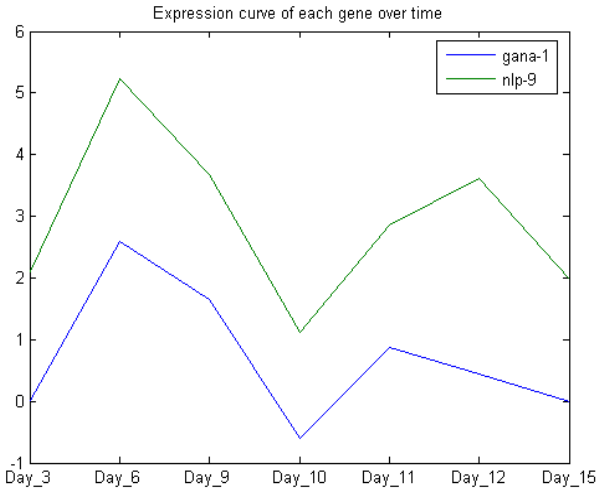


Fig. 5 G2 cluster stability 35: x –axis: time points, y-axis: gene expressions., log2 was applied.

The genes in this 3 clusters have in common the fact that are expressed in body wall muscle; 2 of genes: *mtm-5* and *hen-1* are expressed in pharyngeal muscle. Involved in metabolic process is *gana-1* which is also orthologous of human galactosidase. *mtm-5* has no obvious function in RNAi assays and *nlp-9* has no description of its function so far.

All 4 genes in cluster G9 exhibit an oscillatory gene expression pattern with a down pattern in gene expression at midlife time point, day 10.

Up-regulated pattern: can be noticed in the main cluster G7 and then in the cluster G5 which splits from G7. Cluster G7 has size 14 and stability 20. The cluster G5 has size 11 and stability 3.

Bellow the G7 cluster pattern can be seen.

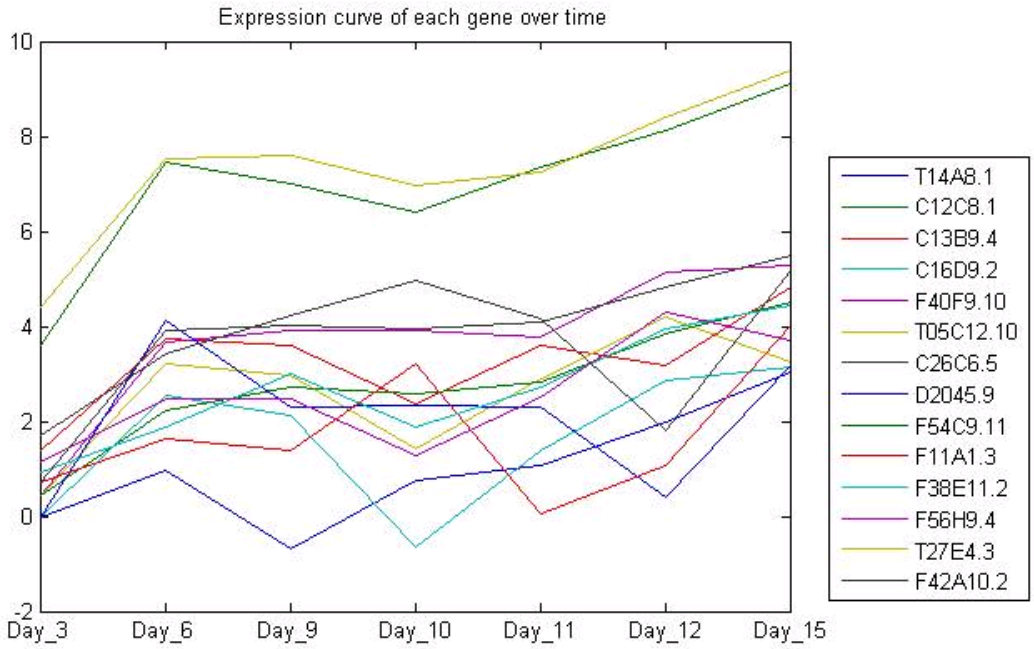


Fig. 6 G7 cluster of size 14 up-regulated pattern: x –axis: time points, y-axis: gene expressions., log2 was applied.

Most of the genes in cluster G7 are responsible for mutants defective in locomotion. Besides G7 contains *daf-12,* which encodes a member of the steroid hormone receptor superfamily homologous to the vitamin D receptor that affects dauer formation downstream of the TGF- and insulin signaling pathways. More important for the sarcopenia signature *DAF-12* together with *DAF-16* affects gonad-dependent adult longevity.

In G7 another human homolog can be found: *nfm-1,* which encodes a merlin/schwannomin (NF2), that when mutated leads to neurofibromatosis.

The genes in cluster G7 might be required to have an increase in gene expression over time as to diminish or delay the sarcopenia signs. A decrease in gene expression over life span of this nematode for the genes in the cluster G7 might induce more accentuated sarcopenia signs. On the other hand, an increase in gene expression over life span of the *C. elegans* for the down-regulated genes in cluster G6 the main cluster with a down regulated pattern might induce early signs of sarcopenia.

So far we discussed the study done on a list of genes identified (using AQL language) to be expressed in muscle cells.

We identified 24 genes with a down –regulated pattern that might have a direct connection with the sarcopenia process. These genes are mainly in the cluster G6. Also we identified 14 genes with an up-regulated pattern, found mainly in cluster G7. These genes might be required to have an increase in gene expression over time as to diminish or delay the sarcopenia signs or might have actually an opposite effect therefore the delay in sarcopenia signs appearances to happen if their expression would be decreasing. If this would be the case then this 14 genes are like 'leftover genes' for the sarcopenia process.

In the next section we performed a detailed study on a list of genes muscle-related which was compiled using Gene Ontology database. We are discussing this study in the next section.

5.4.2 Gene lists specifically involved in biological process, molecular function, or cellular component identified as muscle related.

Compiling gene lists involved in specific biological, cellular or functional muscle related processes.

To compile such a muscle related gene list I used Gene ontology database. The Gene Ontology (GO) project is a collaborative effort to address a consistent description of gene products in different databases.

The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions

In this sense, using Gene Ontology database (GO-term) I searched for genes involved in biological process, molecular function, or cellular component muscle- related.

I identified genes as major structural muscle genes: myosin or actin or genes involved in biological processes as muscle contractions see Table (2) below. We keep in mind that for WormBase, GO annotation is currently a "work in progress" therefore, we've used this list just as a "guiding" list and show no surprise when some searches results were incomplete. Particularly, for cellular component, where a protein is localized within a cell, the search of genes expressed in muscle cells using AQL language outputs many more genes. The advantages of using GO-term search is that it outputs genes based on the involvement in the biological process or molecular function as known in literature.

The number of genes found using GO_term search was 151 genes, but since the same genes were part of several biological or molecular processes, muscle involved, the total number of unique such genes was edited to 117 genes. Out of this, on our array we found a total of 95 genes, which are part of various biological processes as muscle development-the process whose specific outcome is the progression of the muscle over time, from its formation to the mature structure; or even more specific pharyngeal muscle development- the process whose specific outcome is the progression of the pharyngeal muscle over time, from its formation to the mature structure. See Table (2) for such genes.

Table 2: genes identified using GO-term: gene muscle-related involved in development and contraction.

Biological process: **muscle development:**

act-1	T04C12.6	An actin that affects body wall and pharyngeal muscle
eat-1	T11B7.4	eat-1 encodes a homolog of mammalian
hlh-1	B0304.1	hlh-1 encodes a basic helix-loop-helix (bHLH)
mlc-1	C36E6.3	encodes a muscle regulatory myosin light chain
syd-2	F59F5.6	syd-2 encodes alpha-liprin,
tmd-2	C08D8.2	
unc-52	ZC101.2	The unc-52 gene encodes perlecan, a protein orthologues
unc-89	C09D1.1	

Biological process: **pharyngeal muscle development**		
eff-1	C26D10.5	The eff-1 gene encodes a novel, type I transmembrane protein
glp-1	F02A9.6	glp-1 encodes an N-glycosylated transmembrane protein
pop-1	W10C8.2	pop-1 encodes an HMG box-containing protein
myosin:		
biological process: **muscle contraction:**		
egl-2	F16B3.1	egl-2 encodes a voltage-gated potassium channel
itr-1	F33D4.2	itr-1 encodes a putative inositol
jph-1	T22C1.7	jph-1 encodes a junctophilin,
pat-10	F54C1.7	pat-10 encodes body wall muscle troponin C,
twk-18	C24A3.6	twk-18 encodes one of 44 C. elegans TWK
unc-26	JC8.10	
myo-2	T18D3.4	myo-2 encodes a muscle-type specific myosin heavy
myo-3	K12F2.1	myo-3 encodes MHC A, the minor isoform of MHC

Other muscle related genes were found to be implicated in biological process as muscle cell fate specification, which is the process by which a cell becomes capable of differentiating autonomously into a muscle cell in an environment that is neutral with respect to the developmental pathway. Interestingly is that upon specification, the cell fate can be reversed as for example: mls-1 which encodes a T-box transcription factor orthologous to members of the Tbx1 subfamily of T-box transcription factors. MLS-1 is required for fate specification of the eight nonstriated uterine muscle cells generated during postembryonic development. Also ectopic expression of MLS-1 is sufficient for uterine muscle specification in other mesodermal lineages. Besides mls-1 reporter gene expression is detected in uterine progenitors and differentiated uterine muscles, type 2 vulval muscles, the left and right intestinal muscles, and the anal depressor muscle.

Another important biological process I included in my search was muscle contraction which is a process leading to shortening and/or development of tension in muscle tissue. Muscle contraction occurs by a sliding filament mechanism

whereby actin filaments slide inward among the myosin filaments. Major structural muscle genes as myosin and actin.

Besides muscle related genes involved in biological processes, using GO term I identified also genes involved in various biological functions as actin cytoskeleton organization and biogenesis by which we understand the assembly and arrangement of cytoskeletal structures comprising actin filaments and their associated proteins. See table 3 bellow for such genes.

act-1	T04C12.6	An actin that affects body wall and pharyngeal mus
act-4	M03F4.2	An actin that is expressed in body wall and vulval
cap-1	D2024.6	cap-1 encodes an F-actin capping protein alpha sub
cap-2	M106.5	The beta subunit of actin capping protein that reg
cyk-1	F11H8.4	The cyk-1 gene encodes a homolog of Drosophila
fhod-1	C46H11.11	
fhod-2	F56E10.2	
fozi-1	K01B6.1	K01B6.1 encodes a protein with a zinc-finger domain
pfn-1	Y18D10A.20	
pfn-3	K03E6.6	
tag-268	F58B6.2	
unc-53	F45E10.1	UNC-53 encodes at least five large (~1200-1600 residues
	F15B9.4	
	F56E10.3	
	Y48G9A.4	

Table 3 genes involved in actin cytoskeleton organization and biogenesis

Another muscle related biological function would be actin filament organization by which we understand control of the spatial distribution of actin filaments. This includes organizing filaments into meshworks, bundles, or other structures, as by cross-linking. Genes involved in this process would be: *ced-12, die-1, wve-1*

All 117 genes found based on the search using GO-term involved in various biological, molecular or cellular function are presented in several tables -at the beginning of each table I give the definition of the process in which the respective

genes are involved. Note that some genes have no description in these tables as GO-term data base is still a work in progress in *C. elegans* community.

Clustering the gene lists compiled in previous section

In order to identify the gene expression pattern of the 95 muscle related genes found on our chips, we normalized and clustered the genes using the SPC algorithm. For a description on the methods see Chapter 1. We obtained 10 clusters annotated from G1-G9 where G1 contains the entire data to be clustered. The clusters are classified based on size and stability as can be seen in Table 5 bellow:

G1 Size=95

G1(S1)	G2	G3	G4	G5	G6	G7	G8	G9	G10

G2 Stability=9 Size=5
G3 Stability=3 Size=4
G4 Stability=3 Size=10
G5 Stability=4 Size=30
G6 Stability=5 Size=5
G7 Stability=3 Size=7
G8 Stability=13 Size=42
G9 Stability=3 Size=19
G10 Stability=8 Size=47

As in previous analysis we compiled the results of the clustering analysis for an easy access in a web based design. The entire informational content of the web based clustering design is displayed graphically or in tables.

The hierarchical organization found in the data based on SPC algorithm is presented in the dendogram bellow.

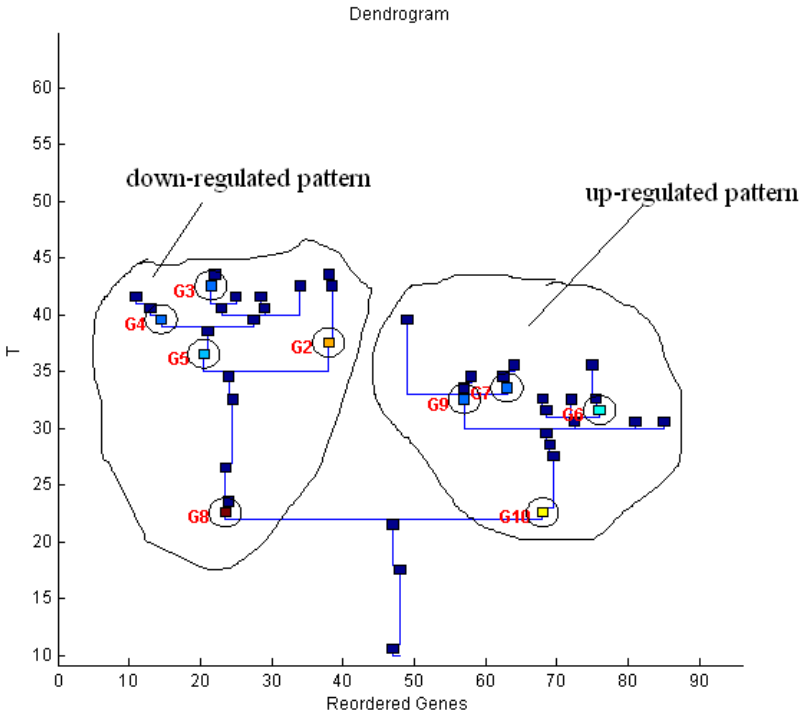
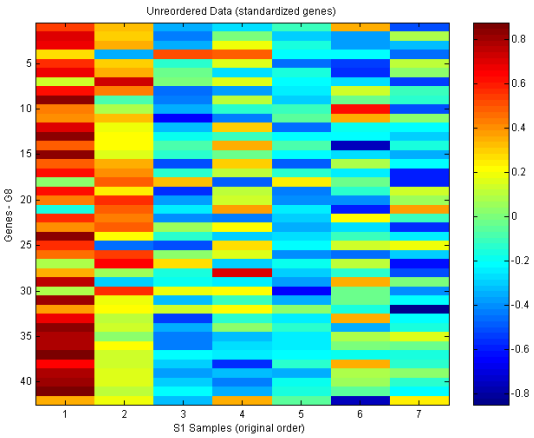


Fig. 7 Dendogram with gene nodes/clusters and patterns

We depicted 2 main patterns in the clustered data. The dendogram above shows the 10 clusters and the pattern category in which the clusters enter.

Most of the structural genes are in the down-regulated pattern found in the cluster G8 of size 42 and stability 13 and all the clusters which split from it: G5,G4,G3,G2.

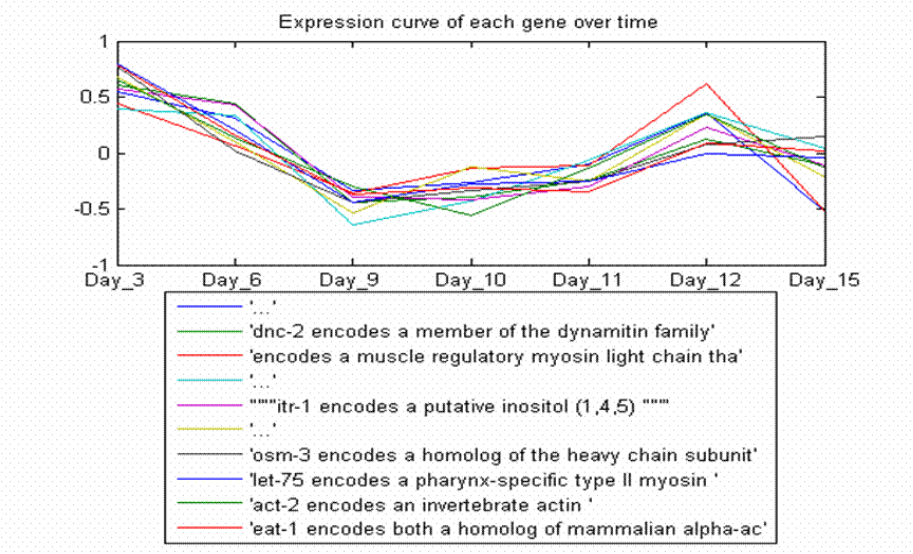


x-axis-7 time points: 1/day3; 2/day6; 3/day9; 4/day10; 5/day11; 6/day12; 7/day15;**y-axis**-42 gene expressions; color bar: from red-high gene expression too dark blue-low gene expressions.

Fig. 8 Heat map cluster G8 -down-regulated pattern

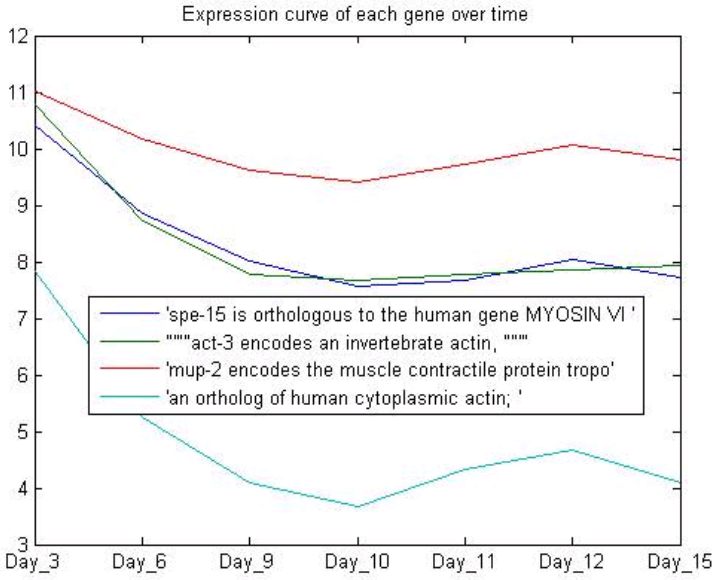
As can be seen in the heat map above of the cluster G8, the gene expression over the first 2 time points (day3, day 6) are higher expressed than the gene expression for the rest of the time points: day 9,10,11,12,15. At the same time it should be mentioned that the gene expression for the rest of the time points 9,10,11,12,15 have an relative steady down -regulated pattern. This pattern is maintained in all clusters that merge from cluster G8 and mentioned before: G5,G4,G3,G2.

Below can be seen the gene expression pattern for some of the clusters that merge from cluster G8: clusters G4 and G3



x –axis: time points, y-axis: gene expressions., log2 was applied.

Fig. 9 G4 cluster gene expression pattern over time points.

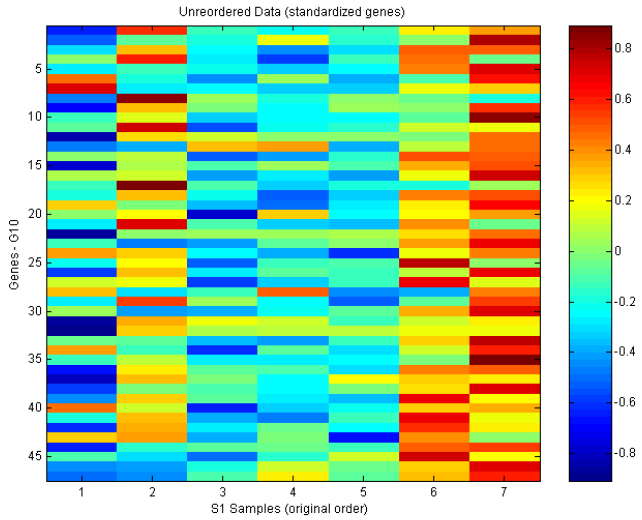


x –axis: time points, y-axis: gene expressions., log2 was applied.

Fig.10 **G3 cluster: gene expression pattern over time points**

As mentioned, most of the structural genes are in down-regulated clusters pattern. **Structural muscle related genes show 'positive connection' with the sarcopenia phenotype, meaning they have the down-regulated gene expression pattern one might expect for a sarcopenia signature.**

In the **up-regulated pattern** found in G10 **most of the unc genes** and some human homologues genes **are included**. The emerging clusters from G10 are: G9,G7,G6. They do maintain same pattern as G10 cluster.



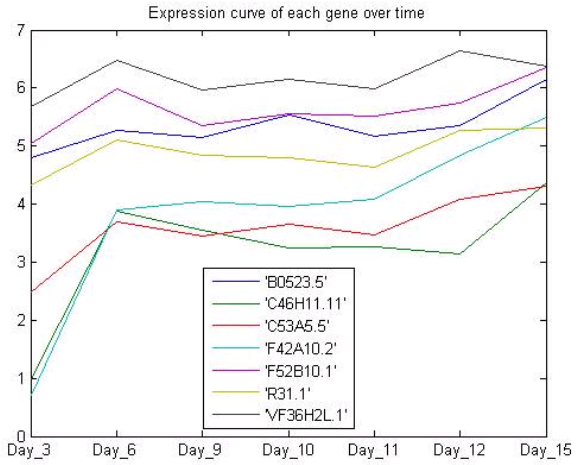
Legend: x-axis-7 time points: 1/day3; 2/day6; 3/day9; 4/day10; 5/day11; 6/day12; 7/day15; **y-axis**-42 gene expressions; color bar: from red-high gene expression too dark blue-low gene expressions.

Fig. 11 Heat map-cluster G10 down-regulated pattern size 47, stability 13.

In the heat map of the cluster G10, gene expression for the last 2 time points (day12, day 15) are expressed higher than the gene expression for the rest of the time points: day 3,6,9,10,11.

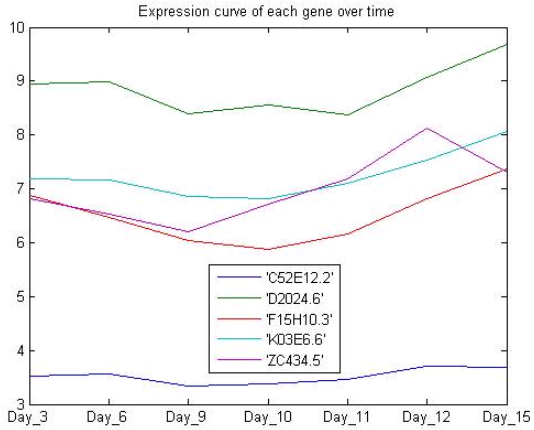
It should be mentioned that the gene expression for the rest of the time points day 3,6,9,10,11 show an almost steady down-regulated pattern. This pattern is maintained in all clusters that merge from cluster G10: G9,G7,G6.

Below the gene expression pattern for the clusters: G7 and G6 can be seen.



x –axis: time points, y-axis: gene expressions, log2 was applied.

Fig. 12 G7 up-regulated cluster: gene expression pattern over time points



x –axis: time points, y-axis: gene expressions, log2 was applied.

Fig.13 G6 up-regulated cluster: gene expression pattern over time points.

The fact that in up-regulated pattern clusters I identified mostly unc genes might be again a signature of sarcopenia. Might be that the unc genes for the fitness of the muscle acts as 'leftover' genes. In this sense, mutants with a decrease in gene

expression for the genes in the up-regulated cluster G10 might slow the sarcopenia signs.

To conclude the GO-term list analysis, the gene lists specifically involved in biological process, molecular function, or cellular components considered to be muscle related can be classified in two categories of gene expression pattern : of up-regulated and down-regulated genes. In the down-regulated group we found more structural muscle related genes like actin or myosin by difference with the up-regulated group where we can see more of unc-related genes as well as human homologues genes.
The decrease in gene expression for structural genes might help promoting sarcopenia signs. In the same time the increase in gene expression for unc genes identified in cluster G10 as far as concerning sarcopenia, might be a sign of 'leftover' genes, in the sense that mutants with a decrease in gene expression for genes in cluster G10 might improve the the reduction in muscle mass and any other sarcopenia signs in general.

5.4.3 Analysis of genes expressed in young muscle

In 2002 an experiment performed by Kim's group identified gene expressed in C. elegans muscle (see Kim, et al. Nature 2002);

In this experiment a poly A binding protein was expressed only in muscle. Muscle messages were then isolated by imunoprecipitation. They used DNA microarrays to analyze the ratio of the mRNA enriched by co-immunoprecipitation with FLAG::PAB-1 relative to the mRNA present in the starting cell-free extract. Fluorescently-labeled probes were then hybridized to DNA microarrays containing 90% of the 19,733 genes currently estimated in the *C. elegans* genome. L1 larvae were studied with 6 repeats for ~19000 genes.

As statistical method used, they've computed a ranking for all genes and a percentile rank for every gene from the 6 repeats. Then, the percentile rank of enrichment for every gene from the six repeats was averaged together.

They considered that genes that are not enriched by mRNA-tagging should have an average percentile rank of about 50%, while genes expressed in muscle should have a rank significantly higher.

After that they performed a Student's t-test and identified 1364 genes that are significantly enriched in the muscle mRNA-tagging experiments for $p<0.001$. See the graphs bellow.

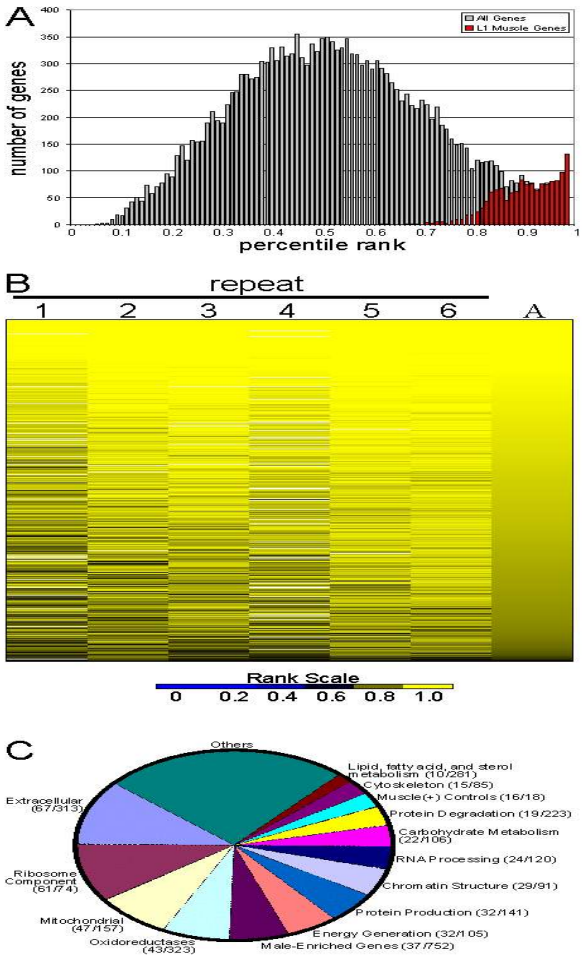


Fig. 14- from Kim et. al. Nature 418: 975-979: 2002;

In the pie chart above can be seen the biological classification of the 1364 genes identified by t-test and enriched by co-immunoprecipitation with FLAG::PAB-in the L1 stage of life. **The 1364 genes are part of almost every known biological function**.

Analysis of muscle enriched genes identified by Kim et.al; changes in adult life

Out of 1364 genes from Kim et. al. experiment we identified 1187 genes on our chips.

We looked for gene expression patterns in the list of 1187 genes and we analyzed the intersection between 1187 muscle expressed genes and the list of 2000 genes from our experiment which have the highest variation across tie points. We identified 111 genes at the intersection of the 2 lists. These 111 are genes expressed in L1 muscle that show greatest variance sometime during adulthood in our experiment.

When we analyzed the gene expression patterns in the list of 1187 using SPC algorithm we identified 28 clusters. (For details on the clustering algorithm see Chapter 1 as well as M. Blatt, S. Wiseman and E, Domany, Neural Computation (1997)).

The Fig. 4 below shows the dendogram that reflects the clustering hierarchy identified in the data as well as the heat map graph for the 1187 genes as found on our arrays.

Blue means low gene expression, red is high gene expression.

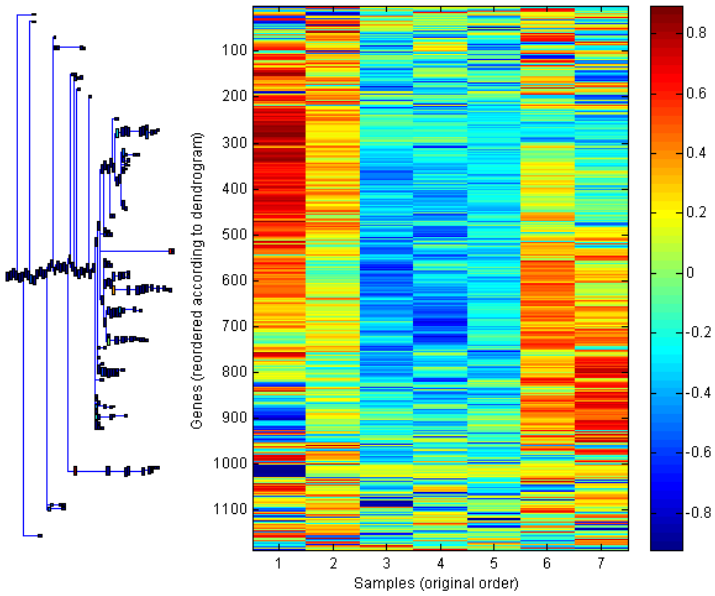


Fig. 15 Heat map and dendogram of the 1187 genes corresponding to the 1364 genes identified by Kim et. al. as expressed in muscle

Given that the 1364 genes identified by Kim group are part of almost every known biological function came as no surprise that same biologically consistency we identified in our list of 1187. In the same time similar patterns identified when analyzed the 2000 list of genes which vary most in our data were found when analyzed the list of 1187 genes.

Fig. 16 bellow shows the dendogram with hierarchical organization of the clusters and the 5 category patterns they enter.

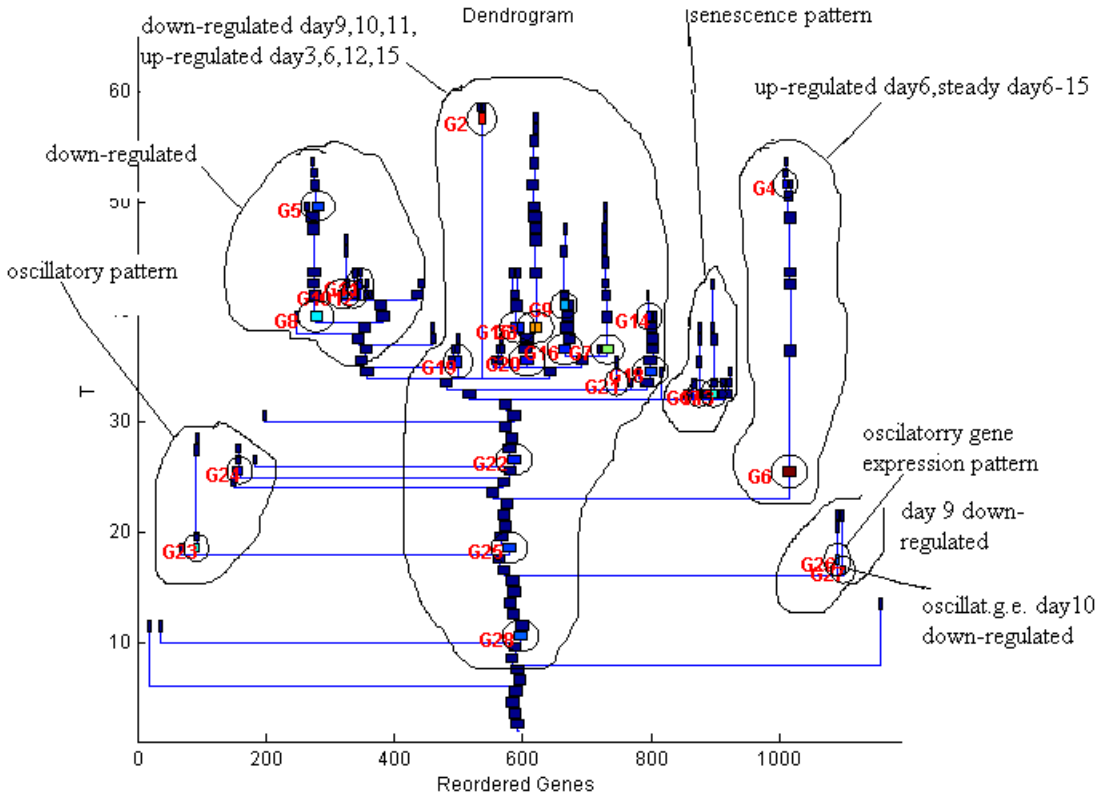


Fig. 16 dendogram – 28 clusters and the 5 category patterns they enter

The 28 clusters can be identified in five gene expression patterns:

1) oscillatory pattern,
2) down-regulated pattern,
3) down-regulated mid life time points: day9,10,11 and up-regulated gene expression pattern for day3,6,12,15
4) "Senescence pattern"-oscillatory low expressed day3-12, up-regulated day12-15
5) a)Up-regulated pattern

   b)"Developmental pattern"-steady state day6-15, up-regulated day3-6

1) oscillatory gene expression pattern (4 clusters with this pattern)- mostly genes with unknown protein function and growth defect.

Cluster G26 has size 8 and stability 6 with a down-regulated peck at day 9. See graph below.

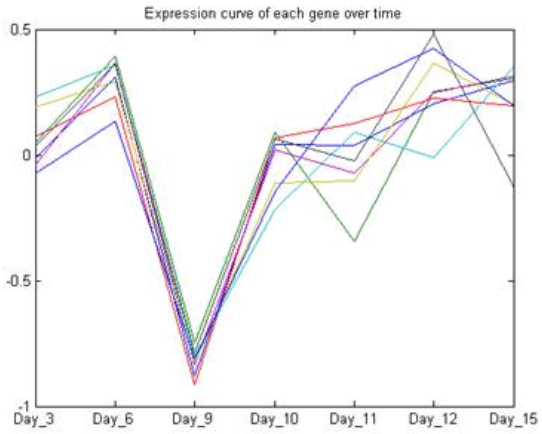


Fig. 17- G26 cluster

In this cluster we have many unknown function genes, genes involved in growth defect as RNAi phenotype and cuticulin component related gene. See below Table with cluster members.

1	'F41G4.7'	F41G4.7 Protein of unknown function
2	'C27B7.6'	C27B7.6 Member of the protein phosphatase protein family
3	'T06A10.1'	T06A10.1/mel-46 ; RNAiphenotype:growth defect
4	'W01H2.2'	W01H2.2 Protein of unknown function
5	'F55C12.4'	F55C12.4 Protein of unknown function
6	'F53F1.1'	F53F1.1 Protein with strong similarity to C. elegans cut-1 (Cuticulin component)
7	'Y71H9A.3'	Y71H9A.3 Member of the stomatin protein family
8	'H13N06.2'	H13N06.2 Protein with weak similarity to C. elegans mup-4 (Member of the EGF-repeat protein family)

Table 4: G26 cluster members.

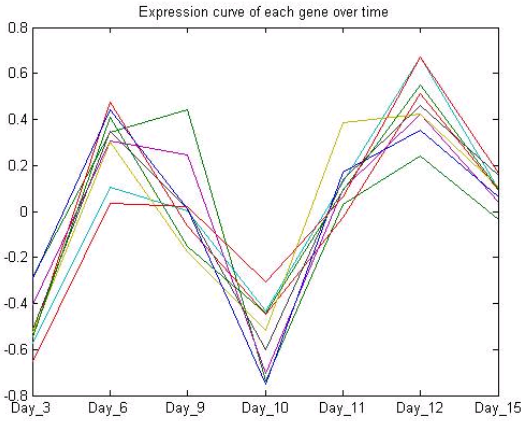


Fig. 18 **G27 cluster, size 10, stability 6.**

Again in this cluster we have mostly genes that have an unknown protein functions, a slow growth gene, and a transcription factor gene. See Table x bellow for cluster members in the G27 cluster.

1	'M02F4.1'	"M02F4.4 Protein of unknown function, has strong similarity to C. elegans R04D3.3 "
2	'ZK54.1'	is on Chromosome X;Protein of unknown function; has SNP's ;
3	'M02F4.1'	"M02F4.4 Protein of unknown function, has strong similarity to C. elegans R04D3.3 "
4	'W08A12.2'	W08A12.2 Protein of unknown function
5	'F28A12.3'	"F28A12.3 Protein of unknown function, has strong similarity to C. elegans F35C5.11 "
6	'ZC334.2'	"ZC334.2 /ins-30;Protein of unknown function, has weak similarity to C. elegans ZC334.3 "
7	'C08E3.1'	C08E3.1 Member of an uncharacterized protein family
8	'E03A3.3'	E03A3.3/his-69; Member of the histone H3 protein family; RNAi phenotype-slow growth
9	'C33D12.1'	"C33D12.1/ceh-31 Homeodomain transcription factor, has similarity over 121 amino acids to D. melanogaster B-H1 (BarH1) homeodomain transcription factor "
10	'C25G4.7'	is on Chromosome IV; "C25G4.7 Protein of unknown function, has strong similarity to C. elegans ZK973_14.D "; has SNP's ;

Table 5- G27 cluster members

Same oscillatory pattern is maintained in the clusters G23 and G24. **In this clusters can be found also member of the chaperonin complex protein family, heat shock proteins.**

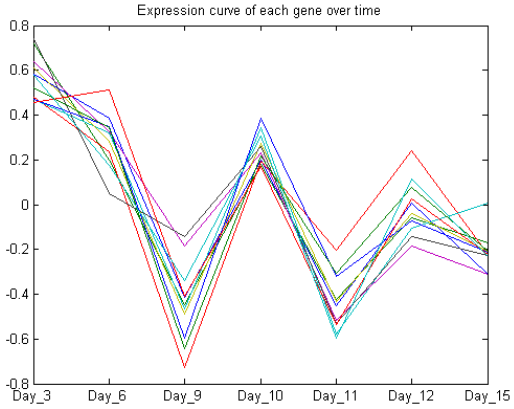


Fig. 19 Cluster G23-0x axis- time points; 0y axis- normalized gene expression

2) down-regulated pattern: pattern seen in 5 clusters-mostly collagen genes

G8

Down-regulated

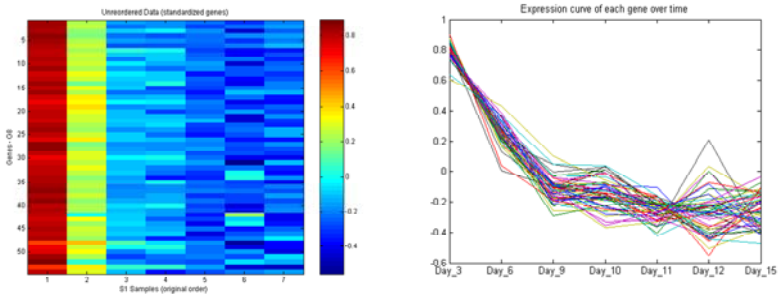


Fig.20 Cluster G8-size 54, stability 10, right: heat map graph, left: normalized gene expressions.

Same pattern is maintained also in clusters G5,10,11,12.

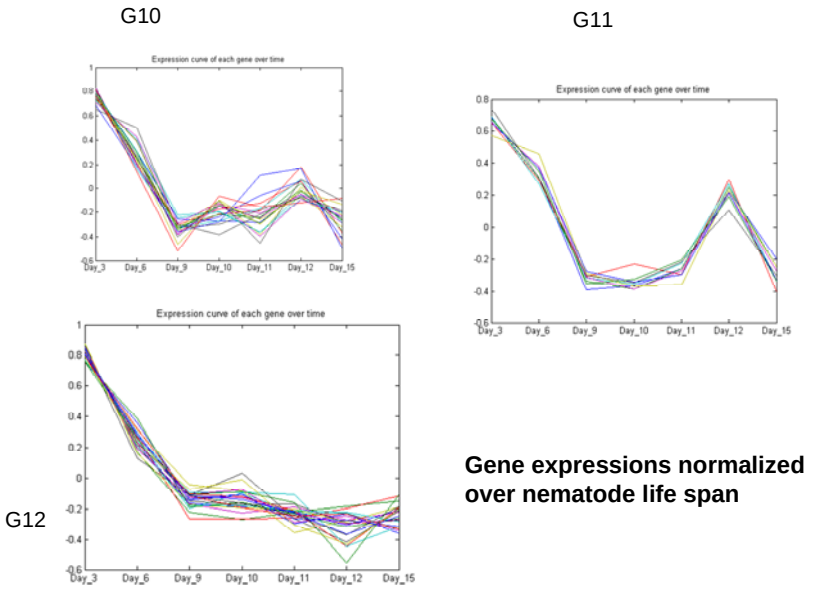


Fig. 21 Cluster pattern for G12,G10,G11

The cluster G8 the biggest size cluster with a down-regulated pattern has mostly collagen related genes, a few genes with uncharacterized protein function and several structural muscle related, no more than 4 genes out of 54 gene expressions, the size of the G8 cluster.

3) **down-regulated mid life time points: day9,10,11 and up-regulated gene expression pattern for day3,6,12,15-pattern found in 15 clusters-mostly ribosomal related genes.**

This pattern can be seen in 15 clusters out of 28 clusters we identified. ~ 50% of clusters have a low expression pattern for day 9,10,11 the mid life time span of this organism.

The overwhelming theme in these clusters is the ribosomal proteins and genes with unknown protein function. The heat map of the cluster G20 which is representative for this pattern is presented bellow:

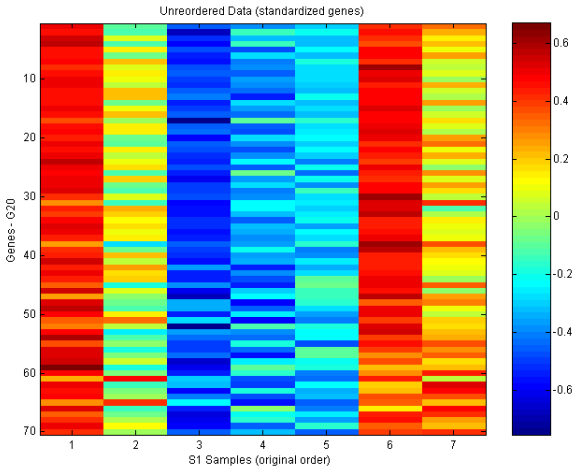


Fig. 22 **Heat map of the G20 cluster; x-axis**-7 time points: 1/day3; 2/day6; 3/day9; 4/day10; 5/day11; 6/day12; 7/day15; **y-axis**-42 gene expressions; color bar: from red-high gene expression too dark blue-low gene expressions.

The lower gene expression for the time points 3,4,5 = day9,10,11 can be clearly seen in the heat map above. Same pattern is maintained in the rest of 14 clusters.

4) Senescence pattern- oscillatory low expressed- day3-12, up-regulated day12-15. Pattern noticed in one cluster-genes mostly involved in growth defect.

Cluster G17 has this pattern**.**

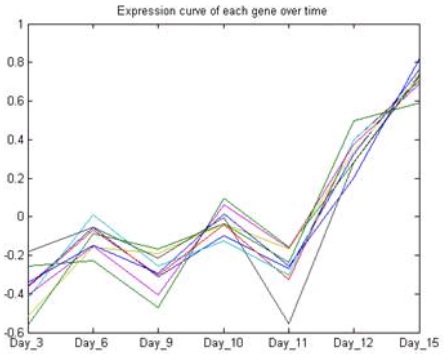


Fig. 23 **Cluster G17 size 9 stability**

Cluster G17 has mostly genes involved in growth rate. See cluster members, in Table 7 below.

1	'F21D5.6'	F21D5.6 Protein of unknown function
2	'C16A3.3'	"C16A3.3 Protein with strong similarity to S. cerevisiae Rrp5p, an essential protein required for processing of pre-rRNA to 18S and 5.8S rRNA "; RNAi phenotype: Egl Emb Gon Gro Lva; biological process:embryonic development gonad development growth (IMP) larval development oviposition (IMP) positive regulation of growth rate (IMP) RNA processing
3	'F33D4.5'	F33D4.5 Protein of unknown function/ RNAi phenotype-embryonic & post-embryonic defect, lethal
4	'C34B2.5'	C34B2.5 Protein with moderate similarity to human TTC1 (tetratricopeptide repeat domain 1);Larval Arrest-Late (L3/L4)
5	'K08F4.1'	"K08F4.1 Protein with strong similarity to S. cerevisiae Ctf18p, which is required for chromosome transmission and maintenance of normal telomere... "; homologies with DNA helicases in H.Sapiens; function:DNA replication
6	'ZK632.3'	ZK632.3 Member of the RIO1/ZK632.3/MJ0444 protein family
7	'M04B2.3'	"M04B2.3 Protein with strong similarity human GAS41/Hs.4029, which is amplified in glioma cells "
8	'F16A11.2'	F16A11.2 Member of the uncharacterized UPF0027 protein family; RNAi phenotype: slow growth; developmental delay
9	'ZK930.2'	ZK930.2 Protein of unknown function

Tabele 6. G17 cluster members.

5) a) Up-regulated pattern

- pattern found in one cluster-size 14-same biological content as G17-mostly genes involved in growth defect

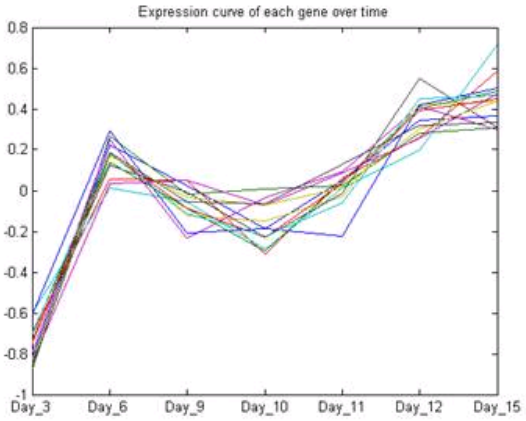


Fig. 24 **Cluster G13 size 14, stability 11.**

Biological theme is the same as for cluster G17. Most of the genes are involved in growth defect, **developmental delay, embryonic &postembryonic defect, larval arrest in L3/L4 stage** as can be seen in Table x bellow. Cluster G13 has size 14 and stability 11.

1	80	'F54F2.7'	F54F2.7 Protein of unknown function; RNAi phenotype: growth defect
2	453	'E04F6.9'	"E04F6.9 Protein of unknown function, has moderate similarity to C. elegans E04F6.8 "
3	540	'W06E11.2'	W06E11.2 Protein of unknown function;RNAi phenotype: growth defect
4	553	'F19B2.5'	F19B2.5 Protein with strong similarity to C. elegans F19B2.G
5	599	'B0212.1'	"B0212.1 Protein of unknown function, has strong similarity to C. elegans F14B8.5 "
6	613	'F42H10.4'	F42H10.4/cal-2; Member of the LIM domain containing protein family;
7	672	'C18E9.1'	C18E9.1 Member of an uncharacterized protein family; RNAi phenotype: embryonic defect
8	718	'R05D11.3'	"R05D11.3 Putative nuclear transport factor, has similarity to human NTF-2 (nuclear transport factor 2) embryonic defect
9	730	'F09F7.7'	F09F7.7 Protein of unknown function
10	961	'T27E4.1'	T27E4.1 Protein of unknown function; is part of cluster aging of 164 genes, see Lund et.al '02
11	1066	'C54E10.6'	"C54E10.6 Small protein containing a CHROMO (CHRromatin Organization MOdifier) domain and a coiled-coil region, has similarity to C. elegans CEC-... "
12	1083	'R05G6.10'	R05G6.10 Protein containing an N-terminal RasGEFN domain

			and a C-terminal RasGEF domain; has similarity over C-terminal half to CDC25-like GDP/G...biological process: intracellular signaling cascade
13	1088	'F45E12.6'	F45E12.6 Protein of unknown function
14	1120	'T22B11.4'	"T22B11.4 Protein of unknown function, has weak similarity to human kinase scaffold protein GRAVIN (Hs.788) "

Table 7. cluster G13 gene members

5) b) **"Developmental up-regulated pattern"-steady state day6-15, up-regulated day3-6.Two clusters are sharing this gene expression pattern.**

Cluster G6 and G4 share this pattern.

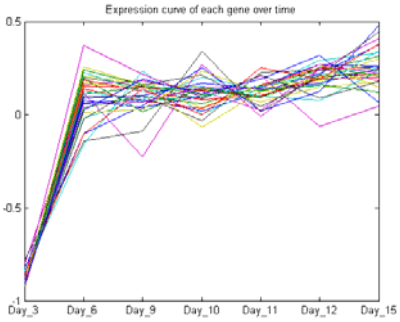


Fig. 25 cluster G6 size 29, stability 28, normalized gene expressions

The cluster G6 has size 29 and very high stability 28.

The gene expression pattern is characterized by high gene expression for day6-day15 and low expression at day3.

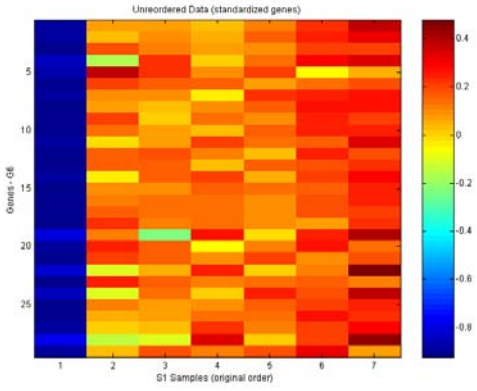


Fig. 26 **Heat map of cluster G6.**

According with expression pattern the gene cluster members might be of importance for the entire life of *C. elegans* right after development time, day3. Interestingly this cluster has a few genes related with proteins known to be human similar. Many of the genes have an unknown protein functions.

Below is the table with gene members in cluster G6.

1	'C40H1.5'	C40H1.5 Member of an uncharacterized protein family
2	'F25H8.5'	"F25H8.5 Putative paralog of dur-1, protein with weak similarity to H. sapiens SNCB (synuclein, beta) "
3	'C08F8.7'	C08F8.7 Ras-related GTP-binding protein of the ras superfamily
4	'F44G4.6'	F44G4.6 Protein of unknown function
5	'W01F3.3'	W01F3.3 Member of the EGF-repeat protein family
6	'F59F5.6'	F59F5.6 Member of liprin (LAR-interacting protein) family of proteins
7	'C13C12.1'	C13C12.1 Calmodulin
8	'F42H10.3'	F42H10.3 Member of the src homology domain 3 protein family
9	'F41B5.1'	F41B5.1 Protein of unknown function
10	'C28H8.2'	C28H8.2 Protein of unknown function
11	'F19F10.9'	"F19F10.9 Putative antigenic peptides, has strong similarity to H. sapiens SART1 gene product [squamous cell carcinoma antigen recognised by T ce... "
12	'C17G10.6'	"C17G10.6 Protein of unknown function, contains a C-terminal ShKt (toxin) domain, has weak similarity over middle region to human TGN51 (trans-Go... "
13	'F27D9.8'	"F27D9.8 Protein with strong similarity to human Hs.172278 protein, beta2-syntrophin "
14	'M05D6.4'	M05D6.4 Member of the esterase protein family
15	'C18B10.3'	"C18B10.3 Protein with similarity to G-protein coupled receptors of an unnamed subfamily, no homolog found in human or D. melanogaster, may have ... "
16	'ZC239.6'	ZC239.6 Member of an uncharacterized protein family
17	'F43C1.3'	F43C1.3 Protein with weak similarity to S. cerevisiae HIT1 (Protein required for growth at high temperature)
18	'C34C6.2'	C34C6.2 Protein of unknown function
19	'K02F3.4'	"K02F3.4 Protein with weak similarity to H. sapiens CEBPG (CCAAT/enhancer binding protein (C/EBP), gamma) "

20	'T13C2.3'	T13C2.3 Protein with weak similarity to C. elegans Y97E10AR.E gene product
21	'C45H4.17'	Y5H2B.F Protein with similarity to cytochrome P450; putative ortholog of C. elegans C45H4.2
22	'VF39H2L.1'	"VF39H2L.1 Protein with weak similarity to human syntaxin 7 (STX7), has weak similarity to C. elegans F36F2.4 "
23	'T14B4.6'	T14B4.6 Collagen of the collagen triple helix repeat (20 copies) family
24	'C17H1.7'	C17H1.7 Member of an uncharacterized protein family
25	'F34H10.1'	F34H10.1 Member of the ubiquitin protein family
26	'Y40B10B.1'	Y40B10B.1 Member of an uncharacterized protein family
27	'F56A3.1'	F56A3.1 Protein of unknown function
28	'F09E8.2'	"F09E8.2 Protein containing EGF-like repeats, has weak similarity to human low density lipoprotein receptors and D. melanogaster TEN-1 (tenascin) "
29	'F54F2.6'	F54F2.6 Protein of unknown function

Table 8. gene members cluster G6.

The same pattern characterized by high gene expression for the time points 2-7, which corresponds to day 6-day 15 and a low gene expression day 3 (see fig. 26) is maintained in the cluster G4 which splits from G6.

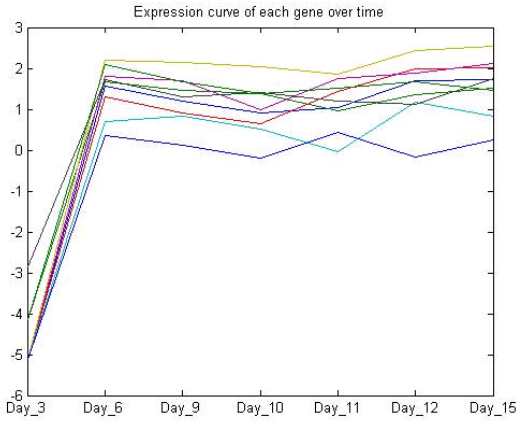


Fig. 27 **Cluster G4, size9, stability3, normalized gene expressions**

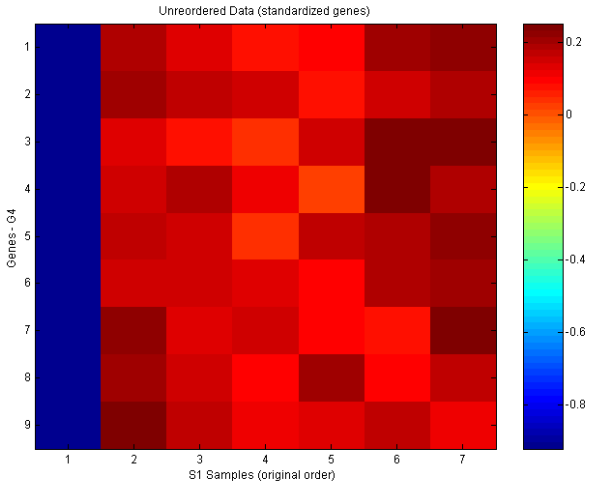


Fig. 28 Heat map cluster G4: x-axis-7 time points: 1/day3; 2/day6; 3/day9; 4/day10; 5/day11; 6/day12; 7/day15; **y-axis**-42 gene expressions; color bar: from red-high gene expression too dark blue-low gene expressions.

Given the expression pattern of cluster G4, the genes in this cluster might be relevant for the time right after development of *C. elegans* from day 6 up to day 15. The high gene expression pattern for most of the life of this organism might imply that this genes play an important role in the maintenance of vital functions for this

nematode. This cluster has again a few genes related with proteins known to be human similar.

1	386	'C08F8.7'	C08F8.7 Ras-related GTP-binding protein of the ras superfamily
2	525	'F59F5.6'	F59F5.6 Member of liprin (LAR-interacting protein) family of proteins; biological process: muscle development
3	666	'C28H8.2'	C28H8.2 Protein of unknown function
4	761	'C17G10.6'	"C17G10.6 Protein of unknown function, contains a C-terminal ShKt (toxin) domain, has weak similarity over middle region to human TGN51
5	787	'F27D9.8'	"F27D9.8 Protein with strong similarity to human Hs.172278 protein, beta2-syntrophin "
6	956	'F43C1.3'	F43C1.3 Protein with weak similarity to S. cerevisiae HIT1 (Protein required for growth at high temperature)
7	970	'C34C6.2'	C34C6.2 Protein of unknown function
8	1005	'C45H4.17'	Y5H2B.F Protein with similarity to cytochrome P450; putative ortholog of C. elegans C45H4.2; biological process:electron transport
9	1012	'T14B4.6'	T14B4.6 Collagen of the collagen triple helix repeat (20 copies) family

Table 9. G4 cluster members

Given that genes that might have an important role in the life of this organism have similarities with human proteins brings again another argument for the importance of using *C.elegans* as animal model for aging studies.

Concluding the clustering analysis the majority of the genes proposed by Kim et.al. and identified on our arrays, a list of 1187 genes, can be found in 2 main pattern categories. The decreasing pattern category and the pattern category with low gene expression over mid life time of the nematode. The decreasing pattern is consistent with what we might consider a sarcopenia signature and consists of mostly collagen- related genes. This pattern is seen in 5 clusters. The pattern of low gene expression over the mid life has mostly ribosomal genes. 14 clusters out of 28 clusters, includes these, we identified from clustering the entire list of 1187 genes ~ 50% out of all clusters. The rest of the genes enter in 3 other pattern categories of 'senescence', 'developmental' and oscillatory.

In general, the cluster patterns found from analysis of the list of 1187 genes proposed by Kim group resemble a lot the cluster patterns we found when analyzing the clusters of 2000 gene list of the 'wild type' *C. elegans* (see Chapter 1 results).

We might conclude that analysis of the experiment performed by Kim et.al brought us valuable information about *C. elegans* in general and development for the larval L1 of stage. For more insights on muscle related genes and sarcopenia as a process new experiments should be designed.

Intersection between 1187 gene list and 2000 gene list

We've looked also at the intersection between Kim list and our list of 2000 genes (see Chapter 1) and we found 111 genes that are both expressed in L1 muscle and have high variation in gene expressions with time.

The lists analyzed in Kim et.al. experiment are schematized in the Fig. x below.

The pie chart in Fig. x shows the biological composition in the list of 111 genes.

41% have collagen related genes, 11% are genes with human homologies, 12% are muscle related genes, and, 36% genes with unknown protein related function.

Analysis performed using data from Kim, et.al 2002. *Chromosomal clustering of muscle-expressed genes* Nature **418**: 975-979;

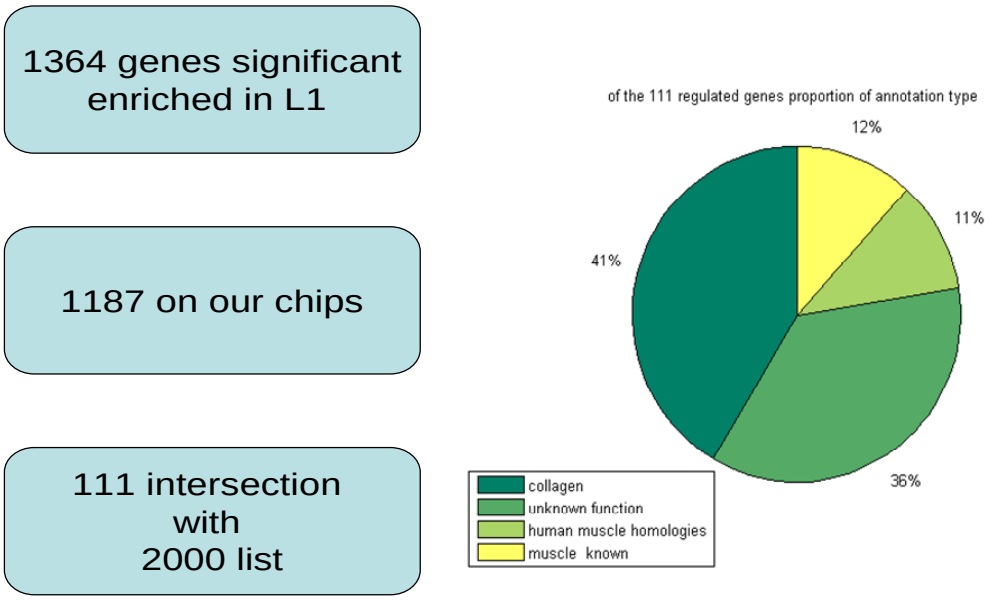


Fig. 29 Lists of genes used in Kim experiment analyzes.

The total genes with muscle or human homologies is 25. See Table 10 below for these 25 genes.

F46H5.3 Member of the arginine kinase, phosphotransferase protein family

F07A5.7 Paramyosin, major component of muscle filaments, structural equivalent of the rod region of myosin heavy chains

F53A9.10 Troponin T, putative paralog of C. elegans F53A9.10 protein; human homologus: Splice Isoform 8 of Troponin T, cardiac muscle
T22E5.5 Putative troponin-T, has strong similarity to C. elegans MUP-2 and D. melanogaster UP (upheld) troponin-T proteins; human homologus-Splice Isoform 8 of Troponin T, cardiac muscle

T25F10.6 Putative paralog of C. elegans UNC-87 which encodes muscle thin filament-associated protein

F42G4.3 ZYX-1 is potentially orthologous to the human gene LIPOMA-PREFERRED PARTNER (LPP, OMIM:600700). [detailsProtein with strong similarity to members of the LIM domain containing protein family

W05G11.6human homolog; Member of the phosphoenolpyruvate carboxykinase protein family

H22K11.1 Probable aspartyl protease and an ortholog of human cathepsin D

Y38F1A.9 Putative member of Immunoglobulin superfamily

F11C3.3 Sarcomeric Myosin Heavy Chain, major component of thick filaments in body-wall muscle

F09B9.4 Protein with weak similarity to S. cerevisiae YDL099W

T20B3.2 Putative troponin-I ; human homologus:Troponin I, cardiac muscle

H14N18.1 Highly similar to mammalian BAG-2, BCL2-associated athanogene 2, a chaperone regulator

W01F3.3/mlt-11; simmilarity with human: Splice Isoform Alpha of Tissue factor pathway inhibitor precursor

F40E10.3 Protein with strong similarity to human CASQ2 protein, a cardiac muscle calsequestrin 2
C16A3.6 Protein with strong similarity to S. cerevisiae Mak16p, an essential nuclear protein required for propagation of M1 double-stranded RNA; human homolog:RNA binding protein

C38C6.4/sre-13 G protein-coupled receptor, member of a subfamily with SRE proteins which are expressed in chemosensory neurons, no homolog found in humans...

F55B11.3 Protein with strong similarity to H. sapiens Hs.169504 gene product [Human mRNA for KIAA0170 gene (GenBank)]

F38C2.5 Zinc finger protein with strong similarity to C. elegans Y57G11C.25 and C. elegans POS-1, a cytoplasmic zinc-finger protein involved in ...

T09A5.6- Component of the Mediator complex required for transcriptional regulation of certain genes human homolog-Hypothetical protein MGC5309

ZC101.2 Muscle protein that is a member of the Immunoglobulin superfamily

C24A3.5 Member of the 4 TM potassium channel protein family

C13C12.1
Calmodulin

F27D9.8 Protein with strong similarity to human Hs.172278 protein, beta2-syntrophin

K07A9.2 Serine/threonine protein kinase, has similarity to human, D. melanogaster, and S. cerevisiae calcium/calmodulin-dependent protein kinase...

Table 10. 25 genes with muscle or human homologues.

The gene expression pattern of the 111 genes on our array is of low gene expression for the mid life time points day 9,10,11 and high gene expression at the beginning of the adult life day3 and the end, day 12,15 of this nematode.

<u>111 genes enriched in muscle with highest variation in our data:</u>
During the midlife time points can be seen a overall trend of decreasing of gene expression to be followed by a slight increase in gene expression.

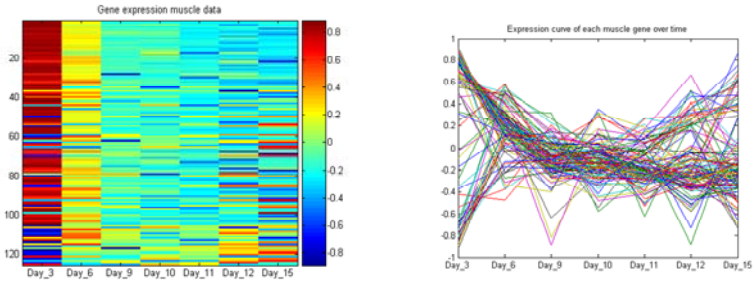


Fig. 30 left -heat map of 111 genes, right –gene expression over the time points day3-day15

5.4.4 *C. elegans* and mammalian muscle homologies

One of the final steps in my sarcopenia analysis was to compile mammalian muscle related genes and find their homologues in *C. elegans*. I will just mention here the lists of genes I compiled. Future analysis remains to be done as comparing the lists I compiled with other experiments. In this sense I will also mention David Miller's recent work on embryonic muscle transcriptome of Caenorhabditis elegans and how his list of human homologues overlaps with my lists.

Using NCBI data base and SQL language I searched for genes mammalian muscle related genes. I identified 4136 genes. Among these genes I searched the genes that have homologies in general and identified 1488 genes. Next step was to look for genes that have *C. elegans* homologues among the 1488 genes. I identified 325 genes. In this list of 325 genes 21 are muscle related genes, 10 aging related genes and the rest are other genes related with various other processes as: cyclin dependent kinase family, abnormal chemotaxis, abnormal cell linage, defective laying eggs, kinase proteins.

Another search I've done was for identifying mammalian muscle genes involved in senescence. I compiled a list of 43 genes. In this list 7 genes have C. elegans homologies, see below.

Below are the 7 genes

1: laminin
K08C7.3a [Caenorhabditis elegans]Other Aliases: K08C7.3Other Designations:
K08C7.3bChromosome: IV
2: *sir-2.1*
yeast SIR related [Caenorhabditis elegans]Other Aliases: R11A8.4Other
Designations: yeast SIR related family member (sir-2.1)Chromosome: IV

3: *pab*-3
PolyA Binding protein [Caenorhabditis elegans]Other Aliases: C17E4.5Other
Designations: PolyA Binding protein family member (pab-3)Chromosome:
4: *hlh*-2
Helix Loop Helix [Caenorhabditis elegans]Other Aliases: M05B5.5Other
Designations: Helix Loop Helix family member (hlh-2)Chromosome:
5: *pmk*-1
P38 Map Kinase family [Caenorhabditis elegans]Other Aliases: B0218.3Other
Designations: P38 Map Kinase family member (pmk-1)Chromosome: IV
6: *ced*-10
CEll Death abnormality [Caenorhabditis elegans]Other Aliases: C09G12.8Other
Designations: CEll Death abnormality family member (ced-10)Chromosome: IV
7: *mpk*-1
MAP Kinase [Caenorhabditis elegans]Other Aliases: F43C1.2Other Designations: MAP
Kinase family member (mpk-1)Chromosome: III

I also compiled a list of 119 mammalian muscle-related genes known to be involved in aging. In this list 20 genes have *C. elegans* homologies. 16 out of these 20 genes are on our chips. The graphs with the 16 genes expression as well as a Table with the 20 genes can be seen in Appendix D. Just four genes out of these 20 genes are expressed in cell muscle and none of them are among our list with 2000 genes with highest variation in our experiment.

The four genes are:

'F10C1.2' ifb-1
'C12D8.10' akt-1
'ZK792.6' let-60
'C29F9.7' pat-4

David Miller group performed also relatively recent (2007) an experiment with the purpose of analyzing the embryonic muscle transcriptome of Caenorhabditis

elegans. They have applied Micro-Array Profiling of Caenorhabditis elegans Cells (MAPCeL) to muscle cell populations extracted from developing Caenorhabditis elegans embryos. Fluorescence Activated Cell Sorting (FACS) was used to isolate myo-3::GFP-positive muscle cells, and their cultured derivatives, from dissociated early Caenorhabditis elegans embryos. Microarray analysis identified 6,693 expressed genes, 1,305 of which are enriched in the myo-3::GFP positive cell population relative to the average embryonic cell. The muscle-enriched gene set was validated by comparisons to known muscle markers, independently derived expression data, and GFP reporters in transgenic strains. This study provides a comprehensive description of gene expression in developing Caenorhabditis elegans embryonic muscle cells. They founded that over half of the muscle-enriched transcripts encode proteins with human homologs suggesting that mutant analysis of these genes in Caenorhabditis elegans could reveal evolutionarily conserved models of muscle gene function with ready application to human muscle pathologies.
I used David Miller's human homologies genes list of 788 genes suggested in this study to compare with my lists. Out of these 788 genes, 593 are in our experiment and all are in the list of 721 genes I compiled. Just as a reminder, the 721 gene list are that genes I identified to be expressed in muscle cell using AQL language.

Also I checked for the intersection of the 593 genes in the list of 2000 genes from our experiment with highest variation and identified 61 genes. A majority of these genes can be found in the cluster G18 mentioned in Chapter 1.

5. 5 Summarizing and conclusions

Using various bioinformatics tools I compiled various gene lists muscle-related. See Fig. 31 below. When cluster the 42 genes that represents genes expressed in cell muscle with high variation in our experiment we identify more the 50% of this genes as having the expected sarcopenia signature of down-regulated genes pattern. I called such pattern and the genes with such pattern as having a 'positive

connection' with the sarcopenia phenotype. All this genes are in the cluster G6 (see section 4.1). The main biological theme for these genes is involvement in growth, developmental and defective in locomotion. In the same time we identified an up-regulated pattern in the cluster G7 (again section 4.1). The genes in this cluster might suggest further bench experiments as I will discuss below.

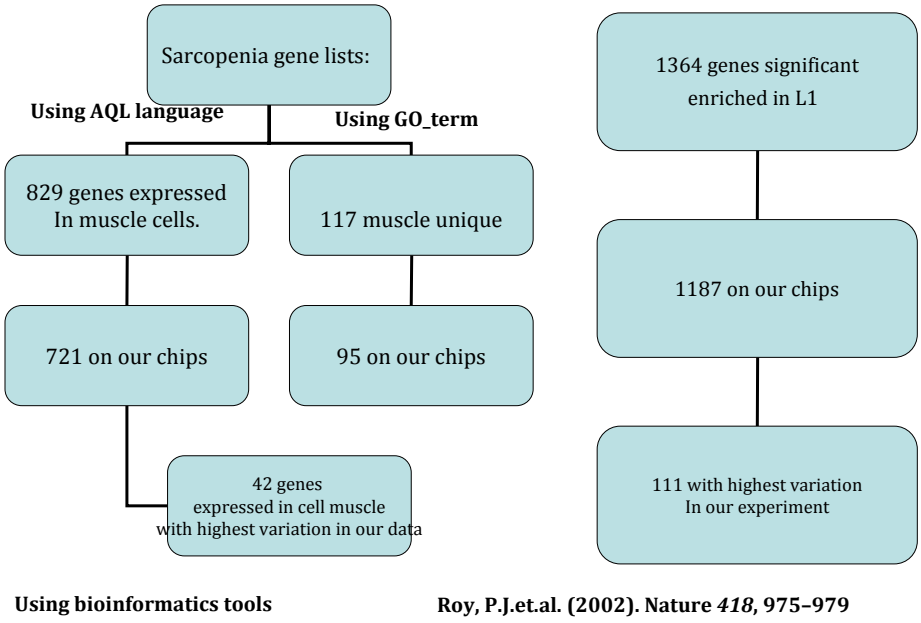


Fig. 31 gene lists used in Sarcopenia analysis

For example, one might want to know what might happen if the genes in cluster G7 introduced in section 4.1 would have a down-regulated pattern instead of an increase in gene expression as these genes show in our analyzes of the wild type of this nematode. One might want to understand if an reverse in gene expression output for the genes in cluster G7 might slow down the sarcopenia signs.

When I analyzed the 95 gene list using the GO-term I identified 2 main patterns of up-regulated and down-regulated genes. The genes in the down-regulated category show same 'positive connection' with the signature of sarcopenia as identified in section 4.1 This genes are mainly structural genes and found in cluster G8 Apart of these findings I also analyzed a list of genes significantly enriched in muscle cells of

young animal (L1 stage). This list was obtained from an experiment performed by Kim's group. I identified 111 genes to have a high variation in our experiment. These genes are mainly collagen related genes. 25 genes have muscle related function, see Table 10.

C. elegans body wall muscle undergoes a process remarkably reminiscent of human sarcopenia. Both have mid-life onset and are characterized by progressive loss of sarcomeres and cytoplasmic volume; both are associated with locomotory decline. To extend understanding of this fundamental problem, I have focused microarray analyses on *C.elegans* muscle aging. Genes expressed in muscle as well as muscle related have been identified and emergent patterns in this list were defined. I surveyed expression of all to describe a profile of transcriptional changes in muscle that transpires during adult life and aging.

This research describes age-associated changes in muscle gene transcription and will constitute the first full-genome profile of sarcopenia in any animal.

References

Apfeld, J., Kenyon, C., 1999. Regulation of lifespan by sensory perception in Caenorhabditis elegans [in process citation]. Nature 402, 804–809.

Arnheim, N., Cortopassi, G., 1992. Deleterious mitochondrial DNA mutations accumulate in aging human tissues. Mutat Res. 275, 157–167.

Beckman, K.B., Ames, B.N., 1998. The free radical theory of aging matures. Physiol. Rev. 78, 547–581.

Corral-Debrinski, M., Horton, T., Lott, M.T., Shoffner, J.M., Beal, M.F., Wallace, D.C., 1992. Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age. Nat. Genet. 2, 324–329.

E, Domany et.al, Neural Computation 9, 1805-1842 (1997)

E, Domany, Phys. Rev. E 57, 3767 (1998)).

Diana David-Rus, Peter J. Schmeissner, Beate Hartmann , Christophe Grundschober , Uri Einav, Eytan Domany, Patrick Nef, Garth Patterson, Monica Driscoll. “A search for mid-life gene expression changes that might influence aging”, in preparation.

Evason, K., Huang, C., et al., 2005. Anticonvulsant medications extend worm life-span. Science 307 (5707), 258–262. Finch, C.E., Kirkwood, T.B.L., 2000. Chance, Development, and Aging. Oxford University Press Inc, New York.

Friedman, D.B., Johnson, T.E., 1988a. A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. Genetics 118, 75–86.

Friedman, D.B., Johnson, T.E., 1988b. Three mutants that extend both mean and maximum life span of the nematode, Caenorhabditis elegans, define the age-1 gene. J. Gerontol. 43, B102–B109.

Garigan, D., Hsu, A.L., Fraser, A.G., Kamath, R.S., Ahringer, J., Kenyon, C., 2002. Genetic analysis of tissue aging in Caenorhabditis elegans. A role for heat-shock factor and bacterial proliferation. Genetics 161, 1101–1112.

Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S. (Eds.), . Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer Science, New York. Gerstbrein, B., Stamatas, G., Kollias, N., Driscoll, M., 2005. In vivo spectrofluorimetry reveals endogenous biomarkers that report healthspan and dietary restriction in Caenorhabditis elegans. Aging Cell 4, 127–137.

Gill, M.S., Walker, D.W., Clayton, P.E., Wallace, D.C., Malfroy, B., Doctrow, S.R., Lithgow, G.J., 2000. Extension of life-span with superoxide dismutase/catalase mimetics. Science 289, 1567–1569.

Golden, T.R., Melov, S., 2004. Microarray analysis of gene expression with age in individual nematodes. Aging Cell 3, 111–124. Harman, D., 1956. Aging: a theory based on free radical and radiation chemistry. J. Gerontol. 11, 298–300.

Hattori, K., Tanaka, M., Sugiyama, S., Obayashi, T., Ito, T., Satake, T., Hanaki, Y., Asai, J., Nagano, M., Ozawa, T., 1991. Age-dependent increase in deleted mitochondrial DNA in the human heart: possible contributory factor to presbycardia. Am. Heart J. 121, 1735–1742.

Herndon, L.A., Schmeissner, P.J., Dudaronek, J.M., Brown, P.A., Listner, K.M., Sakano, Y., Paupard, M.C., Hall, D.H., Driscoll, M., 2002. Stochastic and genetic factors influence tissue-specific decline in ageing C. elegans. Nature 419, 808–814.

Hill, A.A., Hunter, C.P., Tsung, B.T., Tucker-Kellogg, G., Brown, E.L., 2000. Genomic analysis of gene expression in C. elegans. Science 290, 809–812.

Hosono, R., Sato, Y., Aizawa, S.I., Mitsui, Y., 1980. Age-dependent changes in mobility and separation of the nematode Caenorhabditis elegans. Exp. Gerontol. 15, 285–289.

Hulbert, A.J., Clancy, D.J., Mair, W., Braeckman, B.P., Gems, D., Partridge, L., 2004. Metabolic rate is not reduced by dietary-restriction or by lowered insulin/IGF-1

signalling and is not correlated with individual lifespan in Drosophila melanogaster. Exp. Gerontol. 39, 1137–1143.

Jacobs, H.T., 2003. The mitochondrial theory of aging: dead or alive? Aging Cell 2, 11–17.

Johnson, T.E., 1990. Increased life-span of age-1 mutants in Caenorhabditis elegans and lower Gompertz rate of aging. Science 249, 908–912. Katayama, M., Tanaka, M., Yamamoto, H., Ohbayashi, T., Nimura, Y., Ozawa, T., 1991. Deleted mitochondrial DNA in the skeletal muscle of aged individuals. Biochem. Int. 25, 47–56.

Kaiser R, Gottschalk G (1972). Elementare Tests zur Beurteilung von Messdaten, Bibliographisches Institut Mannheim/Wien/Zuerich, pp.18-21)

Kenyon, C., Chang, J., Gensch, E., Rudner, A., Tabtiang, R., 1993. A C. elegans mutant that lives twice as long as wild type [see comments]. Nature 366, 461–464.

Kimura, K.D., Tissenbaum, H.A., Liu, Y., Ruvkun, G., 1997. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans [see comments]. Science 277, 942–946.

Landis, G.N., Abdueva, D., Skvortsov, D., Yang, J., Rabin, B.E., Carrick, J., Tavare, S., Tower, J., 2004. Similar gene expression patterns characterize aging and oxidative stress in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA 101, 7663–7668.

Lin, K., Dorman, J.B., Rodan, A., Kenyon, C., 1997. daf-16: an HNF-3/ forkhead family member that can function to double the life-span of Caenorhabditis elegans [see comments]. Science 278, 1319–1322.

Lund, J., Tedesco, P., Duke, K., Wang, J., Kim, S., Johnson, T., 2002. Transcriptional Profile of Aging in C. elegans. Curr. Biol. 12, 1566. McElwee, J., Bubb, K., Thomas, J.H., 2003. Transcriptional outputs of the Caenorhabditis elegans forkhead protein DAF-16. Aging Cell 2, 111–121.

McElwee, J.J., Schuster, E., Blanc, E., Thomas, J.H., Gems, D., 2004. Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance. J. Biol. Chem. 279, 44533–44543.

Melov, S., Hertz, G.Z., Stormo, G.D., Johnson, T.E., 1994. Detection of deletions in the mitochondrial genome of Caenorhabditis elegans. Nucleic Acids Res. 22, 1075–1078.

Melov, S., Lithgow, G.J., Fischer, D.R., Tedesco, P.M., Johnson, T.E., 1995a. Increased frequency of deletions in the mitochondrial genome with age of Caenorhabditis elegans. Nucleic Acids Res. 23, 1419–1425.

Melov, S., Shoffner, J.M., Kaufman, A., Wallace, D.C., 1995b. Marked increase in the number and variety of mitochondrial DNA rearrangements in aging human skeletal muscle. Nucleic Acids Res. 23, 4122–4126.

Melov, S., Schneider, J.A., Coskun, P.E., Bennett, D.A., Wallace, D.C., 1999. Mitochondrial DNA rearrangements in aging human brain and in situ PCR of mtDNA. Neurobiol. Aging 20, 565–571. Melov, S., Ravenscroft, J., Malik, S.,

Mooradian, A.D., 1990. Biomarkers of aging: Do we know what to look for? J. Gerontol. 45, B183–B186.

Murphy, C.T., McCarroll, S.A., Bargmann, C.I., Fraser, A., Kamath, R.S., Ahringer, J., Li, H., Kenyon, C., 2003. Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. Nature 424, 277–283.

M. Blatt, S. Wiseman, and E. Domany Physical Review Letters 76, 3251, 1996

A. Melendez, Z. Talloczy, M. Scaman, E. L. Eskelinen, D. H. Hall, B. Levine, Essential role of autophagy genes in dauer development and lifespan extension in *C. elegans. Science* 301, 1387-1391 (2003)

Ogg, S., Paradis, S., Gottlieb, S., Patterson, G.I., Lee, L., Tissenbaum, H.A., Ruvkun, G., 1997. The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. Nature 389, 994–999

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. (2001) Bioinformatics 17, 520-5.

Kaiser R, Gottschalk G (1972). Elementare Tests zur Beurteilung von Messdaten, Bibliographisches Institut Mannheim/Wien/Zuerich, pp.18-21)

Vita

Diana David-Rus

2009	Ph.D. in BioMaPs, Computational Biology and Molecular Biophysics , Rutgers University
1994-1999	Diploma in Physics from Bucharest University, Romania
1999-2001	Research assistent, Technical University, Muenchen, Germany
2002-2006	Graduate assistant, Department of Biochemistry
2007-2008	Teaching assistant, Department of Mathematics
2001,2007	Fellowships, Rutgers University
2002-2008 summers	Research Fellowships