Comparative genomics approaches to study species divergence in ageing

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Master in Philosophy

by

Yang Li

September 2010

ABSTRACT

Life, ageing and death have been concepts known to man since time immemorial. To this day, human beings have been fascinated by death and focused on stipulating what occurs after the life ends. In comparison, little attention has been given to ageing, which all experience daily and many take for granted. It is only the recent scientific movement that enabled mankind to shift its attention to studying ageing through the field of biology and gerontology.

The observation that different species show different ageing phenotypes has intrigued many biologists. Comparative biologists, who study genetic, anatomy and behaviour of different species in order to understand the diversity of life, have been trying to explain the different ageing phenotypes through their divergence in their anatomy, natural habitat and behaviour. And since the expansion of the field of genetics, comparative biologists have been employing genomics to study differences in ageing phenotypes at the genome level.

In the last few years, the number of species with their genome sequenced has increased at an impressive rate. However, the number of studies exploiting this wealth of data to study specific phenotypes has remained surprisingly small. Here, we present our work on comparative genomics to study species differences in

ageing, that is, why many species age at different rates. We describe three projects which helped us 1) find a correlation between amino acid usage in mitochondrial proteins and maximal lifespan, 2) detect patterns of selection associated with longevity increases in proteins during mammalian evolution, and 3) compare the genes expression level in two closely related mammalian organisms, the naked mole-rat and the wild-type mouse.

These projects led to several insights about ageing in mammals. We argue that the lack of detection of proteins with anti-oxidant properties in our analyses, coupled with a similar absence observed in other mammalian studies, suggest that contrary to some lower organisms, reactive oxygen species (ROS) may have a lesser impact on ageing in mammals. However, we found evidence that mammalian species may regulate ROS levels through the optimisation of pathways involved, for instance, in the actin cytoskeleton and lipid peroxidation, as well as through the optimisation of amino acid usage in mitochondrial proteins. Additionally, we discovered that genes involved in two pathways, namely lipid metabolism genes and proteasome-related genes, were associated to ageing in both the protein evolution project and the genes expression analysis, suggesting that these two pathways may be important regulators of mammalian ageing. We also discovered a few interesting genes. For instance, the DNA damage-binding protein 1, DBB1, which has a strong selection pattern related to longevity evolution in mammals. Additionally, we found that alpha-2-macroblobulin, A2M, is over-expressed at more

than few hundreds fold in the long-lived naked mole-rat and has previously been associated to ageing.

All in all, we present work that is interesting not only because of the original approaches taken to study mammalian ageing, but also because of the significance of the biological implications obtained. We hope to convince the readers that some of the discoveries made are good candidates for further studies by giving ideas of possible follow up experiments.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor, Dr. João Pedro de Magalhães, who has been of invaluable help. Indeed, Pedro has not only been of great help in the several research projects that I have been working on, but he gave me many sound advices about how to become a good researcher of international calibre. Pedro has been extremely generous with his time and I praise him for that. He has let me grow and learn at my own pace by trusting me and making research feel like a hobby rather than a chore. I look forward to continue working with him in hope that one day, our research will contribute to the invention of interventions that can slow or even halt human ageing.

Next, I would like to thank my parents, who have been supportive, from the very beginning, of my switch from mathematics and theoretical computer science to the much more applied field of comparative genomics. Switching out from a field in which I was comfortable was a difficult and risky choice, and having to move to Liverpool from Montreal for the switch certainly did not help. However, no one rebuked me for my decisions (perhaps not to crush a young man's dream) and I decided to follow my heart.

I'd like to thank Dr. Mathieu Blanchette who has been a great mentor ever since I have met him. His passion in all questions related to biology and computer science is extremely contagious. In every meeting I have had with him, he showed

interested in what I was doing, helped me think through ideas that I had and gave constructive criticisms. Mathieu is a remarkable researcher and teacher, and I believe myself to be incredibly lucky to have met him. I'd also like to thank Dr. Pietro Lió, who though I have only met briefly, acted like an excellent role model and gave me useful advices on being a young researcher.

Last but not least, I would like to thank my friends which have been both motivating and supportive all through my studies. I have yet to meet a group of friends more motivating than the ones I have met at McGill. It was an incredible time, guys. I would like to thank the lab members at the University of Liverpool which made the office life much more bearable and sometimes enjoyable as well as the few friends that I have met in Liverpool. I thank Dr. Chuanfei Yu to have been both a friend and a good colleague to work with in the Church lab where I have learnt a lot about cutting-edge research in biotechnology research. Thanks to Yeting (no, I have not forgotten). Thanks to Andrew for reading my thesis and complaining about my “incorrect” American way of spelling words. Thanks to Louise for reading over my thesis. And lastly, Sipko for the entertainment and for his hospitality.

CONTENTS

Index of Tables and Figures

GLOSSARY

AD	Alzheimer's Disease
BLAST	Basic Local Alignment Search Tool
GASP	Gapped Ancestral Sequence Prediction
ROS	Reactive Oxygen Species
GO	Gene Ontology
GWAS	Genome wide association studies
IGF1/GH	Insulin-like growth factor 1/Growth hormone
A2M	Alpha-2-macroglobulin (gene)
SMC1A	Structural maintenance of chromosomes 1A (gene)
SAT1	Spermidine/spermine N1-acetyltransferase family member 1 (gene)
IGFBP4	Insulin-like growth factor-binding protein 4 (gene)
DDB1	DNA damage binding protein 1 (gene)
MLS	Maximum lifespan same
MLI	Maximum lifespan increase
LSS	Longevity-specific selectivity

INTRODUCTION

Our work is mainly focused on exploiting existing computational techniques as well as inventing novel algorithms that can be used to study a variety of biological phenomenon. One of the most complex (and, arguably, interesting) biological phenomenon that we experience daily is ageing. Of equal interest is why some species seem to show different phenotypes of ageing when compared to others. In fact, the observation that some species show different signs of ageing was already made by Aristotle in 350 BC. A great number of breakthroughs have since been made, however the biological mechanisms underlying ageing is still unknown. For this reason, we embarked into a project aiming to study the phenomenon of ageing by using and developing computational biology algorithms.

There are now hundreds of different theories of ageing (Bengtson et al. 2008), many with experimental support yet none with sufficient evidence to rally all scientists in the field. What makes an unified theory of ageing difficult if not impossible is the fact that certain animals seem to exhibit different phenotypes of ageing or senescence when compared to others (Lindström 1999). Thus, not only is it hard to find a single theory of ageing for all animals (and even just mammals), it is also challenging to find a proper definition for ageing. Therefore, we use Medawar's simple definition of ageing which is the collection of changes that render human beings, or other biological entities, progressively more likely to die (Medawar 1952).

The speed at which new technologies were invented in the last few decades considerably helped biology and specifically genetics and proteomics research by providing the tools to collect novel types of data at an increasing rate (Tyers et al. 2003). The invention of modern personal computer further allowed the analysis of this data which is used both to test old hypotheses as well as generate new ones. As

will be discussed in Chapter 1 and 2, the genotype of a species is thought to heavily influence its ageing phenotype. Thus, the very recent technological advances in genomics including the sequencing of multiple model organisms and second generation expression profiling technologies allow us to study the mechanisms of ageing under a new perspective. Though there is an increasing number of studies making use of the large amount of data available in genomics database such as Ensembl (Flicek et al. 2010), this number remains small and only a handful of them focused on elucidating the mechanisms of ageing. Here, we present three distinct, but complementary projects that take advantage of the early sequencing efforts of the 21^{rst} century to study the genetics of why different species show different ageing phenotypes.

We hope to convince the reader that studying ageing through a comparative genomics perspective is at the same time useful and cost-effective. Experimental projects, especially in ageing, are generally very costly and complicated compared to *in silico* projects. Nevertheless, genomic comparisons are far from trivial and require significant work in statistical and algorithmic design. The bulk of our work consists of scripts of varying length and complexity written in the programming language Python. Many of these scripts are part of a pipeline for data analysis by gluing software such as BLAST, GASP and ClustalW while others implement statistical tests and randomisation algorithms. However, in order to understand the significance of our work, it is important to be familiar with some of the most popular theories of ageing. Therefore, we will start by exploring theories of ageing before plunging into the computational aspect of our work. In this introduction, we will also explore the importance of comparative genomics in biology and ageing research. We will end by discussing some of the classical tools in comparative genomics in order to give an idea of the rigour of comparative genomics and of the interplay between biological problems and computational problems.

Chapter 1: Biology of ageing

The ageing phenotype vary greatly between divergent species and many researchers have been investigating species differences in their ageing process. In fact, even the ageing rates within mammals are markedly different from one species to another. Different animals likely evolved different strategies to respond to environmental stress. They also have different life histories which can influence their ageing phenotype and the selection of their maximal lifespan. Though many differences exist, it is thought that there is a fundamental process underlying ageing or at least mammalian ageing in which the phenotypes are similar albeit at different timings (de Magalhães et al. 2002). By focusing on finding mechanistic differences in the ageing processes between different animals, we may be able to discover the main regulators of ageing.

Many theories of ageing have been postulated, however there is no consensus among researchers on which theory best fits the experimental discoveries that has been made thus far. In this chapter, we present an overview of a few theories of ageing, but for in-depth reviews, see (Arking 2006; Weinert et al. 2003; Kanungo 1994). Some theories speculate that ageing results from the accumulation of damage in different tissues (wear and tear) while others conjecture that ageing is programmed. Consequently, theories can be generally categorised in two main categories: damaged based and programmed.

1.1 Free radical theory of ageing

The free radical theory of ageing, which belongs to the damaged based category, has many proponents. It has first been developed by Denham Harman and discussed in Harman (1956). The free radical theory of ageing suggests that the production of reactive oxygen species (ROS) causes different types of damage to

molecules such as the DNA, proteins and lipids and the accumulation of such damages causes ageing. For example, free radicals such as ROS can contribute to protein misfolding which, in turn, may cause neurodegenerative diseases (Lipton et al. 2007). Furthermore, ROS levels are thought to be altered during calorie restriction, the only intervention that seem to consistently delay ageing in mammals such as mice (Weindruch et al. 1986) and rhesus monkey (Mattison et al., 2003). In fact, glucose restriction has been shown to extend the lifespan of the yeast by inducing mitochondrial respiration and causing ROS formation which is followed by an increase in oxidative stress resistance (Lin et al. 2002). Thus the interplay between ROS and ageing is unclear, but definitely present.

Another evidence that suggests the involvement of ROS in ageing is the enzyme methionin sulfoxide reductase A (MSRA) which catalyses the repair of methionine residues oxidised by ROS. The over-expression of MSRA has shown to increase the longevity in flies (Ruan et al. 2002) and its knock-out was associated with a decrease in longevity in mice (Moskovitz et al. 2001). As discussed in de Magalhães et al. (2006), the existence of this class of enzymes that protects the cell from ROS damage is a strong indicator that ROS are important and potentially dangerous biological molecules. However, most of the studies testing the impacts of ROS on mammals such as mice were inconclusive at best. There are studies showing conflicting effects of feeding antioxidants to mice; some showed an increased average longevity, others showed no such increase, while none showed a delay in ageing (Saito et al. 1998; Harman 1968). Pérez, Van Remmen, et al. (2009) even goes to show that the over-expression of major antioxidant enzymes does not extend the lifespan of mice. One idea is that ROS are mostly involved in the senescence of post-mitotic cells which worms and flies are mostly composed of as opposed to mammals. Interestingly, ROS are associated to pathologies involving post-mitotic cells such as neurons and there is evidence of mitochondrial optimization in the human lineage to delay neurodegeneration (de Magalhães 2005). In sum, although ROS have been implicated in the ageing of many lower organisms

such as yeast, flies and worms, there is no direct evidence that ROS influence the ageing process in mammals. For more in-depth reviews of the free radical theory of ageing, see the work of Finkel et al. (2000) and Muller (2000).

1.2 Damage-linked theories of ageing

DNA is a central molecule of life and its involvement in the basic mechanisms of ageing would be unsurprising. The DNA damage theory (Medawar 1952; Szilard 1959) suggests that it is the accumulation of DNA damage that causes ageing. Since DNA encodes genes which may be important for a variety of biological functions, the mutation or damage in certain coding regions may be a cause of ageing. This is supported by the fact that many diseases with premature ageing phenotypes such as progeroid syndrome, Werner's syndrome, Hutchinson-Gilford's syndrome, xeroderma pigmentosum and Cockayne syndrome are caused by mutations in the genes coding for DNA damage repair proteins (Martin et al. 2000). The lack of DNA damage repair increases the rate at which DNA molecules are damaged and thus may be the cause of the acceleration of some ageing phenotypes to appear in individuals with one of the aforementioned diseases. Furthermore, there is evidence that DNA mutations increase with age in both mice and humans (Vijg 2000; Dollé et al. 1997; Martin et al. 1985). However, there is no convincing proof whether the mutations are the cause of ageing or simply a product of the ageing animal. Cell cycle and its dysregulation have also been associated with ageing from the observation that a deficient cell cycle control can lead to the unstopped replication of mutated DNA following DNA damage (Gu et al. 2005).

Another damage-linked theory of ageing stemmed from the observation that in many model organisms such as mouse and human, protein turnover rate decreases as the organisms age (Chondrogianni et al. 2005). Although it is not clear whether this is a cause or an effect of ageing, researchers discovered that in older organisms, slower protein turnover rate caused proteins with post-translational

modifications to remain longer within the cells. Many of these post-translational modifications hindered the functions of the proteins and the prolonged presence of these damaged proteins in the cell are thought to be detrimental (Farout et al. 2008). Perhaps of interest, a slower proteins turnover rate may be caused by a decrease in proteasome activity in older organisms. The proteasome is the primary machinery for protein degradation in cells and its decline in activity is believed to have an important role in ageing and its activity may the extend of the damage by the degradation of damaged proteins (Vernace et al. 2007).

1.3 Evolutionary theories of ageing

Many other theories of ageing exist, one important subset is the evolutionary theories of ageing including the disposable soma theory and antagonistic pleoitropy. Briefly, the antagonistic pleoitropy theory follows from the observation that selection on phenotypic traits that affect the later life of an organisms, i.e. after the reproductive phase, becomes much weaker (Williams 1957). Therefore, traits which improve an organism's fitness before sexual maturation but which are deleterious for later life are under positive selection (de Magalhães et al. 2005). Though there are some evidence that such pleotropic genes with antagonistic effects exist, it is still unknown if this category of genes can explain why we age.

Another evolutionary view stipulates that ageing is not caused by antagonistic effects of certain genes but simply by the lack of species adaptation for survival after sexual maturity. In his review, Rose (2009) argues that it is possible to extend longevity by increasing the strength of the force of selection on old animals. However, simple mutations that merely increase longevity without antagonistic effects do not exist as a longer health span translates into a better Darwinian fitness and natural selection would have already previously acted upon them. Indeed, Rose (2009) argues that in many experiments showing the mutation of a gene increasing the lifespan of a species, the mutations either retard the onset of maturity or are

deleterious in some way. For examples, studies on dietary restriction, which consists of diminishing the intake of calories while maintaining the bare minimum for survival, has been shown to have life-extension effects in many model organisms (Weindruch et al. 1988). However, there are reports that show a decrease in the effectiveness of the immune system under calorie restriction which may lead to infection and other diseases (Ayres et al. 2009). Thus, we can see a putative trade-off between infection resistance and a longer lifespan in which infection resistance was selected.

There are many other evolutionary views on ageing, such as the disposable soma theory (Kirkwood 1977), which hypothesises that organisms only have a limited amount of energy that has to be divided between reproductive activities and the maintenance of the organism itself or soma. Damage accumulated in the organism can be repaired by the organism, but only at the expense of reproductive capabilities. Thus, the disposable soma theory suggest that ageing is the result of a trade-off between the transmission of the genes and the survival of the individual. This trade-off, or equation, is then optimised according to the environment in order to best guarantee the survival of the species. In the next section, we will discuss how these theories affect our study on mammalian divergence in ageing, that is, why many species age at different rates.

1.4 Hypotheses on divergence in species ageing

We have discussed some of the theories for why we think animals age. However, an useful and complete theory should be able to explain why certain animals show different ageing phenotype in addition to different lifespans. As reviewed by de Magalhães et al. (2002), mammals show little diversity in their ageing phenotype. However, their lifespan can vary greatly from mice living no longer than 4 years to humans living over 120. Thus, an interesting exercise would be to use the different ageing theories to speculate on why such a disparity of

lifespan exists in mammals without having to worry about the effects of other ageing phenotypes. Furthermore, since most of our research is based on hypothesis free computational methods, it is useful to have an idea on what we expect to find.

According to the free radical theory of ageing, the main cause of ageing is the destructive force of the free radicals that roam in our cells. Free radicals cause oxidative damage to different cellular components such as DNA, fatty acids, and proteins (Harman 1956). Consequently, we expect to find differences in genes expression or sequence, as well as differences in pathways related to DNA repair and protection from free radicals. We also expect a stronger resistance to ROS in longer-lived mammals and some pattern of selection in processes which reduce the generation of free radicals either by the optimisation of the mitochondrial respiratory chain or other mitochondrial parts responsible for the generation of ROS (Finkel et al. 2000). In light of other damaged-link theory of ageing, we expect, in addition to the aforementioned pathways, a stronger signal in DNA repair, DNA response and damaged protein degradation pathways optimisation in longer lived mammals. We expect these proteins to be differentially expressed in longer lived mammals and to be under positive selection in their lineages.

Unlike damage-linked theories of ageing, evolutionary theories of ageing do not allow much room for predictions. They do make some predictions such that a species will experience delayed senescence and increased longevity when rates of extrinsic mortality are reduced (Shattuck et al. 2010). However, they do not provide any hypothesis as to which biochemical pathways are responsible for the divergence in lifespans witnessed among species. In this sense, any signal from pathways previously associated to ageing may be expected. We would expect, if evolutionary theories of ageing were to be true, that the signals of longevity evolution in specific pathways are weak since nature is not limited to act on these pathways to find ways to increase longevity.

All in all, most scientific discoveries in the field of ageing supports one or more alternative theories of ageing without falsifying the others. The difficulty which ageing researchers face is that any experimental discovery in one species can be said to be unique to that species alone and thus not representative of ageing as an universal biological phenomenon. As mentioned in section 1.1, ROS seem to have a greater impact on the longevity of lower organisms when compared to mammals and the free radical theory is not supported by strong evidence in mammals. Furthermore, de Magalhães et al. (2002) observes that there is a great difference in ageing phenotypes between mammals and reptiles. All these differences make the study of ageing extremely tricky.

One way to escape this impeding predicament is to restrict our focus to mammals as they appear to have a conserved mechanism of ageing. Another way would be to use comparative biology in order to observe the differences and similarities that different species possess. By finding evidence a particular pathway is related to ageing in multiple organisms, we can be more confident that this pathway is important to the ageing process as its role is conserved across species. In this thesis, we use comparative biology, and more precisely comparative genomics, to study divergence in mammalian ageing by trying to find genes and pathways with longevity-associated signals in multiple mammals.

Chapter 2: Comparative Biology

Comparative genomics is a branch of comparative biology and, thus, in order to exploit it fully, one needs to be aware of the strengths and caveats of comparative biology.

2.1 Comparative biology and definition

Comparative biology is a field that studies species differences and similarities in hope to understand the organismic diversity on Earth. It is often coupled with evolutionary biology as the two are closely intertwined (Futuyma 1997). Different species evolve different phenotypic traits in order to optimise their survival in their habitat or adapt to a changing environment (Rose 2009). One of the early goals of comparative and evolutionary biology was to find out how the different environmental constraints and settings influenced different organisms and their evolution. Different species age at different rates and comparative biologists working on ageing have tried to explain this divergence using evolutionary theories and by studying the different life history traits of animals. Most biologists agree that the rate of ageing has a strong genetic component and is under selection, however, the evolutionary adaptations that influence longevity and the rate of ageing are largely unknown (de Magalhães 2003).

Much of the early work on ageing was conducted by observing different phenotypic traits related to the ageing rate and other life history traits. For instance, it has been found that larger animals tend to live longer than smaller animals (Knut 1984; Austad 2005). Though no one knows why a relationship between body size and maximal lifespan exists for sure, many believe that smaller animals tend to be prone to predation and are expected to have higher extrinsic mortality rates (Shattuck et al. 2010). This could imply a lack of selection of their longevity since

the fitness of animals having high extrinsic mortality rates do not benefit from an increase in longevity as much as would one with low extrinsic mortality rates. As discussed in section 2.5, the correlation between longevity and other traits can be a big problem in comparative biology studies as we often cannot be sure if one difference observed between species with divergent lifespan is directly linked to the ageing mechanism or other traits that are correlated, though unrelated, to ageing (Speakman 2005a). Fortunately, there are established ways to minimize the effects of correlated traits in comparative studies. This is why species selection is crucial for comparative research projects specially in ageing research.

2.2 Selection of model organisms in ageing studies

It is customary for researchers to use model organisms to research ageing. Some examples of model organisms most commonly used in ageing research are yeasts, flies, worms, mice and rats (Kim 2007; Antebi 2007; Kaeberlein et al. 2007). Researchers also use cell lines in some ageing models (Das et al. 1978) but *in vivo* ageing studies are rare in humans. In fact, ageing studies in primates and mammals are rare since ageing studies are usually restricted to short-lived species because of the differences in the length of the studies. However, the study of mammals and specially primates and humans in an ageing context is critical as many studies suggests that ageing models in lower model organisms are not representative of ageing in mammals (Austad 2005; Austad 1997). Apart from common model organisms, few ageing studies have looked at non-traditional model organisms such as certain species of long-lived clams, long-lived fish, the little brown bat and the naked mole-rat which we will discuss further in this thesis. These non-traditional organisms are used in ageing studies mainly for the reason that they either show a higher resistance to senescence or a higher resistance to other types of stress compared to closely related species or species of similar body size (Austad 2009). By comparing these species to related species and species of similar body size, researchers hope to be able to extrapolate differences in genes expression or

differences in protein sequences to explain the divergence in ageing phenotype. Furthermore, by studying closely related species with extreme lifespan divergence, one can minimize the effects of correlated traits in comparative analyses as the signals related to the extreme longevity are expected to be much stronger than the signals related to correlated traits.

In addition to body mass, researchers in the field of ageing study the interplay between the rate of ageing and other life-history traits. In fact, the rate of ageing has been related to extrinsic mortality rate, gestation period and age at maturity (Ricklefs 2010). But it is only recently that researchers have looked directly into the genomes of different species in order to find patterns of selection for longevity.

2.3 Species divergence studies in ageing and the naked mole-rats

Although the mechanisms underlying the different rates at which species age is poorly understood, researchers have been successful in determining several putative pathways that may have contributed to the longevity of long-lived mammals. Austad (1997) suggests four main uses of comparative assessment of mammalian ageing in ageing research which are (1) the formulation and evaluation of hypothesis, (2) the investigation of how widespread a putative ageing mechanism is among mammals, (3) the isolation of key physiological factors regulating ageing, rate and (4) the educated choice of which animal models is most appropriate for a particular study.

One good example of (4) is the recent work by Pérez, Buffenstein, et al. (2009) which uses the naked mole-rats (*Heterocephalus glaber*) as model for successful ageing. The naked mole-rats is the longest-lived rodent known with an expected lifespan of over 28.3 years (Buffenstein et al. 2002). Their natural habitat as well as their eusocial behaviour are some of the reasons why the naked mole-rats

might have evolved such an extreme longevity compared to other rodents (see Figure 1).

Species	Name	MLSP	Coverage
Human	*Homo sapiens*	122.5	Complete
Orangutan	*Pongo pygmaeus*	59	6X
Gorilla	*Gorilla gorilla*	55.4	2X
Rhesus Monkey	*Macaca mulatta*	40	Draft
Chimp	*Pan troglodytes*	59	6X
Bushbaby	*Otolemur garnettii*	18.3	1.5X
Marmoset	*Callithrix jacchus*	16.5	6X
Tarsier	*Tarsius syrichta*	16	1.82X
Mouse Lemur	*Microcebus murinus*	18.2	1.93X
Guinea Pig	*Cavia porcellus*	12	6.79
Mouse	*Mus musculus*	4	Complete
Rat	*Rattus norvegicus*	5	Draft
Squirrel	*Spermophilus tridecemlineatus*	7.9	1.9X
Kangaroo rat	*Dipodomys ordii*	9.9	1.85X
Microbat	*Myotis lucifugus*	34	1.7X
Megabat	*Pteropus vampyrus*	20.9	2.63X
Treeshrew	*Tupaia belangeri*	11.1	2X
Pika	*Ochotona princeps*	7	1.93X
Rabbit	*Oryctolagus cuniculus*	10	2X
Dog	*Canis familiaris*	24	7.6X
Cat	*Felis catus*	30	1.87X
Cow	*Bos taurus*	20	7X
Horse	*Equus caballus*	57	6.79X
Pig	*Sus scrofa*	27	4X
Dolphin	*Tursiops truncatus*	51.6	2.59X
Rock Hyrax	*Procavia capensis*	14.8	2.19X
Armadillo	*Dasypus novemcinctus*	22.3	2X
Hedgehog	*Erinaceus europaeus*	11.7	1.86X
Shrew	*Sorex araneus*	3.2	1.9X

Elephant	*Loxodonta africana*	65	2X
Hedgehog Tenrec	*Echinops telfairi*	19	2X
Sloth	*Choloepus hoffmanni*	37	2.05X
Acalpa	*Lama pacos*	25.8	2.15X

Table 1: Mammalian species, their maximal lifespan (MLSP) in years and their genome coverage (see chapter 4) as of release 54 of the Ensembl database.

We can see here that mammalian species have a wide range of maximal lifespan.

Naked mole-rats live underground and subterranean animals are protected from both climatic extremes and predation, which lower their extrinsic mortality rate (Buffenstein 2005). According to evolutionary theories of ageing, a lower extrinsic mortality rate translates into a stronger selective pressure on the fitness of old organisms and thus the selection of a longer lifespan. A longer lifespan has also been correlated with eusocial behaviour, i.e. animals with group of social living, plausibly because of intergenerational transfer of information and communal responsibilities such as the care of the young and the foraging for food (Buffenstein 2005). In contrast to other rodents, the naked mole-rats show small age-related changes in morphology and maintain many physiological function and activity at old ages (Buffenstein 2008). They also exhibit small age-related changes in mitochondrial mass and efficiency (Csiszar et al. 2007), antioxidant activity (Andziak et al. 2006), membrane composition and lipid peroxidation (Andziak et al. 2006).

Furthermore, compared to mice, the naked mole-rats show attenuated age related change in protein oxidation, resistance to protein unfolding and a sustained proteasomal activity throughout their lifespan. In their work, Pérez et al. (2009) suggest that protein stability, turnover and its resistance to oxidative stress are important players in the naked mole-rats' extreme longevity when compared to lab mice. These phenotypes, as Pérez et al. (2009) argues, may be the principal factors

for the mole-rat's successful ageing.

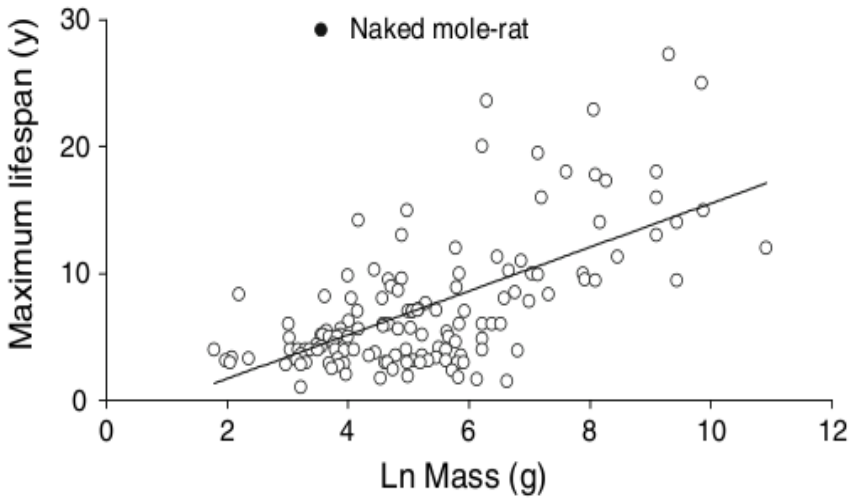

Figure 1: Maximum lifespan as a function of body mass of rodents.

Source: (Rochelle Buffenstein 2008)

In addition to studies on the mole-rats, Salmon et al. (2009) studied differences in protein oxidation levels between long-lived bats and mice. These bats have an extremely long lifespan for their body weight and are used to model successful ageing in mammals. Salmon et al. (2009) found that, relative to mice, Mexican free-tailed bats and cave myotis bats show lower protein oxidation. Moreover, they show that proteins in bats are more resistant to urea-induced protein unfolding compared to mice. Surprisingly, they found that the ubiquitin and proteasomal activity in these bat species were lower than the one found in mice and argue that it is a consequence of the diminished proteins damage.

Already, we can speculate that proteasome activity has an important role in species divergence in ageing. However, divergence in proteasome activity and efficiency alone cannot explain the differences in ageing phenotypes that is witnessed across mammals. In fact, it is likely that different long-lived mammals

evolved distinct mechanisms for successful ageing in line with evolutionary theories of ageing. For instance, (Finch et al. 2004) showed that ApoE is under positive selection in the human lineage and has been proposed to be a meat-adaptive candidate protein in the increase of human lifespan. ApoE is involved in lipid metabolism and has been shown to impact age-related diseases such as neurodegeneration and myocardial infarction (Masliah et al. 1995; Lambert et al. 2000) and to contain polymorphisms linked to human longevity (Schächter et al. 1994).

In lower organisms, (McCarroll et al. 2004) studied genomic expression patterns in the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* and found that genes involved in mitochondrial metabolism, DNA repair, catabolism, peptidolysis and cellular transport tend to have common patterns of change in expression with age. Moreover, (Smith et al. 2008) studied genes previously associated to ageing which are conserved between the nematode *C. elegans* and yeast *Saccharomyces cerevisiae.* They found that many of the genes were conserved including genes from Sir2-family proteins, insulin/IGF-like receptor proteins, and the target of rapamycin (TOR) kinase family which are all conserved in mammals. Thus, they speculate that these genes may also play a role in mammalian ageing.

As many other comparative studies in ageing, we are interested in finding the genes and pathways that are responsible for extreme longevity in mammals which may lead to insights about the mechanisms of human ageing.

2.4 Importance of the mitochondria in ageing

In chapter 1, we saw that the free radical theory of ageing stipulates that the main factor leading to ageing is the accumulation of free radical damage over time (Harman 1956; Hekimi et al. 2003). In recent years, many comparative studies

exploring the evolution of mitochondrial DNA-coded proteins and a possible relationship with oxidative stress and lifespan. This provided a good opportunity to test this theory of ageing. Although correlation between longevity and the evolution of mitochondrial DNA-coded proteins has already been reported (Rotcritenberg 2006), it is not clear whether free radicals and their detrimental effects on mitochondrial proteins can explain this correlation. A relationship between lifespan, oxidative damage and specific protein residues is highlighted in the work of Stadtman (Levine et al. 1996; Stadtman 2006) which suggests that methionine and cysteine may act as antioxidants in proteins due to the reversibility of their oxidation. Stadtman found that the expression level of Msr, a reductase capable of reversing the oxidation of Met to MetO, is positively correlated with resistance to oxidative damage and maximal lifespan. In this line of work, however, a negative relationship between the residue composition of proteins affected by oxidation and lifespan was reported by (Kitazoe et al. 2008) and also by (Moosmann et al. 2008).

In their recent work, Moosmann and Behl conducted a meta-analysis of genome sequences from 248 animal species spanning 10 different phyla in which they report a negative correlation between cysteine encoded in the mitochondrial DNA and longevity. Moosmann and Behl propose that it is the detrimental capacity of cysteine thiyl radicals and their potential to initiate irreversible protein cross-linking that caused a selection against cysteine in mitochondrial DNA-coded proteins. Moreover, they suggest that the uncontrolled oxidation by reactive oxygen species leads to dysfunctional proteins, which results in cellular senescence and organismal ageing. Although the analyses of Moosmann and Behl demonstrate that mitochondrial amino acids can vary widely in closely related species, no correlation between mitochondrial cysteine usage and longevity in mammals was reported. In chapter 5, we will discuss our method to test whether the hypothesis that cysteine or methionine is indeed correlated with the maximal lifespan of mammals. For now, we should note from the discussed studies that the mitochondria and its proteins seem to have at least some role in ageing.

2.5 Correlation of different traits with maximal longevity and phylogenetic dependence

One major difference between an experimental based project and a computational based project in biology is the fact that, more often than not, the experimental based project tests a biological hypothesis that is possibly true (Barnes 2007). On the other hand, a computational based project does not usually test a hypothesis, but rather aims to discover facets of the biology hidden in the data that were previously unknown. This makes the interpretation of the results much harder to validate.

In section 1.1, we saw that the longevity of a species or its ageing rate is correlated to many different traits. Studies using comparative biology in ageing will detect correlations between traits and ageing which are, in reality, not directly related to ageing (Speakman 2005a). For instance, a comparative genomics study between a long-lived mammal and a short-lived one may result in the detection of many genes associated to body size (Speakman 2005b). Furthermore, when using multiple species, a statistical analysis may be biased by the phylogenetic dependence between species since species that are evolutionarily closer tend to have similar ageing phenotype and other phenotypes or genotypes (Speakman 2005a). Thus, one can virtually associate any phenotype or genotype with statistical significance by choosing the right set of species for the analysis.

While it is hard to control for many of the correlated traits, there are few solutions to the aforementioned problems such as the careful selection of the species under study, the correction for body mass and the correction for phylogenetic dependence. As the selection of species in ageing research was covered earlier, we will focus on methods for body mass and phylogenetic dependence correction in Chapter 4.

Chapter 3: Computational biology

The focus of our work is in comparative genomics. Though some other computational biology fields could be of interest to researchers in ageing, we found comparative genomics to be the most suitable choice to study species divergence in ageing. In this chapter, we highlight the importance of computational biology as a supporting field for biology research and explore some of the ways it can help researchers in the biology of ageing.

3.1 Computational biology is an emerging field

We define computational biology as the use of computational techniques in the general field of biology. To us, a complete computational biology project consists of three main parts. First, there must be an underlying biological problem. Next, the biological problem must be translated into a well-structured and well-defined mathematical problem. And finally, the mathematical problem must be solved in an efficient or at least tractable way. Paraphrasing Dr. Mathieu Blanchette, the challenge of computational biology is to find an interesting biological question, to formulate it clearly within a mathematical framework and to use or invent an algorithm able to solve it. By formulating a biological problem in a well-defined mathematical fashion, one can pinpoint the weaknesses of the approaches as well as adjust parameters correctly in an informed way. It is also useful to have a clear mathematical framework as it can also help understanding the output of the computational analyses.

We believe computational biology to be a new and exciting interdisciplinary area of research. Computational biology has recently found a place in most major universities both as a field of its own and as a supporting field for many biology research groups. Since its first use in the mid twentieth century, computational

biology has evolved into a gigantic field comprising of everything from prediction of protein structures to systems biology; from comparative genomics to population genetics. The exponential growth of computational biology has been fuelled by the success of the Human Genome Project in the early twenty-first century along with the impressive speed at which new breakthrough technologies allow researchers to harvest an enormous quantity of data. In fact, computational biology tools are used in virtually all genetics labs and they often help to guide research projects in addition to help in analysing experimental results (Pevsner 2009).

Indeed, computational biology not only serves as a tool to help validating biological hypotheses but also used to generate new ones and guide research. With innovations making high-throughput biology available to almost all research institutions, computational biology makes the concurrent analysis of thousands of genes possible. The community studying the biology of ageing has gathered an impressive amount of data, but many questions remain unanswered. As is, the relevance of computational algorithms that are able to extract useful information from this huge amount of experimental results is evident. These includes algorithms for restriction sites mapping, structural prediction for a range of molecules, culminating to the assembly of the 3 million base pair human genome (Waterman 1995). It is however after the year 2000 that computational biology really exploded in breadth and depth.

3.2 The multiple subfields of computational biology

Computational biology offers a wide range of techniques and algorithms to study biological data. There exists three main classes of computational biology methods classified according to the type of data they can analyse (input) or the type of results it can produce (output). A comprehensive discussion of these subfields can be found in any standard bioinformatics textbook Cristianini et al. 2007; Pevsner 2009). These classes include computational biomodelling, computational

structural biology and computational genomics. Computational biomodelling is the part of computational biology that deals with mathematical and computer models of biology (Swedlow et al. 2006). It usually includes the broad field of systems biology and mathematical biology. The main scope of computational biomodelling is to produce computational models to represent biological systems or processes by usually describing them as mathematical entities like a set of equations with mathematical variables representing quantitative biological measures (Sauer et al. 2007). Computational biomodelling has a great potential to be used in ageing research, for instance, the Kirkwood group focuses on systems biology models of different pathways thought to be related to ageing (Kirkwood 2008; McAuley et al. 2009). Unfortunately, modelling biological pathways is far from trivial and the lack of data and complexity of the biology of ageing make the state of the art models mediocre at best.

Computational structural biology is another important part of computational biology. As its name indicates, computational structural biology deals with the prediction of molecule structures such as proteins, mRNA and DNA (Cozzetto et al. 2008). Some researchers have used structural algorithm methods in order to model mutations in proteins thought to be important to cellular function (Medvedev et al. 2009). Others have use these techniques for drug discovery and design (Weigelt 2010). However, these methods remain low throughput as they are not reliable enough without the curating of experts and thus fail to exploit the vast amount of sequence data available.

In our work, we focus on computational genomics which has been arguably the most successful part of computational biology in the recent years. This is mainly due to the advances in genomics technologies such as microarrays and sequencing technologies allowing ChIP-seq (Pepke et al. 2009; Park 2009), RNA-seq (Marguerat et al. 2010) and genome assembly of short reads (Li, Hu, et al. 2010). Computational genomics is the collection of techniques used to analyse the

genotype of different species. Since a species' phenotype is thought to be largely dependent on its genotype (de Magalhães 2003), the ultimate goal of computational genomics is to assign biological functions or phenotypes to the different genes of a species or individual within a species which first major hallmark was witnessed in the Human Genome Project (Watson 1990) and subsequent analyses (Lander et al. 2001). In the same line of reasoning, computational genomics go so far as to try to predict the effect of a single nucleotide polymorphism on the phenotype of an individual (Rosenberg et al. 2010). The genome of a species provides us with a wealth of data and a lot to work with. Coupled with the recent technological boom, it is our belief that computational genomics is, amongst other subfields of computational biology, the most likely to help the biology of ageing researchers in the near future.

3.3 Comparative genomics and how it can help research in ageing

It is clear that genetics has a major role in the ageing process of any species (de Magalhães 2003; Austad 2005). In fact, there are over hundreds of genetic manipulations that extends the lifespan in model organisms. Comparative genomics studies exploit the fact that different species show different phenotypes of a certain trait. Since ageing is tightly linked with the genome of an individual, by studying the genome of species with different ageing phenotypes, we may be able to find genes or functional regions involved in the molecular mechanisms of ageing. Researchers in ageing have already recognised the importance of exploiting the vast amount of data available in the literature by constructing the Human Ageing Genomic Resource (de Magalhães, Budovsky, et al. 2009), a collection of databases and tools to help researches understand the genetics of human ageing, including a database of ageing- and longevity-associated genes in model organisms (GenAge) and a compilation of data on ageing, longevity, and life history in over 4,000 species (AnAge). Comparative genomics studies in ageing are increasing in number, but we believe that it has the potential to grow much more.

One of the most recent and exciting studies in comparative genomics are genome wide association studies (GWAS) (Ku et al. 2010). In essence, GWAS are high-throughput hypothesis-free studies that examine genetic variation across the genome in order to associate genetic variations to phenotypic traits (Hunter et al. 2008). These studies usually involve two groups of the same species one with a certain phenotypic trait, commonly a disease, and the other without this trait. After analysing the genome of the individuals in each of the two groups, one can construct a set of markers such as single nucleotide polymorphisms (SNPs) for which one variation is significantly enriched in members of one group. These genetic variations are then considered as a pointers to which region of the genome are responsible for the trait differences and may then be further analysed. GWAS were successful in identifying genes associated to numerous diseases. For instance, in 2007 researchers have identified genes associated to type II diabetes in the first major GWAS study (Sladek et al. 2007). Also in 2007, the Wellcome Trust Case Control Consortium carried out genome-wide association studies for seven common diseases including bipolar disorder, coronary artery disease, Crohn's disease and type 1 and type 2 diabetes (WTCCConsortium 2007). The success of GWAS led ageing researchers to carry out their own and targeted genetic differences between centenarians and control individuals. So far, GWAS on centenarians have revealed few lipoproteins associated to long life in humans (Bergman et al. 2007), and the usage of SNPs could predict with up to 77% accuracy exceptional longevity (Sebastiani et al. 2010). Although there is still a lot of work to be done, the initial results obtained from GWAS on centenarians are encouraging. Although GWAS show great potential to discover genes responsible for ageing divergence within a single specie, we were more interested in discovering genes responsible for ageing divergence across multiple species and in particular across mammals.

More in line with our interests are Ka/Ks approaches which considers the evolutionary pressure on different coding regions of the genome across multiple

species. The Ka/Ks ratio is a proxy for the selective pressure on a protein coding gene and is defined by the ratio between the number of non-synonymous substitutions and the number of synonymous substitutions for a gene. A high Ka/Ks ratio suggests a selective pressure and a low one suggests purifying selection. Studies screen the whole genome in order to find genes with significantly higher or lower Ka/Ks ratio compared to a control orthologous gene and attribute these genes with the evolution of new traits or involvement in an important, well-conserved, pathway. This type of method was used in many studies such as a genome-wide survey of pseudogenes (Torrents et al. 2003) and the detection of of rapidly evolving genes in human (Wang et al. 2003) among others. To study the evolution of ageing, (de Magalhães and Church 2007) used a Ka/Ks approach on human-chimpanzee orthologous gene pairs and reported that genes associated with ageing in non-mammalian model organisms and cellular systems appear to be under stronger evolutionary constraints than those associated with ageing in mammal and also provided evidence suggesting the rapid evolution of Werner syndrome gene in the hominids. Though Ka/Ks is a good indicator of selective pressure at the sequence level, Ka/Ks based approaches cannot detect changes in the expression of genes nor account for the different type of amino acid substitutions.

Fortunately, it is possible to use other proxies for evolutionary selection based on substitution matrices that do take into account of the types of amino acid substitutions. The evolutionary landscape of a genome varies greatly from region to region and although the substitution rates for two neutrally evolving regions of sequence are usually well correlated, this correlation decreases rapidly as the genomic distance between the regions increases (Gaffney et al. 2005). Indeed, the rate of sequence mutation can depend on multiple factors such as the location of the region and the nucleotide composition of neighbouring sites (Hardison et al. 2003). Because of this, comparison of evolutionary rates within a single genome can be difficult. Rather, one usually compares the loci of multiple orthologs in a range of species in order to determine the relative mutational rates of the loci as a common

phylogenetic relationship and divergence times can be assumed for orthologs. This is the reason why most studies are based on the alignments of orthologous sequences. Needless to say, for these kind of analyses, constructing an accurate orthologous mapping of regions in different species is of critical importance in order to avoid biases due to sequence location and composition.

3.4 Genes expression profiling

One other important area of comparative genomics is mRNA expression profiling. The DNA is arguably the most important molecule in life, but its transcription and the regulation of its transcription are as essential. The expression activities of the genes in the genome is a key component of the link between the genotype of a species and its phenotypes. In addition, gene expression profiles represent the primary level of integration between environmental factors and the genome, providing the basis for protein synthesis which ultimately guides the implementation of complex phenotypes such as morphology and behaviour (Renn et al. 2004). Therefore, by comparing gene expression profiles, one can study the molecular basis of phenotypic variation.

As evidence of the importance of mRNA expression profiling, a PubMed search for the term “microarray”, which is one method for genes expression profiling amongst many (Schena et al. 1995), yields over 38,000 hits. In mRNA expression profiling, the mRNA levels of two different individuals or species are measured and then compared in order to find differentially expressed genes responsible for divergent traits or states (diseased versus healthy). In fact, gene expression profiling has been crucial to the identification of important genes involved in a number of disease such as many types of cancers (Cooper 2001). Other types of studies explore the differences in genes expression in different tissues in order to discover the key differences in protein composition. Genes expression analyses can also be used to study genes regulation and signalling

pathways such as the IKB signalling pathways (Rao et al. 2010), to study the function of microRNA (Guo et al. 2010) and a wide range of biological phenomenon.

Within the biology of ageing community, differentially expressed genes have been studied in the context of calorie restriction (Lee et al. 1999), the ageing human brain (Lu et al. 2004), and many other ageing tissues in multiple organisms such as human T cells (Remondini et al. 2010), mouse bone marrow mesenchymal stem cells (Wilson et al. 2010) among many others. In addition, meta-analyses have also been conducted such as in de Magalhães, Curado, et al. (2009) and the importance of gene expression analysis is highlighted by the creation of gene expression database for ageing mice AgeMap (Zahn et al. 2007) and by the success of the Gene ageing Nexus, a microarray data repository and data mining tools (Pan et al. 2007).

Like computational biology, comparative genomics is often used as an exploratory tool in a hypothesis-free fashion (Barnes 2007). Its scope is generally to find species differences at a genome level in order to explain differences at the phenotype level. For some studies, researchers start with a good idea about the biological phenomenon at hand and the results can be used to confirm their hypotheses. However, in many ageing studies, researchers are looking to refine the scope of their research by first identifying candidate genes or pathways to study. That said, comparative genomics is often a good way to start studying a particular biological phenotype which little is known. For example, by using different comparative genomics approaches, ageing researchers were able to study SNPs association with exceptional longevity in humans, selective pressure on genes associated with ageing in different lineages as well as genes expression differences in many age-related diseases. Still, there are many more uses of computational biology and specially comparative genomics to study species divergence in ageing.

Chapter 4: Principles of comparative genomics

So far, we gave a brief summary of the theories of ageing and highlighted the field of computational biology. We also discussed some of the potential role of computational biology and comparative genomics in ageing research. Now, we will go through some of the principles of comparative genomics tools that are used in our own analyses.

4.1 Substitution matrices

Substitution matrices in comparative genomics are ubiquitous. They can describe the rate at which one nucleotide or amino acid changes into another over time. They can also represent the physiochemical difference or any other distance metric from one amino acid or nucleotide to another. For instance, in a matrix M, the entry M(A,T) could represent the likelihood that A is substituted by T over a period of time. Another example could be that for another matrix M(C, T) can represent the physiochemical difference between cysteine and threonine. There are many possibilities for substitution matrices and each have their particular use and advantages.

Indeed, the choice of the matrix should entirely depend on the its use. For instance, point accepted mutation (PAM) matrices are calculated by observing the differences in closely related proteins (Dayhoff 1965). As such, PAM matrices often perform poorly when used to align proteins from evolutionarily distant species. The Block substitution matrix (BLOSUM) are computed by looking at blocks of conserved sequences found in multiple protein alignments and has been shown to work better than PAM matrices when aligning evolutionarily divergent protein sequences (Henikoff et al. 1992). Other matrices such as the Grantham matrices uses amino acid physiochemical properties in order to determine the distances

(Grantham 1974), whereas the Naor matrices (Naor et al. 1996) use amino acid interchangeability at spatially, locally conserved regions to define a distance.

4.2 Sequence alignment tools

At the base of the majority of comparative genomics studies lies an alignment between homologous sequences. Aligning DNA or protein sequences is important for three main reasons. Firstly, in an evolutionary point of view, since any two species is thought to have a common ancestor, DNA sequences that are similar may be orthologous to each other which is to say that they stem from the same ancestral DNA sequence. In this perspective, aligned nucleotides are thought to have evolved from the same position in the ancestral sequence either after mutations or conservation and nucleotides that are aligned to gaps are either insertions or deletions in the other lineage. Secondly, DNA and protein sequences that are similar may share the same functions or structure. Thirdly, coding DNA regions are transcribed into mRNA and may then be translated into proteins. Given mRNA sequences, one can construct local alignments in order to find out from which coding region or genes the mRNA come from; which is crucial in genes expression profiling studies. Thus it is clear that sequence aligners have a crucial role in genomics and comparative genomics. As with other algorithms presented, understanding how sequence alignments are generated with different parameters can be of importance to their interpretation in particular when divergence occurs.

There are two general approaches in sequence aligners. There are local aligners and global aligners. When performing a local alignments, one wants to find all regions in the sequences that are similar according to some metric of similarity. The sequences aligned can be much smaller than the initial sequences and they need not be in any particular order. In comparison, when performing a global alignment, the entire sequences are aligned in the same order as the initial sequences with the addition of gaps. Which approach is the best depends greatly on the context of the

initial sequences and the scope of the alignment.

The Needleman-Wunsch algorithm is a pairwise global DNA alignment sequencing tool which takes two DNA sequences as input and outputs the optimal alignments for these 2 sequences according to some parameters. The Needleman-Wunsch algorithm, like many algorithms in computational biology and comparative genomics, aims to maximise an objective function characterising the biological significance of the sequence alignment. Since sequence alignments can be used for different tasks, there exists different objective functions which are mostly based on the different substitution matrices plus extra parameters. In the Needleman-Wunsch case, the parameters are the nucleotide substitution matrix plus the gap opening and extending score and can be tweaked when aligning sequences with different evolutionary distances. For more details and an in depth analysis of the statistics behind aligners, see (Needleman et al. 1970) and (Waterman 1995).

Many of the newer alignment tools are derived from Needleman-Wunsch, Smith-Waterman is an immediate derivative of Needleman-Wunsch and is used for optimal local alignments (Smith et al. 1981). One can even extend the Needleman-Wunsch to iteratively compute an optimal multiple sequence alignment by defining an extended substitution matrix and objective function. Unfortunately, even with the computational power available now, optimal alignments tools are too slow when used for multiple sequence alignments with many sequences or when used in local alignments with whole genome input. This led to a variety of suboptimal alignment algorithms with different heuristics to improve running speed. Perhaps the two most famous algorithms are ClustalW and BLAST.

ClustalW is a multiple sequence aligner which is suboptimal in the sense that it does not guarantee to output the multiple sequence alignment which maximises a specific objective function. ClustalW does a pairwise alignment first with the input sequences and creates a phylogenetic tree based on the pairwise

alignment. It then uses the phylogenetic tree in order to iteratively construct the whole multiple sequence alignment by going from the leaves to the root. The heuristics here is that ClustalW assumes that the initial pairwise alignment will yield the correct (or nearly correct) phylogenetic tree; once the phylogenetic tree is constructed, it will remain unchanged for the rest of the alignments and thus may produce alignment errors due to the wrongly inferred phylogeny.

BLAST is another algorithm using a heuristic which makes the algorithm suboptimal. BLAST takes a DNA sequence as input and aligns it to another DNA sequence which is often a big database containing gigabytes of DNA sequences. For each database, BLAST has a dictionary of seeds sequences of varying length ($l \geq 4$) with their location within the DNA sequences, it then matches the stretches of the input DNA sequence to the seeds and tries to extend the alignments in order to find the stretch of DNA in the database which obtains the highest alignment score with respect to the input sequence, extremely similar to Needleman-Wunsch. The trade-off is between seed size and speed as each seed of length 4 is located on expectation $N/4^4$ times in the database (where N is the size of the database in nucleotides or amino acids) which would slow down considerably the alignment whereas BLAST with a seed length of 20 would not be able to align a sequence with one mutation (compared to the sequences in the database) every 20 nucleotides (Waterman 1995). Most comparative genomics algorithms are fairly sensitive to parameters and it is often crucial to know the type of relationship between the sequences that are to be aligned.

4.3 Ancestral genome reconstruction

As mentioned earlier, comparative biology relies on the fact that any two species share a common ancestor. Ancestral reconstruction algorithms have scope to find the DNA or protein sequence of the ancestral species according to the protein

sequence of extant species. As such, ancestral reconstruction is an important part of comparative genomics. Ancestral reconstruction algorithms generally take as input orthologous multiple sequences alignment and a phylogenetic tree, and outputs the predicted sequences at each node of the phylogenetic tree.

Existing methods for ancestral sequence reconstruction fit in two categories which use a maximum parsimony (MP) or a maximum likelihood (ML) approach. MP approaches do not take into consideration branch lengths when reconstructing the ancestral sequence nor the non-uniform distribution of nucleotide or residue substitutions. For these reasons, ML approaches are generally better than MP approaches. ML ancestral reconstruction approaches take advantage of DNA or protein substitution models such as the Hasegawa, Kishino and Yano (HKY) model for DNA (Hasegawa et al. 1985) and the PAM for proteins in order to predict the most likely ancestral sequence (Dayhoff et al. 1978). Both the HKY and PAM model take into consideration branch lengths and types of substitutions. However, ML approaches normally do not support gaps in alignments while MP offers a simple and elegant way of supporting gapped alignments.

The MP approach is an optimisation algorithm on the parsimony of a phylogenetic tree and the sequences on their node. Given a tree structure and sequences at the leaves of the tree, the maximum parsimony is achieved when the sum of the numbers of substitutions along all branches of the tree is the smallest. It is thus easy to see that the MP approach is very dependent on the tree structure and does not take into consideration branches length or substitution types, but it is also clear that by extending the DNA or amino acid alphabet with the gap character, we can handle alignments with gaps.

The ML approach uses evolutionary models such as the HKY which models nucleotide substitutions. Over a long time span, the nucleotide at a given site might change and the HKY models the change in frequency of a nucleotide substitution

with respect to time. Indeed, depending on the evolutionary time, the substitution probabilities change and can be modelled as a Markov chain. The ML algorithm is an maximisation algorithm as it maximises the likelihood of the the sequences at the inner nodes of the tree given the tree structure and the sequences at the leaves. This likelihood is defined as the probability of getting the sequences at the leaves given the substitution model, the tree structure and the sequences at the inner nodes. For each two branches with different lengths, the substitution matrices will also be different according to the evolutionary model. As such, ML approaches generally perform better than MP approaches on simulated datasets, however, as opposed to in MP approaches, it is not straightforward to analyse insertions and deletions in a ML fashion. For a much more detailed discussion about these two estimation approaches, see (Li 1997).

4.4 Alignments of short reads

With the advent of second generation sequencing, the cost of sequencing has drastically decreased but so has the length of each DNA sequence read. By using second generation sequencing, a tradeoff between read length and coverage is made. While Illumina Solexa and ABI SOLiD offer high output short read lengths, Roche 454 offers lower output with longer read lengths. The decision on which platform to perform sequence often depends on the end goal of the project. For instance, for *de novo* genome assembly in eukaryotes, researchers usually use 454 sequencing as it is extremely hard to handle genome repeats when using short reads. On the other hand, when a good reference genome is available, Illumina Solexa and ABI Solid are preferable as it is generally straightforward to map shorter reads to a good reference genome. One can use the extra few order of magnitude reads to conduct stronger statistical analyses for SNPs discovery, genes expression profiling through RNA-seq or transcription factor activity sites localisation with ChIP-seq (Shendure et al. 2008).

A typical Illumina Solexa run outputs several millions short reads varying from 36bp to over 70bp, it is thus impractical to use local aligners such as Smith-Waterman because of the sheer number of alignments that has to be made. BLAST is too slow for many projects and its sub-optimality also makes it unusable. For this reason, a new wave of algorithms specially designed to map short reads have been invented including MAQ (Li, Ruan, et al. 2008), Eland, SOAP (Li, Li, et al. 2008) and bowtie (Langmead et al. 2009). Most of the new algorithms make trade-offs between speed and quality. Here, we discuss MAQ as it is arguably the most successful short read aligner and is also the one we use.

To have an idea of how MAQ aligns short reads from Illumina Solexa, one first needs to be aware of the outputs of Illumina Solexa sequencing. Illumina Solexa typically outputs millions of reads of the same length which usually vary from over 30 to over 70 base pairs. Along with each read, a quality score for each base pair is included. This quality score, also known as the Phred quality score, measures the probability that base call, i.e. the estimate of the true nucleotide, is wrong. For instance, a Phred score of 10 translates into a base call accuracy of 90% while a score of 40 translates into an accuracy of 99.99%. Much like the Phred score, MAQ assigns a mapping score to each read which estimates the probability that the read is misaligned. To do that, MAQ first maps the reads onto the reference genome by dividing the read into multiple seeds much like BLAST and finding these seeds in a hash to speed up the process. The use of the pigeon-hole principle allows a guarantee mapping reads with 3 mismatches or less as there can be 4 non-overlapping seeds in a read and at least one of these seeds will be error free. Then, MAQ considers all alignments to compute the mapping qualities. The mapping qualities has the properties such that reads falling in repetitive regions of the reference get very low mapping qualities as the reads align to many different regions and low quality base calls lead to low mapping qualities as the read sequence may be wrong. Furthermore, a read alignment mapping quality of 30 or over (Phred scale) usually implies that the quality of the base call is good, that the

best alignment has few mismatches and that the read does not map well to multiple different locations.

As MAQ depends heavily on the reference genome, it is important to make sure that a close reference genome is available. MAQ cannot handle insertions nor deletions with unpaired reads and can only guarantee to find an alignment to a read if there are 3 mismatches or less within the 32 first base pairs of the read. It is also important to note that the quality of the reference also influence MAQ's alignment score and sometimes in a paradoxical way. For instance, a reference genome constructed from short reads cannot efficiently handle repeats that are longer than the length of the reads yielding a lower quality reference genome. Yet, using such reference genome improves the alignment scores for reads coming from the repeat regions. Therefore, one must show care when comparing the results of mappings on two reference genomes of different qualities specially when reads from repeat or common functional sequences are present.

4.5 Phylogenetic independence contrasts and other corrections

Closely related species, in an evolutionary sense, share more similar phenotypes than species that are evolutionarily distant. In fact, for any two animals, we can go up the phylogenetic tree and find a common ancestor. Thus, whenever we find a correlation between two phenotypes in comparative studies involving more than two species, we need to make sure that this is independent from phylogeny, i.e. that phylogeny does not play a role in this correlation. For example, de Magalhães, Costa, et al. (2007) analyses the relationships between metabolism, developmental schedules and longevity and used a statistical method called phylogenetic independent contrasts in order to make sure that the correlations they report were not due to phylogeny.

Briefly, phylogenetic independent contrasts is a statistical algorithm in order

to transform the set of variables in a comparative studies into another set of variables that are statistically independent and identically distributed. For any two adjacent leaves in a phylogenetic tree, we can transform their values into one value that is drawn from a normal distribution with mean 0 and known variance that is independent from any other two adjacent leaves (assuming a Brownian motion governs the evolution of the values from each node to their children). After this step, it is possible to obtain values for inner nodes of the tree by iterating the steps using the new values as leaves values. Once all the values have been computed for both phenotypes under study, one can use the new corrected values for correlation or regression analyses. For more details, see (Felsenstein 1985).

Phylogeny is one type of confounding variable that may alter the result of a regression in comparative biology. However, there are other confounding factors that need to be addressed when studying ageing. As we saw in chapter 2, body size strongly correlates to longevity. Therefore, any phenotype correlated to body size will be also correlated to maximal lifespan without any real biochemical connection. Consequently, there is a need to factor out body size from the correlation.

Fortunately, this task has been studied extensively in the field of statistics. To solve this problem, one can set experimental controls such as case-control studies, cohort studies and stratification methods in order to limit the effect of the confounding variables. However, in many comparative biology studies, it is impossible to decide on the experimental conditions as the data are often obtained from collaborators or online databases. In the latter case, one can use methods in covariate statistics in order to factor out variables that may be confounding (Hennekens et al. 1987).

AIMS

The aim of our work was two-fold. First, our goal was to produce comparative genomics algorithms and tools to help biologists discover interesting patterns in the genomic data that is now available and that will be available in the near future. The amount of data is growing exponentially with the advent of new sequencing technologies and we had this in mind when working on our algorithms. The large quantity of data requires the algorithms to be efficient and also designed in such a way that errors or outliers in the data can be detected without causing too many false positives. All these tools use data generated from two or more species in order to capture differences in their genes and proteins sequences or differences in genes expression that may explain phenotypical differences under study.

Our second goal was to use these techniques in order to study a concrete biological phenomenon which is the molecular mechanisms of ageing in mammals. Since the phenotypes in mammalian ageing are largely similar with exception of its timing (de Magalhães et al. 2002), we thought that studying mammals with different maximal lifespans would be feasible as well as representative of the phenotypic differences that exist in mammalian ageing. By using genomic and expression data in three different contexts, we hoped to discover a common signal which will point to genes or pathways important in the evolution of mammalian ageing. These genes or pathways could give ageing researchers insights about the mechanisms of ageing in mammals and could be putative targets for future experimental studies.

As such, the remainder of this thesis will be divided into two major parts. The “Methods and Results” section will discuss the dataset, methods and algorithm development as well as present some raw results while the “Discussion” section will describe our interpretation of the results in the context of mammalian ageing.

METHODS AND RESULTS

In the introduction, we discussed computational biology and how it has been used to help researchers in ageing provide new insights about the underlying mechanisms of ageing. We also outlined some of the basic principles on which most computational techniques is built upon. In this chapter, we describe the comparative genomics methods based on these basic principles that we employed and original results which provide novel clues about mammalian species divergence in ageing. This chapter will be separated into three parts each consisting of a different project. We will first start with our analyses of residue frequencies in mitochondrial protein. Though the analyses involved in our first project are relatively simple, there are few considerations that will highlight the intricacies of comparative genomics studies. Next, we will explore the detection of signatures of selection in proteins of long-lived mammals showcasing how to exploit the large amount of genomic data available for ageing research. And finally, we end by discussing our work on cross-species comparison of mRNA levels using next-generation sequencing without the reference genome of one of the two species which is, as far as we know, the first of its kind.

Chapter 5: Residue frequencies in mammalian mitochondrial proteins

In section 2.4, we saw that the mitochondria might have an important role in the regulation of ageing. Recent reports suggest correlations between residue usage in mitochondrial-coded proteins and longevity in multiple species. However, it is important to bear in mind that mitochondial DNA (mt-DNA) only codes for 13 proteins, a very small subset of mitochondrial proteins. Furthermore, previous studies did not focus on mammalian species. Thus, we wanted to test the hypothesis that some protein residues are correlated with maximal lifespan by looking at all the mitochondrial proteins of mammals with different lifespans. We wanted to see whether the hypothesis that mitochondrial coded protein cysteine content is negatively correlated with maximal longevity from (Moosmann et al. 2008) still holds when only considering mammals or when phylogeny is taken into account, and whether this correlation extends to all mitochondrial proteins (not only those who are encoded in the mitochondria).

By using simple computational genomics techniques, one can compare the residues of a set of proteins to the residues of one or many orthologous sets and search for correlation between the residue composition and lifespan. Since residues with antioxidant properties are likely to impact on lifespan by acting as a protective buffer against oxidative damage, we wanted to test whether the composition of antioxidant residues such as cysteine and methionine in mitochondrial genes is correlated with lifespan in mammals. To test this, we analysed residues composition in mitochondrial proteins classified in three different sets and used a group of non-mitochondrial proteins as control set. The first set consists of the 13 proteins encoded in the mitochondrial genome, the second set consists of 52 well conserved mitochondrial inner membrane proteins, while the third set consists of 243 well conserved polypeptides that are encoded by nuclear genes and are imported into the

mitochondrion, lastly the fourth set consists of the collection of all 7512 highly conserved non mitochondrial proteins used as a control (see Appendix 1.3).

To be precise, we based our analyses on 10 mammalian species with significantly different lifespans and whose genome has been sequenced at a high coverage. These species included 4 primates (*Pan troglodytes*, *Homo sapiens*, *Pongo pygmaeus* and *Macaca mulatta*), 3 rodents (*Mus musculus*, *Rattus norvegicus* and *Cavia porcellus*), along with *Canis familiaris*, *Bos Taurus* and *Equus caballus*. The orthologous genes set were constructed using InParanoid, a comprehensive database of eukaryotic orthologs (Ostlund et al. 2010) by combining pairwise orthology maps while the protein sequence data was obtained from ENSEMBL. Next, different mitochondrial proteins were classified with their Gene Ontology annotations (Ashburner et al. 2000) into the four aforementioned sets. We then constructed the multiple sequence alignments for all proteins using ClustalW. Only proteins sharing at least 50% identity with the human sequence, including gaps, were considered as valid orthologs and were considered well conserved. Proteins with less than 50% identity or proteins with missing orthologs were removed from the analysis.

The next step was to compute cysteine and methionine frequencies in the 10 mammalian species. After doing that, we constructed a phylogenetic tree for the 10 mammals using the work of Miller et al. (2007) and applied phylogenetic independent contrasts (PIC) to correct for phylogenetic dependence. No correlation between cysteine and maximal lifespan has been found. However, we report a negative correlation between methionine content and (log) maximal lifespan in the first set consisting of proteins encoded in the mtDNA (Pearson's coefficient: -0.685, p-value: 0.086) and second set of proteins located in the mitochondrial inner membrane (Pearson: -0.561, p-value: 0.058). The correlation between methionine usage and MLSP were not significant in the set of mitochondrial protein encoded in the nucleus (Pearson: -0.229, p-value: 0.217) and non mitochondrial proteins

(Pearson: 0.024, p-value: 0.476). We then found that, as expected, methionine usage in mtDNA-encoded proteins (6.03% ± 0.59% SD) was much higher than any other protein group. Interestingly, we also found that within the set of mitochondrial proteins encoded by the nucleus (2.39% ± 0.04% SD), proteins classified in the inner mitochondrial membrane set showed a significantly higher methione frequencies (2.61% ± 0.05%, $p < 10^{-8}$, *t*-test). Methionine content seems to be significantly higher in the mitochondrial proteins than in non mitochondrial proteins ($p < 10^{-12}$). In our joint work with Aledo et al. (Unpublished), they present an even more general correlation extending our initial results using 10 mammals to 168 mammals including 24 different orders. In addition, they invented their own statistical method to test phylogenetic effects and correct them along with basal metabolic rates. Furthermore, they analysed spatial disposition of methionyl residues in mitochondrial proteins to investigate the correlation between differential methionine usages and maximal lifespan, and found that mitochondrial DNA encoded subunit from the short-lived mouse seems to accumulate methionine on the surface of cytochrome b when compared to the one in the long-lived human.

Following the suggestion of an anonymous reviewer, we tested whether methionine was enriched in mitochondrial proteins was higher than any other protein group when taking under consideration the facts that 1) there is a higher proportion of mitochondrial genes that code for membrane proteins than non-mitochondrial genes and 2) membrane proteins tend to have a higher proportion of hydrophobic residues which methionine is. Interestingly, we found that even under such consideration, mitochondrial proteins were still significantly more enriched in methionine than non-mitochondrial proteins. Moreover, we found that methionine residue increased much more, about 15%, in membrane proteins than any other residue. Also following the suggestion of the reviewer, Aledo et al. further developed new statistical techniques examining methionine addition and removal in short-lived and long-lived species respectively in order to test the hypothesis that the adaptation (methionine gain) occurred in short-lived species. They also

performed analyses to determine whether AUA methionine codon was correlated with longevity. They found that methionine gain in short-lived species supported the experimental data better and that AUA methionine codon was indeed negatively correlated with longevity. As discussed in our joint manuscript with Aledo et al. and in chapter 8, this supports our speculation that methionine has antioxidant role in the mitochondria. Also in chapter 8, we discuss that apart from a putative role in prevention against ROS, the antioxidant properties of methionine may have a functional role in mammalian lifespan.

Chapter 6: Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity

Thus far, we have discussed our 10 species comparative genomics study which studied the composition of certain amino acid residues in mitochondrial proteins and their correlation with maximal lifespan. However, there are now more than 30 mammalian species with whole genome sequences finished at over 1.5X coverage and many more are on the way. This provides researchers the opportunity of doing genome-wide comparative studies across a large number of species to identify genes and processes associated with the evolution of longevity (de Magalhães et al. 2010). Few successful comparative genomics studies in ageing (de Magalhães et al. 2007; Jobson et al. 2009) has exploited this wealth of data which provided new insights on the evolution of human and mammalian longevity. Here, we look for selection patterns in genes of lineages in which longevity evolved.

To identify candidate genes and processes underlying the evolution of longevity in mammals, we undertook a phylogenetic based comparative genomics study involving over 30 mammalian species. More precisely, our approach is based on the analysis of accelerated protein evolution in different lineages where longevity evolved. Our results reveal genes and functional groups that are candidate targets of selection in mammalian lineages where lifespan evolved. These include DNA repair genes and the ubiquitin pathway and thus provide evidence that at least some repair systems were optimized in long-lived species.

Since species divergence in lifespan and ageing phenotype is largely determined by genetics, one can expect to find proteins involved in species difference in ageing under positive selection in lineages where longevity evolved. Moreover, we expect these proteins to be under stronger selective pressure in lineages where longevity increased (MLI lineages or branches) compared to

lineages where longevity remained the same (MLS lineages or branches). In MLS branches, it is reasonable to believe that the selective pressure on most proteins contributing to species difference in ageing do not show departure from the neutral rate of sequence evolution. Thus, proteins with a pattern of selectivity specific to MLI branches are candidate proteins responsible for species divergence in ageing. Accordingly, our approach aims to detect proteins that have undergone selective pressure with high specificity to the lineages where longevity evolved.

To find genes and functional groups under selective pressure in phylogenetic branches where maximal lifespan increased (which we call MLI branches) compared to branches maximal lifespan did not increase (MLS branches), ortholog mappings of proteins of 36 mammalian species to *Homo sapiens* were obtained from ENSEMBL resulting in 15,350 proteins with at least one 1:1 ortholog. These 36 species were used to infer the most accurate ancestral protein sequence and few were no longer used in downstream analysis. The phylogenetic tree used was obtained from Miller et al. (2007) and completed with the work of Murphy et al. (2007) for the *Myotis lucifugus* and *Pteropus vampyrus* branches. Using these mappings and proteins multiple sequence alignments along with a reference phylogenetic tree, ancestral protein sequences for the 15,350 proteins were predicted using Gapped Ancestral Sequence Prediction (GASP) (Edwards et al. 2004). GASP uses a likelihood method to fix gap position in the given phylogenetic tree and uses substitution matrices to assign ancestral amino acids. Though GASP has been shown to be less accurate than more sophisticated tools using likelihood estimation, it can handle gapped alignments and was suited for our studies as proteins with divergent regions were discarded. In fact, since any phylogenetic approach aiming to detect selection is highly sensitive to wrongly annotated splice variants, proteins orthologs with more than 10 substitutions out of a sliding window of 20 residues were removed. After this scan, 15,312 proteins had at least one other ortholog.

We then computed an evolutionary pressure score for all proteins in all branches of the phylogenetic tree based on the type and number of amino acid substitutions in each branch. These evolutionary pressure scores measure the strength of the selective pressure on a protein in each lineage.

6.1 Detecting longevity specific selection in proteins

To detect proteins that underwent higher selective pressure in branches where maximal longevity significantly increased, AnAge (de Magalhães and Costa 2009) was used as reference for animal maximal lifespan and 9 closely related species pairs for which their maximal lifespans are significantly different were constructed (see Table 2). In addition to this, 7 control pairs consisting of 2 species with similar maximal lifespans were constructed. That is, the 9 experimental pairs each correspond to species resulting from one MLI lineage and one MLS lineage stemming from a common ancestor and the control pairs to species resulting from two MLS lineages also stemming from a common ancestor. We wanted to detect proteins that have undergone stronger selective pressure in MLI lineages compared to MLS lineages in experimental pairs while exhibiting the same selective pressure in both MLS lineages in control pairs.

For each of the 15,312 proteins, substitutions scores based on the physicochemical properties of the residue substitutions (Grantham 1974) was computed in each branch as a proxy for selective pressure akin to (Zhang et al. 2002) where they use the number of residue substitutions as a measure for evolutionary pressure. It should be noted here that the use of similar matrices as Grantham did not alter the results obtained. For each protein and branch, the expected value of the number of residue substitutions was computed according to the empirical distribution of the residue substitutions in the branch in all proteins.

Longevity divergent pairs (Experimental pairs)	Lineages where longevity evolved	MLSP	Lineages where longevity did not evolved	MLSP
	Choloeapus hoffmanni	37	*Dasypus novemcinctus*	22.3
	Equus caballus	57	*Canis Familiaris*	24
	Myotis lucifugus	34	*Pteropus vampyrus*	20.9
	Loxodonta africana	65	*Procavia capensis*	14.8
	Cavia Porcellus	12	ancestor of *Mus Musculus* and *Rattus Norvegicus*	4*
	Tursiops truncatus	51.6	*Bos Taurus*	20
	Homo sapiens	122.5	*Pongo pygmaeus*	59
	Macaca mulatta	40	*Callithrix jacchus*	16.5
	Erinaceus europaeus	11.7	*Sorex araneus*	3.2
Control pairs	*Tarsius syrichta*	16	*Callithrix jacchus*	16.5
	Procavia capensis	14.8	*Echinops telfairi*	19
	Vicugna pacos	25.8	*Sus scrofa*	27
	Oryctolagus cuniculus	10	*Ochotona princeps*	7
	Rattus norvegicus	5	*Mus musculus*	4
	Spermophilus tridecemlineatus	7.9	*Dipodomys ordii*	9.9
	Otolemur garnettii	18.3	*Microcebus murinus*	18.2

Table 2: Experimental and control pairs used in the computation of the longevity specific scores. The maximal lifespan (MLSP) is measured in years.

*: minimum of MLSP (in years) of *Mus Musculus* and *Rattus Norvegicus*

6.2 Normalisation of evolutionary scores

Since different lineages have different evolutionary rates, one cannot simply compare the number of residue substitutions or the scores computed based on the number of substitutions alone between two lineages sharing the same ancestor. Thus, to compare the evolutionary scores between two lineages, we had to normalise the evolutionary scores specifically for each experimental and control pairs.

In order to normalise the scores properly, let us assume that we are comparing the evolutionary pressure on genes or proteins in two branches (A and B) with different evolutionary rate. We would like to be able to detect whether one gene or protein is under heavier positive pressure in one branch compared to the other. Let $SP_A(p)$ and $SP_B(p)$ denote a measure of selective pressure in branch A and B respectively for a protein p . In our case, it is the sum of Grantham measure of all substitutions between the predicted ancestral sequence and each of the extant species. Since the two lineages represented by the branches may have different evolutionary rate, we can expect a higher measure in one branch, e.g. $SP_A(p)$ higher $SP_B(p)$, for any protein p . Thus, comparing $SP_A(p)$ to $SP_B(p)$ directly would heavily bias the analysis as proteins with high evolutionary pressure in branch A are more likely to be detected as having a significantly higher score than in branch B .

Instead, we need to capture this difference in evolutionary rate and

normalise the two scores $SP_A(p)$ and $SP_B(p)$. The most simple and elegant solution we found was to normalise both scores by the expected number of substitutions each ancestral protein sequences given the protein sequence and linage. We first computed for each lineage *A* and *B,* the empirical probability distribution of amino acid residue substitutions using all the proteins with ancestral sequence prediction. Then, we normalised the selective pressure scores for each protein. Although this expected value is an under-estimate of the true number of substitutions due to the possibility of multiple substitutions at a single residue site, this effect is minimal as the lineages used are short and the proteins are well conserved.

In mathematical terms, let $E(A,B)$ be the set of all ancestral sequences predicted for species *A* and *B*. The empirical probability that amino acid *R* is substituted for another amino acid in branch *A* is:

$$P(R;A,B)=\frac{\sum_{p\in E(A,B)}\sum_{i}\mathbf{1}_{\{aa_{ancestral}(p;i)=R,\,aa_A(p;i)\neq R\}}}{\sum_{p\in E(A,B)}\sum_{i}\mathbf{1}_{\{aa_{ancestral}(p;i)=R\}}}$$

where the first summation runs through all predicted ancestral sequences for species *A* and *B* while the second summation runs through the amino acid in each protein. $aa_{ancestral}(p;i)$ is the amino acid in the ancestral protein sequence *p* at position *i* of the alignment and $aa_A(i)$ is the amino acid sequence of the descendant of protein *p* of species *A* at position *i*.

Thus, the expected number of substitutions the ancestral protein sequence of a protein $p=aa_1aa_2...aa_n$ will have in lineage *A* is

$$N_A(p) = \sum_{i=1}^{n} E\{\mathbf{1}_{aa_{ancestral}(p;i) \neq aa_A(p;i)}\} = \sum_{i=1}^{n} P(aa_{ancestral}(p;i); A, B)$$

while the expected number of substitutions the ancestral protein sequence will have in lineage *B* is

$$N_B(p) = \sum_{i=1}^{n} P(aa_{ancestral}(p;i); B, A)$$

Thus, by normalising $SP_A(p)$ by $N_B(p)/N_A(p)$, we can compare the corrected evolutionary pressure scores $SP_A(p) \cdot N_B(p)/N_A(p)$ and $SP_B(p)$ to infer in which lineage protein *p* was under heavier selection.

6.3 Selectivity criteria and longevity selectivity scores

This expected value was used to normalize the protein score for each branch in order to minimize the effects of different branch length and protein length. After all scores have been normalized, for each of the pairs of species with divergent lifespan (for which an ortholog exist in both branch) and each of the control pairs, a p-value was calculated to measure the relative selective pressure in one lineage versus the other by the binomial test:

$$p(Score_{MLI}+c, Score_{MLS}+c) = \sum_{x=1}^{Score_{MLI}+c} \binom{Score_{MLI}+Score_{MLS}+2c}{x} (\frac{1}{2})^{Score_{MLI}+Score_{MLS}+2c}$$

Where $score_{MLI}$ is the substitution score (evolutionary pressure score) for the protein in the MLI branch in the pair, $score_{MLS}$ is the substitution score for same orthologous protein in the MLS branch of the pair and c is a pseudo-count (c = 3 in this study) which has been chosen by trial and error so that *p* is stable for conserved

proteins. For the control pairs where both branches (*A* and *B*) are MLS branches, only the smallest p-value is kept:

$$p_{control} = min\{p(Score_A + c, Score_B + c), p(Score_B + c, Score_A + c)\}$$

The lower the p-value from the binomial test is, the stronger the selective pressure on the protein in one lineage is compared to the other. Since well conserved proteins may also be responsible for species divergence in ageing, we defined 3 different p-value thresholds to account for the weak selective pressure they show. In other words, the different p-values, 0.05, 0.1 and 0.2, reflect different levels of evolutionary pressure on proteins. For example, proteins undergoing higher evolutionary pressure tend to have species pairs satisfying the 0.05 threshold as their evolutionary pressure scores tend to fluctuate, whereas proteins that are well conserved tend to have low evolutionary pressure scores which are less likely to fluctuate between lineages. Although well conserved proteins tend to have less experimental pairs satisfying the thresholds, they also tend to have less control pairs satisfying them. To see this, consider a well conserved protein with a weak signature of selectivity specific to MLI branches. Under the stringent cut-off, this protein will show no signature of selectivity in any branch at all. Conversely, consider a protein that exhibits a signature of selectivity in many different branches but much stronger in MLI branches. The variability of the strength of selective pressure is big and so many branches will be considered to be under selection with respect to a relaxed cut-off hence yielding a low score because of the lack of specificity to MLI branches. We used this fact to define a "longevity-specific selectivity" for each proteins and each thresholds. By using different selectivity criteria, we were able to detect proteins showing different levels of selection with high specificity towards MLI branches. The longevity-specific selectivity score does not measure how rapid the evolution a protein has undergone in MLI branches.

Rather, it measures the specificity of this selection in MLI lineages compared to MLS lineages. That is to say, proteins with high LSS score may be under positive selection in MLI branches (but not in MLS branches) or under purifying selection in MLS branches (but not in MLI branches). Whether it is the former or latter case, a high LSS score translates into a protein with a signature of longevity selectivity.

Indeed, the fact that different proteins are under different evolutionary pressure was taken into consideration by considering 3 thresholds for significance (*t = 0.05 (stringent), 0.1 (moderate), 0.2 (relaxed)*). And therefore, to assess whether a particular protein is associated to ageing, the number of pairs, N_{MLI} , where the proteins has a p-value such that $p < t$ is recorded. The number of pairs, N_{MLS} , such that $1-p<t$ is also recorded as well as the number of control pairs, $N_{control}$, such that $p_{control}<t$. With these numbers, a normalized "longevity-specific selectivity" score for each protein can be computed as follows:

$$LSS'=N_{divpair}\left[\frac{N_{MLI}}{N_{divpair}}-\frac{\left(N_{MLS}+N_{control}\right)}{\left(N_{divpair}+N_{contpair}\right)}\right]$$

Where $N_{divpair}$ is the number of pairs with divergent lifespans for which a protein ortholog exists in both branches, $N_{contpair}$ is the number of control pairs for which a protein ortholog exists in both branches. Therefore $N_{MLI}/N_{divpair}$ is the percentage of MLI branches in which the protein is selected and $\left(N_{MLS}+N_{control}\right)/\left(N_{diver}+N_{contpair}\right)$ is the percentage of MLS branches in which the protein is selected. It can be seen that $0\le N_{MLI}/N_{divpair}\le 1$ and $0\le\left(N_{MLS}+N_{control}\right)/\left(N_{diver}+N_{contpair}\right)\le 1$ so that $-N_{divpair}\le LSS'\le N_{divpair}$ also holds with equality when the either (1) the protein is selected in all MLI branches and no MLS branch or (2) the protein is selected in all MLS branches in pairs with

divergent lifespans and in 1 branch in all control pairs. Moreover, assuming we have a threshold of $t = 1$, both $N_{MLI}/N_{divpair}$ and $\left(N_{MLS}+N_{control}\right)/\left(N+N_{contpair}\right)$ would be 1 and LSS' would equal 0. In this study, only proteins that have a selection specificity towards MLI branches are of interest so $LSS=max\left(0, LSS'\right)$ was used as the "longevity-specific selectivity" score. It follows that any protein with negative LSS' has a LSS of *0,* i.e. no longevity-specific selectivity. Thus the *LSS* score measures the specificity of the selection in MLI branches.

In sum, we gave each protein three "longevity-specific selectivity" scores computed according to the number of experimental pairs where the protein was under stronger selective pressure in MLI branches, the number of both experimental and control pairs where the protein was under stronger selective pressure in MLS branches, and the total number of divergent lifespan species and control pairs considered. As such, the "longevity-specific selectivity" score encapsulates how specific the selection of the protein is to lineages where maximal lifespan increased (MLI branches) with regard to an arbitrary criteria of stronger selection.

The proteins were then sorted according to their scores separately for each of the three levels of selective pressure defined and we were able to analyse the proteins showing the highest specificity towards MLI branches. The geometric mean does not have any intrinsic meaning, but we used it in order to summarise the data into one single table. For this ranking, no statistical significance test were performed as a random model for protein residue evolution at a genome-wide level, taking into consideration multiple species pairs, was outside the scope of this study. Instead, we focused on the top 25 proteins as candidate proteins related to species difference in ageing with regard to the three cut-offs which correspond the top 0.5% of all proteins analysed.

6.4 Proteins with longevity-specific selectivity

After computing the "longevity-specific selectivity" score for each of the 15,312 proteins in all three selectivity categories, we found that proteins generally obtained a low score with most proteins scoring 0. When applying the relaxed selectivity threshold, 10,182 proteins out of the 15,312 had a score of 0 with only 598 proteins scoring 2 or more and a mere 31 scoring 4 or more, the highest score of 6.0 was attained by only one protein, FAM126B. A typical high scoring protein has a selectivity pattern like DDB1 (rank 9; see fig. 2). Likewise, when applying the moderate selectivity threshold, 10,216 proteins had a score of 0, 339 had a score of 2.0 or more and 6 with a score of 4.0 or more with CPNE5, NUP85 and RSAD2 sharing the top score of 5.0. Lastly, when applying the stringent selectivity threshold, 10,723 proteins had a score of 0, 166 had a score of 2.0 or more. The highest score was obtained by the protein IWS1 scoring the highest with 4.33, followed by HERC4 which shares the score of 4.00 with 7 other proteins. Though some overlap between categories exists as expected, 9 proteins are ranked highly (top 20) in two categories and only one protein, FAM126B, is highly ranked in all 3. Though this ranking bears no statistical significance in itself, the scoring is a rank-order of all proteins and these highly ranked proteins represent the most promising candidates for selection in long-lived species.

6.5 Detecting longevity specific selection in functional categories

After computing the "longevity-specific selectivity" scores for all proteins, GO categories annotation (Ashburner et al. 2000) were obtained in order to score each GO categories by adding the "longevity-specific selectivity" score of each protein within the category. To compute the significance of each GO category, the empirical distribution of the *LSS* scores was obtained, the proteins scores were shuffled and the scores for the GO categories were recomputed 2,000 times. The p-value for each category was computed as the number of times the simulation

yielded a score for a GO category that is larger than the its actual score divided by 2,000 (permutation test). Again, for each GO category, we obtained three enrichment p-values, one for each level of selective pressure, and analyzed the categories with lowest p-values.

In the analysis of GO categories, the high score of one protein could largely influence the p-value of a small GO category thus we only included categories with at least 3 proteins showing specificity of selection towards MLI branches. For instance, in the GO category proteasomal ubiquitin-dependent protein catabolic process we have 5 proteins showing specificity at the relaxed selectivity level including FAF1 (1.85), RAD23A (0.20), TRIM63 (2.80), RAD23B (0.67) and CD2AP (4.86). These 5 proteins give the GO category a "longevity-specific selectivity" score of 10.38 while the expected score is 3.00.

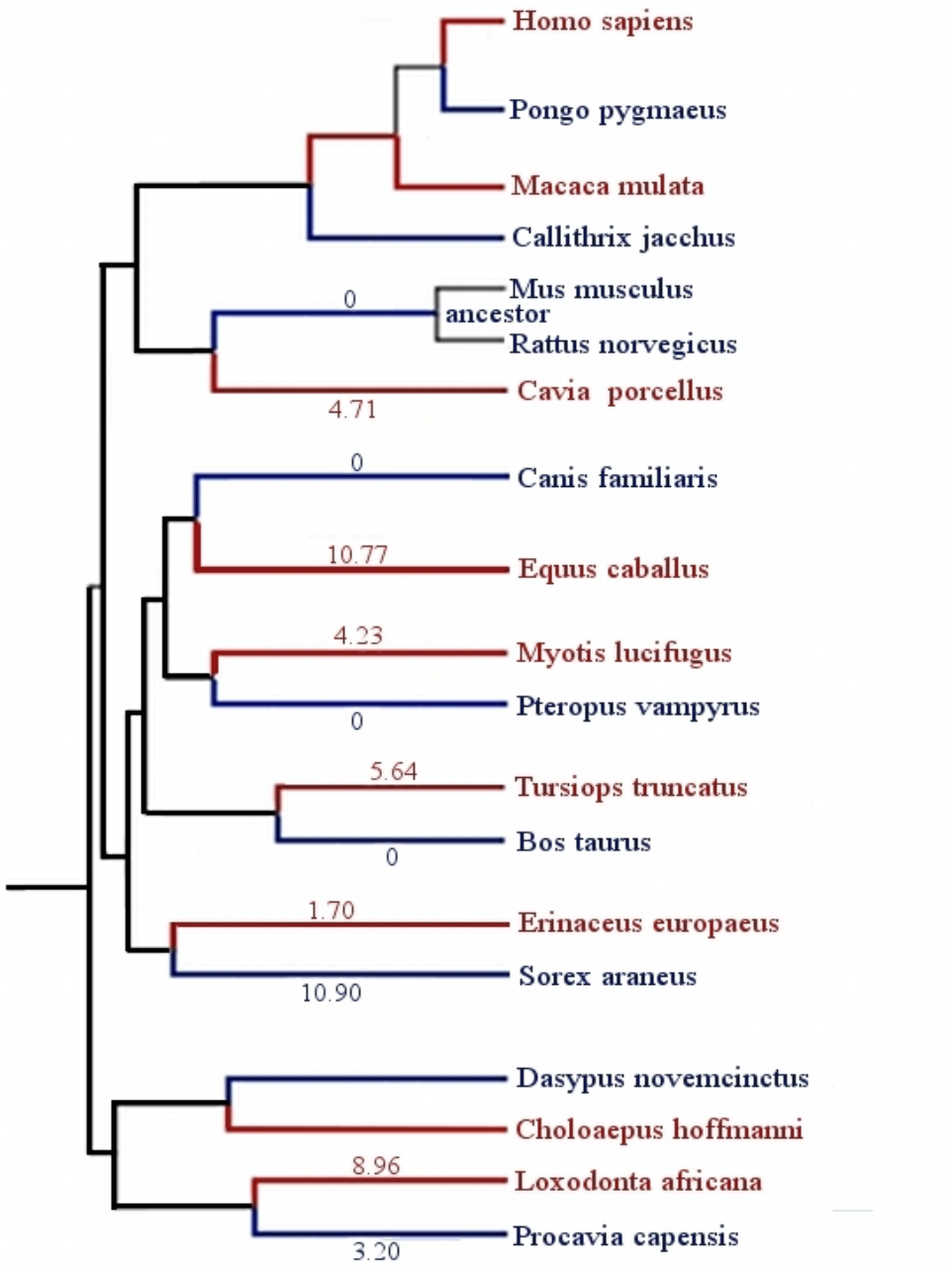

Figure 2: DDB1 under selection in longevity specific lineages.

Longevity drastically increased in red lineages compared to blue lineages. Protein substitution scores are labelled next to branches when available. DDB1 selection was not specific in the lineage leading to Sorex araneus.

We then sorted all categories according to the geometric mean of the p-values for the three thresholds of selectivity. Out of 15,551 GO categories considered, 4,180, 4,396 and 4,443 categories had a non-zero score with respect to stringent, moderate and relaxed selectivity pressure criteria respectively. 3,267 categories had non zero scores with respect to all three criteria, 1,147 with respect to two and 924 had a non zero score with respect to one. We obtained around 150 GO categories with significant p-value with respect to at least one selectivity criteria. This represents less than 1% of all functional groups and many of the categories are closely related to one another (see Appendix 1.2). Because no statistical correction were applied for the multiple testing of GO categories, this ranking, although based on p-values, is a rank order without a specific statistical cut-off. Like in the case of proteins ranking, a mere ranking of functional categories can lead to false positives, however, the categories with lowest p-value represent categories which show the most longevity specific selectivity regardless of the statistical significance. Out of the top 150 GO categories with respect to each selectivity criteria, 33 categories are in the top 150 with respect to all 3, and 76 with respect to 2. Furthermore, many of the top categories are related to each other without necessarily sharing a common proteins with high "longevity-specific selectivity" score. Thus they are likely to be biologically relevant and we grouped these categories into different classes.

Three major classes of functional categories were detected within the highly ranked GO terms in at least one threshold of selectivity. Within the first class, proteins involved in muscle development along with brain development showed the most significant enrichment in high "longevity-specific selectivity" scores when using threshold for moderate and stringent evolutionary pressures. However, when we use a lower threshold for selective pressure, the evolutionary pressure specific to MLI branches were not considered significant any more. Another GO category related to development was spermatid development which was enriched in "longevity-specific selectivity" scores with respect to the three thresholds, but

particularly when the threshold was medium or low. Next, we found that proteins involved in lipid process were also among the statistically significant categories along with cholesterol catabolic process which only show significance at a high selectivity threshold. The last class of proteins comprise of four functional categories involved in the proteasome-ubiquitin system which are protein ubiquitination during ubiquitin-dependent protein catabolic process, proteasomal ubiquitin-dependent protein catabolic process, ATP-dependent peptidase activity and lysosome organization. More surprisingly, the proteins with positive "longevity-specific selectivity" scores detected in these categories were non-overlapping.

We also detected other GO categories with unusually high "longevity-specific selectivity" but with few highly ranked related categories. As such, they are more likely to be false positives. Both actin binding and actin cytoskeleton proteins were ranked highly with regards to all cutoffs. In addition, we found that 1-phosphatidylinositol-3-kinase activity, phosphoinositide 3-kinase complex, response to food and circadian rhythm were all highly ranked with respect to at least one selectivity criteria.

6.6 Checking computational bias

To see whether our approach was biased towards some unwanted protein properties, we looked at the distribution of protein lengths and no correlation with score was found. Furthermore, we looked among the top proteins in each category and found that none of them had an obvious splice variant causing bias in our approach. Many phylogenetic approaches using othologs mappings like ours are highly sensitive to proteins with misannotated orthologs, thus we used an aggressive weeding strategy to remove proteins sequences that are putative splice variants by removing any proteins with more than 10 mismatch in a sliding window of size 20 when compared to its closest ancestral protein sequence prediction. Moreover, phylogenetic approaches need to take into consideration the phylogenetic

dependence within the species considered (Felsenstein 1985), however, our method is unaffected by this dependence as we defined disjoint longevity divergent species pairs with no common evolutionary branch. To further test our approach, we wanted to verify that our selection of parameters did not drastically influence our results. We found that using a different scoring matrix yielded similar results and no significant differences were observed within the proteins and GO categories reported when we used different thresholds for selectivity including 0.01, 0.15 and 0.25.

It is important to note that there is no obvious statistical model for the distribution of longevity selectivity score (LSS) among proteins. We will discuss in the next chapter how this affects the significance of the results and why we believe that our results are interesting despite the fact that we did not show deviation from a random model.

Chapter 7: Genes expression profiling using 2nd generation sequencing in non-traditional model organism

So far, we have looked at amino acid composition and protein evolution across mammalian species. However, it is probable that the mechanisms behind species divergence in ageing lies in non protein coding regions of the genome. For example, the decline in activity of proteasome in many mammals through its life is thought to be a cause for ageing (Friguet et al. 2000; Farout et al. 2008). Among the non protein coding regions of the genome are regulatory elements such as promoters and enhancers that can influence the expression of genes. The mechanism of genes regulation is an extremely complex topic, specially in eukaryotes, worthy of an entire field of researchers. Thankfully, it is not necessary to the mechanisms of genes regulation in order to study genes expression. In this project, we used second generation sequencing data from wild-type mice and naked mole-rats in order to screen for differentially expressed genes in the naked mole-rats that could, at least partly, explain its exceptional longevity.

As mentioned in the introduction, there a few parameters to consider when choosing a model organisms for ageing studies. Due to the relationship between maximal longevity and body size, a good model organism should have an exceptional long lifespan for its size. This is the case for the naked mole-rats (NMR). Indeed, the naked mole-rat has a much longer lifespan than expected for its relative small body size. In fact, the naked mole-rat (*Heterocephalus glaber*) has a record longevity of over 30 years which makes it the longest-lived rodent and thus a prime candidate for comparative genomics as the genomic information of many other rodents such as mouse, rat and guinea pig are presently available. Furthermore, the naked mole-rats have been shown to be extremely resistant to neoplesia. Due to these remarkable differences between the naked mole-rats and their rodent cousins, we hypothesized that the differential expression of certain

genes may be able to explain these different phenotypes.

Since the naked mole-rat genome is unavailable and its closest sequenced genome is the one of the guinea pig, gene expression analyses using micro-arrays are impossible. To design micro-arrays, one needs to have the DNA or RNA sequences of the targets under study. However, without a very close reference genome, micro-arrays technology could be too noisy to be of any interest. The solution to this problem is to use second, or next-generation, sequencing technologies which can directly sequence a pre-constructed library. The decreasing costs of sequencing allowed us to utilise this cost-effective technology for genes expression analysis.

For this study, we were able to obtain Illumina Solexa 39bp and 76bp reads from naked mole rat and wild-type mouse liver mRNA library constructed using a PMAGE like approach (Kim et al. 2007), see Figure 3. The naked mole-rats liver tissues were provided by Dr. Buffenstien of the Barshop Institute for longevity and Aging studies in San Antonio while the experiments were done by Dr. Chuanfei Yu from the Church lab at Harvard Medical School. The choice of liver tissue was made because of the ease if its harvest and its homogeneity, however, one study shows that liver genes expression levels remain constant throughout age and may be a tissue that do not show many ageing phenotypes (Zahn et al. 2007). This, however, is unlikely to affect our analyses as we are comparing the genes expression level of a young naked mole-rat and a young mice. Both naked-mole rats and wild-type mice were in their young adulthood when the tissues were harvested. The scope of this project was to determine genes differentially expressed in the naked mole-rat compared to the wild-type mouse. However, since there is no reference genome for the naked mole-rat, we first had to establish a reliable method to map the NMR reads to their correct genes orthologs.

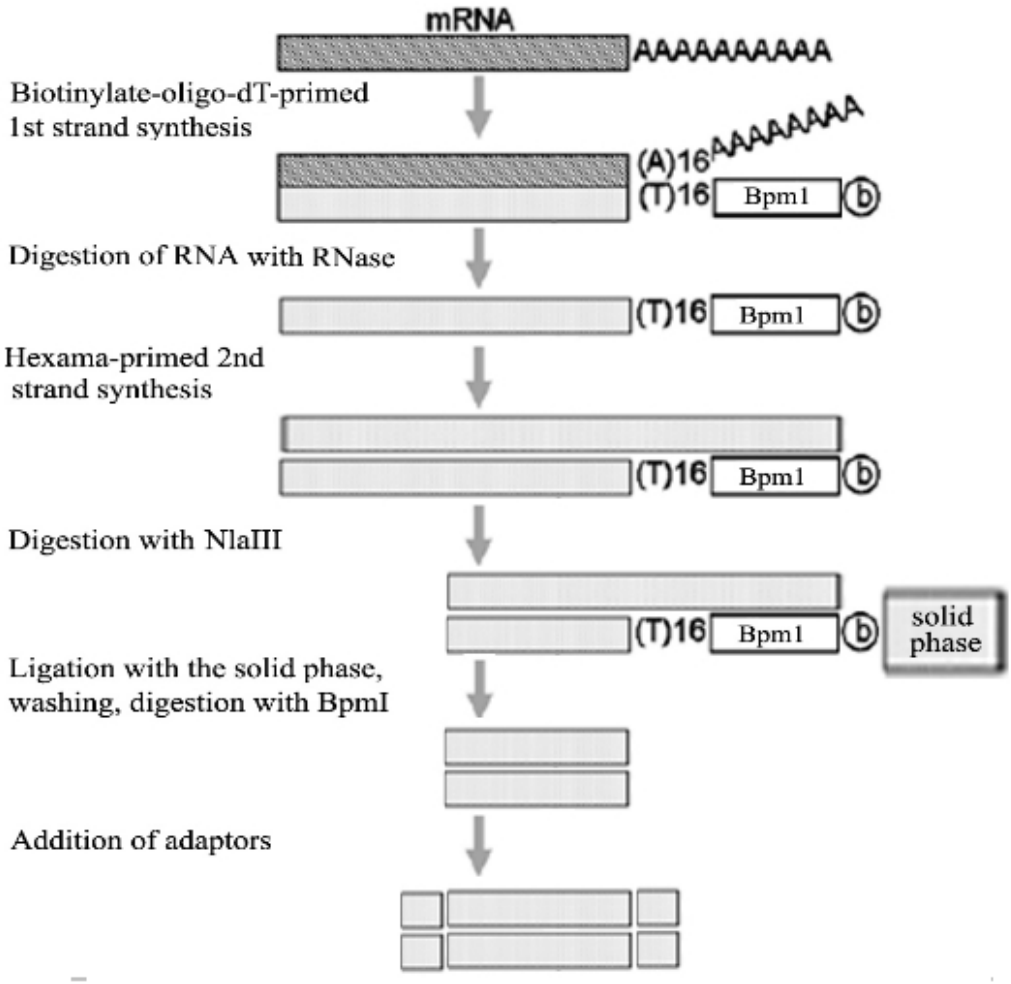

Figure 3: Construction of the mRNA library (done by Dr. Chuanfei Yu)

7.1 Mapping short Solexa sequencing reads

To do that, we have obtained a low coverage assembly of the naked mole-rat transcriptome generated by 454 sequencing complemented by Solexa, from Dr. Platzer group in Jenna, consisting of 77086 contigs. We mapped the 9.2M 39bp and 21.2M 76bp NMR Solexa reads to the 77086 contigs using MAQ. Out of the 21.2M

76bp reads, 10.7M reads (50.2%) mapped successfully and 6.8M had good mapping scores, i.e. a MAQ quality score of over 30 (see Supplemental Material), which can be explained by many poly-T and poly-A reads generated by the PMAGE protocol and the low coverage of the reference. Due to this low coverage, we cannot be sure whether a putative under-expressed genes in the naked mole-rats is under-expressed because of the poor alignments, thus, our plan was to detect genes over-expressed in naked mole-rats.

To determine from which gene a NMR read came from, we employed BLAST to map the contigs of the assembly onto the mouse cDNA and kept the unambiguous mappings. The mapping of 33286 contigs out of the initial 77086 were judged to be unambiguous. We define a mapping to be unambiguous if its BLAST map is either unique and less than 0.05 or the mapping e-value is 10 orders of magnitude smaller than the one with the second smallest. Out of the 6.8M reads with good mapping quality, 5.9M could be assigned to an orthologous genes in mouse or guinea pig.

After these two steps, we obtained a list of genes with their reads number for the naked mole rat Solexa data. We then constructed a reference library for mouse consisting of the 3' end of all mouse cDNA transcript. The wild type mouse solexa reads were then mapped to this reference library and the expression levels of the two were compared after normalisation. After this step, 11.1M out of the initial 24.3M mouse 76bp Solexa reads where successfully mapped with high quality score.

7.2 Normalisation of the read counts

Since two sequencing runs can have a different number of reads output and two samples can have different bias depending on how a library is prepared before sequencing, one needs to normalise the data before being able to tell whether a

genes is over or under expressed in either samples. Many studies simply use as normalisation the ratio of the output of the two sequencing runs, e.g. 21.2M:24.3M in our case or take into account read length for RNA-seq methods. However, it has been shown that these normalisation methods may not work well when a sample has a different mRNA composition. Because we are comparing the genes expression levels of two different species, we wanted a more reliable normalisation constant.

To this end, we used a simple, yet intuitive, method based on (Robinson et al. 2010) to normalise the reads count between naked mole-rats and mouse samples. The method is based on the crucial assumptions that most genes are not differentially expressed, e.g. more than 70%. In mathematical terms, if X_g^1 and X_g^2 are random variables denoting the expression level of genes *g* in detected in sample 1 and sample 2 (e.g. naked mole-rats and mice) respectively, the assumption is that X_g^1 and X_g^2 are identically independently distributed for most genes *g*. Both X_g^1 and X_g^2 depend on sample composition as well as the number of total reads generated by each Solexa run and thus they are not absolute measures of expression. In other terms, the following equation holds:

$$\frac{X_g^1}{X_g^2} \approx c$$

This ratio is also known as the fold differential expression of gene *g* and since neither has been corrected, *c* is the correction factor so that

$$\frac{X_g^1}{\sqrt{c}} / (X_g^2 \cdot \sqrt{c}) \approx 1$$

when the genes *g* is expressed at the same level in the two samples. The problem is to find *c* in order to normalise the read counts so we can compared them across the

samples. In order to do that, we minimise the following:

$$min_c \sum_{g \in G'} \left| \log\left(\frac{X_g^1}{X_g^2}\right) - \log c \right|$$

where G' is the set of genes that are not differentially expressed. By solving the minimisation problem, we find the constant c such that the sum of the differences between c and the real ratios of genes expression levels is smallest. That is, c is our estimate of the ratio X_g^1 / X_g^2 for all genes that are not differentially expressed and thus our normalisation constant. It can be shown (see Appendix 1.1) that c is equal to the median of the real ratios of the genes not differentially expressed, i.e. the median of $\{X_g^1 / X_g^2 : g \in G'\}$.

It now remains to find out what G', the set of genes whose expression are similar, consists of. This task is the trickiest part since if we knew which genes was in G', it would be extremely easy to find the normalising constant. One way we can construct G' is by first guessing a constant c and start removing the genes for which $X_g^1 / (X_g^2 \cdot c)$ is very large or very small (compared to 1) and then re-estimate c and go on until c converges. This class of methods is called expectation maximization (EM) algorithms and are useful when estimating unknown parameters when data is incomplete or has many outliers, for an introduction see (Russell et al. 1995). To get a good idea of the data we are dealing with, we first constructed a histogram of the log expression ratios between mice and naked mole-rats genes. Since we are use log expression ratios, we added pseudo-count of 50 to the reads count of each genes.

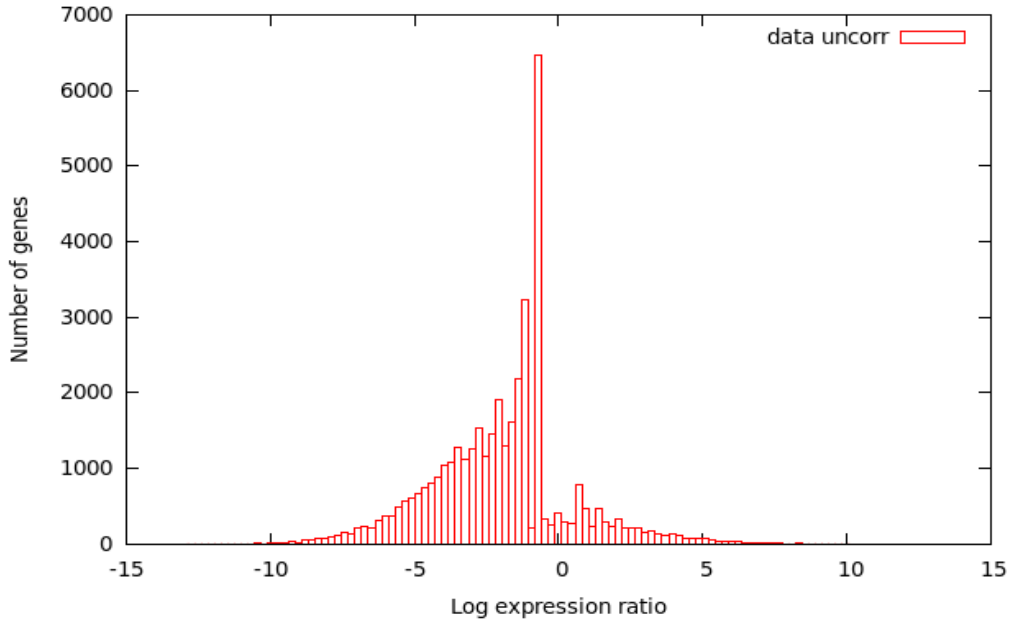

Figure 4: Log expression ratio between naked mole-rats and mouse genes.

We can see a clear bias towards mouse genes. And thus we expect *c* to be less than 1.

From figure 4, a clear bias towards mouse genes can be seen. It turns out that many of those genes have no naked mole-rats read count, possibly because the contigs did not cover the orthologous regions in the naked mole-rats or because the contigs of those genes were not mapped successfully. We thus remove all genes for which no count was observed in one or the other species. The resulting histogram looks much more balanced which indicates that when there is at least one Solexa read mapping to a naked mole-rat contig, the contig is at the right 3' location of the orthologous gene and most of the other naked mole-rats Solexa reads are able to be mapped. We expect the graph to have a normal distribution as because of both the stochasticity in genes expression level at the biological level and at the technological level (sequencing technologies).

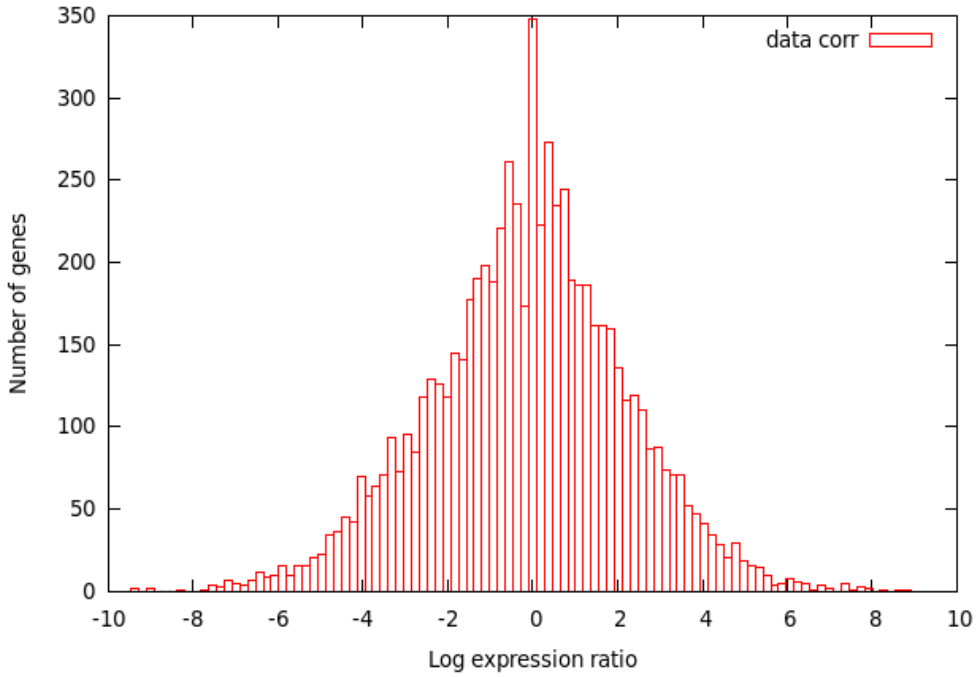

Figure 5: Log expression ratio between naked mole-rats and mouse genes after removing genes where no reads was found in either of the species' sample.

The bias is much smaller compared to the uncorrected data.

After correction, we can make two main observations. First, the distribution of log ratios seems to be symmetrical with mean close to 0. That is, the mean ratio is close to 1. Recall that the absolute output of the naked mole-rats sample is 21.1 compared to 24.3 of the wild-type mice which makes a ratio of around 0.87. The second observation is that number of genes is considerably lesser than without any correction. The shape of the graph was a bit suspicious and we suspected this was the effect of many low count genes for which the log ratio had very high variance compared to high count genes. Thus, we also tried removing all genes which had less than 50 reads in either of the samples which yield the following histogram.

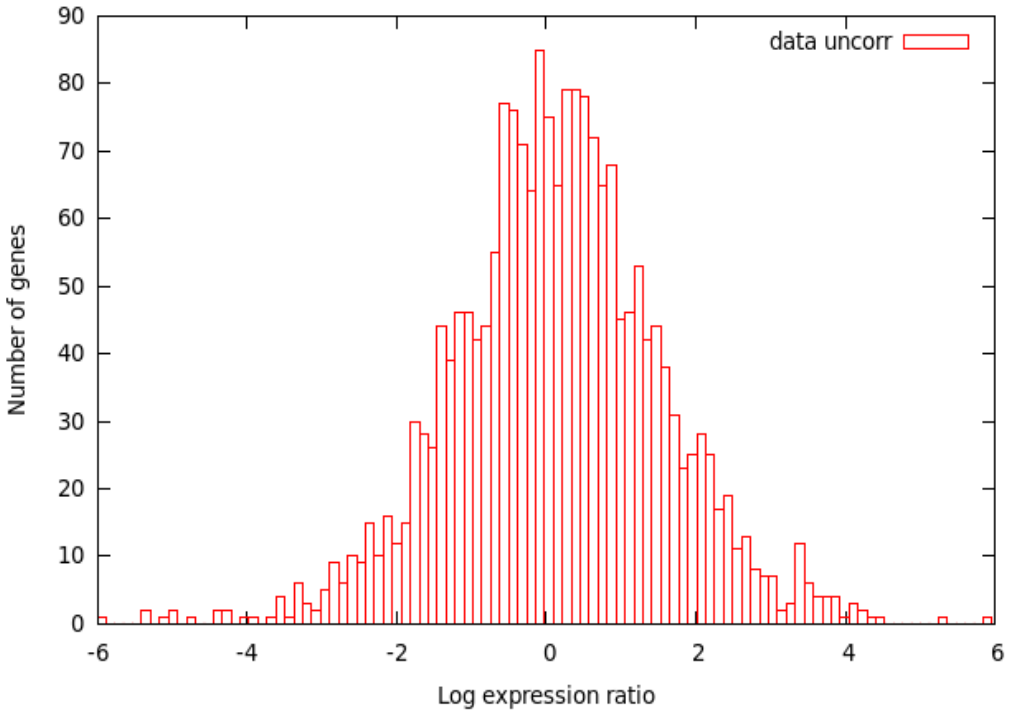

Figure 6: Log expression ratio between naked mole-rats and mouse genes after removing genes where less than 50 reads in either of the species' sample.

The distribution looks more normal now.

Though the distribution seems to peak near 0, we can see that the right side of the curve seems to be heavier. After weeding out 15% of the top outliers according to the aforementioned expectation-maximisation algorithm, we found that c converges near 1.19 which is quite different from our first estimate of 0.87. The histogram after correction looks as follows.

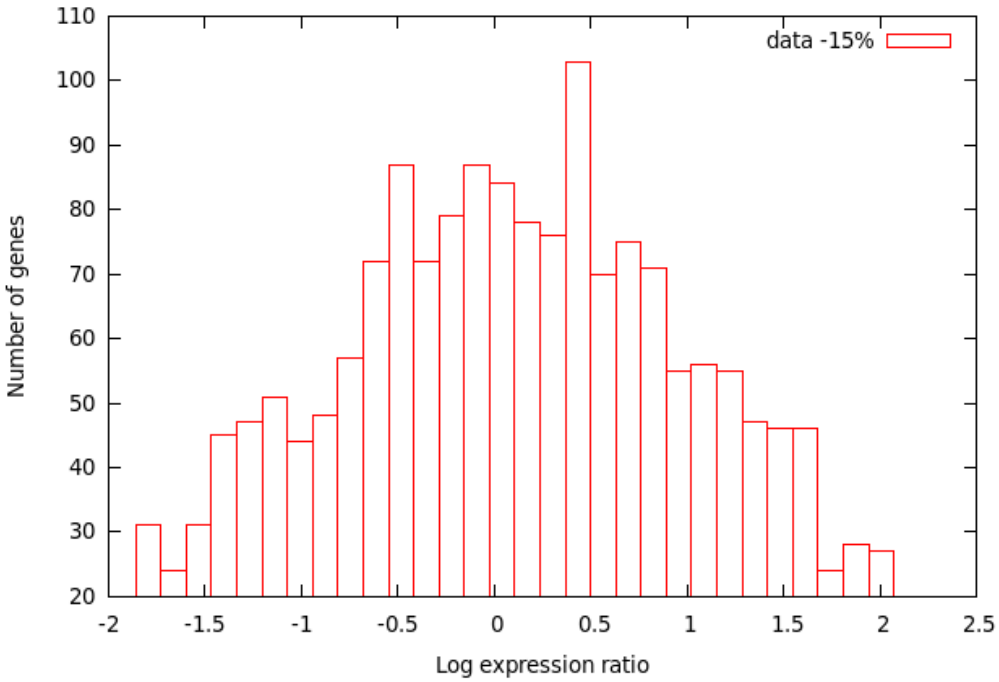

Figure 7: Log expression ratio between naked mole-rats and mouse genes after removing genes where less than 50 reads in either of the species' sample and removing the top 15% outliers

To normalise, we simply used the following equation to compute the expression fold differences of each genes g where X_g^1 and X_g^2 are the expression of the naked mole rats and wild type mice respectively:

$$X_g^1 / (X_g^2 \cdot c)$$

7.3 Functional analysis of genes over-expressed in the naked mole-rats

We then assigned genes to two different categories. The first category consists of all genes, but with a pseudo-count of 50 added to each read count in order to limit the effects of the high variance when read counts are low. Adding a pseudo-count has similar effect as using the binomial where two genes with a big

count and big fold expression difference scores a lower p-value than two genes with small count and the same big fold expression difference. The second category consists of genes where the genes has less than 50 counts in wild-type mouse. Using the first genes category, we were interested in finding a global trend in genes over-expressed in the naked mole-rats. For the second genes category, we were interested in finding about genes that are unexpressed in wild-type mice but expressed in the naked mole-rats, thus we removed all genes with more than 50 reads in wild-type mouse and added a smaller pseudo count of 5.

We then ranked all genes according to their expression fold differences and looked at at the top 50 in each of the two genes categories for interesting candidates (see Table 3). Next, to identify pathways and biological functions that tend to be overexpressed in the naked mole-rats, we chose the arbitrary cut-off for expression fold difference of 15 and looked at all the genes with over-expression fold bigger than 15. In the end, we obtained 273 genes over-expressed 15-fold or over for genes from the first category and 898 for genes from the second category and used the Database for Annotation, Visualization and Integrated Discovery (DAVID) to find out genes functional enrichment. To compute statistical enrichment p-values, one of DAVID's strategy is to use a hyper-geometric distribution on the genes in a provided test dataset against a background dataset (Huang et al. 2007).

Using DAVID and the genes in our analysis as background, we found that in the first genes category, genes were enriched in acetylation (p = 2.9E-14), mitochondrion (3.3E-13), transit peptide (3.3E-10), oxidoreductase (1.1E-10), fatty acid metabolism (2.6E-6), and others (see Supplemental Material). We also found oxidation reduction (9.1E-9), generation of precursor metabolites and energy (2.8E-8) enriched in genes sharing GO terms. For the pathways analysis, we found Parkinson's disease (3.4E-5), oxidative phosphorylation (1.5E-4) and Huntington's disease (1.1E-3) and few others with significant enrichment. Unsurprisingly, for the second category of genes, we found very similar results. Lipid biosynthetic process

(8.3E-9) was highly enriched among GO terms and Proteasome (1.5E-4) was found to be the most enriched in our pathway analysis.

Rank	Corrected Fold difference	Nmr count	Mouse count	Gene Name	Description
1	1070.36	72553	7	Rpl26	ribosomal protein L26 (Silica-induced gene 20 protein)
2	636.0894	42339	6	Rps9	40S ribosomal protein S9
3	2063.14	188996	27	A2m	Alpha-2-macroglobulin-P Precursor
4	609.57	40572	6	Tbrg4	Transforming growth factor beta regulator 4
5	515.71	34317	6	Gsta2	Glutathione S-transferase A2
6	584.39	40980	9	Pafah1b3	Platelet-activating factor acetylhydrolase IB subunit gamma
7	300.30	19962	6	Igfbp2	Insulin-like growth factor-binding protein 2 Precursor
8	329.54	22303	7	D2Bwg1335e	DNL-type zinc finger protein
9	234.32	15565	6	Cth	Cystathionine gamma-lyase
10	247.56	16742	7	Sc4mol	C-4 methylsterol oxidase
11	271.28	18674	8	Hrsp12	Ribonuclease UK114 (Heat-responsive protein 12)
12	224.02	15145	7	Igfbp4	Insulin-like growth factor-binding protein 4 Precursor
13	246.49	16963	8	Crym	Mu-crystallin homolog
14	793.86	88751	44	Apoc2	Apolipoprotein C-II Precursor
15	353.41	26866	14	Cyp3a16	Cytochrome P450 3A16

Table 3: Table showing the genes with highest fold expression differences between naked mole-rats and wild-type mice.

The ratio is computed with 50 pseudo-counts added to both counts. The top genes for the first genes category are almost identical to the ones of the second category.

7.4 Validation of results using microarray data

To validate some of the most interesting candidates found in our study for follow ups, we retrieved expression data from the mouse expression database GeneAtlas (http://www.geneatlas.org/). The database contains microarray intensity of genes expression in many different tissues. For candidate genes, we summarised the expression level data of each gene by using a k-NN approach in order to compare the expression level of the particular gene in liver compared to other tissues. If the read counts for wild-type mouse is very small, we expect the expression level of this particular gene in liver to be small compared to other tissue or small in absolute term. The k-NN approach is better than summarising the data by a ranking with respect to expression in different tissue as a gene might be expressed at a low level in all tissues. On the other hand, using the absolute intensity of expression in liver alone might not be accurate because the genes at hand may be expressed highly in all tissues and may thus suffer from normalisation problems.

DISCUSSION

In the results and methods part, we looked at the comparative genomics methods used in order to study species divergence in ageing and highlighted some of the findings. Here, we discuss the biological significance of the results in more depth. We also discuss the biological limitations of comparative studies in ageing. We argue that contrary to research on ROS in lower organisms, the effects of ROS on longevity seem to be lesser in mammals. On the other hand, our results coupled with previous studies strongly suggest that proteasome-ubiquitin proteins are involved mammalian differences in the ageing phenotype. Akin to proteasome-ubiquitin proteins, our results also seems to indicate that proteins involved in lipid metabolism are important in the mammalian evolution of longevity both at a genotypic level and at an expression level. We show evidence that DNA damage repair and response proteins are important to the evolution of ageing in mammals as well as discuss other pathways that can be of interest for follow up experiments.

Proteins related to longevity in model organisms appear to be well conserved and might even tend to be better conserved than expected by chance, suggesting the genetic mechanisms for longevity regulation within species are not the same that determine species differences in longevity (de Magalhães and Church 2007). Therefore, we wanted to use comparative genomics tools to study differences in the genome of different species as well as genes expression differences across different mammals. In our first project, we decided to look at residue usage in mitochondrial protein and found an interesting negative correlation between methionine and maximal lifespan. For our second project, we wanted to detect selection in proteins with different evolutionary rates but having a specificity towards phylogenetic branches where maximal longevity drastically increased (MLI branches). And for our last project, we looked for genes with different levels of

expression between the wild-type mouse and the naked mole-rat.

The analysis of biological functions enrichment may give us a good idea of the regulatory and expression differences underlying the phenotypic differences between the long-lived naked mole-rat and the shorter lived wild-type mouse.

Chapter 8: Mitochondrial and antioxidant proteins in mammalian ageing

Contrary to findings in lower organisms, we found no strong evidence that supports the idea that ROS causes the divergence in ageing phenotypes observed in mammals. The past work of Moosman and Behl along with the work of Stadtman (Moosmann et al. 2008; Stadtman 2006) suggested a putative correlation between maximal longevity and methionine or cysteine usage in mitochondrial proteins. Briefly, their idea was that since methionine and cysteine are thought to act as antioxidants in proteins (due to the reversibility of their oxidation), an enrichment in methionine or cysteine residue should translate into a stronger resistance to oxidative damage. Stadtman thus hypothesised the existence of a positive correlation between the usage of these two residues and maximal longevity. However, Moosmann and Behl found a negative correlation between cysteine encoded in mtDNA and maximal lifespan of 248 species spanning 10 different phyla and proposed that cysteine thiyl radicals can initiate irreversible protein cross-linking which caused the selection against cysteine residues in mtDNA-coded proteins in long lived species.

8.1 Residue usage and maximal longevity

In our analysis, among mitochondrial inner membrane proteins, we found no correlation between cysteine residue content and maximal longevity while we found that methionine residue content was negatively correlated with maximal longevity in mammals. In our joint work with Aledo et al. (Unpublished), in contrast with Moosmann and Behl's hypothesis, we proposed that the negative correlation was caused by a selection of methionine residues in mtDNA-coded proteins in short-lived species. In fact, supplementary analyses that Aledo et al. conducted showed that the addition of methionine residue in short-lived species correlates with

longevity whereas the removal of methionine residue in long-lived species did not which supports our hypothesis that there was an accumulation of methionine residue in the mitochondrial proteins of short-lived mammals which are subject to high levels of oxidative stress.

Aledo et al. also showed that AUA methionine codon, but not AUG methionine codon correlates with longevity. This, coupled with the work of Bender et al. (2008), which suggests that the recoding of AUA from ileucine to methionine is an adaptive antioxidant response, provide further evidence for the interpretation that the accumulation of methionine in mitochondrial protein was adaptive in the context of oxidative stress. Moreover, we found that mitochondrial proteins are enriched in methionine residue compared to non mitochondrial proteins which seems to indicate that methionine residues have a protective role against oxidative damage. This apparent protective role also supports our hypothesis.

While the correlation between methionine content in mitochondrial proteins and maximal longevity seems to be real, we found no evidence and do not believe that altering the residue content of mitochondrial proteins will result in a longer lifespan in mammals. We hypothesise that, unlike lower organisms, mammals evolved alternative mechanisms of defence against oxidative stress, e.g. cytoskeleton optimisation, or more sophisticated response pathways to molecular damage from other sources. For example, as will be discussed in section 11.5, we discovered that DBB1, a protein of paramount importance to proper damage response and repair after UV damage to the DNA, has been targeted for selection in many mammalian lineages.

8.2 Evolution of actin cytoskeleton may implicate ROS in species divergence in ageing

We found that proteins involved in actin cytoskeleton and acting binding

tended to show a pattern of selection in lineages where longevity evolved. Actin cytoskeleton proteins serve as a physiological regulator of ROS release from mitochondria as well as a key component in the activation of cell death pathways while mutations in actin binding proteins can change cell fate (Gourlay et al. 2005). In yeast, the increase of actin dynamics resulting from a specific actin allele or a deletion of a gene encoding SCP1P, an actin-bundling protein, can increase lifespan by over 65%. Furthermore, this increase is reported to be due to the mutant cells producing lower than wild-type levels of ROS (Gourlay et al. 2004). In yeast, the increase of actin dynamics which is influenced by the actin-bundling protein SCP1P increases maximal lifespan. Furthermore, the mammalian homologue of SCP1P, SM22, was identified in senescence screens by Toussaint et al. (2000) which suggests a well-conserved role of actin in cellular ageing as discussed by Gourlay et al. (2004). The patterns of longevity specific selection in actin related proteins coupled with the fact that actin dynamics might be regulator of ROS production and of cellular senescence (Gourlay et al. 2005) suggests that the selection for a tighter control of ROS production through an optimization of the actin cytoskeleton might be involved in the evolution of longevity in several mammalian lineages.

8.3 The effects of ROS in mammalian divergence in ageing remains to be determined

Still, it is unclear whether mitochondrial proteins have an important role in the regulation of ageing in mammals. Mitochondrial genes were strongly enriched among the genes over-expressed in the naked mole-rats compared to the short-lived wild-type mice. Unsurprisingly, oxidation reduction and generation of precursor metabolites and energy were also found as significantly enriched functional categories in over-expressed genes in naked mole-rats. However, we found no indication that the over-expression of mitochondrial genes was related with oxidative stress prevention rather than with, for example, an optimisation of energy production.

In sum, the effect of ROS on mammalian divergence of ageing seems to be limited as we would otherwise expect proteins with antioxidant properties to be selected or expressed at a higher level in long-lived mammals. On the contrary, we found that methionine usage was higher in the mitochondrial proteins of the shorter lived mammals. Moreover, as in de Magalhães and Church (2007), we failed to detect any significant acceleration in the evolution of proteins in the antioxidant category. It seems, by and large, that either ROS is not a big player in the ageing of mammals or most mammalian species already have mechanisms to deal with ROS efficiently.

Chapter 9: Candidate proteins and functional categories related to longevity

9.1 Detection of longevity-specific selectivity can be caused by longevity correlated traits

As we have discussed in the introduction, a caveat of using comparative studies alone in ageing research is that one can almost never be sure whether a finding is directly linked to the mechanism of ageing or rather a trait that is itself correlated to ageing without any mechanistic role. Ageing is a complex phenotype and the signals that we detect from comparative studies and specially high throughput studies are generally noisy. Furthermore, there are traits such as body weight that are highly correlated to maximal longevity, thus the combination of the noisy data and the existence of many confounding variables makes comparative studies extremely hard to interpret.

For example, in our study of mitochondrial residue usage in mammals, we found that methionine usage is inversely correlated with maximal longevity. Our interpretation, though controversial as we will discuss later, is that since short-lived animals tend to exhibit higher metabolism and thus suffer from an increased ROS level, the increase of methionine, a residue with antioxidant property, may be an evolutionary adaptation to this high level of ROS compared to longer lived mammals. The connection between methionine usage in mitochondrion and the ageing process is thus contingent on the connection between ROS and the ageing process, while a different interpretation might instead strengthen the link between ROS and ageing.

The results of our protein evolution analysis were also open to interpretation. There are few possible reasons why some proteins undergo accelerated selection in

lineages where maximal longevity increased compared to lineages where it stayed the same. In our analysis, we found many proteins involved in development and growth resulting in functional categories such as muscle development, postynaptic density and spermatid development. These results among many others in our list (see Supplemental Materials) are open to interpretation as proteins in functional categories that are selected for phenotypes strongly correlated to ageing will also tend to show high selection specificity towards MLI branches. For example, maximal longevity is well known to be strongly correlated to body size (Austad 2009; de Magalhães et al. 2007; Speakman 2005) ergo a radical increase in lifespan in a particular lineage may very well translate into proteins involved in growth and body size to be undergoing a high evolutionary pressure. Studies have successfully identified phenotypes such as brain size (Sacher 1959; Allman et al. 1993) and energy metabolism to be correlated with mammalian maximal longevity (Speakman 2005b). As discussed by Ricklefs (2010), many of the correlations found between rates of development and longevity are weak when age at maturity is included and although there have been attempts to find mechanistic explanations to the correlations, there exists no strong evidence showing a causality relationship. Correlation without causation, as Speakman (2005a) warns us, will remain a big caveat of comparative studies and phylogenetic approaches to study ageing.

Lastly, in our comparison of expression between wild-type mice and naked mole-rats, the problems with confounding traits correlated with longevity vanishes as there are only 2 species. However, any phenotypic difference other than maximal lifespan or ageing can be the cause of the observed genes differential expression. For example, the naked mole-rat has been shown to be extremely resistant to cancerous growth (Liang et al. 2010). There exists an intricate relationship between cancer and ageing (Finkel et al. 2007) which makes the interpretation of the results of our analysis even harder. Due to these aforementioned problems, we urge the reader to exert caution when interpreting our results. Nevertheless, we follow with a discussion concerning the possible links between our findings and the divergence of

ageing mechanisms in mammals.

Despite the previously discussed caveats of comparative approaches, we found, in our search for pattern of selection for longevity among proteins, proteins with high "longevity-specific selectivity" scores that might have contributed to the evolution of longevity.

9.2 Protein evolution analysis reveals proteins with longevity specific selection

In the protein evolution analysis, many of the top proteins were poorly annotated including IWS1 (rank 1; stringent) which, as far as we know, has not been studied in mammals. A recent study in *Arabidopsis* shows that IWS1 is involved in plant steroid hormone and a loss of function mutations in AtIWS1 lead to overall dwarfism (Li, Ye, et al. 2010) thus may have been detected due to a putative role in the evolution of body size. Examples with a score of ``longevity specific score" of 4.0 (see section 6.3) include LGALS3 which has been associated with early embryogenesis (Fukushi et al. 2004), HERC4 a probable E3 ubiquitin-protein ligase by sequence similarity (Wu et al. 2006) and NUP85 a component of the nuclear pore complex thought to be required for nuclear pore complex assembly (Harel et al. 2003) and maintenance and to play a role in spindle assembly during mitosis (Orjalo et al. 2006) as well as in phosphatidyl-inositol-3-kinase dependent pathways (Terashima et al. 2005).

CPNE5 (rank 1; medium) is a poorly studied protein which exhibits calcium-dependent phospholipid binding properties by similarity (Wu et al. 2006) while RSAD2 (rank 1; moderate) is involved in antiviral defence (Wang et al. 2007). DDHD1 was another high scoring protein, scoring 4.36 (rank 5; medium), and is a probable phospholipase that hydrolyzes phosphatidic acid by similarity (Wu et al., 2006). CAPNS1, scoring 4.0 (rank 7; moderate), belongs to a well-conserved family of calcium-dependent, cysteine proteases whose link to cellular senescence

and DNA damage response has been studied in (Demarchi et al. 2007) and (Demarchi et al. 2010).

COL3A1 (rank 4; relaxed) is a collagen type protein whose expression is significantly decrease with age (de Magalhães, Curado, et al. 2009). Next, TAOK3 (rank 7; relaxed) is a serine/threonine-protein kinase thought to inhibit basal activity of JNK/SAPK (Tassi et al. 1999) and its over-expression may activate ERK1/ERK2 and JNK/SAPK (Zhang et al. 2000). Furthermore, TAOK3 is thought to be phosphorylated upon DNA damage possibly by ATM or ATR (Dephoure et al. 2008; Matsuoka et al. 2007). And finally, the damage-specific DNA binding protein DDB1 (rank 11; relaxed) is a well studied protein and is a subunit of the DDB1-CUL4-X (DCX) box which can form many different complexes that are involved in different DNA damage response pathways.

Other examples of high scoring proteins comprise of PIK3C2A, scoring 3.0 (rank 26; stringent), and SMC1A, scoring 2.77 (rank 74; stringent). PIK3C2A is a protein belonging to the PI3/PI4-kinase family and is believed to play a role in the EGF signaling pathway (Arcaro et al. 2000). Moreover, Didichenko et al. (2003) showed that *Homo Sapiens* PIK3C2A is phosphorylated upon exposure of cells to UV irradiation and is also target of stress-induced phosphorylation during the G2/M transition of cell cycle by the JNK/SAPK pathway. They further found that the phosphorylation seems to lead to the proteasome-dependent degradation of PIK3C2A. SMC1A is a protein involved in chromosome cohesion during cell division as well as in DNA repair. More precisely, SMC1A is related to the cohesion between sister chromatids during DNA replication and, at least in yeast, the cohesin complex also has functions in DNA repair and is essential for efficient double-strand break repair in mitotic cells (Sjögren et al. 2001).

9.3 Few over-expressed genes in the naked mole-rats are candidate regulators of

ageing

One of the most over-expressed genes in the naked mole-rat is the serum pan-protease inhibitor, alpha2-macroglobulin (A2M). While A2M is only barely expressed in wild-type mice liver (27 reads out of 24.3M), we detected almost 190,000 (out of 21.2M) reads in naked mole-rats liver which makes it the third most over-expressed gene. Interestingly, A2M is listed as a candidate protein relevant to the human ageing process in GenAge (de Magalhães, Budovsky, et al. 2009), a database of ageing and longevity associated genes identified in model organism. A2M is known to interact with ApoE and, unsurprisingly, is associated with Alzheimer's disease (Blacker et al. 1998). It has been found that in a study on long lived individuals in Germany that a particular allele in A2M, which was previously associated with AD, was depleted (Flachsbart et al. 2010). Furthermore, A2M was determined to be a biomarker for ageing in vivo as its mRNA expression level was showed to be positively correlated with age (Ma et al. 2004). Though the exact role of A2M in the mechanism of ageing is unknown, it seems that perhaps, like ApoE as discussed by Christensen et al. (2006), A2M is mainly related through its association with age-related disease such as AD and other neurodegenerative diseases. Yet, we believe that the function of A2M as proteinase inhibitor is of further interest in the context of protein turnover regulation in the naked mole-rats which Pérez, Buffenstein, et al. (2009) think may be a key contributor to the extreme longevity of the naked mole-rats.

Two other genes over-expressed in the naked mole-rats that caught our attention was SAT1 (3142 versus 6 reads in naked mole-rats and wild-type mice liver respectively; rank 72) and SAT2 (1769 reads versus 8; rank 177) which are spermidine/spermine N1-acetyltransferases. SAT1 and SAT2 have been both associated to ageing through its interaction with hypoxia-inducible factor 1. It has been shown that both SAT1 and SAT2 binds to HIF-1alpha and promotes its ubiquitination and degradation through 2 different but complementary mechanisms

(Baek et al. 2007) while HIF-1 modulates dietary restriction-mediated lifespan extension and its deficiency results in extended lifespan (Chen et al. 2009). Thus, it is not impossible that the over-expression in SAT1 and SAT2 in the naked mole-rats translates into the degradation of HIF-1 and its deficiency, resulting in a longer lifespan compared to the mice. Like A2M, the study of SAT1 and SAT2 in naked mole-rats could be of interested, this time, in the context of HIF-1 and dietary restriction.

The insulin-like growth factor-binding protein 2 and 4 precursor, IGFPBG-2 (19962 reads versus 6; rank 9) and IGFPBG-4 (15145 reads versus 7; rank 17), have been shown to be involved in bone density. Amin et al. (2004) found that higher levels of IGFBP-2 are associated with lower bone mineral density. In humans, IGFBP-2 is known to be increasing with age and, like all insulin-like growth factor-binding protein, binds to IGFs to modulate their action. Among others, IGFBP-2 and IGFBP-4 were shown to inhibit IGF action by binding to them and preventing the binding of IGFs to IGF receptors (Jones et al. 1995)**.** In another study, Mohan et al. (1995) showed that IGFBP-4 also increased with age and stipulates that this increase could be a factor of bone conditions such as osteoporosis seen in aged individuals. Another less direct link between IGFPBG and ageing is through its interaction with IGF. It is known that insulin-like growth factor-1, or IGF-1, is related to ageing. According to (Shimokawa et al. 2002), mutations that lower IGF-1 levels in mice can extend lifespan. Furthermore, (Bonafè et al. 2003) reports that IGF-1 response pathway genes such as IGF-IR (IGF-I receptor) and PI3KCB (phosphoinositol 3-kinase) among others play a role in human longevity. We believe that the interaction between IGFPBG-2, IGFPBG-4 and IGF-1 may be of interest to future research on ageing in the naked mole-rats.

In fact, a possible follow up experiment could be to increase the expression level of these genes in mouse cell lines and look for changes in markers for ageing. By the transfection of plasmids containing the target gene as done by Yáñez et al.

(2002), one can increase the expression of this gene in transfected cell lines. Conversely, in the naked-mole rats cell lines, one can reduce the protein levels of a target gene by introducing exogenous double-stranded RNA as Montgomery et al. (1998) describes or by using RNAi as Seluanov et al. (2009) did. In these two models, we can study the effects of certain differentially expressed genes on ageing phenotypes. For instance, one can analyse the effect of reducing the expression level of SAT1 and SAT2 in naked mole-rat cell lines on HIF-1alpha levels. If, indeed, a connection is found between SAT1, SAT2 and HIF-1alpha in naked mole-rats, then we would strengthen our hypothesis that the differential expression of these genes are related to the ageing process.

9.4 Comparative studies reveals candidate functional categories related to ageing

It is not unexpected that few proteins have evidence of being under selection in all long-lived lineages. Rather, it would be surprising if the evolution of longevity in all mammalian lineages could be explained by the positive selection or the over-expression of the same few proteins. A more intuitive explanation of the evolution of longevity is the selection or over-expression of proteins in common pathways and biological processes. In fact, we found functional categories, which have been previously associated to ageing, showing specificity of selection in MLI branches such as actin cytoskeleton (Gourlay et al. 2005; Gourlay et al. 2004), 1-phosphatidylinositol-3-kinase activity, phosphoinositide 3-kinase complex, response to food and circadian rhythm . For example, recent work from Wyse et al. (2010) established a connection between circadian rhythm and lifespan in laboratory mouse strains and in other mammals. Furthermore, 1-phosphatidylinositol-3-kinase activity is thought to play a critical role in DNA repair and cell cycle checkpoint. In fact, both mTOR and ATM show 1-phosphatidylinositol-3-kinase activity and are two proteins implicated in the molecular mechanism of ageing. Phosphoinositide 3-kinase has been studied in relation with the IGF response pathways and ageing by Bonafè et al. (2003).

However, as discussed in section 9.1, these categories might all have been detected in our analysis because of their involvement with a phenotypic trait that is associated to maximal longevity but is not related to the cellular mechanism underlying the ageing process. We present, in the next two chapters, the other functional categories that specially caught our attention. First, the proteasome-ubiquitin system and then, cellular response to damage and repair pathways. While these categories also suffer from the possibility that their selection was not biochemically or physiologically related to the ageing process, we discuss what we believe are the most logical interpretations of their patterns of selectivity.

Chapter 10: Lipid metabolism, ubiquitin-proteasome and damage response pathways differences may be able to explain a large portion of mammalian divergence in ageing

Apart from the pathways and functional categories discussed in chapter 9, we found pathways that are far more interesting and that are present in both our protein evolution and differential expression analyses. In fact, our finding of phospholipid metabolic process proteins corroborates the findings of a similar protein evolution study by Jobson et al. (2009). Interestingly, both fatty acid metabolism and lipid biosynthetic process functional groups figured among the top categories enriched in the oxer-expressed genes of naked mole-rats. This may emphasise the importance of these functional groups in mammalian divergence of ageing as we witnessed a marked change in both protein sequences and genes expression between species with significantly different lifespans.

10.1 Putative links between lipid metabolism, cholesterol catabolism and age-related degeneration

According to Hulbert (2008), membrane fatty acid composition is correlated with the maximal lifespans of mammals through the reduction of oxidative damage caused by products of lipooxidation. Proteins belonging to the lipid biosynthetic process were also identified in our analysis which Jobson et al. (2009) link to the peroxidative damage via increased saturation and to the control of mitochondrial ROS production by reducing membrane potential and increasing the efficiency of ETC uncoupling (Kua 2006). Moreover, cholesterol catabolic process related proteins were also identified and these findings fit studies of the apolipoprotein E (ApoE) well.

Although the protein ApoE does not show longevity specific selectivity,

Finch (2010) highlights the importance of ApoE in the clearance of triglycerid-rich lipoprotein components as well as its importance in cholesterol transport in the brain. In addition, ApoE and its role on the cholesterol catabolic process was shown to influence Alzheimer's disease (AD) progression (Evans et al. 2004). In fact, Christensen et al. (2006) observed that human genetic studies have shown that common polymorphisms in ApoE influence lifespan, probably mainly through their association with disease such as AD. Interestingly, we found in our expression analysis that over-expressed genes in the naked mole-rats were enriched in genes involved in AD. The connection between the extreme longevity of naked mole-rats, ApoE and the over-expression of genes involved in AD is hard to understand. However, there are speculations concerning the effects of the evolution of ApoE in the human longevity. The evolutionary pressure detected on proteins involved in lipid metabolic process and cholesterol catabolism along with the over-expression of genes involved in AD and fatty acid metabolism in the naked mole-rats lead us to believe that lipid metabolism and cholesterol catabolism may have been important players in the evolution of longevity in other mammalian lineages.

These results suggest that it may be worthwhile to investigate the levels of lipid peroxidation in each experimental pairs of the protein evolution analysis. One can do this by using a lipid peroxidation assay such as the one used by Garcia et al. (2005). A link between lipid peroxidation levels and longevity selection in lineages could provide a stronger evidence that lipid peroxidation is involved in ageing specially if the levels of lipid oxidation is lesser in species for which longevity evolved compared to their paired species. In contrast to the discovery that an increase in methionine usage is an evolutionary adaptation to high oxidative stress in short-lived species by Aledo et al. (Unpublished), the evolutionary pressure detected on genes involved in lipid metabolism processes could be an adaptation to oxidative damage by the optimisation of lipid metabolism or other cellular components such as the actin cytoskeleton. Since we compare the levels of lipid peroxidation of species in our experimental pairs, the oxidative stress should be a

fixed variable and a decrease in lipid peroxidation may entail a better mechanism for lipid peroxidation protection.

10.2 Many proteasome related genes are over-expressed in the naked mole-rats and proteins involved in the proteasome-ubiquitin show pattern of longevity-associated selection.

We found that genes involved in proteasome are over-expressed in the naked mole-rat and proteins involved in the proteasome-ubiquitin system have high "longevity-specific selectivity" scores. More precisely, we found that genes involved in the proteasome pathway were enriched in the set of genes over-expressed in the naked mole-rat while not expressed (less than 50 reads) in the wild-type mice. These genes include many proteasome subunit such as PSME3 (323 reads in the naked mole rats versus 6 in wild-type mice), PSMD1 (3174 reads versus 7), PSMD4 (236 versus 7), PMSD6 (322 versus 9), PSMD13 (417 versus 6), PSMA1 (600 versus 6), PSMA7 (1047 versus 7) among others. In the protein evolution analysis, we found 4 highly ranked GO categories related to the proteasome-ubiquitin system which are protein ubiquitination during ubiquitin-dependent protein catabolic process, proteasomal ubiquitin-dependent protein catabolic process, ATP-dependent peptidase activity and lysosome organization. Interestingly, these GO categories do not share a single protein that contributed to their score which provides strong evidence that the system was the target of evolutionary pressure in lineages where longevity increased.

The proteasome-ubiquitin system operates in 2 steps by first selecting and labelling proteins for degradation via ubiquitylation by a serial of enzymatic reactions (E1 to E3 ubiquitin ligase) and then by degradation in the proteasome complex. The proteasome has been linked to ageing time and again, for instance, oxidative damage in proteins is believed to be a causal mechanism of ageing (Levine et al. 2001) and the proteasome may limit the extend of the damage by

degrading damaged proteins. Further evidence from Chondrogianni et al. (2000) and Friguet et al. (2000) connect proteasome decline with the accumulation of damage and in ageing. Interestingly, Pérez, Buffenstein, et al. (2009) found that, compared to mice, naked mole-rats show resistance to protein unfolding and attenuated accumulation of ubiquitinated proteins and a sustained proteasomal function during ageing. Pérez, Buffenstein, et al. (2009) further propose that these mechanistic differences may contribute to species divergence in ageing and that the maintenance of protein stability is of great importance to successful ageing. The over-expression of genes involved in the proteasome pathways may be related to this observed phenotype.

The importance of the proteasome and the resistance of the naked mole rats to oxidative stress hypothesised by Pérez, Buffenstein, et al. (2009) is supported by their observation that naked mole-rats show higher levels of lipid peroxidation and DNA oxidative damage compared to mice even at a young age despite their much longer lifespan. Our results show that many proteasome sub-units are over-expressed in the naked mole-rats liver compared to the wild-type mice during young adulthood. As discussed in section 9.1, it may also be interesting to experimentally increase the expression these proteasome sub-units and observe the protein turnover rates in senescent cell lines.

10.3 Proteasome may be involved in DNA damage response and repair activity

Proteasome activity was also shown to participate in DNA repair at different levels reviewed by Brégégère et al. (2006), one of which is through the response of UV-mediated DNA damage. The degradation of the replication factor CDT1 (Kondo et al. 2004), repair factor repressor ZBRK1 (Yun et al. 2003) and the transcription factor p21 are thought to help UV damage repair (Bendjennat et al. 2003) and may be critical to proper cellular response to DNA damage by activating DNA repair mechanisms or cell cycle arrest. The evolutionary significance of the

proteasome-ubiquitin system to DNA damage response and repair is further strengthened by the strong pattern of selection exhibited by DDB1 in lineages where longevity evolved. In sum, the ubiquitination process and the proteasome complex are active components of cellular response to stress and damage and their evolutionary selection might have contributed to a lifespan increase in mammals.

10.4 The protein evolution analysis reveals that many proteins related to damage response and repair are selected in long-lived lineages

Though, we have found no obvious genes or categories related to damage response and repair pathways in our analysis of differentially expressed genes. We found many proteins with signature of longevity-associated selection which are involved in damage response and repair pathways.

In our protein evolution analysis, many high ranked proteins were involved in cellular response to damage at different levels of selection. With respect to the stringent cutoff, PIK3C2A (rank 26) and SMC1A (rank 74) are two proteins thought to be sensitive to external stress or DNA damage. PIK3C2A is phosphorylated in response to to UV irradiation and is involved in UV-induced damage response (Didichenko et al. 2003), PIK3C2A is part of the phosphoinositide-3-kinase family which also has been associated to ageing in the context of IGF-1 regulation (Bonafè et al. 2003). What is interesting about SMC1 is that it has been shown that it is an essential component of the IR ATM DNA damage response network (Yazdi et al. 2002). Experimental evidence indicates that ATM responds to IR damage by activating the S-phase checkpoint which slows down the DNA replication in two different pathways one of which depending on SMC1. A defect in SMC1A can thus result in radioresistant DNA synthesis, or a defective S-phase checkpoint, which has been identified in cancer-prone patients. Though newer evidence show that radioresistant DNA synthesis does not cause ataxia telangiectasia (Sasaki et al. 1994), a disease some argue is characterised by signs of premature ageing (Pesce et

al.), radioresistant DNA synthesis has been shown to be tightly coupled with ataxia telangiectasia (Painter 1981; Mohamed et al. 1986). It is thus possible that SMC1A is related to some of the premature ageing phenotype witnessed in patients with ataxia telangiectasia.

The identification of CAPNS1 caught our attention as it has been connected to DNA damage response. Indeed, Demarchi et al. (2010) shows that CAPNS1 depletion is coupled to a reduction in DNA damage induced histone H2AX phosphorylation, a well-established marker of DNA damage response (Hanasoge et al. 2007). It is possible that in response to DNA damage, CAPNS1 mediates the phosphorylation of H2AX which helps in opening stretches of DNA to DNA repair proteins. Demarchi et al. (2010) also provides evidence that CAPNS1 is necessary for a lasting phosphorylation of H2AX triggered by oncogenic Ras and genotoxic stress. Furthermore, the phosphorylation of H2AX has been associated with genome stability (Chanoux et al. 2009).

10.5 The DNA damage-binding protein, DDB1, show strong pattern of longevity-associated selection

In the relaxed cutoff category, we have TOAK3 (rank 7) and DDB1 (rank 9; fig. 2) which have been shown to respond to DNA damage in some way. However, we were most interested in DDB1 and the DDB1-CUL4-X (DCX) box as it is involved in many distinct DNA response and DNA repair pathways. For example, DC(DDB2) ubiquitinates histone H2A at the sites of UV lesions in a DDB2-dependent manner which has been shown to be important for proper UV damage response (Guerrero-Santoro et al. 2008; Sugasawa et al. 2005). DC(ERCC8) also known as the Cockayne syndrome A complex is thought to play a role in transcription-coupled repair (Groisman et al. 2003) and DC(ROC1) has been shown to be a histone ubiquitin ligase that participates in the cellular response to DNA damage by the ubiquitylation of H3 and H4 histones (Wang et al. 2006). DDB1 also

binds to SKP2 and plays a role in the ubiquination of CDKN1B, a cyclin-dependent kinase inhibitor (Nishitani et al. 2006) and may recruit nucleotide excision repair proteins in order to repair DNA damage (Li et al. 2006). Deficiency in DDB1 is strongly associated to xeroderma pigmentosum (Hwang et al. 1996; Hwang et al. 1998; Kapetanaki et al. 2006) which is marked by signs of premature skin ageing (Andrews et al. 1978). Mutational inactivation of DDB1 is also associated to the Cockayne syndrome (Groisman et al. 2003) which is characterized by premature and accelerated ageing in addition to neurodegeneration (Weidenheim et al. 2009). All this seem to suggest that our approach was able to detect proteins that are connected to the ageing process. We speculate that the evolutionary pressure on proteins involved in DNA repair or DNA damage response could be an optimization leading to a better regulation of damage, cell cycle and genome stability and hence to a longer lifespan.

Since we have a poor assembly of the naked mole-rat complementary DNA, we can possibly predict the protein sequence of DDB1 in the naked-mole rats. We can then determine whether DDB1 has been selected in the naked-mole rats. Since DDB1 is strongly associated to UV-dependent DNA damage response, a strong selection in the naked-mole rats lineage could suggest that DDB1 has a role in mammalian ageing in a UV-independent pathways as the naked-mole rats live underground (Buffenstein et al. 1991), perhaps through its role in the ubiquitin-proteasome pathway. Other putative genes or proteins related to ageing can be analysed in the context of the extreme longevity difference between naked mole-rats and wild-type mice. In fact, the de Magalhães lab and the Church lab among others are collaborating in a project which aims to clone genes from the naked mole-rats and use homologous recombination techniques in order to replace a wild-type mice version of a genes by its orthologous version from the naked mole-rat in embryonic stem cell. One of their long-term project is to develop a high-throughput method for homologous recombination of genes across species and upon its completion, it will be possible to test many putative ageing genes predicted via computational methods

in cell lines or *in vivo*.

CONCLUSIONS

In this thesis, we presented three original projects in which we use comparative genomics in order to exploit the wealth of genome data available to researchers in biology. We showed that not only is it possible to discover simple correlations such as residue usage and maximal lifespan, but it is also possible to detect more complex patterns of longevity-specific accelerated evolution at the protein level. The analysis of protein evolution used many fundamental techniques of comparative genomics such as ancestral genome reconstruction as well as sequence divergence estimation. Although these techniques are widely used, our method is significantly novel in the way it glues the different algorithms together and the way it uses control pairs for outliers detection.

We also showed that it is possible to make use of second generation data in order to study non-traditional organisms for which no annotated reference genome is available. We believe that this line of research paves the way for many future research projects using non-traditional model organisms of great interest to the field of ageing.

Our work allowed us to discover many interesting evolutionary trends related to species divergence of ageing and longevity. Our findings supports the idea that ROS exert a lesser influence in mammalian ageing than in lower organisms, perhaps due to new mechanisms dealing with oxidative stress in mammals. We identified many pathways that might be of importance to ageing such as lipid metabolism, proteasome-ubiquitin, damage response and damage repair pathways, the optimisation of which might have contributed to an increase in maximal lifespan in several mammals. Furthermore, both lipid metabolism and proteasome systems were among the top hits in two of our independent studies. Indeed, both lipid metabolism and proteasome systems were enriched in proteins with longevity-

specific selection and enriched in genes over-expressed in the long-lived naked mole-rats.

Finally, perhaps of greater interest, we have identified single candidate genes that are of interest for future studies in the context of ageing. The DNA damage-binding 1, DDB1, shows a strong signature of longevity-specific selection and has been associated with many diseases associated with ageing such as xeroderma pigmentosum and the Cockayne syndrome. CAPNS1, another protein with longevity-specific selection, has been implicated in DNA damage response by phosphorylating H2AX which helps in opening stretches of DNA to DNA repair proteins. We found A2M, SAT1, SAT2, IGFBP2, IGFBP4 and GSTA2 to be over-expressed in the liver of naked mole-rats when compared to wild-type mice. It would be of great interest for future experiments to test the effect of mutations within these genes on the lifespan of mice. It would also be interesting to observe the effects of replacing the mice promoters of genes over-expressed in the naked mole-rats by its naked mole-rat ortholog on biomarkers of ageing.

At this point, we hope to have convinced the reader that comparative genomics is not only an useful tool for research in biology and ageing, but it can stand as a field on its own by both both generating hypotheses and testing them.

APPENDIX

Supplemental Material

1.1 $c = median\{\log \frac{X_g^1}{X_g^2} : g \in G'\}$ **is the solution to** $min_c \sum_{g \in G'} \left| \log(\frac{X_g^1}{X_g^2}) - c \right|$ **.**

Proof: Let $Y_g = \log \frac{X_g^1}{X_g^2}$, $A(c) = \{g \in G' : Y_g < c\}$ and $B(c) = \{g \in G' : Y_g > c\}$.

Then we have, $\sum_{g \in G'} |Y_g - c| = \sum_{g \in A(c)} (c - Y_g) + \sum_{g \in B(c)} (Y_g - c)$ and taking the derivative with respective to *c* yields $\sum_{g \in A(c)} 1 + \sum_{g \in B(c)} -1$, which is 0 when $|A(c)| = |B(c)|$, this happens when *c* is the median of $\{Y_g : g \in G'\}$. Since the function $f(c) = \sum_{g \in G'} \left| \log(\frac{X_g^1}{X_g^2}) - c \right|$ is bounded below by 0 and is monotone decreasing when *c* is such that $|A(c)| < |B(c)|$ and monotone increasing when *c* is such that $|A(c)| > |B(c)|$, *f(c)* must only have one local (and global) minimum, thus the solution of $f'(c) = 0$ which is $c = median\{\log \frac{X_g^1}{X_g^2} : g \in G'\}$ is the solution to $min_c \sum_{g \in G'} \left| \log(\frac{X_g^1}{X_g^2}) - c \right|$. QED.

1.2 Table of GO categories enriched in proteins with high longevity-selectivity scores.

Statistical cutoff	0.05				0.1				0.2			
GO categories	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
postsynaptic membrane	.001	40.000	24.100	47.31	1.001	46.000	28.400	59.84	6.001	46.000	31.400	61.54
synapse	.001	56.000	40.000	63.38	0.001	69.000	47.100	83.79	1.001	64.000	52.000	84.17
spermatid development	.040	9.000	5.700	10.25	8.006	12.000	6.700	15.40	6.003	12.000	7.400	17.92
response to food	.053	3.000	1.700	4.08	3.007	4.000	2.000	7.167	7.003	5.000	2.200	8.583
actin binding	.004	54.000	36.600	53.19	1.040	51.000	43.100	56.04	5.007	53.000	47.600	68.86
phospholipid metabolic process	.026	11.000	4.500	9.61	1.026	9.000	5.300	11.08	3.002	12.000	5.900	17.47
postsynaptic density	.007	17.000	10.500	19.67	5.009	17.000	12.400	22.78	6.052	15.000	13.700	21.37
fatty acid beta-oxidation	.004	6.000	3.700	10.00	0.006	9.000	4.300	11.77	9.153	7.000	4.800	7.505
ATP-dependent peptidase activity	.021	6.000	2.000	5.767	7.004	5.000	2.300	8.162	2.044	3.000	2.600	6.200
muscle development	.004	21.000	13.600	25.52	3.008	21.000	16.000	27.08	7.162	15.000	17.700	22.86
1-phosphatidylinositol-3-kinase activity	.009	4.000	1.400	5.188	8.034	4.000	1.700	4.650	0.018	4.000	1.800	5.957
phosphoinositide 3-kinase complex	.011	5.000	1.700	5.552	2.016	5.000	2.000	6.013	3.043	4.000	2.200	5.957
peptidase activator activity	.067	2.000	.900	2.400	0.002	3.000	1.000	5.600	0.068	3.000	1.100	3.164
proteasomal ubiquitin-dependent protein catabolic process	.185	6.000	2.300	3.67	3.030	7.000	2.700	6.477	7.004	5.000	3.000	10.37
protein ubiquitination during ubiquitin-dependent protein ca	.039	3.000	2.300	5.41	7.015	5.000	2.700	7.167	7.042	4.000	3.000	6.893
condensed chromosome kinetochore	.007	10.000	6.200	14.00	4.012	8.000	7.300	14.96	5.452	7.000	8.100	8.547
circadian rhythm	.042	8.000	5.400	9.845	.010	10.000	6.300	13.59	6.096	9.000	7.000	11.21
lipid biosynthetic process	.107	5.000	4.000	6.651	1.016	7.000	4.700	10.87	3.036	6.000	5.200	10.45

(a) p-value, (b) number of proteins with LSS score bigger than 0 (c) expected sum of LSS score (d) actual s

Table 4: GO categories and their longevity-specific selection significance

1.3 Website supplemental data

www.cs.mcgill.ca/~yli142/thesis_supplementary.tar

BIBLIOGRAPHY

Allman, J., McLaughlin, T., and Hakeem, A. (1993). Brain weight and life-span in primate species. Proceedings of the National Academy of Sciences of the United States of America 90, 118-22.

Amin, S., Riggs, B. L., Atkinson, E. J., Oberg, A. L., Melton, L. J., and Khosla, S. (2004). A potentially deleterious role of IGFBP-2 on bone density in aging men and women. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research 19, 1075-83.

Andrews, A. D., Barrett, S. F., and Robbins, J. H. (1978). Xeroderma pigmentosum neurological abnormalities correlate with colony-forming ability after ultraviolet radiation. Proceedings of the National Academy of Sciences of the United States of America 75, 1984-8.

Andziak, B., and Buffenstein, Rochelle (2006). Disparate patterns of age-related changes in lipid peroxidation in long-lived naked mole-rats and shorter-lived mice. Aging cell 5, 525-32.

Antebi, A. (2007). Genetics of aging in Caenorhabditis elegans. PLoS genetics 3, 1565-71.

Arcaro, A., Zvelebil, M. J., Wallasch, C., Ullrich, A., Waterfield, M. D., and Domin, J. (2000). Class II phosphoinositide 3-kinases are downstream targets of activated polypeptide growth factor receptors. Molecular and cellular biology 20, 3817-30.

Arking, R. (2006). The biology of aging: observations and principles (Oxford University Press).

Ashburner, M. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25, 25-9.

Austad, S N (1997). Comparative aging and life histories in mammals. Experimental gerontology 32, 23-38.

Austad, Steven N (2009). Comparative biology of aging. The journals of gerontology. Series A, Biological sciences and medical sciences 64, 199-201.

Austad, Steven N (2005). Diverse aging rates in metazoans: targets for functional genomics. Mechanisms of ageing and development 126, 43-9.

Ayres, J. S., and Schneider, D. S. (2009). The role of anorexia in resistance and tolerance to infections in Drosophila. PLoS biology 7, e1000150.

Baek, J. H., Liu, Y. V., McDonald, K. R., Wesley, J. B., Zhang, H., and Semenza, G. L. (2007). Spermidine/spermine N(1)-acetyltransferase-1 binds to hypoxia-

inducible factor-1alpha (HIF-1alpha) and RACK1 and promotes ubiquitination and degradation of HIF-1alpha. The Journal of biological chemistry 282, 33358-66.

Barnes, M. (2007). Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data (Wiley).

Bender, A., Hajieva, P., and Moosmann, B. (2008). Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria. Proceedings of the National Academy of Sciences of the United States of America 105, 16496-501.

Bendjennat, M., Boulaire, J., Jascur, T., Brickner, H., Barbier, V., Sarasin, A., Fotedar, A., and Fotedar, R. (2003). UV irradiation triggers ubiquitin-dependent degradation of p21(WAF1) to promote DNA repair. Cell 114, 599-610.

Bengtson, V., Gans, D., Putney, N., and Silverstein, M. (2008). Handbook of Theories of Aging, Second Edition (Springer Publishing Company).

Bergman, A., Atzmon, G., Ye, K., MacCarthy, T., and Barzilai, N. (2007). Buffering mechanisms in aging: a systems approach toward uncovering the genetic component of aging. PLoS computational biology 3, e170.

Blacker, D. et al. (1998). Alpha-2 macroglobulin is genetically associated with Alzheimer disease. Nature genetics 19, 357-60.

Bonafè, M. et al. (2003). Polymorphic variants of insulin-like growth factor I (IGF-I) receptor and phosphoinositide 3-kinase genes affect IGF-I plasma levels and human longevity: cues for an evolutionarily conserved mechanism of life span control. The Journal of clinical endocrinology and metabolism 88, 3299-304.

Brégégère, F., Milner, Y., and Friguet, Bertrand (2006). The ubiquitin-proteasome system at the crossroads of stress-response and ageing pathways: a handle for skin care? Ageing research reviews 5, 60-90.

Buffenstein, R, and Yahav, S. (1991). Cholecalciferol has no effect on calcium and inorganic phosphorus balance in a naturally cholecalciferol-deplete subterranean mammal, the naked mole rat (Heterocephalus glaber). The Journal of endocrinology 129, 21-6.

Buffenstein, Rochelle (2008). Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology 178, 439-45.

Buffenstein, Rochelle (2005). The naked mole-rat: a new long-living model for human aging research. The journals of gerontology. Series A, Biological sciences and medical sciences 60, 1369-77.

Buffenstein, Rochelle, and Jarvis, J. U. M. The naked mole rat--a new record for the

oldest living rodent. Science of aging knowledge environment : SAGE KE 2002, pe7.

Chanoux, R. A., Yin, B., Urtishak, K. A., Asare, A., Bassing, C. H., and Brown, E. J. (2009). ATR and H2AX cooperate in maintaining genome stability under replication stress. The Journal of biological chemistry 284, 5994-6003.

Chen, D., Thomas, E. L., and Kapahi, P. (2009). HIF-1 modulates dietary restriction-mediated lifespan extension via IRE-1 in Caenorhabditis elegans. PLoS genetics 5, e1000486.

Chondrogianni, N, Petropoulos, I., Franceschi, C, Friguet, B, and Gonos, E S (2000). Fibroblast cultures from healthy centenarians have an active proteasome. Experimental gerontology 35, 721-8.

Chondrogianni, Niki, and Gonos, Efstathios S (2005). Proteasome dysfunction in mammalian aging: steps and factors involved. Experimental gerontology 40, 931-8.

Christensen, K., Johnson, T. E., and Vaupel, J. W. (2006). The quest for genetic determinants of human longevity: challenges and insights. Nature reviews. Genetics 7, 436-48.

Cooper, C. (2001). Applications of microarray technology in breast cancer research. Breast Cancer Res 3, 158-175.

Cozzetto, D., and Tramontano, A. (2008). Advances and pitfalls in protein structure prediction. Current protein & peptide science 9, 567-77.

Cristianini, N., and Hahn, M. W. (2007). Introduction to computational genomics: a case studies approach (Cambridge University Press).

Csiszar, A., Labinskyy, N., Orosz, Z., Xiangmin, Z., Buffenstein, Rochelle, and Ungvari, Z. (2007). Vascular aging in the longest-living rodent, the naked mole rat. American journal of physiology. Heart and circulatory physiology 293, H919-27.

Das, N. K., and Murphy, D. G. (1978). National Institute on Aging cell-line repository. Experimental aging research 4, 321-31.

Dayhoff, M. (1965). Atlas of protein sequence and structure.

Demarchi, F., Cataldo, F., Bertoli, C., and Schneider, C. (2010). DNA damage response links calpain to cellular senescence. Cell cycle (Georgetown, Tex.) 9, 755-60.

Demarchi, F., and Schneider, C. (2007). The calpain system as a modulator of stress/damage response. Cell cycle (Georgetown, Tex.) 6, 136-8.

Dephoure, N., Zhou, C., Villén, J., Beausoleil, S. A., Bakalarski, C. E., Elledge, S. J., and Gygi, S. P. (2008). A quantitative atlas of mitotic phosphorylation. Proceedings of the National Academy of Sciences of the United States of

America 105, 10762-7.

Didichenko, S. A., Fragoso, C. M., and Thelen, M. (2003). Mitotic and stress-induced phosphorylation of HsPI3K-C2alpha targets the protein for degradation. The Journal of biological chemistry 278, 26055-64.

Dollé, M. E., Giese, H., Hopkins, C. L., Martus, H. J., Hausdorff, J. M., and Vijg, J (1997). Rapid accumulation of genome rearrangements in liver but not in brain of old mice. Nature genetics 17, 431-4.

Edwards, R. J., and Shields, D. C. (2004). GASP: Gapped Ancestral Sequence Prediction for proteins. BMC bioinformatics 5, 123.

Evans, R. M., Hui, S., Perkins, A., Lahiri, D. K., Poirier, J., and Farlow, M. R. (2004). Cholesterol and APOE genotype interact to influence Alzheimer disease progression. Neurology 62, 1869-71.

Farout, L., and Friguet, Bertrand (2008). Proteasome function in aging and oxidative stress: implications in protein maintenance failure. Antioxidants & redox signaling 8, 205-16.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. Am Nat 125, 1.

Finch, C. E. (2010). Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. Proceedings of the National Academy of Sciences of the United States of America 107 Suppl , 1718-24.

Finch, C. E., and Stanford, C. B. (2004). Meat-adaptive genes and the evolution of slower aging in humans. The Quarterly review of biology 79, 3-50.

Finkel, T, and Holbrook, N. J. (2000). Oxidants, oxidative stress and the biology of ageing. Nature 408, 239-47.

Finkel, Toren, Serrano, M., and Blasco, M. A. (2007). The common biology of cancer and ageing. Nature 448, 767-74.

Flachsbart, F., Caliebe, A., Nothnagel, M., Kleindorp, R., Nikolaus, S., Schreiber, S., and Nebel, A. (2010). Depletion of potential A2M risk haplotype for Alzheimer's disease in long-lived individuals. European journal of human genetics : EJHG 18, 59-61.

Flicek, P. et al. (2010). Ensembl's 10th year. Nucleic acids research 38, D557-62.

Friguet, B, Bulteau, A. L., Chondrogianni, N, Conconi, M., and Petropoulos, I. (2000). Protein degradation by the proteasome and its implications in aging. Annals of the New York Academy of Sciences 908, 143-54.

Fukushi, J.-ichi, Makagiansar, I. T., and Stallcup, W. B. (2004). NG2 proteoglycan promotes endothelial cell motility and angiogenesis via engagement of galectin-3 and alpha3beta1 integrin. Molecular biology of the cell 15, 3580-90.

Futuyma, D. J. (1997). Evolutionary Biology (Sinauer Associates).

Gaffney, D. J., and Keightley, P. D. (2005). The scale of mutational variation in the murid genome. Genome research 15, 1086-94.

Garcia, Y. J., Rodríguez-Malaver, A. J., and Peñaloza, N. (2005). Lipid peroxidation measurement by thiobarbituric acid assay in rat cerebellar slices. Journal of neuroscience methods 144, 127-35.

Gourlay, C. W., and Ayscough, K. R. (2005). The actin cytoskeleton: a key regulator of apoptosis and ageing? Nature reviews. Molecular cell biology 6, 583-9.

Gourlay, C. W., Carpp, L. N., Timpson, P., Winder, S. J., and Ayscough, K. R. (2004). A role for the actin cytoskeleton in cell death and aging in yeast. The Journal of cell biology 164, 803-9.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science (New York, N.Y.) 185, 862-4.

Groisman, R., Polanowska, J., Kuraoka, I., Sawada, J.-ichi, Saijo, M., Drapkin, R., Kisselev, A. F., Tanaka, Kiyoji, and Nakatani, Y. (2003). The ubiquitin ligase activity in the DDB2 and CSA complexes is differentially regulated by the COP9 signalosome in response to DNA damage. Cell 113, 357-67.

Gu, Jian, Spitz, M. R., Zhao, H., Lin, J., Grossman, H. B., Dinney, C. P., and Wu, X. (2005). Roles of tumor suppressor and telomere maintenance genes in cancer and aging--an epidemiological study. Carcinogenesis 26, 1741-7.

Guerrero-Santoro, J., Kapetanaki, M. G., Hsieh, C. L., Gorbachinsky, I., Levine, A. S., and Rapić-Otrin, V. (2008). The cullin 4B-based UV-damaged DNA-binding protein ligase binds to UV-damaged chromatin and ubiquitinates histone H2A. Cancer research 68, 5014-22.

Guo, Huili, Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature 466, 835-840.

Hanasoge, S., and Ljungman, M. (2007). H2AX phosphorylation after UV irradiation is triggered by DNA repair intermediates and is mediated by the ATR kinase. Carcinogenesis 28, 2298-304.

Hardison, R. C. et al. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome research 13, 13-26.

Harel, A., Orjalo, A. V., Vincent, T., Lachish-Zalait, A., Vasu, S., Shah, S., Zimmerman, E., Elbaum, M., and Forbes, D. J. (2003). Removal of a single pore subcomplex results in vertebrate nuclei devoid of nuclear pores. Molecular cell 11, 853-64.

Harman, D. (1956). Aging: a theory based on free radical and radiation chemistry. Journal of gerontology 11, 298-300.

Harman, D. (1968). Free radical theory of aging: effect of free radical reaction inhibitors on the mortality rate of male LAF mice. Journal of gerontology 23, 476-82.

Hasegawa, M, Kishino, H, and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of molecular evolution 22, 160-74.

Hekimi, S., and Guarente, L. (2003). Genetics and the specificity of the aging process. Science (New York, N.Y.) 299, 1351-4.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89, 10915-9.

Hennekens, C. H., and Buring, J. E. (1987). Epidemiology in Medicine (Lippincott Williams & Wilkins).

Huang, D. W. et al. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome biology 8, R183.

Hulbert, A. J. (2008). Explaining longevity of different animals: is membrane fatty acid composition the missing link? Age (Dordrecht, Netherlands) 30, 89-97.

Hunter, D. J., Altshuler, D., and Rader, D. J. (2008). From Darwin's finches to canaries in the coal mine--mining the genome for new biology. The New England journal of medicine 358, 2760-3.

Hwang, B. J., Liao, J. C., and Chu, G (1996). Isolation of a cDNA encoding a UV-damaged DNA binding factor defective in xeroderma pigmentosum group E cells. Mutation research 362, 105-17.

Hwang, B. J., Toering, S., Francke, U., and Chu, G (1998). p48 Activates a UV-damaged-DNA binding factor and is defective in xeroderma pigmentosum group E cells that lack binding activity. Molecular and cellular biology 18, 4391-9.

Jobson, R. W., Nabholz, B., and Galtier, N. (2010). An evolutionary genome scan for longevity-related natural selection in mammals. Molecular biology and evolution 27, 840-7.

Jones, J. I., and Clemmons, D. R. (1995). Insulin-Like Growth Factors and Their Binding Proteins: Biological Actions. Endocrine Reviews 16, 3-34.

Kaeberlein, M., Burtner, C. R., and Kennedy, B. K. (2007). Recent developments in yeast aging. PLoS genetics 3, e84.

Kanungo, M. S. (1994). Genes and Aging (Cambridge University Press).

Kapetanaki, M. G., Guerrero-Santoro, J., Bisi, D. C., Hsieh, C. L., Rapić-Otrin, V., and Levine, A. S. (2006). The DDB1-CUL4ADDB2 ubiquitin ligase is

deficient in xeroderma pigmentosum group E and targets histone H2A at UV-damaged DNA sites. Proceedings of the National Academy of Sciences of the United States of America 103, 2588-93.

Kim, J. B., Porreca, G. J., Song, L., Greenway, S. C., Gorham, J. M., Church, G. M., Seidman, C. E., and Seidman, J. G. (2007). Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science (New York, N.Y.) 316, 1481-4.

Kim, S. K. (2007). Common aging pathways in worms, flies, mice and humans. The Journal of experimental biology 210, 1607-12.

Kirkwood, T. B. L. (1977). Evolution of ageing. Nature 270, 301-304.

Kirkwood, Thomas B L (2008). A systematic look at an old problem. Nature 451, 644-7.

Kitazoe, Y., Kishino, Hirohisa, Hasegawa, Masami, Nakajima, N., Thorne, J. L., and Tanaka, M. (2008). Adaptive threonine increase in transmembrane regions of mitochondrial proteins in higher primates. PloS one 3, e3343.

Knut, S.-N. (1984). Scaling: Why is Animal Size so Important? (Cambridge University Press).

Kondo, T. et al. (2004). Rapid degradation of Cdt1 upon UV-induced DNA damage is mediated by SCFSkp2 complex. The Journal of biological chemistry 279, 27315-9.

Ku, C. S., Loy, E. Y., Pawitan, Y., and Chia, K. S. (2010). The pursuit of genome-wide association studies: where are we now? Journal of human genetics 55, 195-206.

Kua, C.-H. (2006). Uncoupling the relationship between fatty acids and longevity. IUBMB life 58, 153-5.

Lambert, J. C. et al. (2000). Independent association of an APOE gene promoter polymorphism with increased risk of myocardial infarction and decreased APOE plasma concentrations-the ECTIM study. Human molecular genetics 9, 57-61.

Lander, E. S. et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10, R25.

Lee, C. K., Klopp, R. G., Weindruch, R, and Prolla, T. A. (1999). Gene expression profile of aging and its retardation by caloric restriction. Science (New York, N.Y.) 285, 1390-3.

Levine, R. L., Mosoni, L., Berlett, B. S., and Stadtman, E R (1996). Methionine

residues as endogenous antioxidants in proteins. Proceedings of the National Academy of Sciences of the United States of America 93, 15036-40.

Levine, R., and Stadtman, E R (2001). Oxidative modification of proteins during aging. Experimental Gerontology 36, 1495-1502.

Li, Heng, Ruan, J., and Durbin, Richard (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research 18, 1851-8.

Li, Jinyou, Wang, Q.-E., Zhu, Q., El-Mahdy, M. A., Wani, G., Praetorius-Ibba, M., and Wani, A. A. (2006). DNA damage binding protein component DDB1 participates in nucleotide excision repair through DDB2 DNA-binding and cullin 4A ubiquitin ligase activity. Cancer research 66, 8590-7.

Li, L., Ye, H., Guo, Hongqing, and Yin, Y. (2010). Arabidopsis IWS1 interacts with transcription factor BES1 and is involved in plant steroid hormone brassinosteroid regulated gene expression. Proceedings of the National Academy of Sciences of the United States of America 107, 3918-23.

Li, Ruiqiang, Li, Yingrui, Kristiansen, K., and Wang, Jun (2008). SOAP: short oligonucleotide alignment program. Bioinformatics (Oxford, England) 24, 713-4.

Li, W.-H. (1997). Molecular Evolution (Sinauer Associates).

Li, Yingrui, Hu, Y., Bolund, L., and Wang, Jun (2010). State of the art de novo assembly of human genomes from massively parallel sequencing data. Human genomics 4, 271-7.

Liang, S., Mele, J., Wu, Y., Buffenstein, Rochelle, and Hornsby, P. J. (2010). Resistance to experimental tumorigenesis in cells of a long-lived mammal, the naked mole-rat (Heterocephalus glaber). Aging cell 9, 626-35.

Lin, S.-J., Kaeberlein, M., Andalis, A. A., Sturtz, L. A., Defossez, P.-A., Culotta, V. C., Fink, G. R., and Guarente, L. (2002). Calorie restriction extends Saccharomyces cerevisiae lifespan by increasing respiration. Nature 418, 344-8.

Lindström, J. (1999). Early development and fitness in birds and mammals. Trends in Ecology & Evolution 14, 343-348.

Lipton, S. A., Gu, Z., and Nakamura, T. (2007). Neuroinflammation in Neuronal Death and Repair (Elsevier).

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., and Yankner, B. A. (2004). Gene regulation and DNA damage in the ageing human brain. Nature 429, 883-91.

Ma, H., Li, Renzhong, Zhang, Z., and Tong, T. (2004). mRNA level of alpha-2-macroglobulin as an aging biomarker of human fibroblasts in culture.

Experimental gerontology 39, 415-21.

Magalhães, J. P. de (2005). Human disease-associated mitochondrial mutations fixed in nonhuman primates. Journal of molecular evolution 61, 491-7.

Magalhães, J. P. de (2003). Is mammalian aging genetically controlled? Biogerontology 4, 119-20.

Magalhães, J. P. de, Budovsky, A., Lehmann, G., Costa, J., Li, Yang, Fraifeld, V., and Church, G. M. (2009a). The Human Ageing Genomic Resources: online databases and tools for biogerontologists. Aging cell 8, 65-72.

Magalhães, J. P. de, and Church, G. M. (2007). Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. Mechanisms of ageing and development 128, 355-64.

Magalhães, J. P. de, and Church, G. M. (2006). Cells discover fire: employing reactive oxygen species in development and consequences for aging. Experimental gerontology 41, 1-10.

Magalhães, J. P. de, and Church, G. M. (2005). Genomes optimize reproduction: aging as a consequence of the developmental program. Physiology (Bethesda, Md.) 20, 252-9.

Magalhães, J. P. de, and Costa, J. (2009b). A database of vertebrate longevity records and their relation to other life-history traits. Journal of evolutionary biology 22, 1770-4.

Magalhães, J. P. de, Costa, J., and Church, G. M. (2007). An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. The journals of gerontology. Series A, Biological sciences and medical sciences 62, 149-60.

Magalhães, J. P. de, Curado, J., and Church, G. M. (2009c). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics (Oxford, England) 25, 875-81.

Magalhães, J. P. de, Finch, C. E., and Janssens, G. (2010). Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions. Ageing research reviews 9, 315-23.

Magalhães, J. P. de, and Toussaint, Olivier (2002). The evolution of mammalian aging. Experimental gerontology 37, 769-75.

Marguerat, S., and Bähler, J. (2010). RNA-seq: from technology to biology. Cellular and molecular life sciences : CMLS 67, 569-79.

Martin, G. M., and Oshima, J. (2000). Lessons from human progeroid syndromes. Nature 408, 263-6.

Martin, G. M., Smith, A. C., Ketterer, D. J., Ogburn, C. E., and Disteche, C. M. (1985). Increased chromosomal aberrations in first metaphases of cells isolated

from the kidneys of aged mice. Israel journal of medical sciences 21, 296-301.

Masliah, E., Mallory, M., Ge, N., Alford, M., Veinbergs, I., and Roses, A. D. (1995). Neurodegeneration in the central nervous system of apoE-deficient mice. Experimental neurology 136, 107-22.

Matsuoka, S. et al. (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science (New York, N.Y.) 316, 1160-6.

Mattison, J. A., Lane, M. A., Roth, G. S., and Ingram, D. K. (2003). Calorie restriction in rhesus monkeys. Experimental gerontology 38, 35-46.

McAuley, M. T., Kenny, R. A., Kirkwood, Thomas B L, Wilkinson, D. J., Jones, J. J. L., and Miller, V. M. (2009). A mathematical model of aging-related and cortisol induced hippocampal dysfunction. BMC neuroscience 10, 26.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C.-S., Jan, Y. N., Kenyon, C., Bargmann, C. I., and Li, Hao (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nature genetics 36, 197-204.

Medawar, P. (1952). An Unsolved Problem of Biology (London: H.K. Lewis).

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nature methods 6, S13-20.

Miller, W. et al (2007). 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Research 17, 1797-1808.

Mohamed, R., and Lavin, M. F. (1986). Ataxia-telangiectasia cell extracts confer radioresistant DNA synthesis on control cells. Experimental cell research 163, 337-48.

Mohan, S., Farley, J. R., and Baylink, D. J. (1995). Age-related changes in IGFBP-4 and IGFBP-5 levels in human serum and bone: implications for bone loss with aging. Progress in growth factor research 6, 465-73.

Montgomery, M. K., Xu, S., and Fire, A. (1998). RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans. Proceedings of the National Academy of Sciences of the United States of America 95, 15502-7.

Moosmann, B., and Behl, C. (2008). Mitochondrially encoded cysteine predicts animal lifespan. Aging cell 7, 32-46.

Moskovitz, J., Bar-Noy, S., Williams, W. M., Requena, J., Berlett, B. S., and Stadtman, E R (2001). Methionine sulfoxide reductase (MsrA) is a regulator of antioxidant defense and lifespan in mammals. Proceedings of the National Academy of Sciences of the United States of America 98, 12920-5.

Muller, F. (2000). The nature and mechanism of superoxide production by the electron transport chain: Its relevance to aging. AGE 23, 227-253.

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research 17, 413-21.

Naor, D., Fischer, D., Jernigan, R. L., Wolfson, H. J., and Nussinov, R. (1996). Amino acid pair interchanges at spatially conserved locations. Journal of molecular biology 256, 924-38.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48, 443-53.

Nishitani, H. et al. (2006). Two E3 ubiquitin ligases, SCF-Skp2 and DDB1-Cul4, target human Cdt1 for proteolysis. The EMBO journal 25, 1126-36.

O Dayhoff, M., M Schwartz, R., and C Orcutt, B. (1978). A model of evolutionary change in proteins. Atlas of protein sequence and structure 5, 345-352.

Orjalo, A. V., Arnaoutov, A., Shen, Z., Boyarchuk, Y., Zeitlin, S. G., Fontoura, B., Briggs, S., Dasso, M., and Forbes, D. J. (2006). The Nup107-160 nucleoporin complex is required for correct bipolar spindle assembly. Molecular biology of the cell 17, 3806-18.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic acids research 38, D196-203.

Painter, R. B. (1981). Radioresistant DNA synthesis: an intrinsic feature of ataxia telangiectasia. Mutation research 84, 183-90.

Pan, F. et al. (2007). Gene Aging Nexus: a web database and data mining platform for microarray data on aging. Nucleic acids research 35, D756-9.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nature reviews. Genetics 10, 669-80.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nature methods 6, S22-32.

Pesce, K., and Rothe, M. J. The premature aging syndromes. Clinics in dermatology 14, 161-70.

Pevsner, J. (2009). Bioinformatics and Functional Genomics (Wiley-Blackwell).

Pérez, V. I. et al. (2009). Protein stability and resistance to oxidative stress are determinants of longevity in the longest-living rodent, the naked mole-rat. Proceedings of the National Academy of Sciences of the United States of America 106, 3059-64.

Pérez, V. I., Van Remmen, H., Bokov, A., Epstein, C. J., Vijg, Jan, and Richardson, A. (2009). The overexpression of major antioxidant enzymes does not extend the lifespan of mice. Aging cell 8, 73-5.

Rao, P. et al. (2010). IκBβ acts to inhibit and activate gene expression during the inflammatory response. Nature 466, 1115-1119.

Remondini, D. et al. (2010). Complex patterns of gene expression in human T cells during in vivo aging. Molecular bioSystems 6, 1983-92.

Renn, S. C. P., Aubin-Horth, N., and Hofmann, H. A. (2004). Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BMC genomics 5, 42.

Ricklefs, R. E. (2010). Life-history connections to rates of aging in terrestrial vertebrates. Proceedings of the National Academy of Sciences of the United States of America 107, 10314-9.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 11, R25.

Rose, M. R. (2009). Adaptation, aging, and genomic information. Aging 1, 444-50.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nature reviews. Genetics 11, 356-66.

Rottenberg, H. (2006). Longevity and the evolution of the mitochondrial DNA-coded proteins in mammals. Mechanisms of ageing and development 127, 748-60.

Ruan, H. et al. (2002). High-quality life extension by the enzyme peptide methionine sulfoxide reductase. Proceedings of the National Academy of Sciences of the United States of America 99, 2748-53.

Russell, S. J., and Norvig, P. (1995). Artificial Intelligence: A Modern Approach (Prentice Hall).

Sacher, G. A. (1959). Ciba Foundation Symposium - The Lifespan of Animals (Colloquia on Ageing, Vol. 5) G. E. W. Wolstenholme and M. O'Conner, eds. (Chichester, UK: John Wiley & Sons, Ltd).

Saito, K., Yoshioka, H., and Cutler, R. G. (1998). A spin trap, N-tert-butyl-alpha-phenylnitrone extends the life span of mice. Bioscience, biotechnology, and biochemistry 62, 792-4.

Salmon, A. B., Leonard, S., Masamsetti, V., Pierce, A., Podlutsky, A. J., Podlutskaya, N., Richardson, A., Austad, Steven N, and Chaudhuri, A. R. (2009). The long lifespan of two bat species is correlated with resistance to protein oxidation and enhanced protein homeostasis. The FASEB journal : official publication of the Federation of American Societies for Experimental

Biology 23, 2317-26.

Sasaki, M. S., and Taylor, A. M. (1994). Dissociation between radioresistant DNA replication and chromosomal radiosensitivity in ataxia telangiectasia cells. Mutation research 307, 107-13.

Sauer, U., Heinemann, M., and Zamboni, N. (2007). Genetics. Getting closer to the whole picture. Science (New York, N.Y.) 316, 550-1.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.) 270, 467-70.

Schächter, F., Faure-Delanef, L., Guénot, F., Rouger, H., Froguel, P, Lesueur-Ginot, L., and Cohen, D. (1994). Genetic associations with human longevity at the APOE and ACE loci. Nature genetics 6, 29-32.

Sebastiani, P. et al. (2010). Genetic Signatures of Exceptional Longevity in Humans. Science (New York, N.Y.) . [Epub ahead of print]

Seluanov, A., Hine, C., Azpurua, J., Feigenson, M., Bozzella, M., Mao, Z., Catania, K. C., and Gorbunova, V. (2009). Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. Proceedings of the National Academy of Sciences of the United States of America 106, 19352-7.

Shattuck, M. R., and Williams, S. A. (2010). Arboreality has allowed for the evolution of increased longevity in mammals. Proceedings of the National Academy of Sciences of the United States of America 107, 4635-9.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology 26, 1135-1145.

Shimokawa, I., Higami, Y., Utsuyama, M., Tuchiya, T., Komatsu, T., Chiba, T., and Yamaza, H. (2002). Life span extension by reduction in growth hormone-insulin-like growth factor-1 axis in a transgenic rat model. The American journal of pathology 160, 2259-65.

Sjögren, C., and Nasmyth, K. (2001). Sister chromatid cohesion is required for postreplicative double-strand break repair in Saccharomyces cerevisiae. Current biology : CB 11, 991-5.

Sladek, R. et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445, 881-5.

Smith, E. D. et al. (2008). Quantitative evidence for conserved longevity pathways between divergent eukaryotic species. Genome research 18, 564-70.

Smith, T. F., and Waterman, M S (1981). Identification of common molecular subsequences. Journal of molecular biology 147, 195-7.

Speakman, John R (2005a). Body size, energy metabolism and lifespan. The Journal of experimental biology 208, 1717-30.

Speakman, John R (2005b). Correlations between physiology and lifespan--two widely ignored problems with comparative studies. Aging cell 4, 167-75.

Stadtman, Earl R (2006). Protein oxidation and aging. Free radical research 40, 1250-8.

Sugasawa, K. et al. (2005). UV-induced ubiquitylation of XPC protein mediated by UV-DDB-ubiquitin ligase complex. Cell 121, 387-400.

Swedlow, J. R., Lewis, S. E., and Goldberg, I. G. (2006). Modelling data across labs, genomes, space and time. Nature cell biology 8, 1190-4.

Szilard, L. (1959). On the nature of the aging process. Proceedings of the National Academy of Sciences of the United States of America 45, 30-45.

Tassi, E., Biesova, Z., Di Fiore, P. P., Gutkind, J. S., and Wong, W. T. (1999). Human JIK, a novel member of the STE20 kinase family that inhibits JNK and is negatively regulated by epidermal growth factor. The Journal of biological chemistry 274, 33287-95.

Terashima, Y. et al. (2005). Pivotal function for cytoplasmic protein FROUNT in CCR2-mediated monocyte chemotaxis. Nature immunology 6, 827-35.

Torrents, D., Suyama, M., Zdobnov, E., and Bork, Peer (2003). A genome-wide survey of human pseudogenes. Genome research 13, 2559-67.

Toussaint, O, Dumont, P., Dierick, J. F., Pascal, T., Frippiat, C., Chainiaux, F., Magalhaes, J. P., Eliaers, F., and Remacle, J. (2000). Stress-induced premature senescence as alternative toxicological method for testing the long-term effects of molecules under development in the industry. Biogerontology 1, 179-83.

Tyers, M., and Mann, M. (2003). From genomics to proteomics. Nature 422, 193-7.

Vernace, V. A., Schmidt-Glenewinkel, T., and Figueiredo-Pereira, M. E. (2007). Aging and regulated protein degradation: who has the UPPer hand? Aging cell 6, 599-606.

Vijg, J (2000). Somatic mutations and aging: a re-evaluation. Mutation research 447, 117-35.

Wang, H., Zhai, L., Xu, J., Joo, H.-Y., Jackson, S., Erdjument-Bromage, H., Tempst, P., Xiong, Y., and Zhang, Y. (2006). Histone H3 and H4 ubiquitylation by the CUL4-DDB-ROC1 ubiquitin ligase facilitates cellular response to DNA damage. Molecular cell 22, 383-94.

Wang, H.-Y., Tang, H., Shen, C.-K. J., and Wu, C.-I. (2003). Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. Molecular biology and evolution 20, 1795-804.

Wang, X., Hinson, E. R., and Cresswell, P. (2007). The interferon-inducible protein viperin inhibits influenza virus release by perturbing lipid rafts. Cell host & microbe 2, 96-105.

Waterman, Michael S. (1995). Introduction to Computational Biology: Maps, Sequences and Genomes (Chapman & Hall/CRC Interdisciplinary Statistics) (Chapman and Hall/CRC).

Watson, J. D. (1990). The human genome project: past, present, and future. Science (New York, N.Y.) 248, 44-9.

Weidenheim, K. M., Dickson, D. W., and Rapin, I. (2009). Neuropathology of Cockayne syndrome: Evidence for impaired development, premature aging, and neurodegeneration. Mechanisms of ageing and development 130, 619-36.

Weigelt, J. (2010). Structural genomics-impact on biomedicine and drug discovery. Experimental cell research 316, 1332-8.

Weindruch, R, Walford, R L, Fligiel, S., and Guthrie, D. (1986). The retardation of aging in mice by dietary restriction: longevity, cancer, immunity and lifetime energy intake. The Journal of nutrition 116, 641-54.

Weindruch, Richard, and Walford, Roy L. (1988). The Retardation of Aging and Disease by Dietary Restriction (Charles C Thomas Pub Ltd).

Weinert, B. T., and Timiras, P. S. (2003). Invited review: Theories of aging. Journal of applied physiology (Bethesda, Md. : 1985) 95, 1706-16.

Williams, G. C. (1957). Pleiotropy, Natural Selection, and the Evolution of Senescence. Evolution 11, 398.

Wilson, A., Shehadeh, L. A., Yu, H., and Webster, K. A. (2010). Age-related molecular genetic changes of murine bone marrow mesenchymal stem cells. BMC genomics 11, 229.

WTCCConsortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-78.

Wu, C. H. et al. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic acids research 34, D187-91.

Wyse, C. A., Coogan, A. N., Selman, C., Hazlerigg, D. G., and Speakman, J R (2010). Association between mammalian lifespan and circadian free-running period: the circadian resonance hypothesis revisited. Biology letters. [Epub ahead of print]

Yazdi, P. T., Wang, Y., Zhao, S., Patel, N., Lee, E. Y.-H. P., and Qin, J. (2002). SMC1 is a downstream effector in the ATM/NBS1 branch of the human S-phase checkpoint. Genes & development 16, 571-82.

Yun, J., and Lee, W.-H. (2003). Degradation of transcription repressor ZBRK1 through the ubiquitin-proteasome pathway relieves repression of Gadd45a upon DNA damage. Molecular and cellular biology 23, 7305-14.

Yáñez, R. J., and Porter, A. C. G. (2002). Differential effects of Rad52p overexpression on gene targeting and extrachromosomal homologous

recombination in a human cell line. Nucleic acids research 30, 740-8.

Zahn, J. M. et al. (2007). AGEMAP: a gene expression database for aging in mice. PLoS genetics 3, e201.

Zhang, J., Webb, D. M., and Podlaha, O. (2002). Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. Genetics 162, 1825-35.

Zhang, W., Chen, T., Wan, T., He, L., Li, N., Yuan, Z., and Cao, X. (2000). Cloning of DPK, a novel dendritic cell-derived protein kinase activating the ERK1/ERK2 and JNK/SAPK pathways. Biochemical and biophysical research communications 274, 872-9.