

Durham E-Theses

TRANSCRIPTOMIC STUDIES IN AGEING AND SENESCENCE

LAHAT, ALBERT

How to cite:

LAHAT, ALBERT (2019) *TRANSCRIPTOMIC STUDIES IN AGEING AND SENESCENCE*, Durham theses, Durham University. Available at Durham E-Theses Online:
<http://etheses.dur.ac.uk/13129/>

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a [link](#) is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full Durham E-Theses policy](#) for further details.



TRANSCRIPTOMIC STUDIES IN AGEING AND SENESCENCE

Department of Biosciences 2019

A thesis submitted in accordance with the requirements for the degree of Doctor of
Philosophy

Albert Lahat

ACKNOWLEDGEMENTS

I could not have done this research without the supervision and tutelage of my PhD supervisor Sushma Nagaraja Grellscheid. I would also like to thank David Dolan at Durham University for help with day to day computational biology.

I would like to thank collaborators at Newcastle University who supported me through this research: Thomas von Zglinicki for many helpful discussions on ageing and senescence, and members of his laboratory Diana Jurk and Mikolaj Ogrodnik for helping in the preparation of the crossover datasets, and Satomi Miwa for collaboration on the postfed datasets and mitochondria proteomics data.

The computational resources and expertise provided by Anton Enright and Stijn van Dongen of the European Bioinformatics Institute were pivotal for my research.

I would also like to thank everyone who supported me non-academically especially my dear wife Rachel whose patience is greater than mine.

DECLARATION

For this thesis RNASeq datasets (in fastq format) were contributed by Thomas von Zglinicki's lab including work done by Sushma Grellscheid, Diana Jurk, Mikolaj Ogrodnik and Satomi Miwa which include the mice liver datasets studied in 0. I have then studied and analysed these datasets according to Chapter 2 and Chapter 4.

Publicly available IMR90 datasets were contributed by Peter Adam's lab which I have then studied and analysed these datasets according to Chapter 3 and Chapter 4.

Publicly available WI38 datasets were contributed by Tamir Chandra's group from the Babraham Institute. I have then studied and analysed these datasets according to Chapter 3 and Chapter 4.

ABSTRACT

Senescence is known as an irreversible departure from the cell cycle and is considered a leading factor in the ageing phenotype, including age related diseases. With limited exception (such as negligible senescence and biologically immortal organisms), all cells reach a senescent state and all organisms age.

The advent of Next-Generation Sequencing (NGS) technology has enabled the study of the cellular transcriptome in a highly intimate manner. From NGS experiments it is possible to infer both the quantity and isoform of a transcript of interest. Currently, an ever-growing body of easily accessible NGS experimental data has allowed researchers to foster collaborative endeavours by utilizing datasets from diverse experiments to ask new questions. Here, we study datasets from mouse tissue samples across the lifespan under a normal ad libitum diet or under 40% dietary restriction, and human cell lines which have undergone replicative senescence or irradiation induced senescence. The overall aim of this study was to investigate ageing from a transcriptomic point of view.

Here, we focus on a series of parallel projects to study the landscape of changes occurring in ageing and senescence. We investigate differential gene expression, differential exon usage, and differential lncRNA expression. To further understand the biological relevance of the landscape changes, we utilized gene ontology (GO and Reactome) enrichment for differential expression changes, and we present a novel tool (MAItESERS) to understand the biological significance of alternative exon usage.

We found that there was post senescence plasticity, meaning that both expression and splicing were altered after senescence induction. Our data suggests that senescence in mouse hepatic tissue was induced suddenly and catastrophically. We also observed three systems (immune/inflammation, chromatin structure, and energy metabolism) being strongly altered and each system can strongly induce changes in the other, which may strengthen the irreversibility of senescence.

TABLE OF CONTENTS

Acknowledgements	2
Declaration	2
Abstract.....	3
Table of Contents	4
Chapter 1 Introduction to ageing, Senescence, and RNA Biology	6
Ageing.....	6
Introduction	6
Senescence	9
Dietary Restriction pathways	9
RNA and Splicing Biology.....	13
Transcription	13
RNA Splicing.....	13
Objectives	15
Review of Bioinformatic Methods for RNA seq and Splicing Analysis	17
Quality Control of the Sequencing Output	18
Aligning	19
Gene Counting.....	21
Quantifying Differential Gene Expression.....	21
Gene network and clustering	24
Network and Cluster Fishing.....	24
Pathways analysis	25
Heterogeneity Analysis	26
Detection of Alternative Splicing	28
Protocol Management.....	29
Conclusion	32
Chapter 2 Age and Caloric Restriction Effect on Transcriptome	33
Introduction and Background.....	34
Results.....	41
Splicing in Mice Ageing and Diet	57
Conclusion	61
Senescence Blooming	63
Chapter 3 Senescence Signatures In human cell.....	67
WI38 oncogene induced senescence	67
Introduction	67
Methods	68
Results	70

Splicing	74
Conclusion	75
IMR90 ionizing radiation induced senescence	78
Introduction	78
Methods	78
Results	79
Splicing	80
Conclusion	81
Chapter 4 MAltESERS	82
Introduction	82
Algorithm.....	83
Testing with publically available data	84
Testing with mice ageing data	87
Testing with IMR90 data.....	89
Testing with WI38 data.....	90
Conclusion	91
Chapter 5 Conclusions.....	92
Senescence Trifecta.....	92
Ageing is an Event or a Process?	92
Future Research Questions.....	94
Post Senescence Plasticity.....	94
Senescence Trifecta Modelling.....	95
Appendix	97
References	112
External Files	127

CHAPTER 1 INTRODUCTION TO AGEING, SENESCENCE, AND RNA BIOLOGY

Chapter objectives:

- To illustrate the causes and effects of ageing
- To provide an overview of distinct theorized models that contribute to the ageing phenotype
- To summarize the basis of the central dogma
- To review genetic splicing and RNA metabolism

AGEING

INTRODUCTION

Ageing is a gradual physiological deterioration with age. Ageing leads to increased mortality, decreased stamina ¹, and has been found to increase morbidity in nearly every multicellular or asymmetrically dividing unicellular organism from yeast to mammals (with some notable exceptions including; *Hydra ssp* ², *Turritopsis ssp* jellyfish ³, lobsters ⁴, and planarian flatworms ⁵).

There is a diverse repertoire of biological reasons and mechanisms that cause ageing in organisms. Three major pathways have been shown to contribute to ageing; TOR (Target Of Rapamycin) pathway ^{6,7}, insulin/IGF-1-like pathway ⁸, and mitochondrial activity ⁹⁻¹³. These three pathways act independently with dissimilar downstream effects. However, experimental alterations to these pathways provide cumulative effects ¹⁴. Additional factors implicated in ageing are telomere shortening ¹⁵ (see senescence), genetic instability ¹⁶, metabolic waste ^{17,18}, and wear and tear ¹⁹.

One of the early theories affecting ageing was the Hayflick limit proposed in 1961. Due to the mechanistic nature of the DNA replication fork, the 5' ends of chromosomes shorten during each replication, as Okazaki fragments cannot be initiated. To guard against the loss of precious genetic material, 5' chromosomal ends are capped with a long repetitive sequence (TTAGGG in vertebrates). These repetitive non-coding sequences found at the 5' end of chromosomes, are known as telomeres. Telomeres are shortened during each replication cycle. However, in the

germline, telomeres are extended by the enzyme telomerase. Telomere length decreases with age. Human new-borns have approximately 11kb of telomeric caps²⁰, while at an advanced age (80+ years) telomere caps have shortened to approximately 4kb²¹. Telomere length at birth relates to the average lifespan in different dog breeds²². Telomeric attrition through cell replication leads to the cell reaching the replicative limit and becoming senescent in a pathway mediated by the enzymes p53 and p16²³. Replicative senescence resulting from telomere shortening is known as the Hayflick limit; named after Leonard Hayflick who showed that non-immortalized cell lines become senescent after a limited number of cell divisions^{15,24,25}.

An extensive study of the effects of ionizing radiation on animals was conducted during the Manhattan project in 1947 which showed that low dose radiation (doses which did not cause radiation sickness symptoms or induce tumours) seem to lower the lifespan of the organism and increase age like pathologies at an earlier age²⁶. This experiment showed the importance of DNA integrity even before we knew the actual structure and mechanisms of DNA function, this effect has also been found to affect humans. Apollo astronauts who left the magnetosphere (the magnetosphere protects the earth surface from solar ionizing radiation) had higher incidences of cardiovascular disease²⁷. Early research showed that individuals with improved DNA damage repair had longer lifespans²⁸. Therefore, reduction in lifespan is thought to be caused by accumulated cellular mutations, which cause cells to exit the cell cycle and enter an irreversible state of senescence.

Not all DNA insults are the same, external oxidative stressors such as UV, ionizing radiation, and oxidative stress tend to damage a single base which is then repaired by the base excision repair (BER) pathway and nucleotide excision repair (NER). DNA mismatch repair (MMR) is involved in repairing DNA damage during DNA replication. Regardless of the repair mechanisms, a permanent epigenetic modification is left on the repaired region. The most prominent change is the Ser-139 phosphorylation of a histone 2 variant H2AX (γ H2AX)²⁹ in regions up to a hundred nanometers in diameter around the original DNA damage insult (visible in an optical microscope through γ H2AX staining)³⁰. γ H2AX recruits DNA repairing factors MDC1 and 53BP1³¹. Accumulation of γ H2AX is a reliable marker of cellular senescence especially on telomeric loci³².

As the correct repair of individual DNA damage is stochastic (there is a chance that the wrong base is inserted and therefore a mutation is induced) the persistent DNA damage foci serve as a tally of the mutations received. With increased mutations, the likelihood of a cell to become tumorigenic increases. Therefore, a biological tally of DNA damage enables the cell to exit the cell cycle if it is at a high risk of inducing tumours.

Ageing is not only affected by nuclear DNA mechanism. Nearly all anabolic activity in a cell is coupled to the energy released by the conversion of adenosine triphosphate (ATP) to adenosine diphosphate (ADP). The majority of ATP is regenerated within the mitochondria, a cellular organelle which houses the citric acid cycle. The citric acid cycle utilizes derivatives of glucose and lipids as starting metabolites. These metabolites can be completely reduced to water and CO₂, using oxygen to regenerate the ATP. To do so, the mitochondria maintains a high proton concentration within its inner membrane. Protons are then pumped through ATP synthase. As mitochondrial membranes are 'leaky', escaped protons contribute to the synthesis of reactive oxygen species (ROS) such as the super acid anion radical and hydroxyl radical. Unlike other organelles, the mitochondria has its own DNA with its own genes.

It has been theorized by Dr Denham Harman in 1965 that the ROS leaking from mitochondria will cause gradual damage to the cell and nuclear DNA leading to ageing. This theory was supported by the findings that aged animals and tissues have greater ROS than young tissues³³⁻⁴⁰ and superoxide overproduction in mice leads to premature ageing⁴¹.

An extension to Harman's theory, postulated by Loeb et al⁴² suggests that free radicals from the electron transport chain damage the mitochondrial genome. These mutations cause the mitochondria to be less efficient, increasing the generation of ROS, thus exacerbating the mutation rate until the entire cell dies.

Other theories of ageing include biological by-products that cannot be digested metabolically. These by-products accumulate extracellularly and intracellularly within lysosomes. Metabolic waste may not decrease lifespan directly. However, there is evidence to suggest metabolic waste plays a role in specific age-related diseases and symptoms such as Alzheimer's (accumulation of β -amyloid in the central nervous system), Atherosclerosis (lipid accumulation in the wall of blood vessels), and arterial stiffening (due to spontaneous undigestible protein glycosylation⁴³). It is important to note that age prolonging treatments such as dietary restriction can delay the onset of Alzheimer's disease and β -amyloid build-up⁴⁴ as well as atherosclerosis⁴⁵. Thus, research shows that the metabolic-waste consequences are downstream to ageing in general and not a cause of ageing. However, autophagy, the process by which lysosomes and their content are cleared, is attenuated by ageing and activated by dietary restriction after and other age prolonging treatments⁴⁶.

SENESCENCE

Senescence is defined as an irreversible cell cycle arrest. A senescent cell remains metabolically active and develops a pro-inflammatory phenotype.

There are two main pathways for senescence induction, telomere-based, and cyclin-dependent kinase inhibitor (CDKN2A or P16). The telomere pathway induces senescence by telomere shortening (telomeres shorten each cell cycle; page 6) known as the Hayflick limit. The Hayflick limit is the induction of senescence due to replicative stress. The cyclin-dependent kinase inhibitor pathway for senescence can be triggered by a variety of stressors ^{47,48} and is a potent anticancer mechanism ^{49,50}. Senescence is involved in ageing and ageing related pathologies such as osteoarthritis ⁵¹, atherosclerosis ⁵², atherogenesis ^{53,54}, and prostatic hyperplasia ^{55,56}.

Senescent cells secrete cytokines and reactive oxygen species (Senescence-Associated Secretory Phenotype; SASP) which induces senescence in proximal unstressed cells ⁵⁷ the so called 'bystander effect'. It is possible, that cellular senescence induction in tissues will not be gradual but catastrophic as the more senescent cells in a tissue the more likely the otherwise healthy cells will turn senescent.

There are various cellular markers for senescence. One of the early methods to stain senescent cells was with the use of senescence associated β -gal. It has been observed that senescent cells hydrolyse β -gal producing a visible blue precipitate. Telomeres are unable to properly treat DNA damage (possibly due to Shelterin, a protein that maintains the telomeres and might block DNA damage response) making the telomeres of senescent cells rich in γ H2AX phosphorylation (a marker of DNA damage) ⁵⁸. Enlarged nuclei has been show to appeared in hepatocyte senescence ⁵⁹, this is due likely to the ability of hepatocyte to undergo polyploidy during development and age ⁶⁰.

DIETARY RESTRICTION PATHWAYS

During the 1930s, Clive Maine McCay theorized that slower growth rate leads to a longer lifespan. In his research, he decreased growth rate in rats by calorific restriction. The results seemed to agree with his hypothesis as lifespan increased from 33 to 44 months in male rats, and from 37 to 48 months in female rats. However, in one of his experiments, he controlled for calorific restriction by inducing a calorific restricted diet after puberty. DR-induction after puberty resulted in similar ageing benefits as DR prepuberty. Therefore DR was the main factor in life extension rather than growth retardation ⁶¹.

Since McCay's seminal study, DR has been shown to increase lifespan in a variety of organisms across biological kingdoms. A meta-analysis showed up to 43% lifespan increase in yeast ⁶² and a 25% increase in maximal lifespan in *C. elegans* ⁶³. Mice experiments revealed that calorific restriction can increase lifespan up to 53 months (a remarkable age for mice, world record for oldest mice being 60 months ⁶⁴) ⁶⁵.

Experiments on larger and longer-lived organisms are harder from a practical standpoint. An ageing study in yeast and worms might be conducted in a span of days with thousands of individual organisms. In contrast, studying larger animals such as primates might take decades with only a hundred individual organisms and a very high cost. A long-term experiment with rhesus monkeys (*Macaca mulatta*) was set up in the University of Wisconsin-Madison (76 monkeys) and the National Institute on Aging (121 monkeys) in 1987. Rhesus monkeys can live an average of 40 years in captivity but can live up to 26 years in the wild ⁶⁶. This is an ongoing study, which may end after 2030. So far, the Rhesus monkeys seem to benefit from

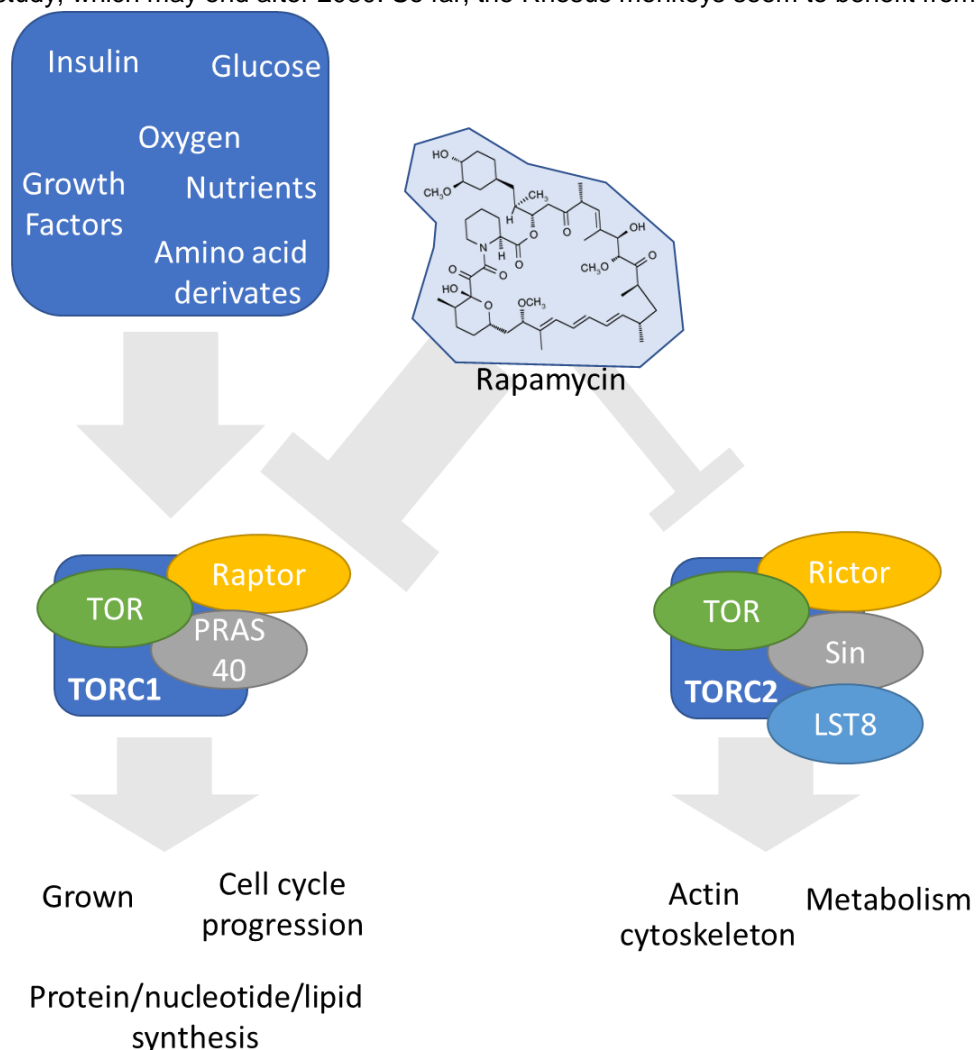


Figure 1-1. Simplified diagrammatic representation of the targets of Rapamycin.

DR when compared to control groups ⁶⁷. Recent studies have revealed that the spliceosome and therefore splicing is alternatively regulated during dietary restriction in Rhesus monkeys (not the same Rhesus monkeys from the Coleman experiments)⁶⁸. Calorie restriction studies in humans must be done on surrogate evidence for which none has been validated ⁶⁹. It would be nearly impossible and highly unethical to control the diet of a cohort of humans throughout their life and the project alone may exceed a century. However, a meta-analysis of human studies showed that many of the metabolic, hormonal, and pathological effects observed in smaller mammalian models were also seen in humans ⁷⁰.

The TOR pathway regulates cell metabolism and is therefore involved with caloric restriction. Rapamycin was originally discovered in the Polynesian island of Rapa Nui (Easter Island) as a new antifungal drug in 1970. Rapamycin was also found to have immunosuppressant activity in 1977 and was shortly thereafter used as a drug to prevent host-rejection in transplant patients. Between 2003 and 2012, it was shown that Rapamycin extends lifespan in yeast ⁷¹ (2003), mice ⁶⁴ (2009), fruit flies ⁷² (2010), and nematodes ⁷³ (2012). Rapamycin interacts with the TOR pathway (target of Rapamycin). Specifically, Rapamycin inhibits TORC1 (TOR complex) and with longer exposure TORC2 is also inhibited ⁷⁴. The downstream effects and regulators of TORC1 and TORC2 are different. TORC1 acts as a nutrient sensor and controls growth, cell cycle progression, and protein/nucleotide/lipid synthesis ⁷⁵. TORC2 acts as a regulator of the actin cytoskeleton ⁷⁶. Rapamycin inhibits TORC1, and it is thought that this inhibition leads to lifespan extension. TORC2 has a more nuanced interaction and is inhibited in a tissue dependant manner during prolonged exposure, leading to insulin insensitivity⁷⁴.

The insulin pathway is also strongly involved with ageing and caloric restriction. Glucose tolerance declines with age, largely due to increased insulin resistance ⁷⁷. Type II diabetes is an age-related disease (not to be confused with type I diabetes mellitus).

However, insulin has been shown to be a hormonal regulator of ageing. *C.elegans* worms can halt their development by entering dauer life stage when food availability is low. Within this stage, the worms can live up to 8 times longer than normal worms. Worms return to their normal life cycle when food availability is sufficient ⁷⁸. Mutants have been identified which have defects in dauer formation, or dauer like characteristics outside the dauer stage. The genes found to be affected have been characterized as homologues to genes in the mammalian insulin and insulin-like growth factor signal transduction cascade ^{79–81}. In addition, mice with a deleted insulin receptor in adipose tissue were resistant to obesity, insulin insensitivity, and were longer lived ⁸². In mice, overexpression of Klotho (a transmembrane enzyme

that recycles insulin and insulin growth factor-1(IGF-1)), extended life up to 31% longer than in controls ⁸³.

In humans, defects in insulin signalling results in diabetes or insulin resistance. However, there is some evidence in humans, such as Laron syndrome, that decreased insulin signalling may increase life span. Laron syndrome is a type of dwarfism caused by the lack of production of IGF-1. Those afflicted exhibit a decreased stature, resistance to type-2 diabetes, and have lower incidence of cancer ⁸⁴. Type-2 diabetes and most cancers are age related pathologies, however, Laron syndrome does not significantly alter life expectancy.

However, previous experiments have also suggested that insulin signalling is positively correlated with increased life span. Polymorphisms in insulin and insulin-like signalling genes are more prominent in Italian centenarians communities⁸⁵ and the Leiden 85-plus cohort (a study of octogenarians in Leiden, Netherlands)⁸⁶. The same Leiden 85-plus study found a correlation between insulin and insulin-like signalling with improved survival age ⁸⁶, and reduced cognitive decline ⁸⁷. In addition to this, Ashkenazi centenarians have higher serum levels of IGF-1 ⁸⁸. Single nucleotide polymorphisms(SNPs) in IGF1R (part of the insulin cascade) were associated with longevity in three Caucasian cohort studies⁸⁹. Variants in FOXO3A (part of the insulin cascade) were found to be correlated with longevity in four Caucasian cohort studies ⁸⁹⁻⁹¹, three Chinese cohort studies⁹², and a Japanese cohort study ⁹³. Most of these centenarian and octogenarian cohorts have also shown that socialization, activity, and diet are strongly related with longevity.

With few exceptions, research supports the insulin pathway as a regulator of growth and lifespan. However, the effect the insulin pathway has on human life span is not well understood as sometimes it does seem to improve aging prognosis (as in centenarian studies) and sometimes it does not (Laron's syndrome).

Ageing is the primary factor in a large variety of pathologies affecting every anatomical system from cancers to senility. It is perhaps the single most common disease. The cause of ageing is multifactorial, and any improvement in its prognosis will significantly decrease morbidity and suffering in most of the population. The primary objective of anti-ageing research is not to extend ageing caused morbidity by prolonging lifespan, but to extend healthy lifespan by decreasing ageing-related ailments.

RNA AND SPLICING BIOLOGY

DR experiments in Rhesus monkeys have shown that RNA splicing is affected by diet ⁹⁴ and mice experiments show that splicing dysregulation negatively affects ageing ⁹⁵, it is possible that splicing is responsible for the beneficial effects of dietary restriction. Therefore, it is important to study and understand the processing of mRNA.

TRANSCRIPTION

One of the most important bioprocesses of all living organisms is the synthesis of proteins from DNA, a process known as the central dogma (Figure 1-2). During protein synthesis, protein-coding regions of the DNA are transcribed into messenger RNA (mRNA) and then ribosomes translate the mRNA into protein. In some circumstances such as certain RNA viruses, DNA can be transcribed from RNA by a viral reverse transcriptase protein.

The basic premise of the central dogma holds true for all known forms of life (the notable exception being the RNA theory of the origin of life ⁹⁶ which states that early life used RNA and not proteins/DNA as both an enzymatic and genetic medium).

RNA SPLICING

While genes in prokaryotic organisms are single units of gene expression for a single gene, eukaryotic life developed a more complex system for protein synthesis using alternative splicing, where genes are split, with expressed coding regions (exon) and interrupted by in-expressed regions (intron). Splicing involves a selective extraction of pre-mRNA (introns) from a mRNA transcript consisting of ligated expressed regions (exons). The spliced mRNA is then translated into proteins. This process allows variation in proteins as a gene can be alternatively spliced enabling the production of functionally different proteins from the same gene. This is important as up to 90% of human genes are spliced ⁹⁷. Mutations in non-coding (intronic) regions

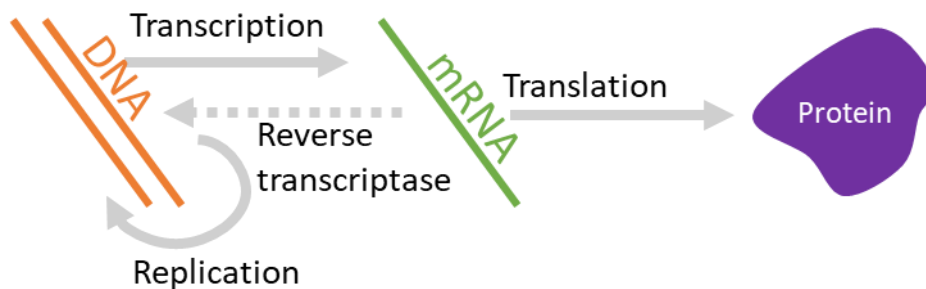


Figure 1-2. Canonical representation of the central dogma, showing replication, transcription and translation.

site is ligated to the splice donor site through transesterification making a lariat. The free exon in the spliced donor site then gets ligated to the splice acceptor site through transesterification. The intronic lariat remains in the nucleus and its nucleotides are recycled.

The remaining RNA is then capped on its 5' end with a guanine in a non-standard 5' to 5' triphosphate link. The 3' of the RNA is poly-adenylated. The RNA is now mature mRNA which is exported out of the nucleus to be translated.

Non-spliceosomal splicing occurs through self-splicing in some RNA ribozymes and the tRNA splicing pathways.

There are many types of splicing that can occur; Skipped exon (SE) involves an isoform that is just missing an exon, retained intron (RI) involves an isoform that includes expresses an intron, alternative 3' and 5' splice site (A3SS, A5SS) involves an alternative end or start of gene expression, multiple exclusive exons (MXE) involves isoforms that each includes an exon that the other doesn't. There are other reported forms such as MXE sets and non-adjacent sets which we will not discuss here as they are part of MXEs.

Resulting protein isoforms are functionally distinct, for example the *Drosophila* DSX gene encodes two isoforms, one using the exons 1, 2, 3, 5, and 6 which creates a transcription factor for male development, and another isoform containing exons 1, 2, 3, and 4 which encodes a transcription factor for female development⁹⁹. Splicing alterations by silent mutations can lead to pathological effects such as a mutation c.1824C > T in the LMNA gene which does not change the amino acid sequence but activates a cryptic splice site resulting in an isoform lacking exon 11. This LMNA isoform results in the Hutchinson–Gilford progeria syndrome, a rare but debilitating condition leading to severe rapid ageing and a lifespan rarely reaching above the age of fifteen¹⁰⁰.

The basis of the central dogma discovered by Watson, Crick, and Rosalind, explains how all proteins are encoded within genes. As RNA can be spliced, the variety of proteins encoded within the DNA greatly increases, because one gene can encode functionally distinct transcripts. These alternative transcripts can be metabolically, functionally, and pathologically different.

OBJECTIVES

Here we attempt to understand why and how dietary restriction increases lifespan and decreases morbidity. Using RNAseq to interrogate hepatic tissues of dietary restricted and control ageing mice, we can obtain precise snapshots of their transcriptome at many ages.

We will explore the splicing changes with both diet and age, as well as which metabolic pathways are altered and are likely to have downstream health effects.

REVIEW OF BIOINFORMATIC METHODS FOR RNA SEQ AND SPLICING ANALYSIS

Chapter objectives:

- To present an informative guide on modern bioinformatic practices from raw FASTQ file interpretation to biological significance.
- To provide a comprehensive overview of the limitations and caveats used in current bioinformatic techniques.
- To review and discuss up to date management tools for complex multistep bioinformatical analysis.

Next generation sequencing of RNA provides the researcher with very intimate information on how tissues (or single cells) behaved. However, there are many caveats that need to be understood in order to attain biologically significant results from sequencing data. A very basic oversimplified pipeline [sequencing → quality control → aligning → gene/exon counting → statistical analysis] is followed or varied upon, and the tools used at each step need to be carefully selected and used depending on the biological questions.

QUALITY CONTROL OF THE SEQUENCING OUTPUT

During sequencing, it is common that the read quality of the DNA sample is decreased during each cycle (each machine cycle reads one base at a time). This causes the 3' end of the DNA transcript to accumulate base pair errors and become lower quality than the 5' end. Left unchecked, the lower quality 3' end will impair mapping of the FASTA reads to the reference genome. When aligning for splice junctions (page **Error! Bookmark not defined.**), FASTA reads are split into two and aligned onto both sides of the junction. However, if the 3' quality is too low, the read will be unable to bind to the other end of the junction. Kraken¹⁰¹ is a multifunctional tool used to reduce the problems caused by read 3' mismatching. Kraken utilizes pair-end reads to improve mapping quality (Figure 1-4) and remove reads with too low quality. If this step is skipped, reference genome alignment is still feasible. However, fewer read matches will be found due to reduced read quality, and splice junction discovery will not be possible. Kraken can also be used for demultiplexing (using Kraken's reaper tool), tallying (using Kraken's Tally tool), and infer if the 3' adapter sequence used for sequencing is still in the data (using Kraken's minion tool).

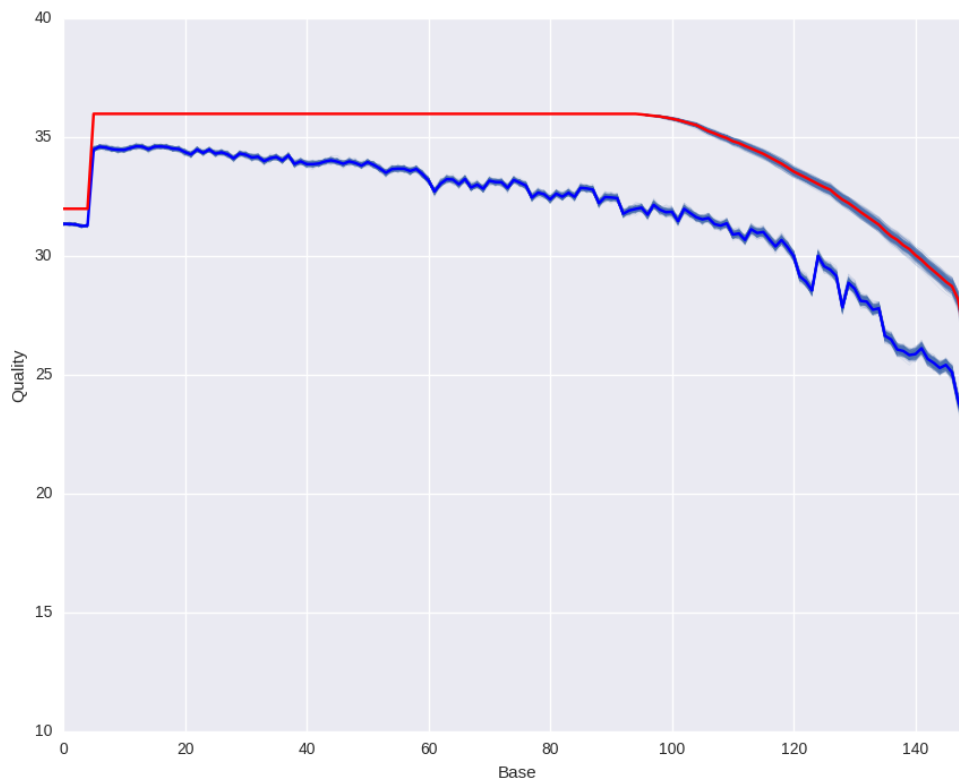


Figure 1-4. Line plot showing the read quality of a FASTQ before (blue) and after (red) Kraken quality control trimming. Sample quality was improved after Kraken analysis.

ALIGNING

To study sequencing data, it is necessary to know where FASTA/FASTQ reads originate from within the genome. Assuming an n sized genome (with bases randomly assigned), the probability of uniquely matching (binding to the correct genomic sequence and not to any other randomly identical region) a read of i length to the genome is $p = \left(1 - \frac{1}{4^i}\right)^{1+n-i}$. This means that in the human transcriptome ($n=2,000,000,000$ base pairs; according to hg18 RefSeq $n=2,011,862,672$ ¹⁰²) the probability of uniquely matching a 20-base pair length is only 99.82%. With each base pair added to the read length, the aforementioned probability approaches 100% exponentially. This is without taking into account the possibility of errors in the sequencing and expected mismatches during alignment. Unlike a perfectly randomized string of characters, the human genome has many repeated conserved domains. For this reason, a length of at least 50 base pair reads is recommended for alignment.

Earlier high throughput sequencing could only output reads up to 50 base pairs (bp) in length. Below the 50bp length it becomes less likely to uniquely map reads unto a reference genome (Figure 1-5). Unlike unspliced reads, spliced junction reads do not fit directly unto the genome, this is due to the read having parts from two separated loci.

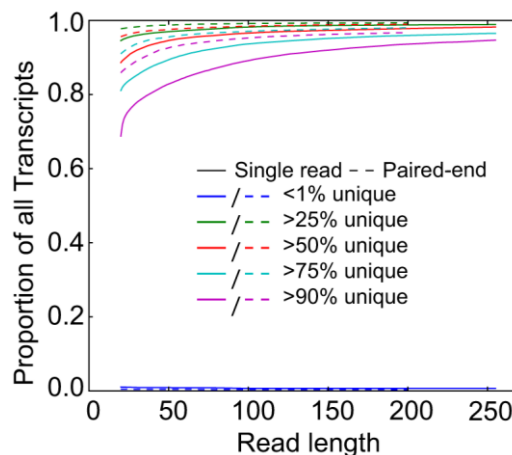


Figure 1-5. Proportion of unique binding sites by read length. Reads below 50 base pairs are harder to align to unique sites. ¹⁹².

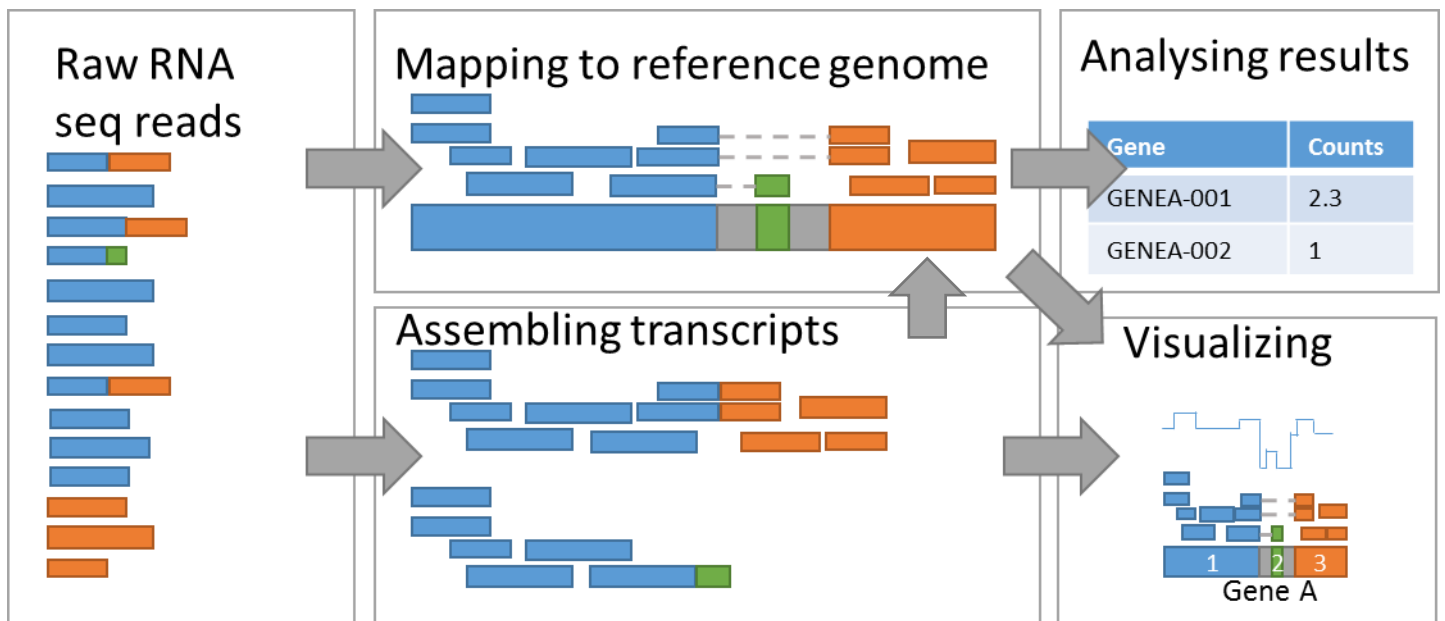


Figure 1-6. Basic pipeline summary of the steps required for general purpose splicing analysis. The colour represents exons. First reads are either aligned to a reference genome or assembled, this step must account for splicing as spliced reads do not map directly (there is no continuous blue green or blue orange reads in reference, and alternatively spliced reads may not be assembled with other assembled reads). The results can then be visualized and analysed.

As next-generation sequencing (NGS) is capable of outputting reads larger than 100bp, it is possible to align junction reads. The basic algorithm involves dividing the unmapped reads into smaller reads, where some will map to different junction sites (Figure 1-6). Because part of a spliced junction will map to a reference genome, the rest of the read can be mapped onto the other splice junction. Variations of this algorithm are used by TopHat ¹⁰³, MapSplice2 ¹⁰⁴, and STAR ¹⁰⁵.

TopHat ¹⁰³ is a command line tool from the Trapnel lab at Johns Hopkins University, available online through a web interface in usegalaxy.org (no need for command line interface usage). TopHat uses a multistep pipeline to report the isoforms present. TopHat first uses bowtie (a conventional aligner) to align the reads to the genome. Next, unmapped reads are set aside as initially unmappable reads, and the mapped reads are assembled into exons creating a list of exons present. Afterwards, TopHat creates possible splice sites using the exons found, and finally initially unmappable reads are aligned to constructed splice sites. Bowtie uses the Burrows-Wheeler transform to compress data (also used in zip2 compression) so even if bowtie loads the genome into memory (RAM), it still only needs a couple of gigabytes of memory to run. TopHat returns the results of mapped reads in a binary BAM (binary alignment map) format, a junction bed file with the junctions found, and a deletions bed/insertions BED (Browser Extensible Data) indicating deletions and insertions found. TopHat must load the reference genome, so depending on the genome size the minimum RAM needed to run TopHat changes (up to 4Gb). The amount of RAM required to run TopHat also depends on the number of reads given. This means that

Tophat could run on a desktop for small read counts, but it may need up to 30Gb in larger experiments (still possible on a high end desktop ¹⁰⁶).

Spliced Transcripts Alignment to a Reference (STAR) ¹⁰⁵ from Thomas Gingeras lab in Cold Spring Harbor NY uses a seed algorithm to map reads to the genome. Rather than attempting to insert entire reads onto the genome, STAR only uses a small portion of the reads to match. This seed is the maximal length required for uniquely binding onto the genome. Once a seed has been mapped to the genome, the rest of the read can be realigned with another seed and propagated to the rest of the genome. Therefore, aligning will easily find splice sites and junctions when different seeds from the same reads do not match to contiguous regions. This algorithm is very fast and does not require a lot of RAM (ca. 30Gb for aligning to the human genome) and runs orders of magnitude faster than any other current aligner¹⁰⁷ (as of 2018) enabling alignments to be performed in desktop computers.

An alternative method to finding splice sites involves simply mapping the RNAseq data to the transcriptome rather than a reference genome. This has the advantage of giving quicker alignments but will be unable to detect novel splice sites.

GENE COUNTING

Once the genome has been aligned and mapped, the mapped reads need to be counted to proceed with downstream analysis. One tool available for this step is htseq-counts ¹⁰⁸. This python module enables scriptable analyses to be conducted in python but also provides a standalone “HTSeq-count” for counting the reads using only an aligned BAM (or SAM (sequence alignment map)) file and a GTF (Gene Transfer Format) file. This will output an integer value for the number of reads located on each gene. It is important to note that HTSeq-count works better if the aligned BAM/SAM file is sorted by name rather than by position, to avoid buffer overflow and appropriating excessive memory.

Some analyses require specific counting algorithms. DEXSeq¹⁰⁹ requires the alignment files to be counted to an exonic resolution to analyse differential exon expression. DEXSeq comes bundled with its own counting tools to aid in this process.

QUANTIFYING DIFFERENTIAL GENE EXPRESSION

In order to compare experimental samples to which genes were significantly differentially expressed, we need to use the DESeq2 R package¹¹⁰.

Even though it might be tempting to simply run a statistical test to compare the gene counts in different experiments, this is not recommended for a variety of reasons. One problem is that the reads must be normalized. If a sequencing lane gave greater aligning depth than other lanes, genes sequenced in that lane will show much more upregulation. This is one example of a batch effect. Additional batch effects can be caused by the proportion of GC base pair content, or gene length. Due to high sensitivity, batch effects can arise simply by sequencing samples on different sequencing runs due to

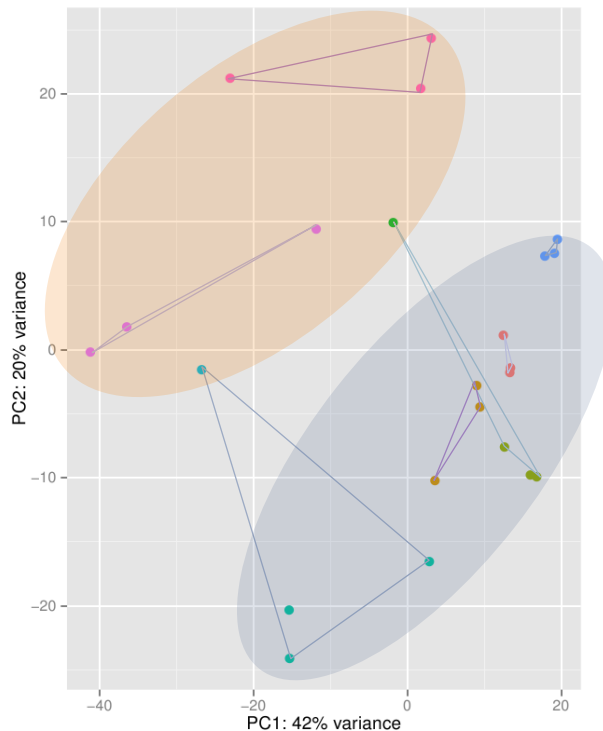


Figure 1-7. PCA plot showing an example of batch effects. Biological replicates have been connected by vertices. There have been two sequencing batches (highlighted by the ellipses). This plot shows a large effect caused by the sequencing batch alone, even separating biological replicates.

minute variations on the implementation of the same sequencing preparation protocol. This means that in order to study the read counts across separately sequenced samples, the batch effects must either be modelled or reduced as much as possible and the data must be normalized.

There are a handful of tools to rid samples of batch effects. DESeq will by default model batch effects in its design. To do so the batch information should be provided in the design model as shown in the code below:

```

coldata <- data.frame(
  experiment= c('3 months','3 months', '3 months',
               '6 months','6 months', '6 months'),
  batch = c('1','1','2','1','1','2'),
  row.names=colnames(countdata)
)
# Here the coldata dataframe holds the metadata of the experiment including
# the experiments names ('3 months', and '6 months'), as well as the batch
# ('1', and '2')
dds <- DESeqDataSetFromMatrix(countData = countdata,
                              colData = coldata,
                              design = ~batch+experiment)
# the design formula provides the variables to the model. In this case, it should
# try to rid of the batch effect

```

One caveat of DESeq is that it cannot compare different samples sequenced in

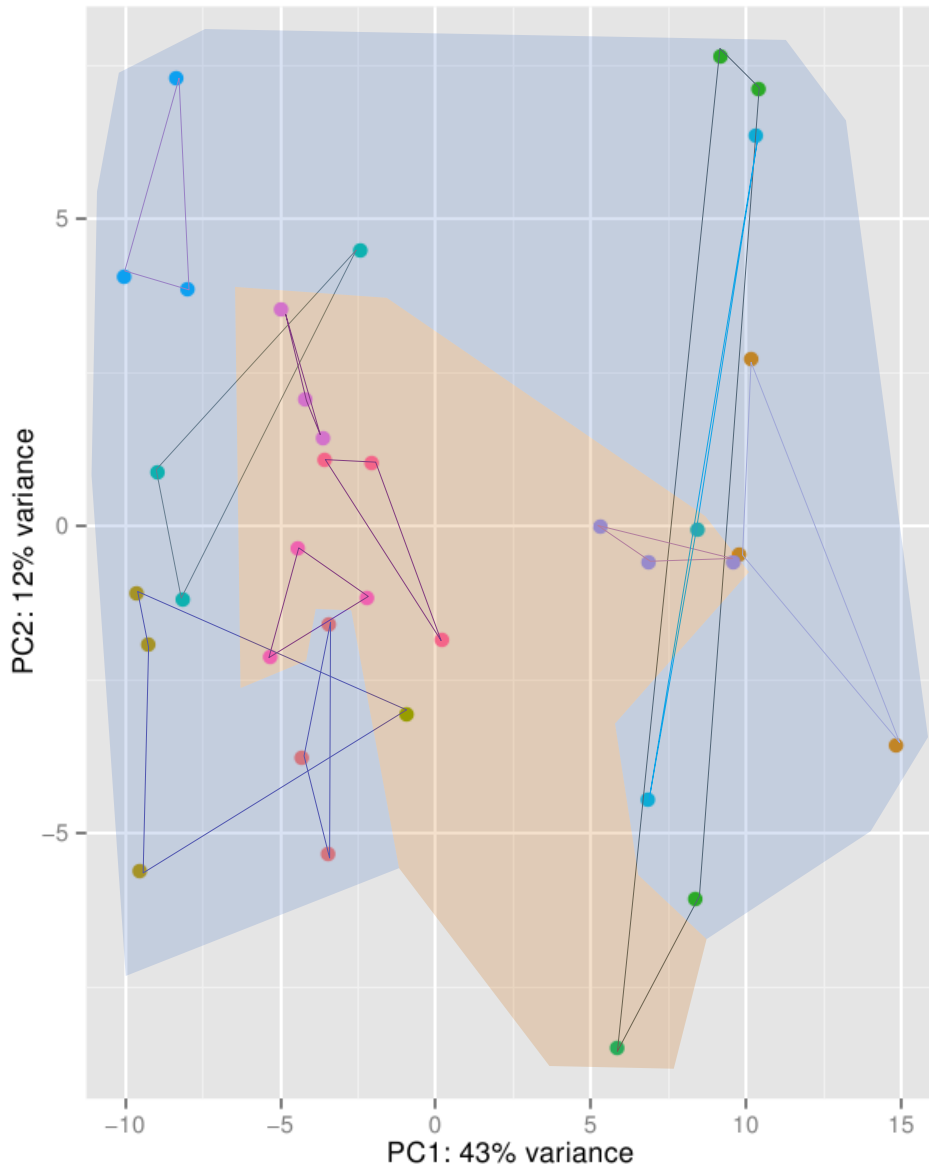


Figure 1-8. PCA plot showing an example of batch removal with Combat. Biological replicates have been connected by vertices. This figure shows two sequencing batches (highlighted by blue and brown shadow). The same data was used in Chapter 2. The batch effect was less prevalent than in Figure 1-7 and biological replicates related to themselves more closely than to the batches.

different batches (a sample must be present in each batch for comparison). Therefore, it is recommended in the experimental design not to sequence all the samples of one experiment together and then sequence another experimental sample in another batch. It is preferable to include a part of each sample in each batch as much as possible. If more samples are sequenced posteriori to the first batch, it is necessary to include technical replicates of the original experiment to model the batch effect.

If there is no overlap between the samples and batches, a more drastic approach is needed. Combat¹¹¹ uses empirical Bayes inference to try to remove batch effects (Figure 1-8). Even though this method is very robust, it removes genes with low variability between samples (regardless of significance) and can remove around 40% of genes with batch effects.

GENE NETWORK AND CLUSTERING

Bioinformatic analysis conducted for the comparison of two samples is relatively straightforward. Differential expression algorithms are utilized to study genes that have been unchanged, upregulated, or downregulated. However, for each additional experimental sample, the total number of comparisons needed grows exponentially (akin to the classic handshake problem). The possible number of expression patterns grows exponentially with each new sample ($possible\ patterns = 3^{\frac{n(n-1)}{2}}$). For 5 samples, the possible pattern space (59049 possible patterns) is more than twice the number of human genes. This means simple one-on-one comparisons is a strong method for very specific questions but a weak method for studying many experimental samples (such as; genes upregulated with an inducer but not altered when treated with repressors to said inducer). A more complex arrangement of expression patterns is also feasible (rather than up/down regulated and not significantly changed) making this estimate extremely conservative.

Gene network analysis can generate a network based on the gene expression correlation, without any experimental assumption. In doing so we can explore the expression topology of many biological samples¹¹². Gene network analysis enables investigations into gene co-expression. Visualization of the gene network facilitates observation of differential gene expression between samples. For downstream analysis, it is often useful to separate what appears to be a continuous cluster, but is actually comprised of separated clusters of co-expressed genes using MCL clustering¹¹³. These clusters can then be tested for enrichment of gene ontology, Reactome, or transcription factors.

NETWORK AND CLUSTER FISHING

Empirical phenotypic data from experimental samples can be inserted into gene expression datasets as a pseudogene. By performing network analysis on a dataset for co-expressed genes, genes that co-express with the empirical measurements will be isolated. It is also possible to investigate an inverted version of the empirical

pseudogene in question by inverting z-score¹ values (multiplying by -1) which will isolate genes that expressed in an opposite fashion to the empirical measurements.

When studying a multitude of factors, it may be preferable to use MCL clustering, to get discreet gene expression clusters as the empirical pseudogenes can co-cluster.

Once a list of genes is isolated with the same pattern of expression as the empirical measurement, downstream analyses such as Reactome enrichment can be performed.

PATHWAYS ANALYSIS

Reactome¹¹⁴ is a manually curated database of biological pathways and associated genes. Using gene enrichment tools such as ReactomePA¹¹⁵(an R library to enrich gene sets) it is possible to infer biological pathways that are overrepresented in a gene list of interest (by means of a Fisher's exact test). Annotations in Reactome are interconnected and range from highly specific pathways containing a few genes (e.g.: ANG-binds F-actin containing five genes) to their parent's categories (eg: cell-cell communication containing 130 genes). Due to the parental relationships between Reactome annotations, when one Reactome annotation is enriched, it is common that less specific Reactome parent annotations are enriched as well. Therefore, it is useful to filter Reactome enrichment results to remove redundancy.

A set of rules to remove Reactome redundancy are:

1. Remove annotations whose constituent genes are present in a more encompassing Reactome with more genes.
2. If two or more Reactome annotations have the same constituent genes, keep the annotation with the lowest P value and remove the rest.

The results can now be plotted externally by any plotting software such as matplotlib.

Alternatives to Reactome are Gene Ontology ¹¹⁶ (GO) and Kyoto Encyclopaedia of Genes and Genomes ¹¹⁷ (KEGG). KEGG offers similar tools but with independently curated annotations. GO uses a hierarchical classification system. KEGG, on the other hand, was once popular but is based on a subscription service while its free tools are no longer updated.

¹ Z-score: a measurement related to the number of standard deviations (above or below) a sample is away from the mean.

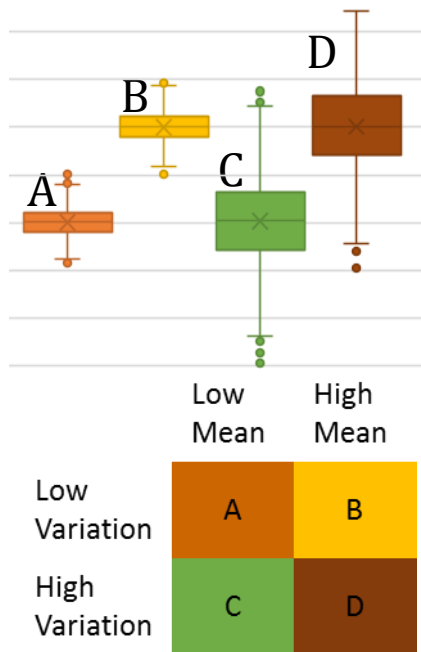


Figure 1-9. Diagrammatic representation of sample distributions. Each box plot represents a different type of distribution with low means (A, C), high means (B, D), low variation (A, B), and high variation (C, D). If comparing a distribution of A with B there would be significant difference, but if comparing A with A, C, or D there would be no significant difference.

HETEROGENEITY ANALYSIS

Differential expression analysis, like any statistical test, relies on comparing the dispersions of two groups. A strong P-value (below 0.05) means that there is likely to be a significant difference between the samples within a certain probability. A weak P-value does not mean that the sample distributions are the same or have the same mean (**Error! Reference source not found.**). For example, if one of the samples has high heterogeneity, the resulting P-value will be weak. This is important to note that two samples with the same distribution (two identical distributions) or one or more samples with high heterogeneity (dysregulation) are very different biological results which might have equally weak P-values.

Variances naturally follow Fisher-Snedecor distributions (or simply f-distribution) rather than Gaussian distribution. In order to compare the distribution of variations, either a non-parametrical statistical test is needed, or the data can be converted into a normal statistical distribution.

A Box-Cox transformation (Figure 2-8) can be used to transform Non-normal data into a normal distribution. The transformation of non-parametric data into parametric data facilitates downstream analysis such as simple t-tests for sample comparison.

It is also possible to analyse non-normal data using non-parametric statistical tests such as Mood's or Kruskal-Wallis tests. Mood's test is a variant of a Pearson's chi-squared test and compares medians rather than means. Kruskal-Wallis test, on the other hand, is a variance test that compares the distribution between sample sets.

Another simple option is to study the coefficient of variance difference of the same genes across experimental samples. The mean of this distribution should refer clearly to the difference in variation. If the difference is very low (close to 0 once the distribution has been log normalized), it is possible to assume non-significant heterogeneity differences between the distributions.

It is important to note that a considerable number of statistical tests can be used to investigate a data set. As different statistical tests can return different P-values, it is possible to perform various statistical tests until a desired result is achieved. This practice is known as P-hacking, and involves running an array of statistical tests, until the desired P-value is achieved, and justifying the usage of the test. P-hacking enables a researcher to justify previous assumptions regardless of the actual underlying scientific facts.

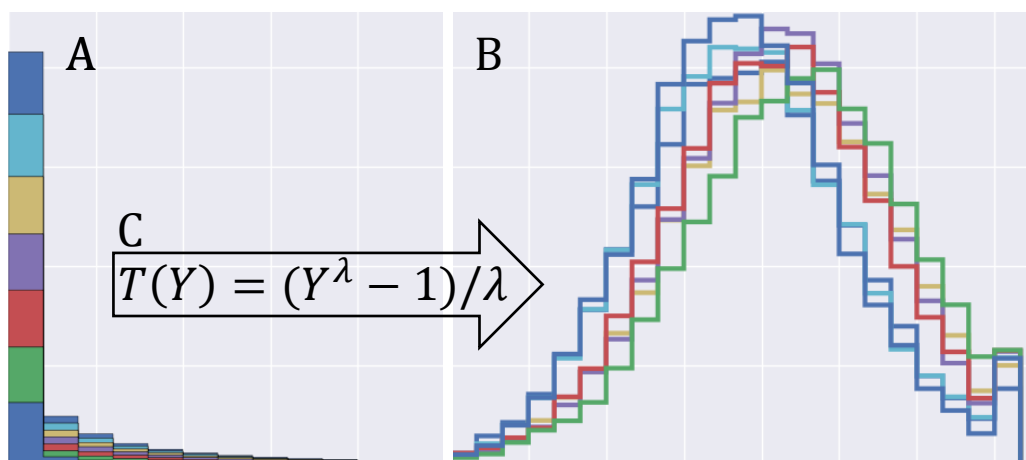


Figure 1-10. (A) a series of different non-parametric f-distributions. (B) the same distributions as A after being boxcox normalized, now fitting a parametric normal curve. (C) Box-Cox transformation where Y is the response variable and the λ is the power parameter.

DETECTION OF ALTERNATIVE SPLICING

There are three different methodologies for quantifying differential splicing. One simple option is to assemble the transcriptome before mapping and then quantify the amount of each transcript found. There is a variety of assemblers each with their subtle differences, however, they will not be discussed in depth here: Cufflinks ¹¹⁸, Trinity ¹¹⁹, Trans-ABYSS ¹²⁰...

Another possibility is to run differential expression analyses at exonic resolution rather than at gene level. DEXSeq ¹⁰⁹ is an implementation of the DESeq ¹¹⁰ algorithm that focuses on exonic level resolution rather than gene level resolution. It is a slight oversimplification to state that DEXSeq just runs DESeq using exon resolution as it must normalize each exonic read, otherwise, a significantly upregulated (or downregulated) gene will show as all its exons being significantly included (or excluded). DEXSeq is not the only tool to quantify alternative exon expression, SplicingCompass ¹²¹ which can detect differential exon expression as well. CuffDiff ¹²² is part of the tuxedo suite that includes Tophat, and Cowtie. CuffDiff cannot find novel splice sites or exons, however, CuffDiff outputs files of isoforms, genes, coding sequences, and primary transcript in FPKM (Fragment per Kilobase per Million fragments Mapped), raw counts, and differential expression tests formats by analysing BAM/SAM mapped files and a GTF/GFF file. MISO ¹²³ can provide a similar analysis as CuffDiff with similar limitations.

A third possibility is to quantify the amount of each type of splice event (skipped exon, alternative 5'/3' splice site, mutually exclusive exon, and retained intron). In order to quantify distinct splice events, a more complex set of statistical tools are needed. rMATS uses a Bayesian hierarchical model to find and annotate splice events. Unlike DEXSeq or assembly techniques it provides information on each type of splice event. rMATS requires all the reads to be trimmed to the same length (regardless of FASTA file or BAM file input). Ideally, the largest possible lengths are preferred for trimming. However, this leads to shorter reads being lost. In order to maintain as much data as possible, the quantity of information should be obtained for various trim lengths. The option with the most data retained should then be used for further analyses.

Pseudocode for finding the best trimming length:

```
listOfLengths=[0,0,0,0,0...,0,0,0,0]
for line in bamFile:
    #add one to the appropriate lengths
    listOfLengths[length of sequence of line] += 1
for i in [1,2,3,4,5...largest length]:
    print i,i*Σ(listOfLengths[i:])
#This should output something like:
#1 100
#2 200
#3 300
#...
#80 5000
```

```

#81 6100
#82 6143 <- linear growth until the longest reads are being trimmed
#83 6198
#...
#90 9180
#91 9207
#92 9170 <- Largest amount of information preserved with trimming length 92
#93 9120 <- the amount of information is decreasing now
#...

```

Python code:

```

#! /usr/bin/env python
import fileinput
    # module to handle STDIN
    # this enables us to use samtools piping to read straight from a bam file
MAXLEN = 500 #use an arbitrarily large number as the read size limit
lengths = [0]*MAXLEN #unlikely to have reads longer than MAXLEN )
for line in fileinput.input(): #reads stdin one line at a time
    lengths[len(line.split()[9])+1
            #splits the sam string,
            #gets the 9th item (sequence) measures its length,
            #adds one to the item in the list with the index of the sequence length

for i in range(MAXLEN): # iterates through all the lengths
    print (i,i*sum(lengths[i:]))
    #prints the length
    #prints the amount of sequences larger and equal to the
    #iterated length times the iterated length.

```

Should run like:

```
samtools view [bam files] | python optimalTrimm.py
```

Using this extremely short program (only 7 lines without comments), we can choose the best trimming length to minimize information loss. Another similar tool is SpliceR¹²⁴. SpliceR is an R bioconductor tool to analyze. It is designed to work with specifically with cufflinks. It can find and categorize alternative splicing events (single exon skipping exclusion/inclusion, multiple exon exclusion/inclusion, intron retention/inclusion, alternative 3'/5' splice sites, alternative transcription start/end site, mutually exclusive exons). It is also capable of discovering de novo splice sites as it does not rely on annotation files to find exons. SpliceR can then generate annotation GTF files which can be used as input for many other analysis pipelines and visualized into genome browsers.

PROTOCOL MANAGEMENT

Interpreting biologically relevant events from a FASTQ file outputted by the sequencing process requires a fair number of steps and various software programs. These programs and tools are modular allowing the researcher to choose the latest or best-suited sequencing mapping tool or quality control tools. One large inconvenience involves the tenuous work of managing all the files and steps. There is no need for a tool to manage protocols if all the steps and commands are annotated and carefully organized.

Makefile ¹²⁵ is a tool originally designed to automate the building of executable programs and compile their libraries. Makefile is quite possibly one of the oldest programming languages currently in widespread use (published in 1976) apart from C (1972; Makefile was built mainly for compiling C projects). Make relies on simple rules with a target and dependencies:

```
target: dependencies
    command to create target
```

To executes Makefile in order to generate the target:

```
make target
```

Makefile will check for dependencies to see if they exist (if dependencies do not exist then Makefile will try to create them if there are rules to create them in the Makefile) or if the dependencies have been updated, Makefile will run the command to create the target. This is useful for the management of complex protocols and custom commands:

```
1 counts = counts/Pro_NGSs.1.count counts/Pro_NGSs.2.count counts/Pro_NGSs.3.count
2
3 counts/%.count: aligned/%.Psorted.bam #the % sign acts as a wildcard between the target and the dependency
4     mkdir -p counts ; \                # makes the folder for the counts to be saved
5     echo counting $* ; \                # $* stands for the %
6     python -m HTSeq.scripts.count \    # This runs the counting program
7         -f bam \
8         aligned/$*.Psorted.bam \
9         genome/Homo_sapiens.GRCh38.84.gtf > counts/$*.count && \ # this create the target
10     chmod 555 counts/$*.count && \ # changes priorities of the target in order for it not to be easily
deleted
11
12 counts/merged.csv: $(counts)
13     python -c 'import glob,re,pandas; \ # python command
14         files = glob.glob("counts/*.count"); \
15         df=pandas.concat([pandas.Series.from_csv(x,sep="\t") for x in files],axis=1);\
16         # this reads the files and concats them in a dataframe
17         df.columns = [re.findall("./(.)\.",x)[0] for x in files]; \ # this renames the header
18         df.to_csv("counts/merged.csv")' # this saves the new dataframe
```

In this Makefile snippet, the 'make counts/merged.csv' command will try to create a file named 'counts/merged.csv' (line 12) but before running a python command to merge all the counts files, it will first check to see if the files in the variable \$(counts) exist (line 1). If the files in the \$(counts) variable do not exist, it will run the command to create them (by first checking if there is a dependency BAM file) using HTSeq-count (line 3). This snippet is not part of the entire Makefile (it's a simplified example) but the entire Makefile can take raw FASTQ files, run quality control, index the genome for alignment, align the samples and then proceed all the way to standard analyses such as differential expression or splicing analysis.

Some major advantages of using a Makefile for the management of computational protocols is that a script is available for all analysis performed, including all the parameters used allowing for replicability and troubleshooting. Another advantage is that when the input files and the Makefile are not deleted, every file in the analysis

can be recovered allowing peace of mind in case of an accident. This also significantly diminishes the number of files that need to be archived (by default Makefile will delete intermediary files once the script has finished). The main disadvantage of Makefile involves the learning of a reasonably complex computational language, and with little experience troubleshooting can be a daunting process. The reason Makefile can seem an intentionally esoteric language is because during its development the computational zeitgeist was different and included the need for short single letters commands rather than humanly understandable commands (due to restrictive computational memory). Learning to use Makefile can be harder than most computer languages as it is designed and implemented for the use of large complex and highly technical computational projects. Scant 'non-computational expert' help for Makefile troubleshooting is available (unlike most computational languages where help is available and plentiful for all skill levels).

There are many modern alternatives to Makefile such as `snakemake`¹²⁶ (which offers more pythonic syntax), `cmake`¹²⁷, and `SCons`¹²⁸ (which also offers a more pythonic syntax). Although all Makefile variations offer similar functions, the modern alternatives provide improved grammar and more advanced features.

CONCLUSION

Sequencing data is highly sensitive to a variety of issues, from batch effects to low-quality reads. However, sequencing data can also provide a wealth of information when properly curated and can be used in a menagerie of bioinformatical analysis.

Careful choice needs to be made at each analysis step to prevent the introduction of errors into the pipeline. It is important to take into account that splicing studies require greater quality control than general purpose differential expression.

As there are multiple steps to analyse data and each step has a variety of tools with their own strength and weaknesses, the specific pipeline used must be carefully chosen depending on the experimental question. A standard analysis pipeline does not exist. However, the basic outline of [sequencing → quality control → aligning → gene/exon counting → statistical analysis] holds true for most biological questions.

CHAPTER 2 AGE AND CALORIFIC RESTRICTION EFFECT ON TRANSCRIPTOME

Chapter objectives:

- To investigate the biological pathways altered during dietary restriction and how this may affect ageing in mice.
- Bioinformatic analysis of the progression of ageing in mice under normal diet and short-term or long-term form of dietary restriction.
- To investigate RNA splicing involvement in ageing and DR.

Chapter Hypothesis:

- DR seems to slow down the onset of ageing phenotype, therefore we would expect to see DR affecting senescence pathways or counteract to a certain degree the effect ageing has on the transcriptome.
- With use of RNAseq technology and DR mice time series (with a normal diet control) we should be able to probe and see the mechanisms of increased longevity through DR

INTRODUCTION AND BACKGROUND

Since the 18th century, mice have been a popular mammalian model organism. We (humans) share almost 99% of our genes with mice and due to the short lifespan of mice, they enable us to study mammalian biology which is relevant to human biology. Mice ageing aging is not directly correlated to humans.

Mice age at different rates depending of the life stage (Figure 2-1). Important to note that the Figure 2-1 ignores differences in gender or breed but is rather only a summary ¹²⁹. As with humans and animals in general, mice tissues accumulate DNA damage as seen by γ -H2AX loci staining and become senescent ^{130,131}.

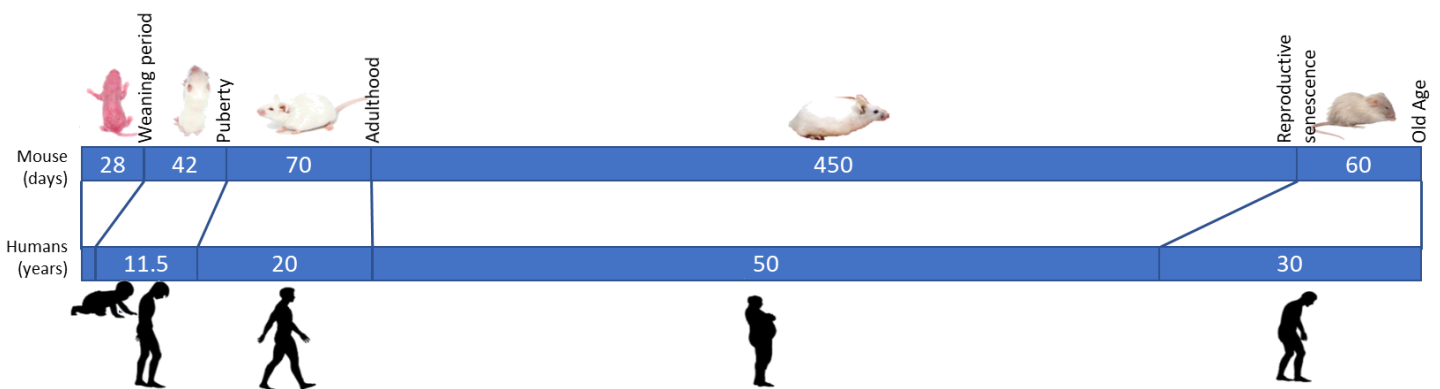


Figure 2-1. Diagrammatic representation of the ageing rate differences between mice and humans. Adapted from ¹²⁹.

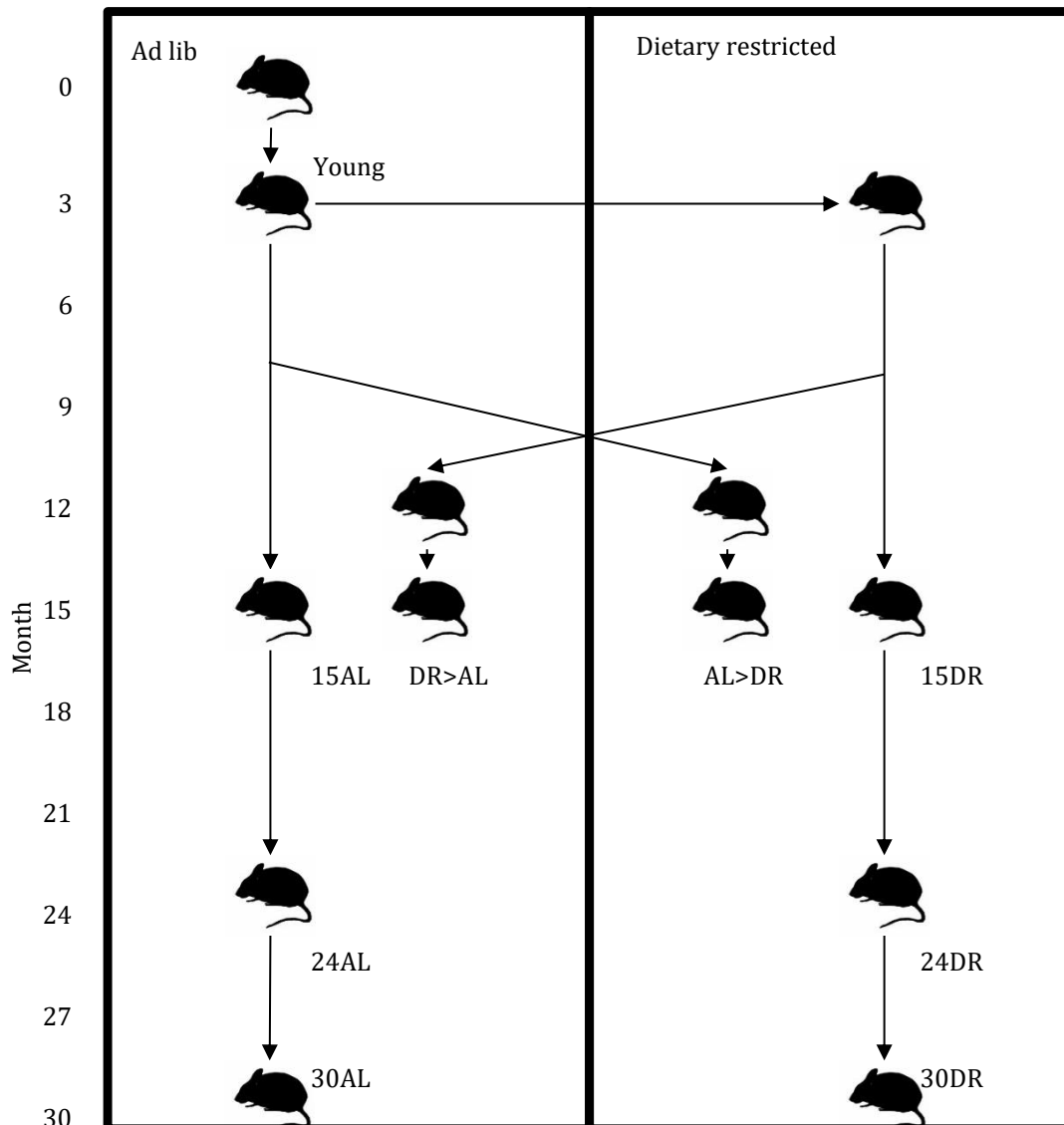


Figure 2-2. Diagrammatic representation of mice growth and sampling of our main study. Each mice diagram represents a point where three mice were sacrificed for sequencing.

The Institute for Ageing and Health at Newcastle University, UK conducted a mouse ageing cohort study to investigate the effect of diet on mouse health ¹³² (This study was conducted and completed before the PhD projects commencement, and the author of this thesis did not have any input or interaction with the original mice study). In this study (Figure 2-2) a cohort of C57/BL6 mice were fed a normal chow diet (Ad libitum; AL) until they reached three months of age. Then, a subgroup of these mice was placed on a dietary restricted (DR) regime¹³². Finally, three mice from both AL and DR cohorts were sacrificed at 3 (only AL) 15, 24, and 30 months of age. At 12 months, a subgroup of mice from AL and DR diets were placed on the opposite diet and sacrificed at 15 months to study the effects of short sudden AL and DR dietary changes (AL>DR and DR>AL). Mice were chosen for this study as they are a well-known and study mammalian subject to study with a short lifespan (compared with the long lived rhesus monkeys whose dietary experiments started in the late 80s and are still ongoing ⁶⁷).

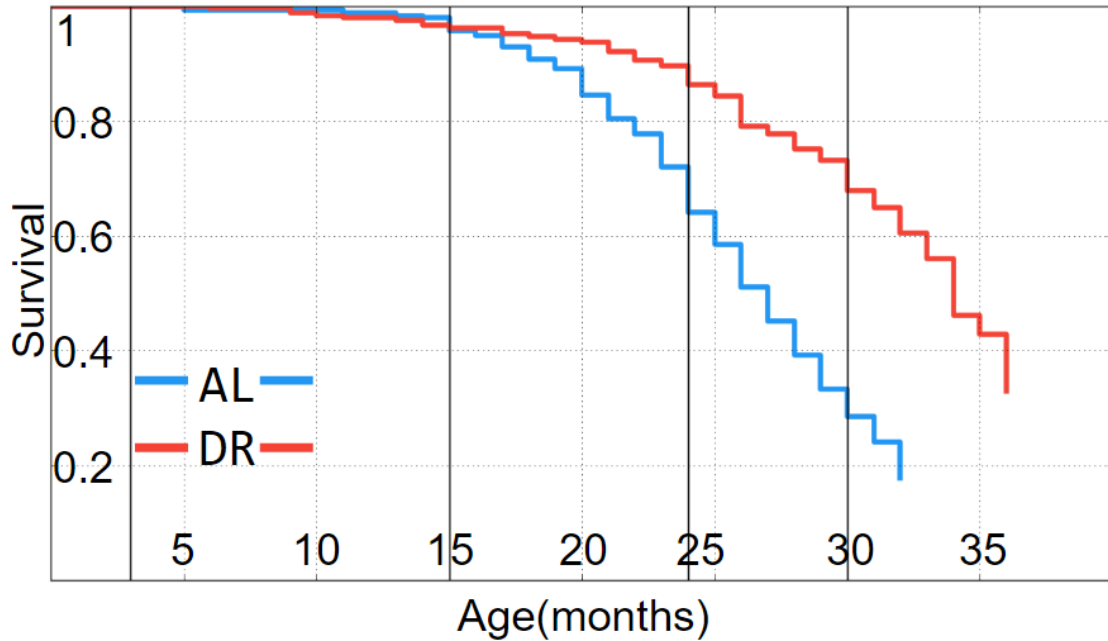


Figure 2-3 Kaplan Meier survival curve of ageing mice ¹³². Showing survival rates of mice under different dietary regimes. Survival rates decline between 15 to 24 months and remain low past 24 months.

Mice were fed standard rodent pellet chow (CRM (P) from Special Diets Services, Witham, UK). DR mice were fed 60% of the amount AL mice eat on a standard diet. Mice were housed in same-sex cages in groups of 4 to 6 mice. Each mouse was individually tagged by earmarks and maintained with an ambient temperature of $20^{\circ} \pm 2^{\circ}$ C. Mice hepatic tissues were collected from mice of all groups at 3, 15, 24, and 30 months of age.

Hepatic tissues were chosen for this study as they are key regulators of nutrient intake, storage, and usage. All nutrient intake from the gut goes through the hepatic portal vein into the liver for processing by the hepatic tissues. The liver also has endocrine by regulating insulin growth factors which affects aging.

AL mice had food available throughout the day, while DR mice were fed once per day circa 9:00 AM (Mice were not fed for six hours prior to being sacrificed. This was to facilitate dissection of the mice, as their digestive tracts were empty). Therefore, differences in mouse metabolism between AL and DR could be due to a circadian rhythm affect, and not diet. To control for transcriptomic changes induced by the 6 hours fast, a subgroup of mice from both diets at 15 months of age were sacrificed after feeding (post fed mice; PF).

Both AL and DR mice were healthy and had low mortality rates until 15 months ¹³² (Figure 2-3). From 15 to 24 months of age, both AL and DR mice had an increased mortality rate. The increase in mortality rate remained constant between 24 to 30 months. DR mice, however, had a lower mortality rate compared to AL ¹³².

AL mice had a heavier body mass than DR mice, but there was an interesting trend regarding the crossover mice (Figure 2-4 A). AL>DR mice quickly lost weight once switched to a DR diet at 12 months but when returned to an AL diet at 15 months, they did not reach the same weight as their AL diet peers in the same period of time under AL diet. DR>AL mice increased in weight during AL and, unlike AL>DR, reached and retained the same weight as their DR peers when returned to a DR diet.

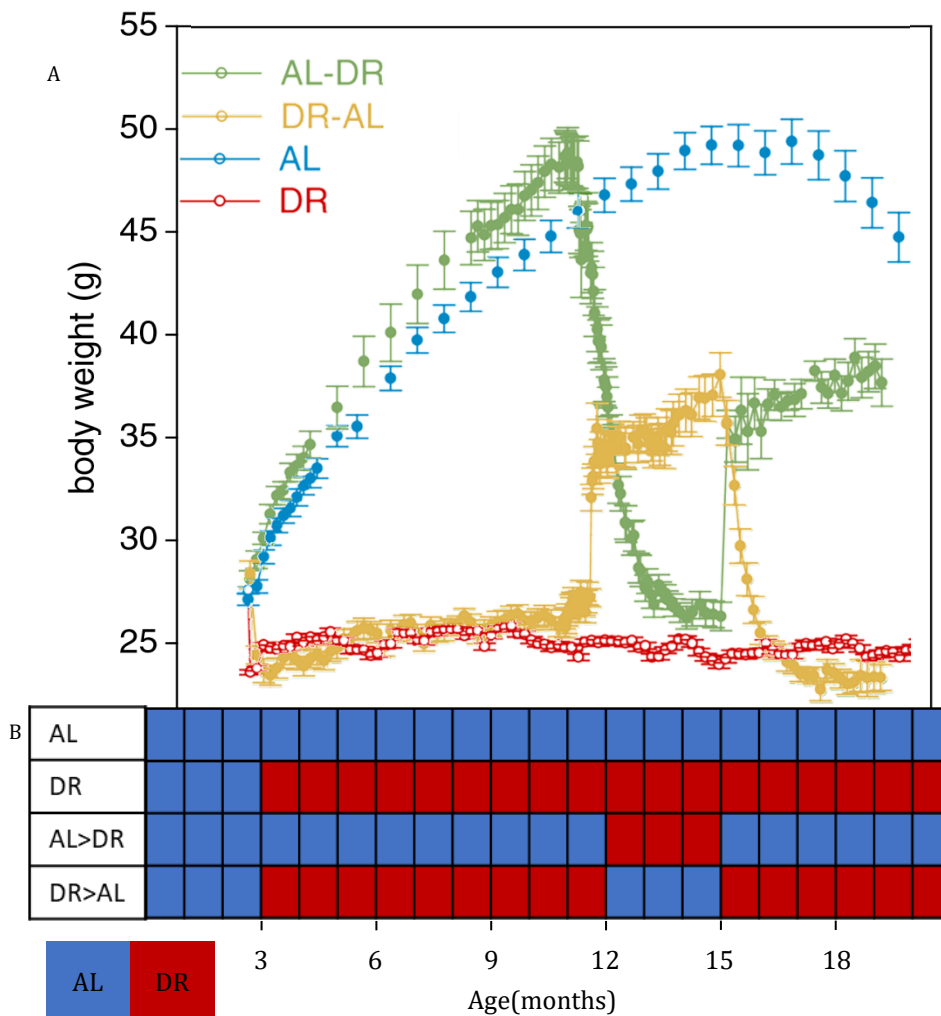


Figure 2-4.(A) Mouse weight with age, Dotted lines represent sequencing time points. Adapted from ¹³². (B) Mice diet schedule, red being DR and blue being AL.

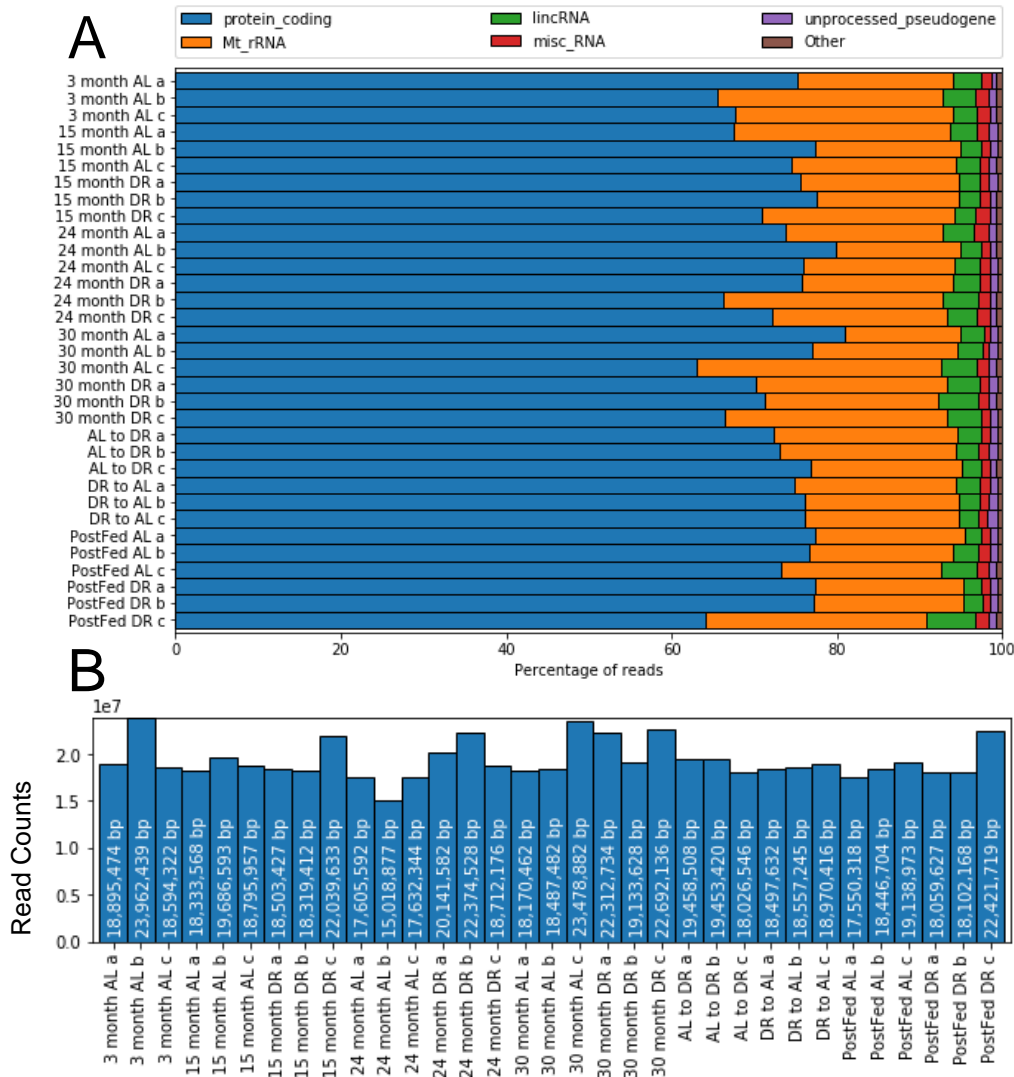


Figure 2-5. (A) Gene biotypes of different samples showing most protein coding genes as well as a large number of mitochondrial genes. Protein_coding genes transcripts are translated into protein, Mt-rRNA are shuttled into the mitochondria to form mitochondrial ribosomes, lincRNA are long (longer than 200 bases) intergenic non coding RNA, misc_RNA are uncategorized non coding transcripts, unprocessed_pseudogene are pseudogenes that are not translated, and others are an amalgamation of all other categories that accounted for less than 1% of transcripts. (B) Read counts of different samples.

Strand-specific paired-end libraries for RNA-seq were generated from DNase treated total RNA using NuGene (Y, 15AL, 15DR, 24AL, 24DR, 30AL, 30DR) or ScriptSeq (crossovers, postfed, one 24AL repeat and one 24 DR repeat), then run with an Illumina 2500 sequencer to obtain 100 base pair-end reads.

Up until this point all the work was performed prior of me joining the lab. My research focused on studying the FASTQ files obtained from the sequencer.

The sequenced FASTQ files were filtered to remove low-quality reads by Kraken¹⁰¹. Reads were aligned using Tophat2¹³³ (GRCm38 reference genome) without multi-mapping or novel junctions, returning a high percentage of protein-coding RNA and mtRNA (Figure 2-5). Sequence alignment files were sorted by name¹³⁴ for HTSeq-count¹⁰⁸ to read gene counts. Protein coding counts were normalized and compared sample by sample using DESeq2¹³⁵.

The expression datasets were averaged, filtered to remove low variability (variability less than 10% of the mean ;from 43628 genes down to 10920 genes), then used to generate an expression network with Biolayout express¹¹² using a minimum Pearson correlation of 0.7 and clustering coefficient of 0.95. Clusters were extracted from the network using Markov cluster algorithm¹¹³ with an inflation coefficient of 2.2 and pre-inflation coefficient of 3, clusters with less than 30 components were removed (leaving only 26 clusters containing a total of 5419 genes). The clusters were then analysed for Reactome annotations overrepresentation using ReactomePA¹³⁶. As the network was ring-shaped, the clusters were manually renamed to reflect their position on the ring rather than the cluster size (default sorting). As expression values were averaged before clustering, clusters where intra-sample variation was greater than 0.1 means were filtered out. Reactome annotations were filtered by p adjusted value of less of 0.05 and redundancy was removed by filtering Reactomes where all the composing genes were contained in another significantly enriched Reactome annotation. In case of two Reactome annotations with the same composing genes, the one with the lowest p adjusted value was kept.

In order to isolate genes with similar expression patterns as senescence markers, empirical karyomegaly and telomere-associated foci measurements were z-normalized and inserted into the gene expression data. The z-normalized measurements were inverted and inserted in the expression dataset as well, to evaluate genes that showed anti-clustering with senescence markers. Expression values with senescence markers in it were clustered using a 0.9 correlation coefficient and a minimum Pearson correlation of 0.7 using BiolayoutExpress. Genes that were positively corelated with senescence markers were extracted. As both karyomegaly linked genes and telomere-associated foci linked genes were similar and had the same enrichments (same for the negatively correlated clusters), these lists were intersected to study the strongest senescence and anti-senescence genes.

Splicing analysis was conducted with rMATS¹³⁷. rMATS is a statistical tool to find quantify and categorize alternative splicing events.

Locus analysis was conducted by counting the number of significantly up/downregulated genes within specific loci (determined using UCSC gap tables).

Loci then were filtered by loci which contained at least 10 genes of which at least two thirds of those genes were significantly upregulated for upregulated loci or two thirds of those genes were downregulated for downregulated loci.

RESULTS

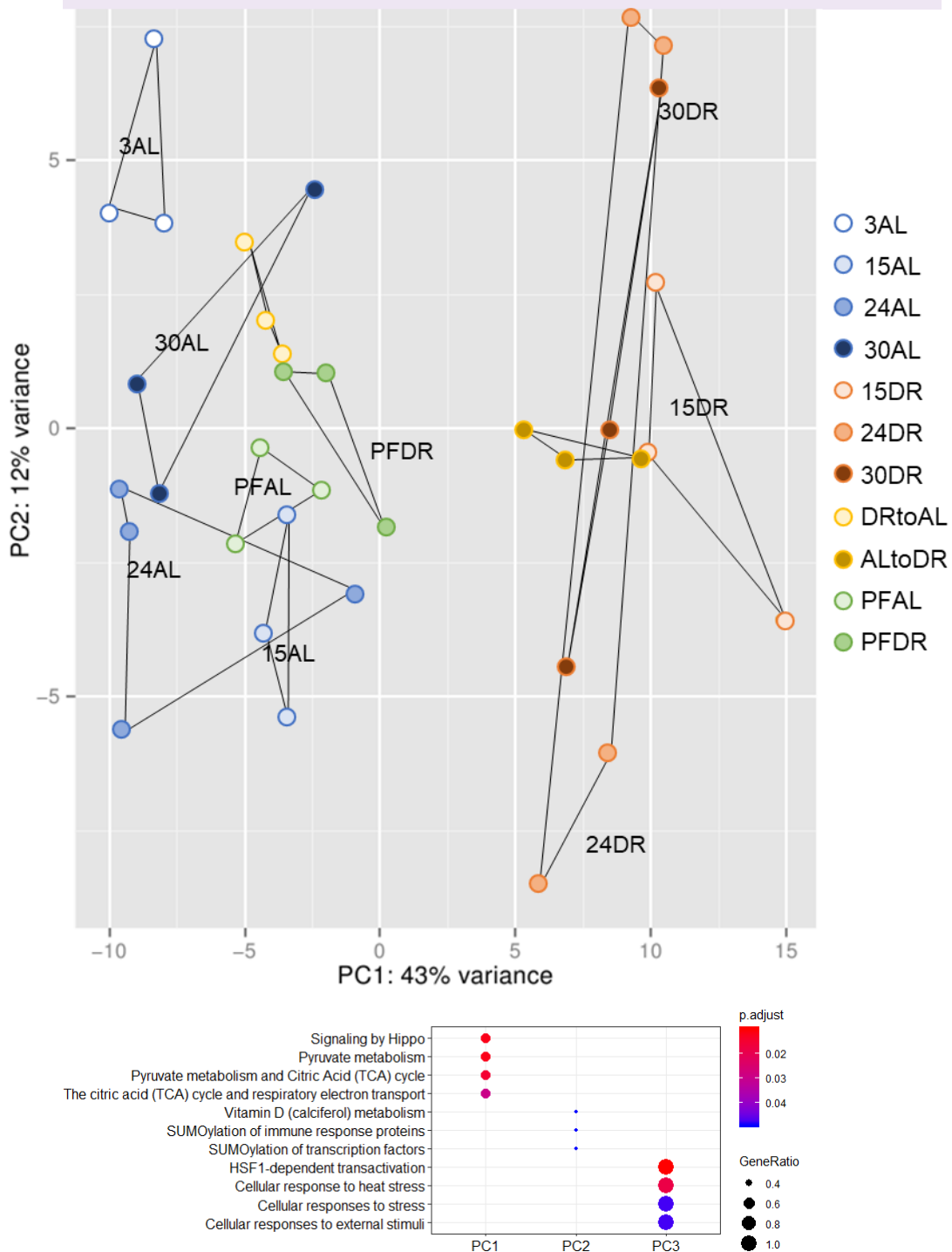


Figure 2-6. (A) PCA plot showing mice samples after batch removal with Combat. Biological replicates have been connected by vertices. (B) Reactome overrepresentation analysis of the top most influential genes for principal components one, two, and three (PC1, PC2, & PC3).

Principal component analysis (PCA) analysis (Figure 2-6, A), performed using the DESeq2 module (see page **Error! Bookmark not defined.**), showed a clear separation between 15 to 24 months of age and a smaller separation between 24 to 30 months of age. Diet, on the other hand, had a large separation within the PC1

representing 43% of the variance. Interestingly there was a large intrasample PC2 variance within 24 and 30 months DR.

Differential expression analysis (Figure 2-7;file [1]) was in agreement with the PCA analysis (Figure 2-6) and network analysis (Figure 2-14), that the largest changes occur between 15 and 24 months (Figure 2-7 II & VII) regardless of diet (Figure 2-7 X). In contrast, there was a small change between 3AL and 15AL (Figure 2-7 I) and very little change between 24 and 30 months (Figure 2-7 III, IX & XI). However, diet still had a strong effect (Figure 2-7 IV, V & VI) which was consistent regardless age (Figure 2-7 VII).

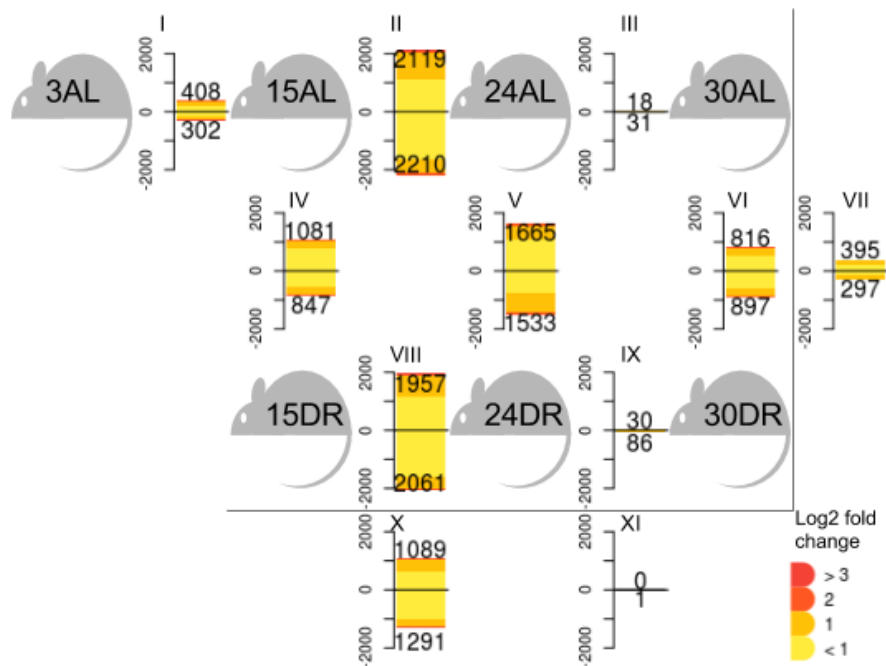


Figure 2-7. Diagram showing the up/downregulated genes between ageing comparisons (left to right) and diet comparisons (up/down). Each bar plot represents the number of genes upregulated (upper bar) and downregulated (lower bar) between the treatment groups (shown as mice) it is placed between (e.g.: plot I is placed between 3AL and 15 AL and represents the significantly changed genes between 3AL and 15AL). Largest changes occur between 15 to 24 months regardless of the diet (II, VIII) with most of these gene changes occurring regardless of the diet (X). Diet also had significant effects on gene expression (IV, V, VI) with many gene expression changes occurring regardless of the age (VII). Altered genes that have been up/downregulated in common across age or diet (X, and VII). Changed between 24 and 30 (XI) have been omitted from Reactome enrichment as not enough genes were commonly changed between those comparisons.

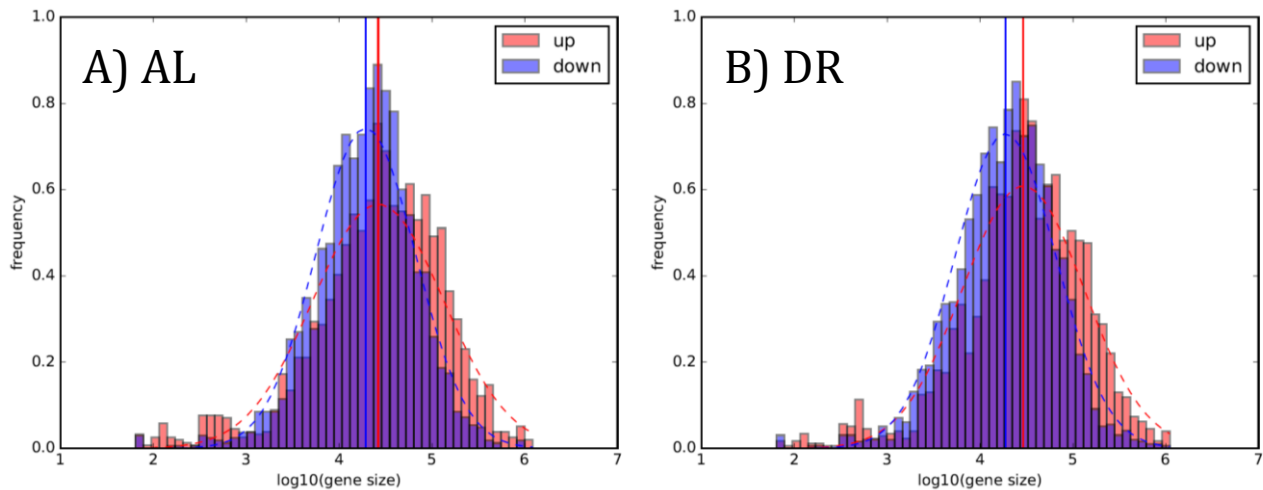


Figure 2-8. Distribution of gene length for downregulated and upregulated genes between 15 to 24 months in AL (A) and DR (B) showing that upregulated genes were on average 3% longer in AL ($p=4.7E-21$) and 4% longer in DR ($p=1.7E-49$). The geometric mean has been marked with vertical lines with their respective colours.

When checking if there was consistency of gene length distribution with up/down regulation (Figure 2-8), we saw a slight but significant difference in size of up and downregulated genes showing a bias of upregulating longer genes and downregulating shorter genes with ageing ($p=4.7E-21$ for AL, and $p=1.7E-49$ for DR). This is contrary to previously published results, which theorized that longer genes are more susceptible to mutations (due to size) and so will be less likely to be expressed¹³⁸.

As a large amount of differentially expressed genes existed between many samples, we looked into which regions (defined as inter-heterogeneous loci by UCSC gap table browser¹³⁹) showed prevalent differentially expressed genes (at least twice as many significantly upregulated genes compared to significantly downregulated gene loci and vice versa; Figure 2-9). There were 14 upregulated and 20 downregulated inter-heterogeneous loci between 15 to 24 months. Age had downregulatory effects

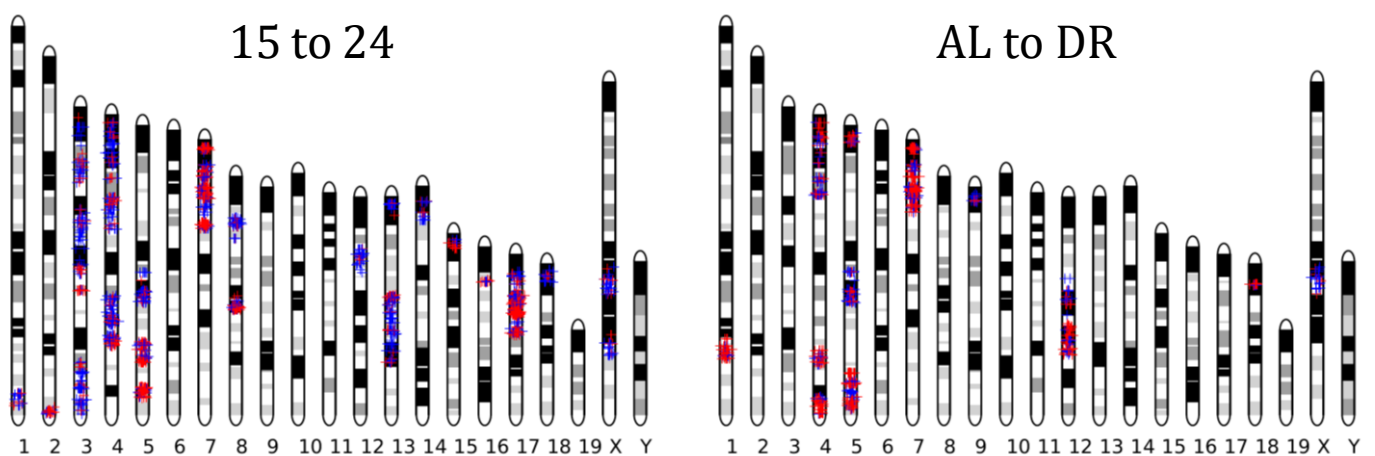


Figure 2-9. Karyomere showing locus of genes upregulated (red), or downregulated (blue) between 15 to 24 months and AL to DR.

in the pericentromeric region on chromosomes 13, chromosome 14 and chromosome 18. Diet, on the other hand, upregulated regions near telomeres in chromosome 4 (both ends), chromosome 5 (both ends), and chromosome 7 (pericentromeric end). Excluding telomeric effects, there were only five downregulated loci compared with fourteen upregulated loci between AL to DR. Some regions were suppressed with age but not diet (centromeric end of chromosome 4 and chromosome 18). These changes were indicative of chromatin remodelling with both age and diet as large regions of chromatin were being activated and inactivated. No chromatin regions were specifically up/down-regulated between 24 to 30 months.

There are a few ways to explain why so few significant changes in genes were found between 24 and 30 months across both AL and DR groupings. One possibility is that minimal transcriptomic changes occurred between 24 months and 30 months timepoints, and the results shown accurately reflect the biology. Another possibility is that heterogeneity increases with age, and therefore statistical tests would be less likely to show significance with higher variability. To test that possibility the variations have been normalized (Figure 2-10 A) and variations have been compared (Figure 2-10 B). No significant differences on variances were found between 24 and 30 months of age (Mann Whitney U test $p=0.8$).

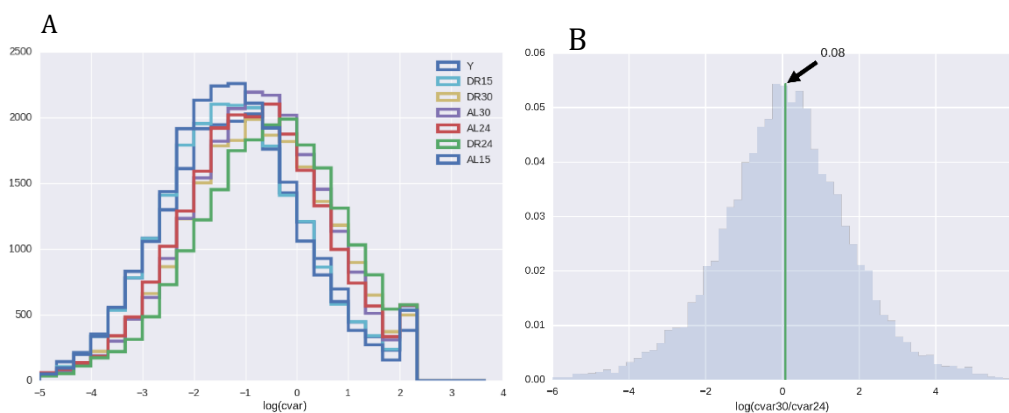


Figure 2-10. (A) histogram of the log of the coefficient of variations for each gene of different samples. The distributions seem to overlap. (B) histogram of the log of the coefficient of variation at 30 months divided by the coefficient of variation at 24 months.

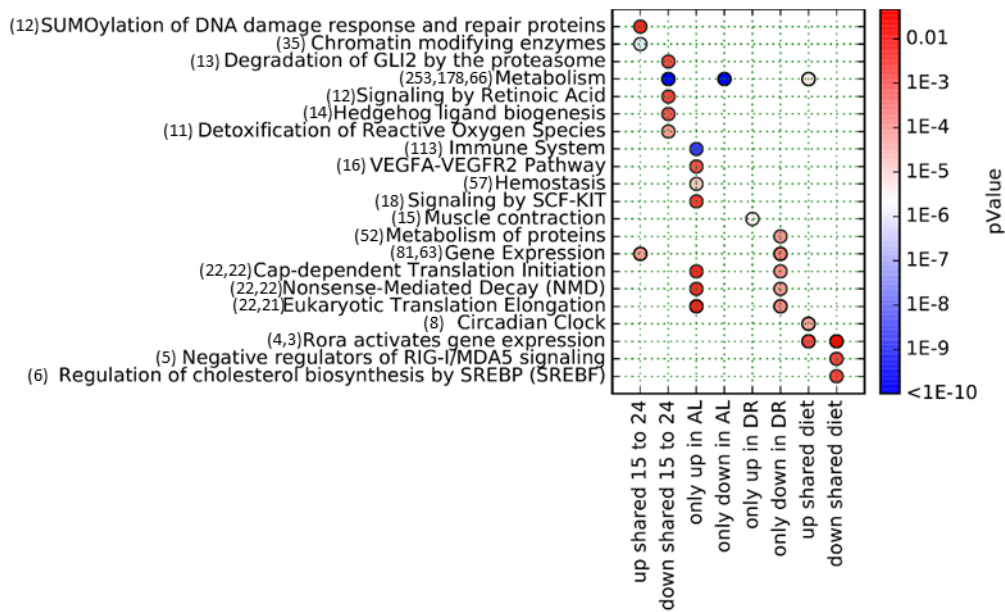


Figure 2-11. Reactome enrichment of differentially expressed genes common or unique between comparisons (only 5 most significant annotations for each group are shown; In brackets are the gene counts for each annotation).

Reactome enrichment (Figure 2-11;file [2]) revealed that both chromatin organization and gene expression (including splicing) were upregulated with age (between 15 and 24 months). During the same time there was a broad downregulation of metabolism including: metabolism of amino acids and derivatives, metabolism of lipids and lipoproteins, the citric acid cycle and respiratory electron transport, detoxification of reactive oxygen species, metabolism of vitamins, bile acid metabolism, and glucose metabolism. Other significant changes included downregulation of insulin pathway genes between 15 and 24 months, even though insulin pathway genes were found in the enrichment for Regulation of AMPK (regulates lipid oxidation and inhibits lipid uptake; not shown in the plot) activity via LKB1 (liver kinase B1; main upstream activating kinase for AMPK).

The effects of DR were distinct from ageing, DR tissues showed upregulated circadian clock genes, and phospholipid metabolism. RORA activated gene expression regulates circadian rhythms (R-MMU-136809) genes were both up and downregulated between AL to DR (Figure 2-7 VII). Regulation of cholesterol biosynthesis by SREBP (R-MMU-1655829), and negative regulators of RIG-I/MDA5 signalling (R-MMU-936440; regulation of the immune response) pathways were significantly enriched among downregulated genes with DR.

Between 15 and 24 months of age, we found the largest numbers of differentially regulated genes under both dietary regimens. Moreover, many of these changes occurred similarly under AL and DR (Figure 2-7 X). Immune system regulatory genes

(R-MMU-168256) were significantly upregulated in AL (between 15 to 24 months) but not significantly altered in 15 to 24 months DR.

Two genes (Bcl2l13 and Bcl2l15) from the Bcl2 (an apoptotic regulator) family were found to be upregulated in DR. This may be relevant as Bcl family genes have been found significantly upregulated in centenarians¹⁴⁰ and suspected to have tumour suppressing roles in ageing.

Some metalloproteins and metal related enzymes have been found to be significantly altered. Therefore, it is possible that during senescence the metal landscape of a tissue is altered (Appendix 1).

A simple question when considering ageing and DR is what genes are up/down through ageing and their relationship with DR (Figure 2-12). We see 131 genes that

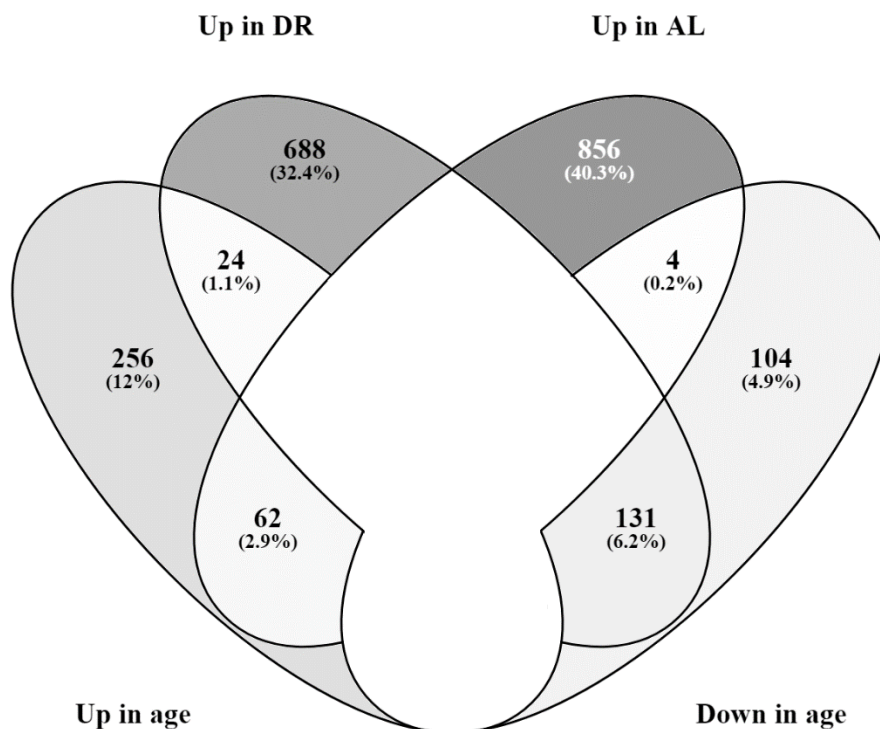


Figure 2-12. 4-way Venn diagram representing the genes that have been downregulated between 15 months AL to 15 months DR (Up in DR), upregulated between 15 months AL and 15 months DR (Up in AL), upregulated between 15 months AL to 30 months AL (Up in age), and downregulated between 15 months AL to 30 months AL (Down in age). All changes are significant and at least $\log_2 > 1$. Regions where by definition we expect to find nothing have been blank out for clarity. A 4-way Venn diagram was chosen instead of 4 individual 2-way Venn diagram.

were upregulated in DR but downregulated with age, 64 genes downregulated with DR but up in age, 24 genes that went up in DR and up in age, and only 4 genes that was down in DR and down in age.

To get the landscape of changes regarding transcriptomic changes, the expression data was networked and clustered into an unbiased distinct pattern of gene clusters using Markov Clustering implemented in Biolayout Express (file [3]). The visual cluster was graphed using a fast multipodal multilevel method (FMMM) graph with a minimum correlation coefficient of 0.95 (Correlation coefficient declares how related is the expression between two genes; correlation of 1 indicates identical expression). The Markov chain (used to declare distinct clusters) was made using an inflation coefficient of 2.2 and a preinflation coefficient of 3 (These two parameters define the statistical distribution of the clusters). Cluster with less than 30 genes were filtered out.

The dimensionless network distances genes depending on how similar the expression patterns were. Genes that were expressed in similar ways across samples were connected. The clustering attempts to segregate the connected network into distinct expression patterns. Individual clusters were then tested for enrichment for Reactome pathway annotations with a 0.05 P-value cut off.

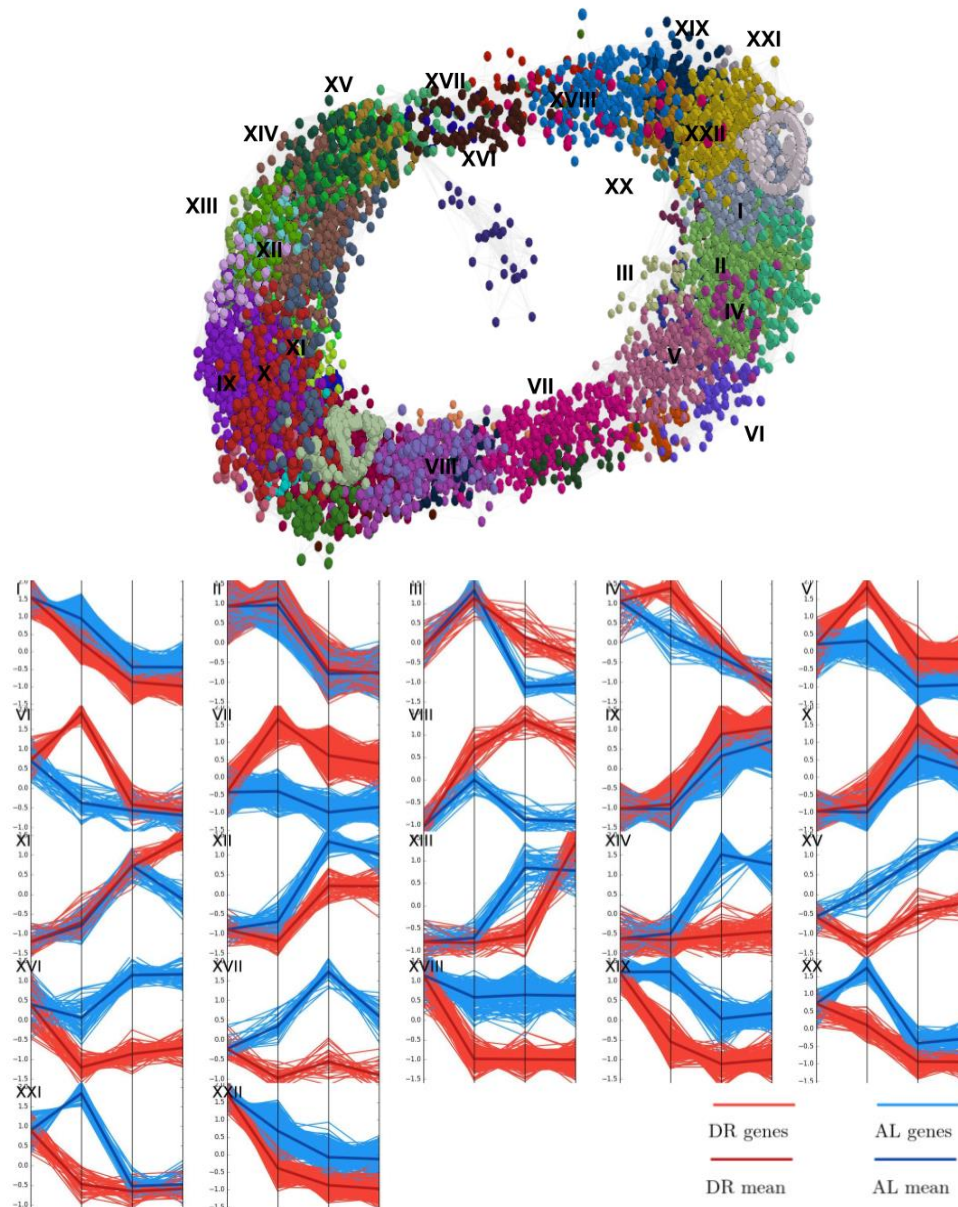


Figure 2-13 Expression network of mice ageing samples (top). Expression curves of different distinct MCL clusters shown in clockwise order of location in the network (bottom).

Construction and visualization of the co-expression network (Figure 2-13, top) shows a genus 1 topology (doughnut/ring shape). After clustering, the resulting clusters were ordered by position in the torus (by default clusters were ordered by size) which interestingly showed that the network ring corresponds to an expression pattern continuum (Figure 2-13, bottom).

It is possible to visualize how the network of clustered genes changed with respect to diet and age (Figure 2-13) by graphing the node size of each experimental point by the expression amount. Displaying the data this way is useful to explain the genus 1 topology of the expression network. The torus shape was predominately caused by major expression changes between 15 to 24 months, with notably fewer changes occurring between all other age comparisons (AL 3 \approx AL 15, AL 24 \approx AL 30, DR 24 \approx DR 30). Even though the data is comprised of seven experimental points, there are only two data points showing major changes which occurred across diet and age between 15 to 24 months. Consequently, the continuum was pulled across two axes. In addition to the expression changes between 15-24 months, there is a common 'young' expression (right of the network), an 'old' expression (left of the network), an 'AL' expression (top of the network), and a 'DR' expression (bottom of the network). These four expression types define the network morphology.

It is also interesting to note that the expression network analysis was practically unchanged between 24 to 30 months of age. This implies that almost all age induced transcriptomic effects on mouse hepatic tissue occurred prior to 24 months.

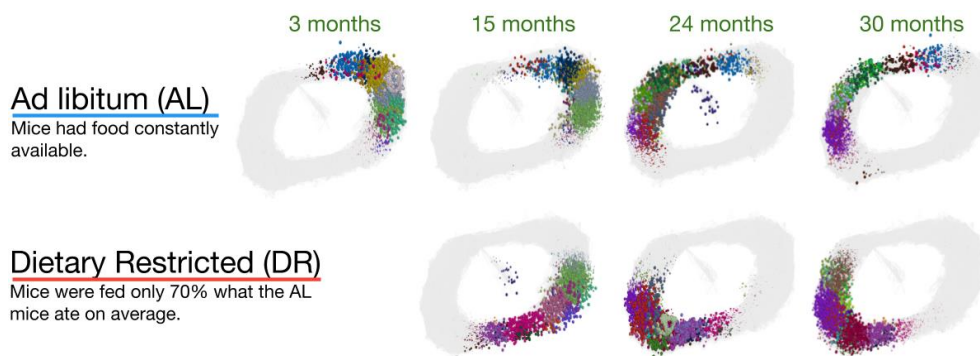


Figure 2-14 Network showing the genes expressed in each diet and age.

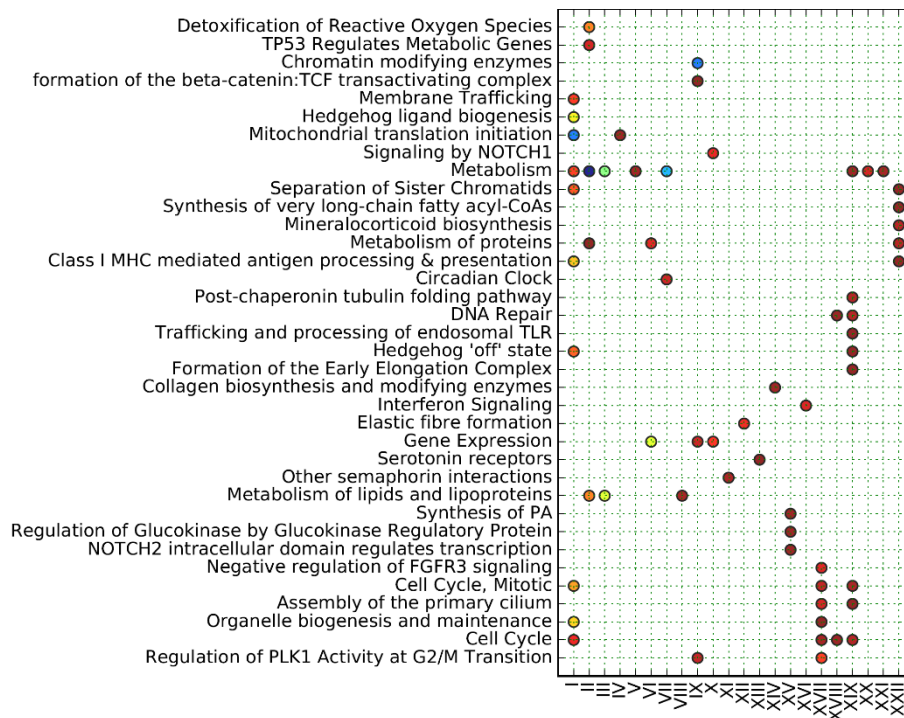


Figure 2-15. Reactome enrichment of discrete clusters.

Due to the continuum in gene expression patterns within the clusters (Figure 2-13; file [4]) and the division between neighbouring clusters being somewhat arbitrary, some Reactome enrichments (Figure 2-15) span multiple clusters. This effect is most prevalent in metabolism (R-MMU-1430728.1) (together with metabolism of amino acids and derivatives (R-MMU-71291.1) as its most enriched subcategory) spanning from cluster XIX to cluster VII (these clusters were upregulated at 3 and 15 months regardless of diet). Similarly, clusters II and III showed a pronounced increase of metabolism of lipoproteins (R-MMU-556833). The area most upregulated at 30 months was gene expression (R-MMU-74160.1; cluster XI and X). DR was shown to induce down-regulation of cell cycle genes (R-MMU-1640170) and DNA repair (R-MMU-73894; cluster XVII to XIX) across all ages (The complete list of genes for each category are in file [4]).

It was apparent from the gene network topology (Figure 2-14) that the largest and most significant change across age grouping occurred between 15 to 24 months of age.

Both cell cycle (R-MMU-1640170) and immune system (R-MMU-168256) genes were found enriched in the expression network (Figure 2-15) and DESeq analysis (Figure 2-11) associated with ageing. This could be indication of senescence activation as these are known markers of senescence. Therefore, we integrated well accepted experimental measurements of senescence; karyomegaly and telomere-associated foci (TAF). These were measured by Mikolaj Ogrodnik in a collaboration with the Newcastle University Institute for Ageing, UK ¹⁴¹ (Appendix 2). We extracted all the genes that followed the same coexpression patterns as the senescent markers of TAF, Karyomegaly and oil droplets, we call this method network fishing (page **Error! Bookmark not defined.**). Karyomegaly and TAF were both upregulated at 15 months (Figure 2-16) compared to 3 months. However, this difference was minimal when compared to the marker upregulation shown in AL at 24 months (DR was still upregulated at 24 months compared to 15 months). Oil lipids were highly increased in AL at 15 months but were reduced at 24 months. This pronounced increase at 15 months was due to hepatic steatosis. This decrease at 24 months may be due to a survival effect of the steatosis resistant mice. Many of the same genes were found in the karyomegaly and TAFS marker expression network. These shared genes were found to be overrepresented by the immune system (R-MMU-168256) and hemostasis (R-MMU-109582) Reactome clusters (Figure 2-16.c). Genes that cluster in an inverted pattern to both karyomegaly and TAF were found to be enriched for respiratory electron transport chain (R-MMU-1428517) and metabolism (R-MMU-1430728) Reactome clusters.

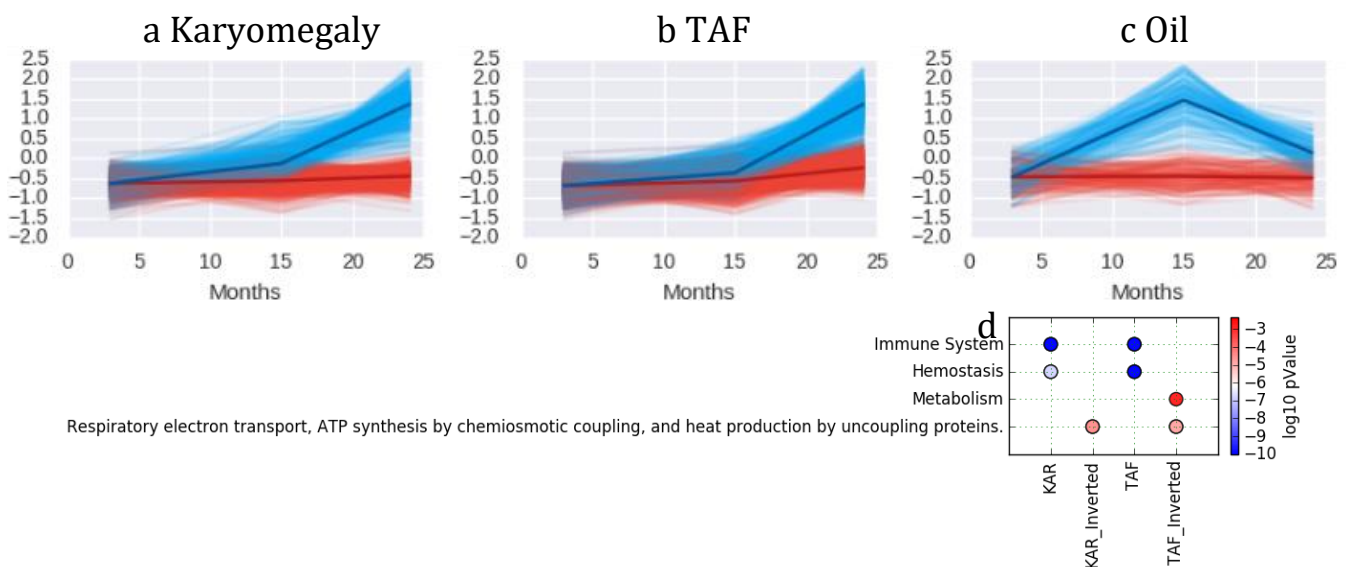


Figure 2-16. Network fishing of genes with same expression pattern of karyomegaly (a), TAF (b) or intracellular lipid droplets (c) for AL (blue) and DR (red). (d) Reactome enrichment of karyomegaly and TAFs and the genes which followed an expression opposite to them (inverting the z-scored values across the mean).

The crossover mice (AL>DR and DR>AL) provided a glimpse into the effects (or lack thereof) of short-term dietary changes. This study focused on four mice treatments. AL mice spent their entire lifespan on an AL diet, and DR mice were placed on a dietary restricted diet after 3 months of age (DR mice were fed as AL mice from 0 to 3 months). The AL>DR mice were placed on a DR diet from an AL diet at 13 months. The DR>AL mice were on an AL diet until 3 months of age when they were placed on a DR diet; then at 12 months they reverted to an AL diet.

Oil red lipid staining (Figure 2-17 A, B ;Appendix 2Error! Reference source not found.) shows that AL mice developed steatosis (accumulation of fat in the liver),

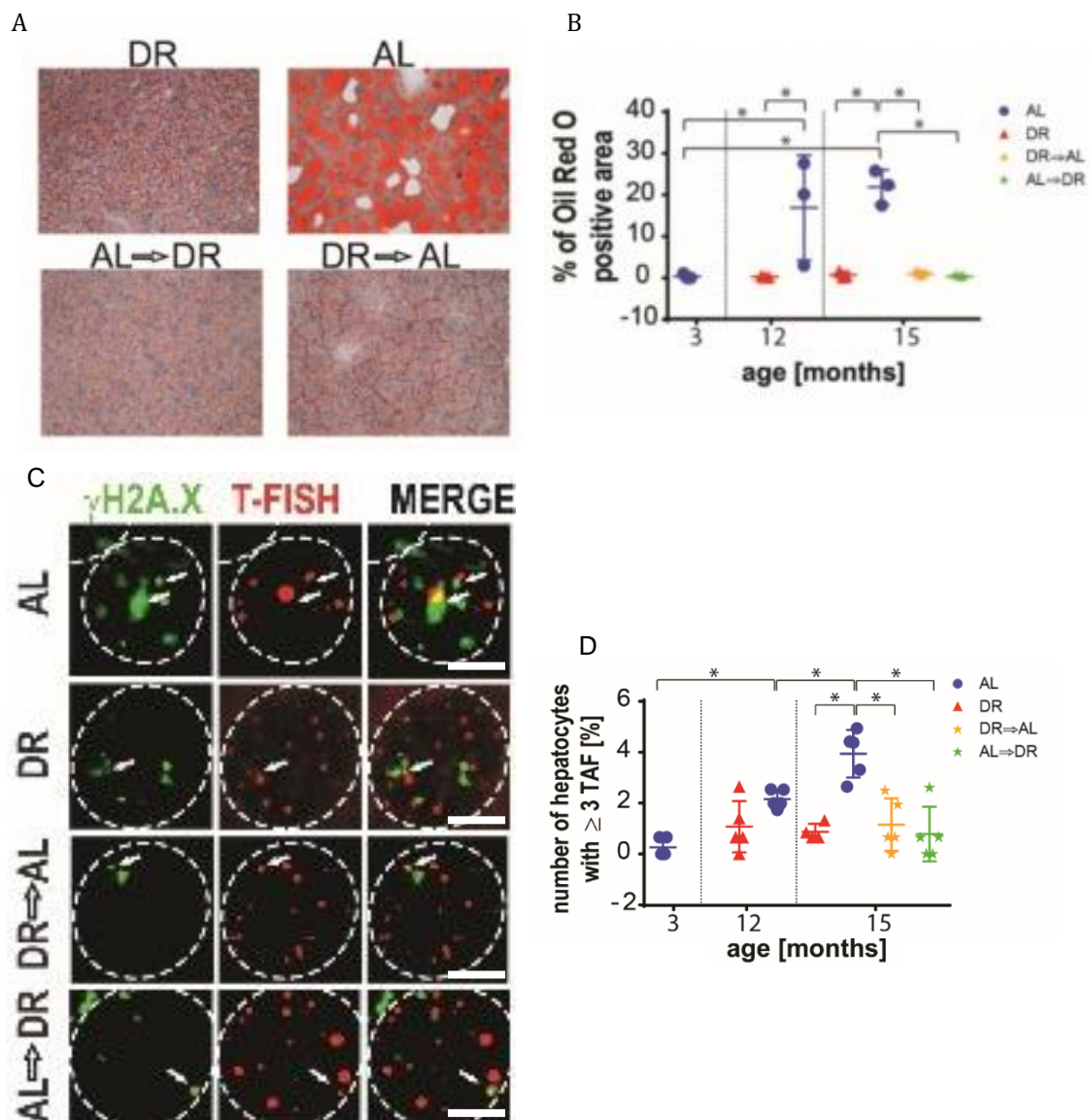


Figure 2-17. (A) Lipid staining of liver sections. (B) AL livers were steatotic while any form of DR had protective effects (B). Senescence marker staining of γ H2A.X and telomere fluorescent in situ hybridization. (C) Representative microscopy samples of γ H2A.X labelling (green) and telomere *in situ* hybridization (red) in mice hepatocytes of different treatments (scale bar is 4 μ m). (D) Percentage of hepatocytes with ≥ 3 TAF. Adapted from Ogronnik (2017) ¹⁴¹.

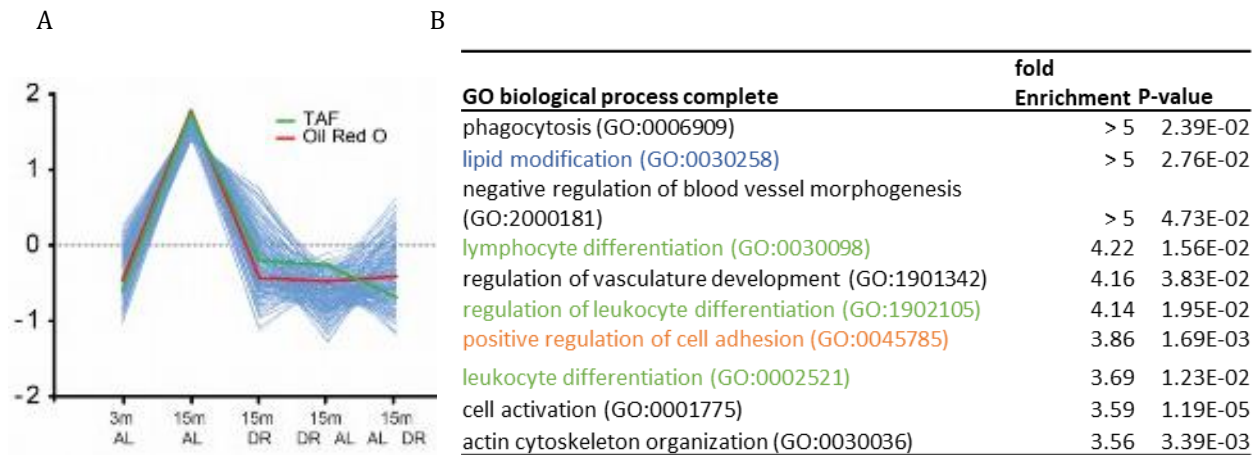


Figure 2-18.(A) Cluster fishing of senescence markers Oil staining and TAFs provides 709 genes that co-clustered with the aforementioned senescence markers. (B) GO enrichment of genes that co-clustered with senescence markers coloured by immune system annotations (green), cell-cell binding annotations (orange) and lipid annotations (blue).

while all forms of DR (DR, AL>DR, and DR>AL) did not show signs of hepatic steatosis ($P=0.0045$). This effect was supported by changes to telomere-associated foci in mice hepatic tissue (TAF; Figure 2-17 C, D). These results suggest that an AL diet induces senescence in the liver and is correlated with steatosis. Any form of DR induces a degree of protection against both senescence and steatosis.

Empirical measurements of senescence markers (intracellular lipids, and TAFs) and RNAseq of similar tissues at the same age and treatments, were combined to assess gene expression networking and clustering. This allowed for the isolation of genes that co-clustered with senescence markers. This was achieved by inserting empirical measurements into a normalized gene expression dataset of the samples. The Network was produced with BiolayoutExpress¹⁴² using a Pearson coefficient of 0.7 and a correlation value of 0.95. This network was clustered using MCL implementation of Markov Cluster Algorithm¹¹² using an inflation coefficient of 2.2 and a pre-inflation coefficient of 3.0. This returned a cluster which contained both TAF and lipid measurements as well as 709 other genes. Fished genes were found to contain a large amount of GO annotations for cell-cell binding, lipid modification, and immune system function (Figure 2-18 B; Appendix 3). There was also a large number of significant annotations related to immune system genes, cell-cell binding, and lipid modifications.

The results of the network fishing analysis have been published (Ogrodnik, 2017)

141.

A possible source of experimental error might result from the DR mice daily feeding (ca 9:00), while AL mice had food available throughout the day. It has been demonstrated in previous research that altering feeding schedules affects biological circadian rhythms ¹⁴³. In addition to feeding schedules other studies have demonstrated that calorie content (sans feeding schedule alteration) ¹⁴⁴ and age also effects circadian rhythms ¹⁴⁵. It is possible that the dietary affects in this study could also be caused by circadian changes. This problem is further confounded by the fact that most factors in this study are known to affect circadian rhythms (age, diet, and calorie content). However, this study does not attempt to understand the role and biology of the circadian clock in ageing. Therefore, we have assumed that the circadian changes are downstream of the dietary treatments (whether or not ageing benefits are downstream of circadian changes will remain outside of the scope of this study).

Within the datasets to study circadian effects, there were stark changes in the transcription of different mitochondrial tRNAs. The changes in different amino acids mt-tRNA might in turn affect the expression of different mitochondrially expressed proteins. This hypothesis was modelled and explored (Appendix 4). However, there was not enough empirical data to obtain a concrete conclusion.

To control for the effects of circadian rhythm, mice samples were sacrificed after being fed rather than 6 hours after feeding.

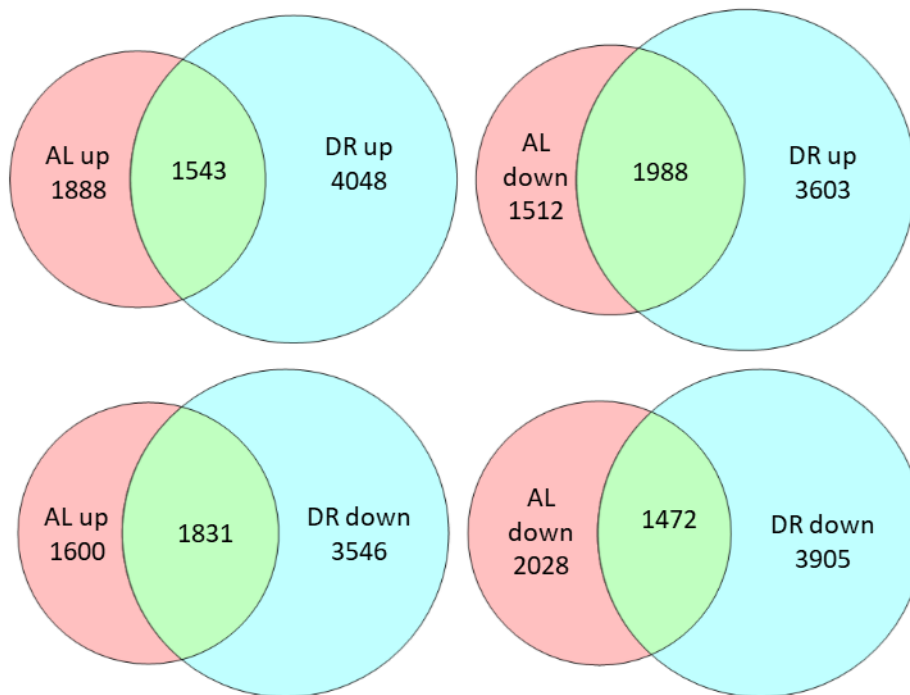


Figure 2-19. Venn diagram showing upregulated and downregulated genes when compared with post fed sequencing for both AL and DR.

The PCA plot comparing all treatments (Figure 2-6) showed that post fed mice gene expressions were closer to each other (AL and DR) than to 15 months AL and 15 months DR. Genes significantly up or down-regulated with post fed sequencing (Figure 2-19) show some interesting inconsistency. Genes upregulated and downregulated with AL and DR, compared with their respective post fed sequencing show that half of the genes upregulated in AL are also upregulated in postfed DR (while the other half is in fact downregulated in DR postfed). A similar effect was present with downregulated genes in AL (Figure 2-19). DR had, however, the most genes that were only upregulated or downregulated with DR postfed and were not significantly altered when comparing AL to AL postfed. This means that postfed DR had more transcriptomic alterations affecting genes not affected in postfed AL.

Reactome enrichment for genes altered with post fed mice (Figure 2-20; file [5]) showed that apoptosis (R-MMU-109581) was upregulated with DR postfed (when compared with DR) and mitosis Reactomes were downregulated. These results implied that DR tissues were more apoptotic and less mitotic after feeding than 6 hours after feeding. However, many genes were found altered only in DR postfed, including upregulation of DNA repair (R-MMU-73894).

These results suggested that sporadic feeding may act as a stressor in DR diets. In contrast to AL mice (who ate throughout the day as food was always available) which did not show dietary induced changes to mitotic or apoptotic gene expression patterns.

It is possible that the dietary stress of sudden food availability would cause a short-term metabolic trauma, albeit DR mice were healthier than AL mice. This might provide some evidence of a hormesis effect ¹⁴⁶ which states that low levels of stress can provide beneficial effects.

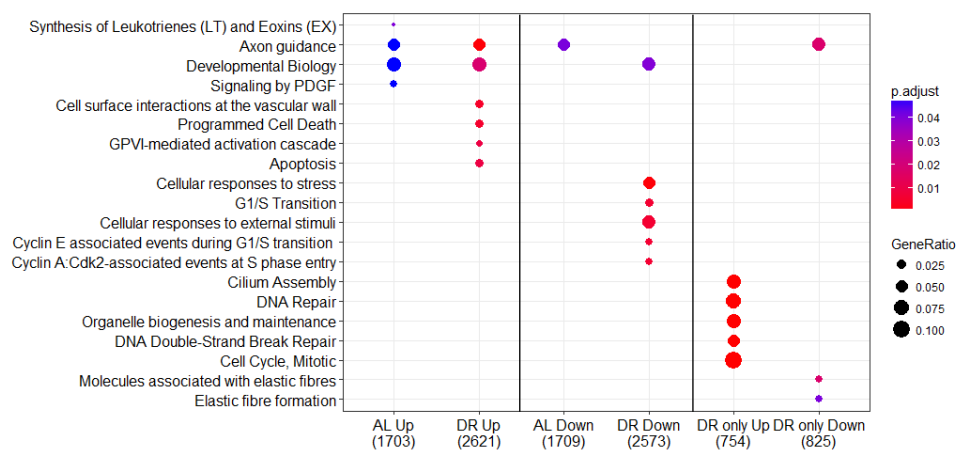


Figure 2-20. Reactome enrichment of significant genes when compared with postfed mice of the same treatment. Also included are genes that only appeared upregulated or downregulated in DR.

SPLICING IN MICE AGEING AND DIET

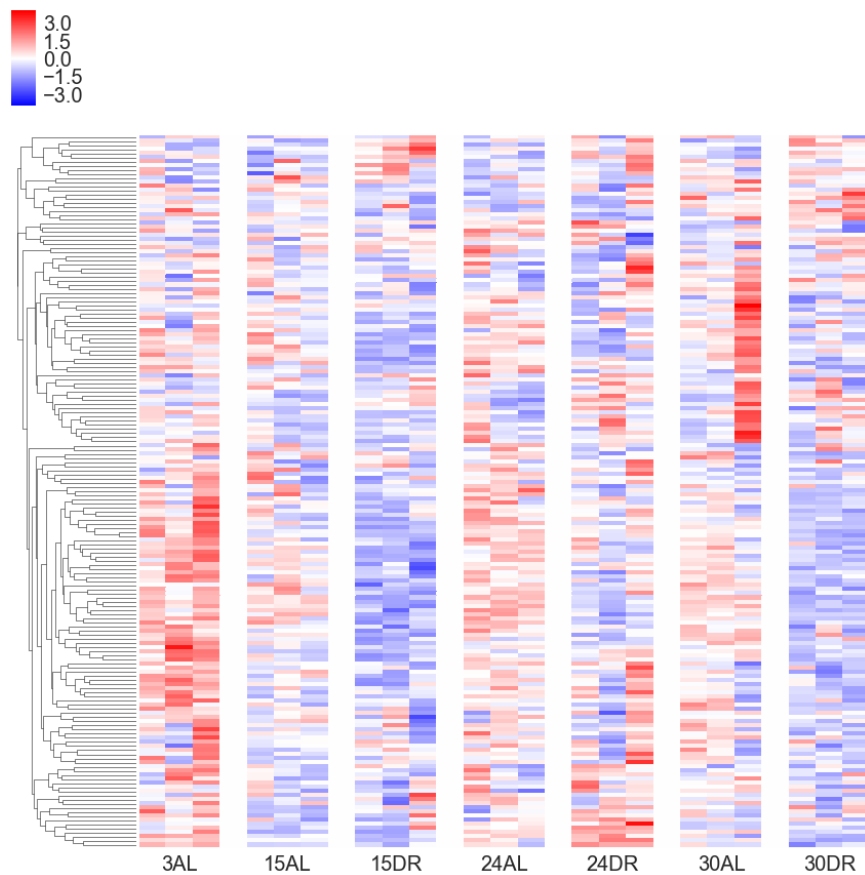


Figure 2-21. Z-scored clustered heatmap showing strong age and dietary effects spliceosome genes (R-MMU-72172; 252 genes).

While studying RNAseq data, we can not only look at the transcriptomic changes, but also at the splicing changes. Using the differential expression data we see strong regulation of the spliceosomal genes (Figure 2-21; splicing example in Appendix 6). Therefore, after studying in detail differential gene expression, we focused on alternative isoform expression, and expression of long non-coding RNA expression (Appendix 5).

Further analysis has been conducted using MAltESERS, a novel tool to identify changes in functional domains (page 87).

Unlike differential gene expression, differential splicing changes were found across all timepoints (Figure 2-22). The number of spliced events increased with age (1532 events between 3 months to 15 months; 2207 events between 15AL to 24AL; and 1871 events between 24AL to 30AL).

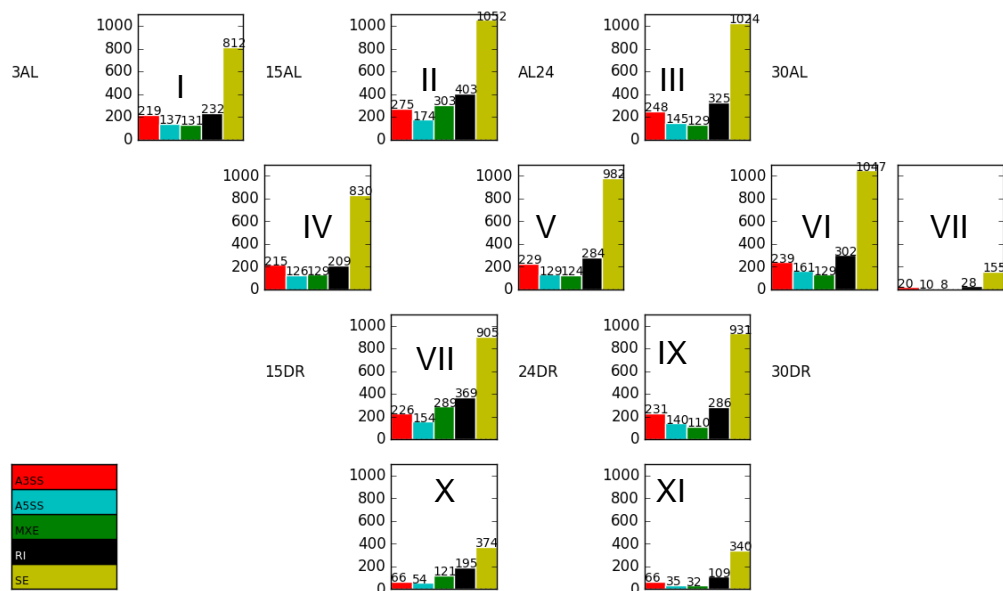


Figure 2-22. Significant alternative splicing changes among age and diet. Significant splicing between AL to DR at all timepoints. Unlike differential expression, there are many events in all age point comparisons and few of the events occur with age regardless of diet (XI & XII) or with diet regardless age (VII). A3SS and A5SS is Alternative 3' or 5' Splice site respectively, MXE is Multiple Exclusive Exon, RI is Retained Intron, and SE is SplicedExon.

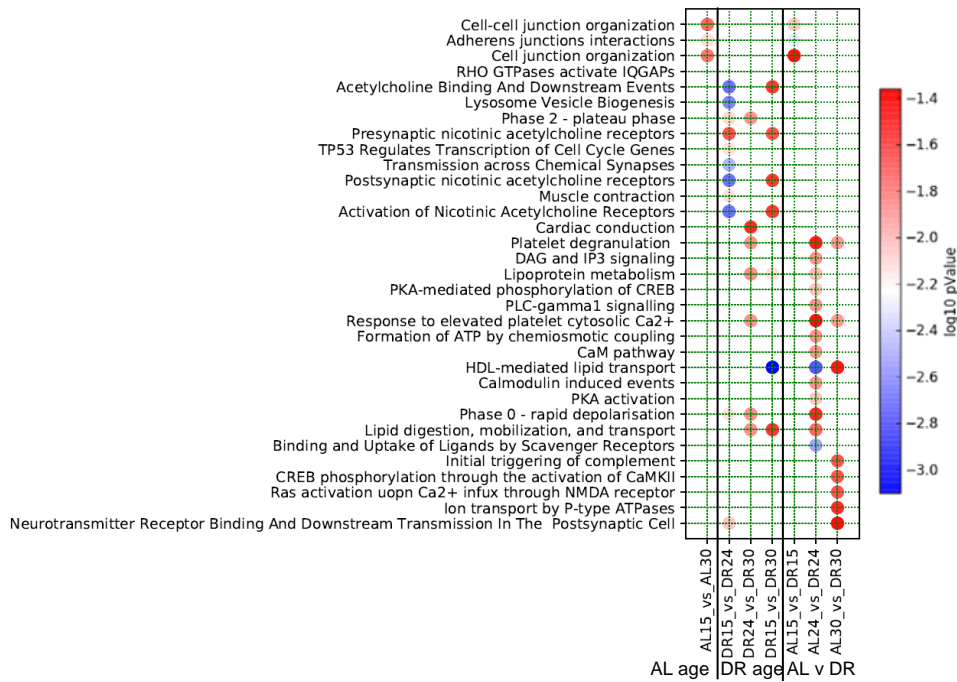


Figure 2-23. Reactome enrichment of splice events between treatments (file [14]).

When comparing all splicing events (filtering for 0.001 P adjusted and absolute inclusion level difference of at least 50%), circa 2000 splice events occurred in every comparison (file [6]). Approximately half of the splicing events were spliced exons (compared to alternative 3' or 5' polyadenylation, mutually exclusive exons, or retained introns) and a third of the events were conserved regardless of diet (Figure 2-22 X, XI), and around a tenth of splicing events were conserved in ageing (Figure 2-22 VII). When tested for enrichment of Reactome annotations (Figure 2-23), cell-cell junctions organization (R-MMU-446728) genes were found alternatively spliced both between AL15 to AL30 and between AL15 to DR15. Reactomes for “lipid digestion, mobilization and transport” and “HDL-mediated lipid transport” were found alternatively spliced during ageing in DR, but also alternatively spliced in DR at 24 months.

Reactome enrichment analysis without merging event types (Figure 2-24) found significant changes in mRNA splicing (Transport of Mature mRNA derived from intron-Containing Transcript (R-MMU-159236) & mRNA Splicing – Major Pathway (R-MMU-72163)) in spliced exons and multiple exclusive exons between 24 to 30 months and at 30 months between diets (Figure 2-24). In addition, Reactome enrichment found diet dependant alternative splicing of 5' splice sites in genes important for formation of senescence foci (R-MMU-2559584) at 15 months. Chromatin modifying enzymes (R-MMU-3247509) were also found alternatively spliced with a 5' spliced site between 24 months DR and 30 months DR.

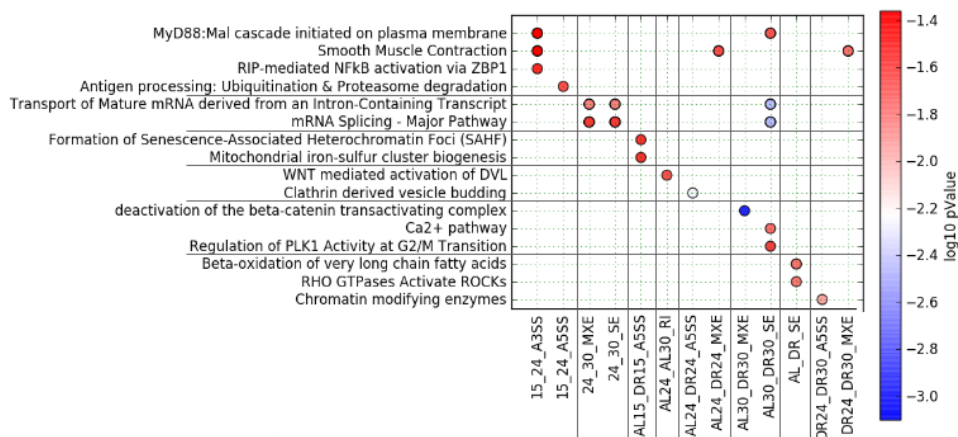


Figure 2-24. Reactome enrichment of spliced events without merging splice types (file [12]).

CONCLUSION

CHROMATIN STRUCTURE

Reactome annotations for chromatin modifying enzymes were shown to be upregulated by differential expression analysis between 15 to 24 months (Figure 2-11), downregulated between 15 to 24 months by network analysis (Figure 2-15), and alternatively spliced between 24 DR to 30 DR (Figure 2-22).

Genes coding for chromatin modifying enzymes were significantly upregulated in DESeq and network clustering analysis (Figure 2-25). This was expected as both DESeq and network clustering are essentially two different methods for detecting changes in expression. However, genes coding for chromatin modifying enzymes were also found increasingly alternatively spliced in DR between 24 to 30 months (Figure 2-25). The increased alternative splicing of genes encoding chromatin modifying enzymes was not found in DESeq or network analysis enrichment. This might mean that DR affects chromatin structure through splicing at a later age to compensate/ameliorate the ageing effects initiated between 15-24 months of age.

RESPIRATION

Differential expression shows a strong downregulation of metabolism genes between 15 to 24 months. However, metabolic genes were found upregulated with DR treatments and further downregulated with AL treatment (Figure 2-11). In addition to differential expression, clustering analysis showed similar transcriptomic changes to metabolic pathways across age and diet (Figure 2-13 clusters I, II, III, V, VII, XIX, XX, XXI). Clustering analysis also found respiratory associated genes to be anti-clustered with senescence markers (Figure 2-16). The decrease in metabolic

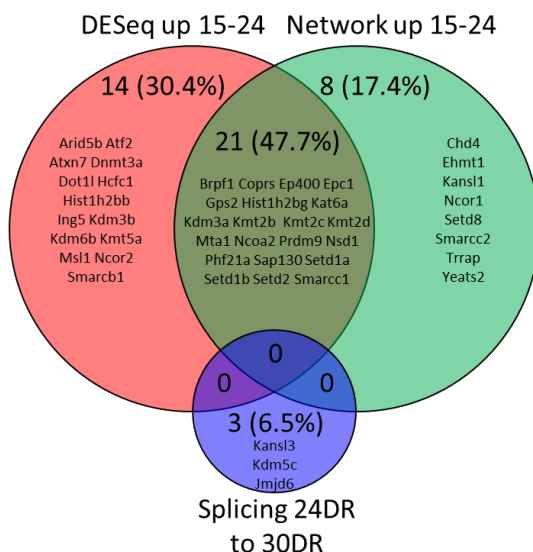


Figure 2-25. Venn diagram of genes enriched for chromatin modifying enzymes (R-MMU-3247509) across multiple analysis. This Venn shows overlapping genes found by differential expression and clustering. However, these genes were not found to be significantly spliced between 24 to 30 months DR.

genes with age noted by clustering and differential expression studies here, has been previously reported in mice metabolic analyses ^{147,148}.

Decreased stamina is a prominent symptom of ageing. Decreased stamina with ageing has been attributed to Aconitase1 (ACO1) downregulation ¹. ACO1 is found in both the mitochondria and cytosol. In the mitochondria, ACO1 is a crucial part of the citric acid cycle, where it catalyses the isomerization of citrate to iso-citrate. In the cytosol, ACO1 acts as a post-transcriptional regulatory factor for iron homeostasis. As intracellular iron concentrations decrease, ACO1 binds to iron-responsive elements at the 5' end of ferritin mRNA, repressing ferritin translation (ferritin is a protein that stores iron). Our data show that ACO1 was indeed downregulated with ageing between 15-24 months (suddenly not gradually; Figure 2-26). However, ACO2 was strongly upregulated with DR. This might provide a compensatory mechanism that improved the DR mice stamina with age.

ACO1 and ACO2 have distinct functions and are not enzymatically identical. ACO2 lacks iron regulatory functions, and ACO2 has been shown to prevent oxidative damage in mitochondrial DNA ¹⁴⁹. This means that the DR use of ACO2 may function to increase oxidative damage protection to the mitochondrial DNA, further improving stamina.

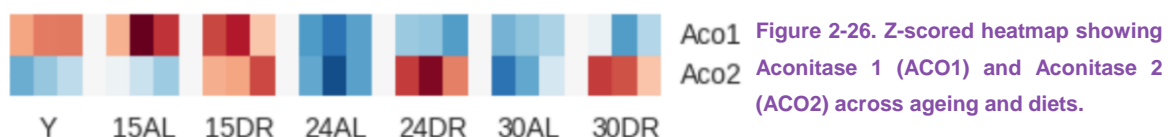
INFLAMMATION

Differential expression showed an upregulation of immune system (R-MMU-168256) genes with AL, as well as decreased inflammation in DR (negative regulators of RIG-I/MDA5 signalling; Figure 2-11). Immune system genes were also strongly co-clustered with senescence markers (karyomegaly and TAF; Figure 2-16).

There was a clear link between inflammation, immune system activation, and obesity ¹⁵⁰⁻¹⁵². However, it is important to stress that AL mice were not obese. However, DR was shown to decrease inflammation in mice.

SENESCENCE TRIFECTA

Transcriptomic changes found throughout the mice ageing experiments have shown three motifs to be affected by both age and diet: respiration, chromatin structure, and immune system. These three factors are likely to be involved in the ageing phenotype, and seem to be ameliorated to a certain extent by DR.



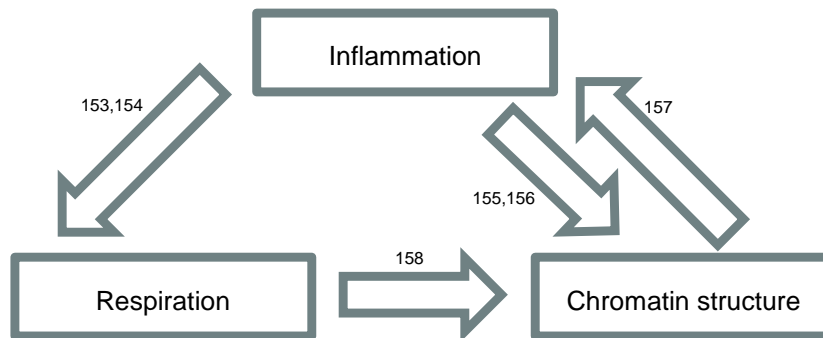


Figure 2-27. Diagrammatic representation of how the senescent trifecta factors interact with each other.

These three factors appear to influence each other (which is not surprising given the interdependence of many biological systems) in the following ways:

- Inflammation decreases respiration.
 - Inflammation leads to a decrease in mitochondrial activity ^{153,154}.
- Inflammation affects chromatin structure.
 - Inflammation has been shown to induce chromatin remodelling ^{155,156}.
- Chromatin structure changes induce inflammation
 - Reversible progeroid models show that when chromatin structure is returned to a non-pathogenic state, inflammation decreases¹⁵⁷.
- Respiration affects chromatin structure.
 - Packaging chromatin by SIRT2 requires NAD⁺. As respiration increases NAD⁺ availability, this can lead to an increase in SIRT2 activity ¹⁵⁸.

These three factors are highly interconnected (Figure 2-27) and proper modelling might reveal in depth interactions of the ageing trifecta. One possibility is that each factor of the trifecta can activate and enforce other factors, and once activated they are very stable. However, by studying this trifecta we might be able to pinpoint specific triggers that activate the trifecta which may be of clinical importance. It is also likely that these three interacting pathways act as three locks which prevent reversal of the cell cycle. In this case, all three pathways have to be returned to normal to reverse senescence thus preventing mutations affecting a single pathway from reversing senescence.

SENESCENCE BLOOMING

As seen in differential expression studies (Figure 2-7), there was a large change in gene expression between 15 to 24 months, followed by little changes thereafter. A similar result was found using network analysis (Figure 2-13). Between 15 to 24 months, there was a stark increase in senescence markers (TAF and karyomegaly;

Figure 2-16). During the same 15 to 24 months period the survival rate changed in both AL and DR (less in DR; Figure 2-3). A likely explanation for the sudden transcriptomic and survival rate changes could be a sudden senescence induction in cells during this period. This means that the mouse hepatic tissues were found to undergo induction of senescence in a very short time window.

A possible mechanism for sudden induction of senescence is that senescent cells secrete proinflammatory factors in the senescence-associated secretory phenotype (SASP). SASP paracrine signaling can induce healthy cells to transition into senescence through the bystander effect ⁵⁷.

It is likely that once some cells become senescent in the liver tissue (be it from the accumulation of mutational insults, telomeric attrition, oxidative stress, or other reasons), these cells then secrete SASP factors. SASP secretion will induce healthy cells to commit to the senescence transition and in turn, push a large fraction of cells into senescence. This senescence bloom can disproportionately age the mice in a very short term. The question thus arises as to whether the more gradual induction of ageing phenotypes can happen after a sudden senescent induction. It has been shown that eliminating senescent cells can significantly delay ageing-related pathologies ¹⁵⁹. In doing so senescent blooming might be delayed.

Previous mouse ageing studies have shown a sudden change in metabolism after 18 months ¹⁶⁰ and ageing biomarkers suddenly altered at 12 months of age ¹⁶¹. These might confirm the hypothesis that ageing involves a sporadic senescent bloom.

The results of this research point to the conclusion that ageing is a sudden event. During a reasonably short period of time tissues and therefore organs (and perhaps systems), undergo a sudden senescent blooming. As a consequence of the senescent induction, specific age pathologies are more likely to take hold.

This study only focused on female mice hepatic tissues, so it is not known whether these effects are indeed extrapolatable unto other tissues, organs, and/or species.

An alternative explanation of the effects observed could be a survival bias, we only see the healthier older mice as the unhealthy ones succumb to age related illnesses. Therefore, showing us older mice who are healthy vs younger mice with more variety in their health status. However, the mice were genetically identical and had the same treatment and environment. Another technical issue could arise from batch effects (see gene counting page: **Error! Bookmark not defined.**) as consequence of sequencing the different age mice on different batches. However, if the gene expression changes are due to batch effects, up/downregulated genes would be stochastic (or derive from a technical reason such as specific nucleotide distribution,

gene length...) and therefore, make it unlikely to get Reactome enrichments. Another experiment could be performed were tissues are biopsied from the same animal in a longitudinal study. This study would be too hard to perform on hepatic tissue as it would require a liver biopsy on the same animal every few months. Perhaps looking at dermal tissues using small skin biopsies would be possible. However, we might first need validation that the senescence blooming effect happens on skin tissues.

In order to validate the observation and interpretations of this study, we should study if the effect is replicated in another longitudinal study, and at the same time study if the effect is replicated in other tissues. Increased temporal resolution would show how quick and in which order senescence occur on an organ. There is a need for longitudinal ageing experiments with many time points, there are plenty of RNA-seq ageing experiments, however they usually compare a defined young age (between 1 to 6 months in mice) to an old age (between 12 to 18 months in mice), and with only two timepoints the effects seen here become impossible to observe (There are however development studies with greater temporal resolution for developmental studies but those tend to stop after the animal reach maturity). More research with more timepoints is needed to accept the observations made in this chapter.

The most direct way to assert this theory would be to retrieve temporal samples from a model organism longitudinal study (ideally short-lived chordates such as Zebrafish or mice). Rather than sacrificing the animal at each time point, by using skin biopsies, we could make a longitudinal study on the same individuals. Instead than sequence tissues (high cost), it will be more informative to stain for β -Galactosidase (senescent cell staining), γ H2AX, Ki67, BrdU (former two are negative markers) and count senescent and non-senescent cells by flow cytometry. Plotting the results in a

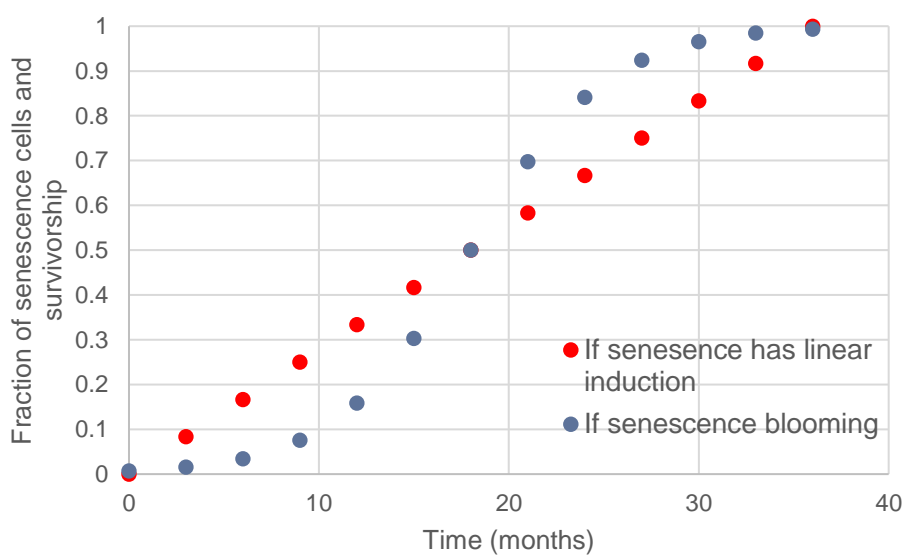


Figure 5-3. Timeline of the percentage of senescent cells if senescence induction is linear (yellow) or senescence induction is sudden (blue), overlaid is a Kaplan Mayer survivorship line.

timeline (Figure 5-3) could either corroborate or dismiss this hypothesis. These methods need to be conducted in as many timepoints as possible to determine if senescence blooming is or not real.

CHAPTER 3 SENESCENCE SIGNATURES IN HUMAN CELL

Chapter objectives:

- To showcase the results of an in-depth bioinformatic study on post senescence transcription.
- To review and investigate replication induced senescence.
- To investigate the plasticity of senescent cells.
- To provide an overview and bioinformatic study on ionizing radiation induced senescence.

Hypothesis:

Cells become senescent through many mechanisms (such as ageing, replicative limit, and IR). In Chapter 2 we studied age induced senescence in mice tissues. However, using public published datasets, we can interrogate the senescence phenotype derived from different methods for senescence induction.

WI38 ONCOGENE INDUCED SENESCENCE

INTRODUCTION

In a study conducted by Tamir Chandra's group in the Babraham Institute UK, WI38 lung fibroblasts cells were cultured into and past oncogene induced senescence. The aim of the study was to investigate chromatin structure under senescence¹⁶². The sequencing data obtained in Chandra's fibroblast study has been repurpose here to study the post senescence transcriptome. All cell culture and sequencing was performed by Tamir Chandra's lab¹⁶³, and the data is publicly available (Gene Expression Omnibus GSE95021).

In this section we study the effects of senescence on the transcriptome, and splicing, by interrogating WI38 datasets that were driven to oncogene induced senescence.

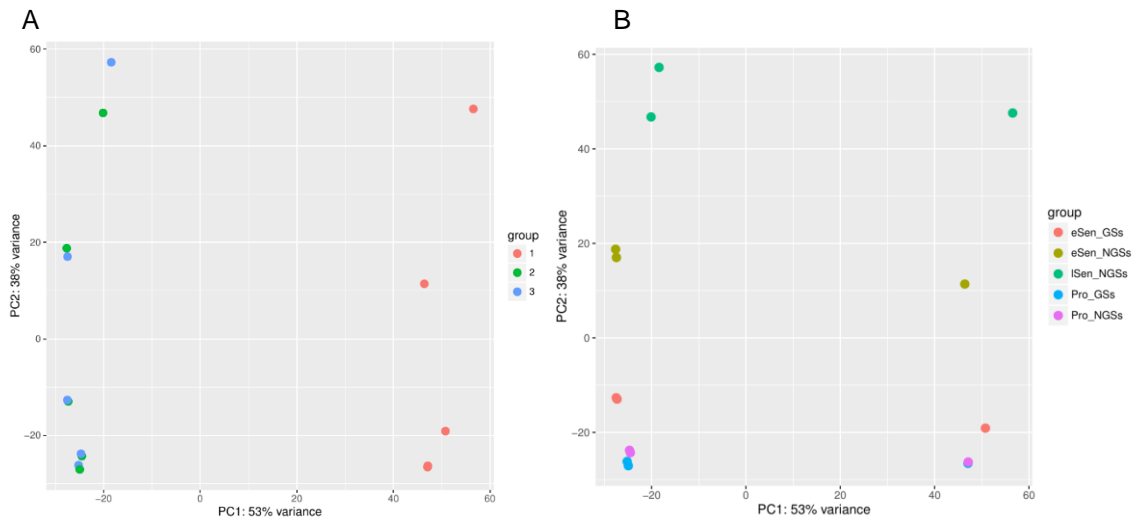


Figure 3-1. PCA plots of WI38 samples coloured by batch number (A), and by biological sample (B). The biological samples are early senescence (eSen), late senescence (ISen) and proliferating (Pro) as well as whether they have been treated with glucocorticoids (*_GSs) or without (*_NGs); glucocorticoids treatments are not relevant for our study but were used to normalize and remove batch bias.

METHODS

WI38 data was aligned using STAR (page **Error! Bookmark not defined.**) however due to ample sequencing being conducted in different batches, with one experimental replicate per batch, there was a strong batch effect (Appendix 8) affecting the principal component 1 of the PCA plot (Figure 3-1).

One solution to reduce batch effect was to run the DESeq recommended SVA correction (Appendix 9). This normalization method decreased the batch effect. However, batch effect was still prevalent in the principal component 2 (Figure 3-2).

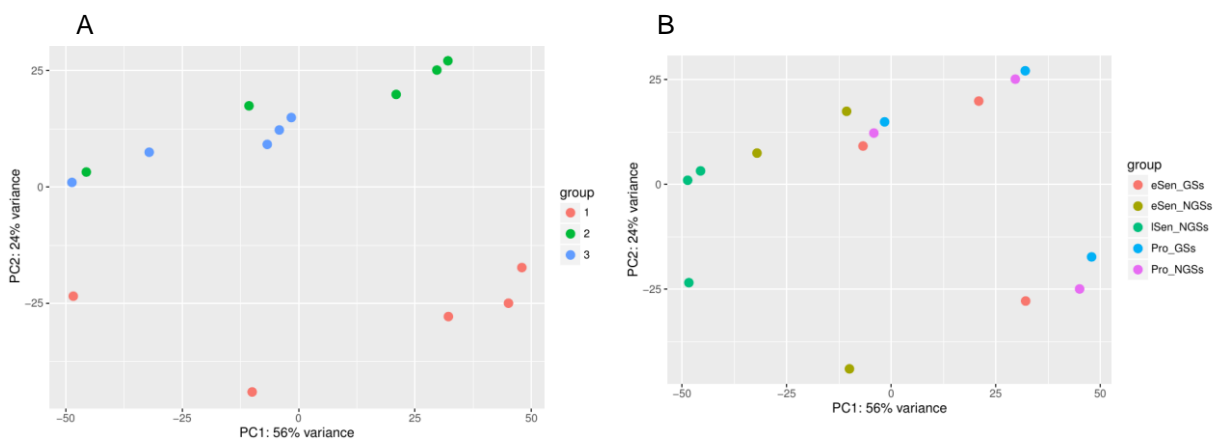


Figure 3-2. PCA plots of WI38 samples after SVA normalization coloured by batch number (A), and by biological sample (B).

A more powerful approach to reducing batch effect is running Combat batch removal (Appendix 10). This method provided a strong batch effect treatment as now the biological samples clustered without any batch bias (Figure 3-3 A and B). However, when Combat batch effect removal was applied only to the non-glucocorticoid treated samples, an even better batch effect removal was achieved (Figure 3-3 C).

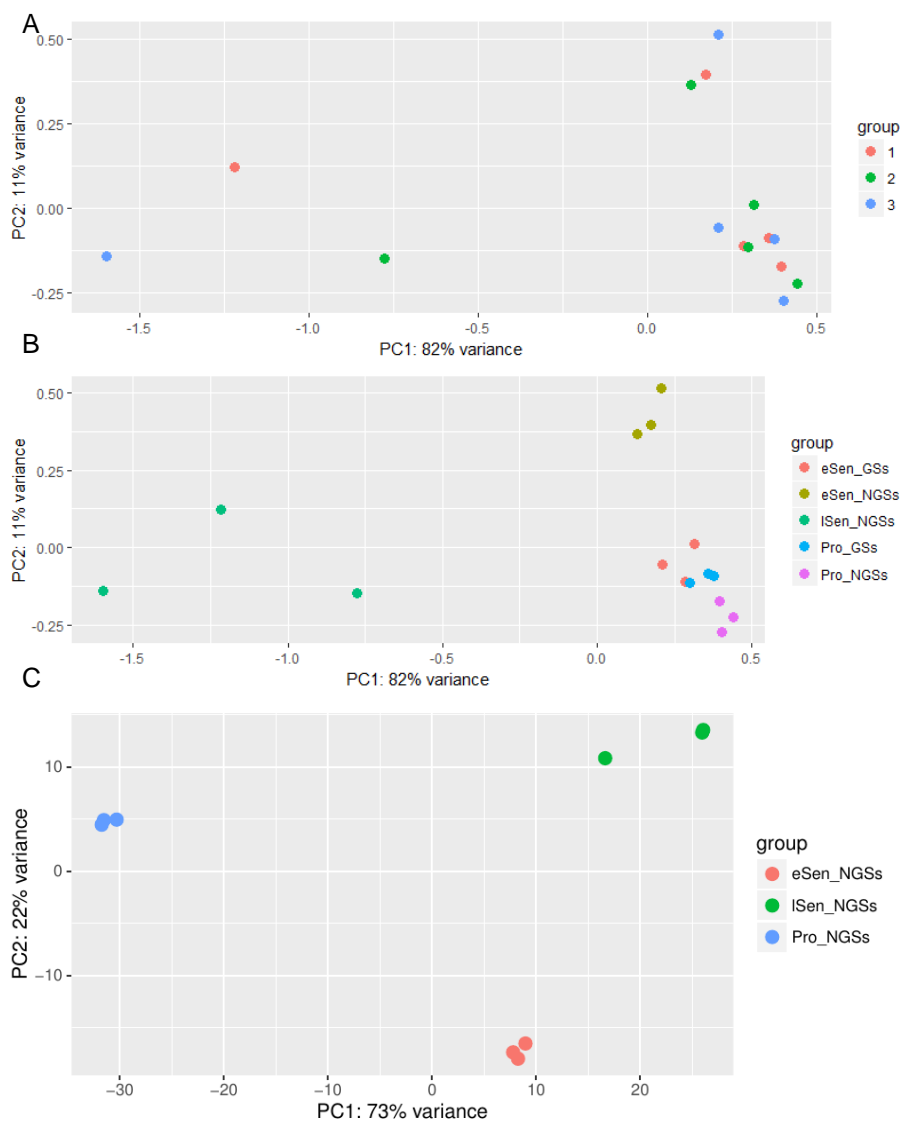


Figure 3-3. PCA plots of WI38 samples after Combat normalization coloured by batch number (A), and by biological sample (B). The same batch extraction performed without glucocorticoid treated samples showed an even better batch effect removal (C)

RESULTS

Principal components analysis (Figure 3-3 C) of WI38 experiment showed that experimental samples distinctly separated with small intra-sample variation. Principal component one explains 73% of the variation and places the samples in a temporal sequence with a wide gap between the proliferating and senescent cells.

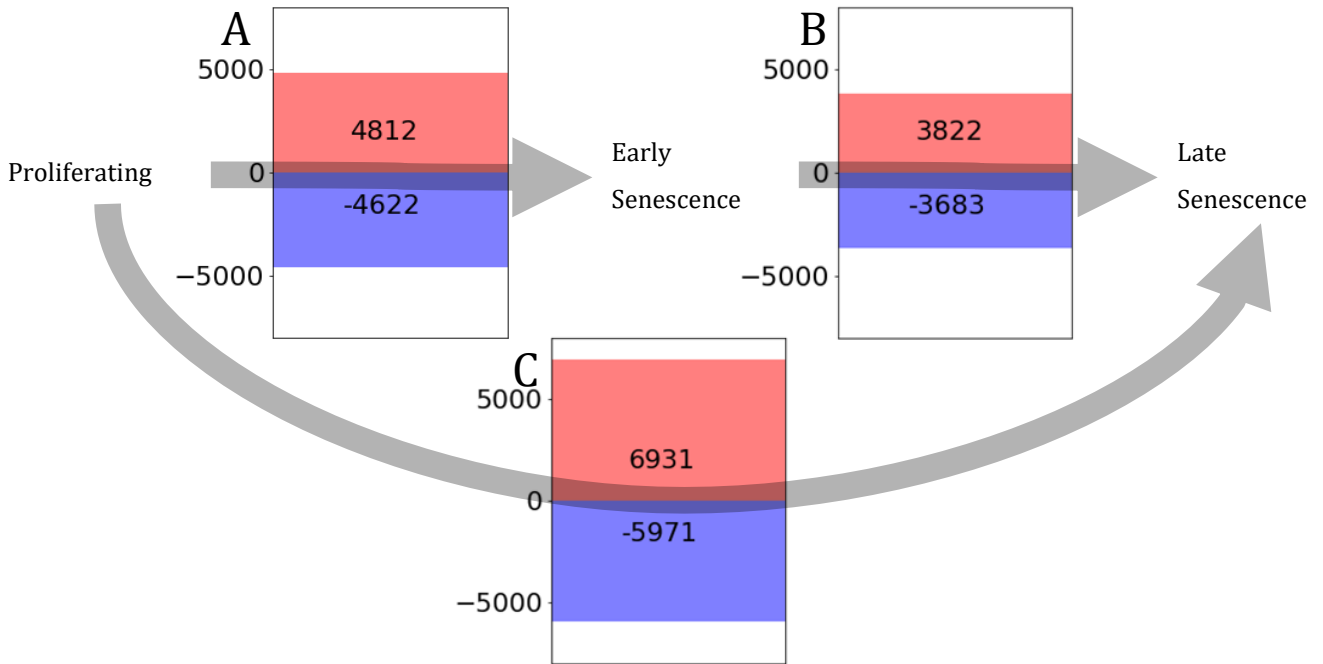


Figure 3-4. Bar charts representing the up regulated (red; $P\text{-adjusted} \leq 0.05$, $\text{lod2} > 1$) and down regulated (blue; $P\text{-adjusted} \leq 0.05$, $\text{lod2} < -1$) genes between time points, showing (A) genes altered between proliferating and early senescence, (B) early senescence and late senescence, and (C) proliferating to late senescence (Files [16]).

Differential expression of the WI38 studies show how many genes were up/downregulated between individual comparisons. Figure 3-4 shows over 9000 genes altered between proliferating and early senescence, circa 7000 genes altered between early senescence and late senescence, and circa 13000 genes altered between proliferating and late senescence. This means that there were still many differential expression changes occurring between early and late senescence.

If there is a large overlap between the genes altered between proliferating and early senescence, and between early senescence and late senescence, it would seem that changes between early and late senescence could be attributed to the number of senescent cells increasing. However, if the genes altered are different, a biologically distinct post senescent effect would explain the transcriptomic changes between early and late senescence.

Genes up/down-regulated between proliferating and early senescent were shared with up/down-regulated genes in proliferating vs late senescence (Figure 3-5 A) with 23.4% and 24% of significantly altered genes being upregulated and downregulated respectively in both comparisons. This indicated that altered genes between proliferating and early senescence tend to remain altered at late senescence.

Surprisingly, when comparing proliferating vs early senescence and early senescence vs late senescence (Figure 3-5 B), the percentage of genes being up/down-regulated in both comparisons fell to 7.9% and 16.1% respectively. This indicated that although senescence is irreversible, transcriptomic changes were still taking place.

Genes that were altered between proliferating and early senescence remained altered in late senescence (Figure 3-6 A, B). A variety of immune system genes were both upregulated and downregulated. Interestingly, apoptosis (R-HSA-109581) was found to be upregulated (not senescence) in late senescent cells.

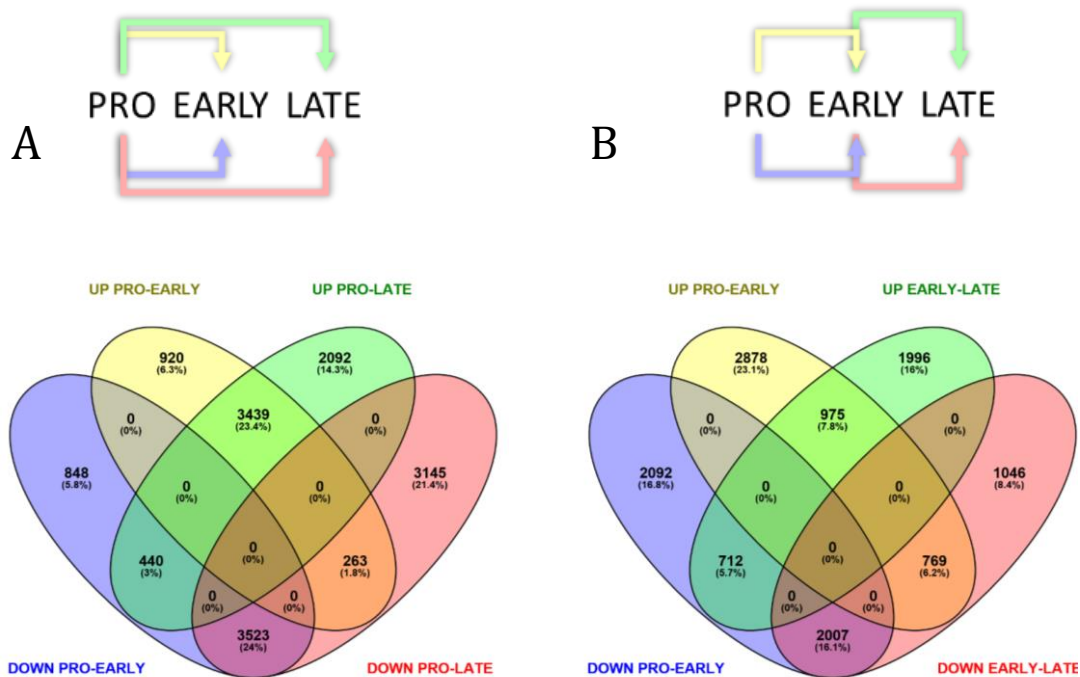


Figure 3-5. (A) Venn diagram showing the differences in upregulated and downregulated genes between proliferating cells (PRO) vs early senescent cells (EARLY) and proliferating cells vs late senescent cells (LATE). Each circle represents the significantly upregulated (top circles) or downregulated (bottom circles) genes of a specific comparison. (B) Venn diagram showing the difference between proliferating cells vs early senescent cells and early senescent vs late senescent cells. We see that the overlaps between upregulated and downregulated genes of the same category have by definition zero genes. Venn diagrams made using Venny¹⁹³.

Genes that were altered between proliferating to early senescence and early senescence to late senescence (genes that continued being upregulated/downregulated after early senescence; Figure 3-6), were found to correspond to nearly all Reactome enrichment from genes that were altered between proliferating and early senescence and remained altered in late senescence (Figure 3-6), this is because by definition they have to be included. However, WNT signalling (R-HSA-195721), and G-protein coupled receptors (R-HSA-500792) pathways appear to be misregulated as they appear both up and downregulated. Downregulation of platelet degranulation (R-HSA-114608) was also apparent between proliferating to late senescent cells.

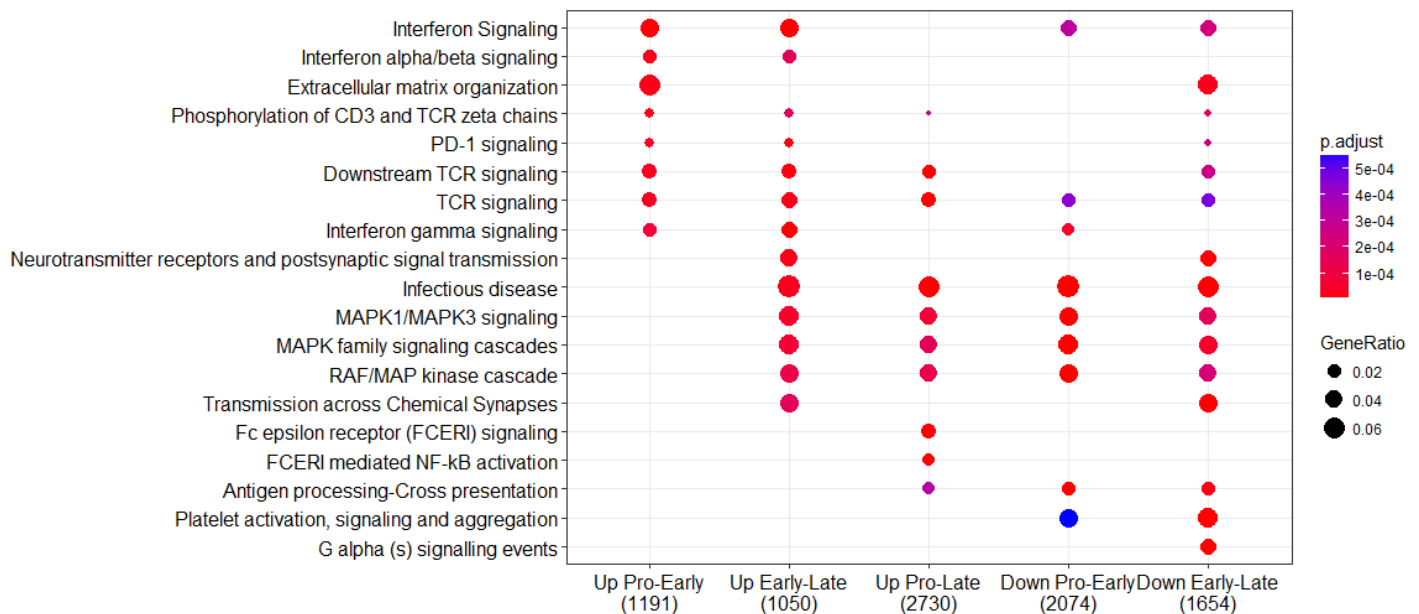


Figure 3-6. Reactome enrichment of genes that were upregulated in proliferating (Pro) vs early senescence (Early) and late senescence (Late). (File [10]).

When looking at splicing factors genes of WI38 (Figure 3-7; R-HSA-72203), we see a similar effect as we saw with IMR90. We see that some genes are downregulated going from proliferating to early senescence (Figure 3-7 B; SRRT, SRSF2, SNRPG, SRRM1, SNRPB, HNRNPR, DHX9, LSM4, HNRNPA1, HNRNPA3, TRA2B, POLE2B, HNRNPM, POLR2L, GCFC2, SRSF7, WBP11, NCPB1), and some genes are upregulated during late senescence (Figure 3-7 A; U2AF1L4, CHERP, SNRNP25, PRPF4, PLRG1, PCF11, PPIL6, TFIP11, BUD31, SRSF9, PCBP2, SNRPD3, SNRPB2, CSTF3, CSTF2T, CLP1, RMB22).

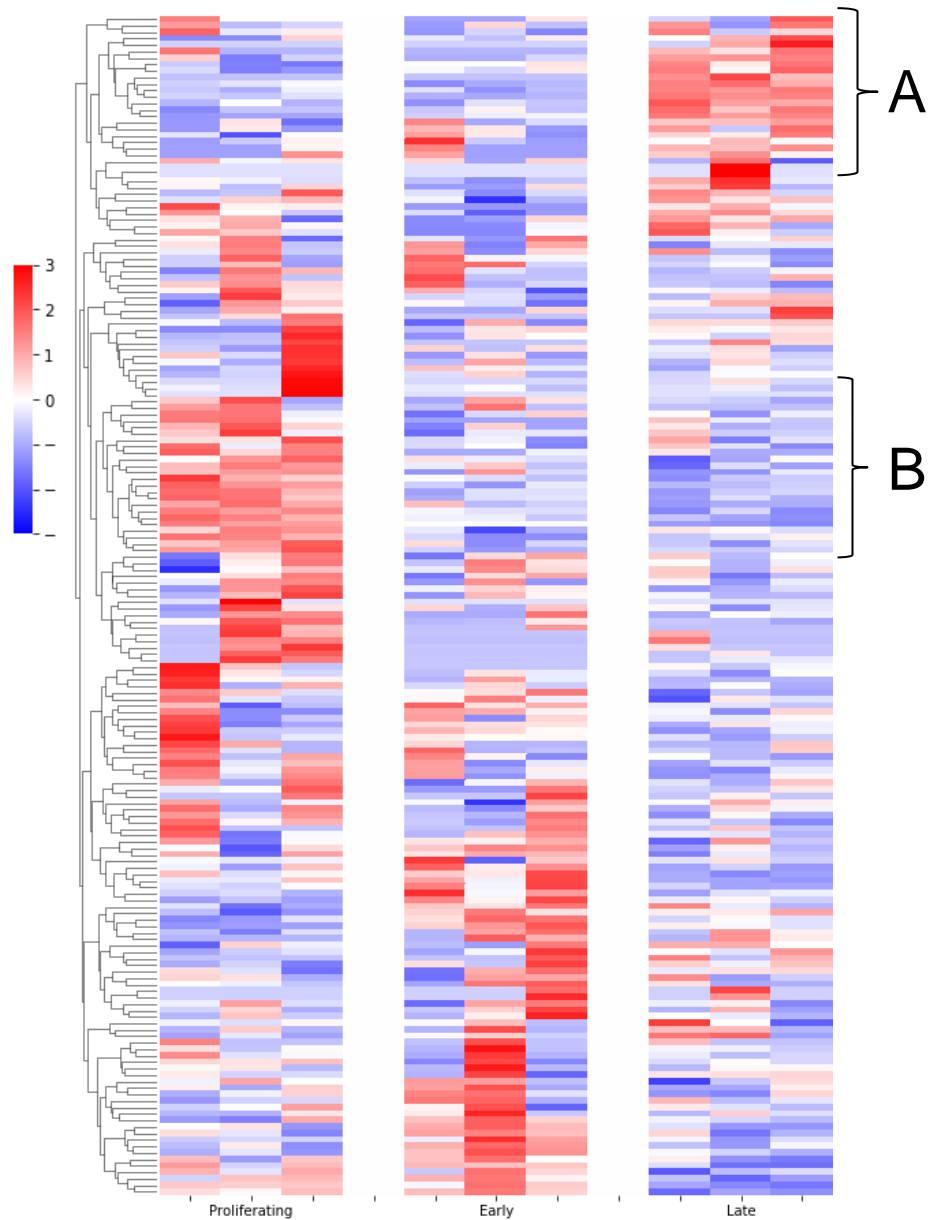


Figure 3-7. Z-scored heatmap showing splicing genes (R-HSA-72203) during proliferating, early senescence (Early) and late senescence (Late). Some genes are downregulated going into senescence (B) and some are upregulated during late senescence (A)

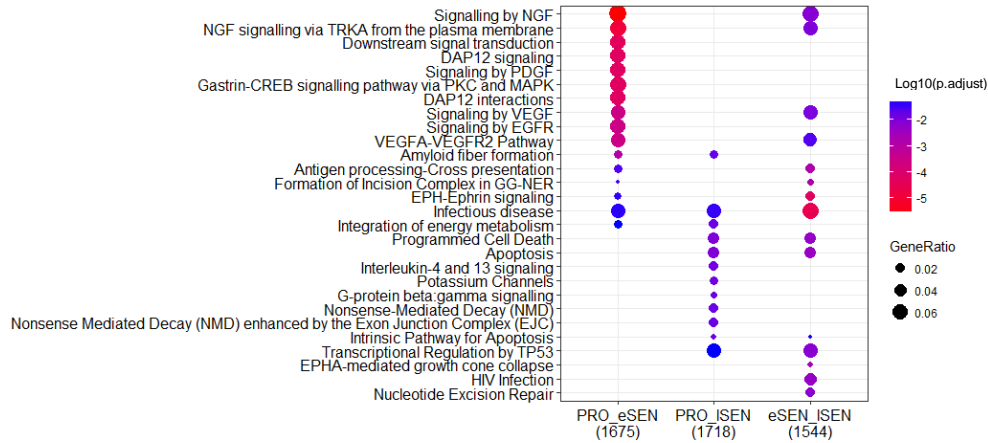


Figure 3-9. Reactome enrichment of spliced genes of WI38 cells (at least 0.05 P value and 0.05 inclusion level difference).

SPLICING

After seeing evidence of differential expression of splicing factors and spliceosome (Figure 3-7), we focused on splicing analysis of the WI38.

Splicing analysis of WI38 cells revealed a total of 382 significantly spliced events (Figure 3-8). But only 12 events were conserved between early senescence and late senescence: ZNF598, EDA2R, AFTPH, ARF1, ERBB2IP, ILKA, MIPOL1, CTSB, APBB3, FOPNL, DOCK7, and ENDOV. This list was too short to find significant gene ontology or Reactome annotations. These results showed that the splicing events which occurred between proliferating and early senescent cells were rarely conserved into late senescence.

Reactome enrichment of the rMATS results (Figure 3-9) showed that immune system annotations (Infectious disease (R-HSA-5663205)) were enriched in all comparisons. This meant that immune system genes were not only alternatively spliced and continued to be alternatively spliced after early senescence, but that their

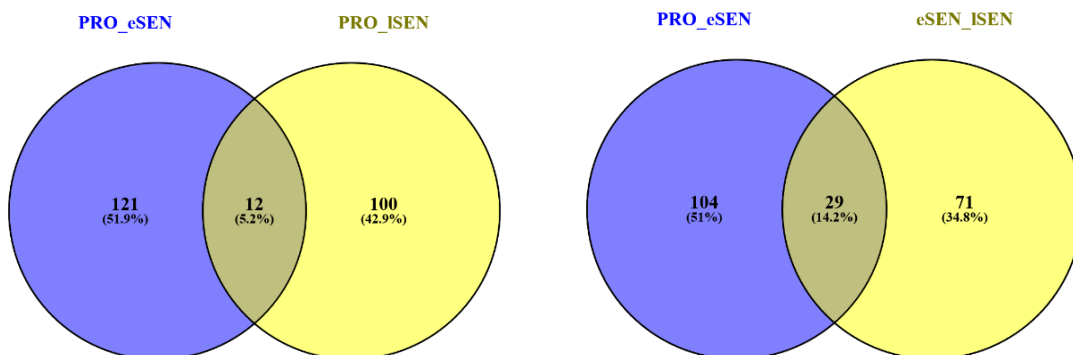


Figure 3-8. Venn diagram showing how many genes were significantly spliced between paired comparisons; proliferating to early senescence (PRO_eSEN) and proliferating to late senescence (PRO_ISEN). As well as proliferating to early senescence and early senescence to late senescence (eSEN_ISEN) in WI38 cells.

alternative splicing increased in late senescence (showing stronger enrichment in early senescence vs late senescence).

Energy metabolism was found to be enriched in both proliferating to early senescence and proliferating to late senescence, but not in early to late senescence. These results showed that alternative splicing of genes involved in energy metabolism splicing was mainly occurred in early senescence and was not further regulated between early to late senescence.

Splicing of apoptotic genes and nonsense-mediated decay was not strongly significant when compared with early senescence but was found to be strongly significant when compared with late senescence cells.

Splicing dysregulation was found in a variety of pathways at the induction of senescence, but not in late senescence. The pathways affected (NGF, DAP12, PDGF, Gastring-CREB, VEGF, and EGFR) showed an increase in isoform switch between proliferating and late senescence, which was then subdued in late senescence. It is possible that alternative splicing of genes in these pathways played an important role in cell changes during early and late senescence. This finding supports a hypothesis that a strong yet ephemeral splice alteration heralds senescence induction.

CONCLUSION

Postsenescence plasticity raises many questions about the nature of cellular senescence, as we tend to assume (and perhaps oversimplify) that senescence is a reasonably static terminal cellular state. If senescence is not static and there is still plasticity in the post senescence phenotype, this would raise questions about the nature of senescence. For example, do senescent cells/tissues/organs/systems behave differently after they have been senescent for a certain length of time? Do late senescence transcriptional changes affect the cells progression towards a certain outcome? Do post senescent transcriptional changes increase or decrease cellular homeostatic abilities?

DESeq analysis showed that late senescence was distinct from early senescence. This difference, however, is subtle, and perhaps consists mostly in increased deregulation of immune system genes. This hypothesis was supported by immune system Reactome annotations which were both upregulated and downregulated across the timeline past early senescence.

There seems to be a splicing effect as some splicing factors are differentially affected in different time points.

Unfortunately, this experiment did not contain enough time points to assess if increased biological replicates would cluster depending on early/late senescence or on a chronological progression. Such an analysis would distinguish between two distinct hypothetical possibilities (Figure 3-10). One possibility is that there is a distinct post-senescent pathway, and experimental repetition would show distinct clustering of proliferating, early senescence, and late senescence (Figure 3-10 A). However, if late senescent changes were due to senescence led misregulation, clustering would be segregated by proliferating and senescent cell types (Figure 3-10 B). However, senescence plasticity is a contributing factor in both hypotheses, in contrast to the current model where senescent cells become 'transcriptionally frozen'.

As there are very few studies showing late senescence alterations, it is not possible to determine how prominent and what role post senescence plasticity plays in senescent cells. More research is needed to determine if senescent plasticity only affects WI38 cells or cells across all eukaryotes (as seen in cotton leaves ¹⁶⁴). Further understanding into the role and prevalence of senescent plasticity is also medically relevant. For example, does a senescent tissue's function further deteriorate through post senescence plasticity? Could post senescence plasticity be

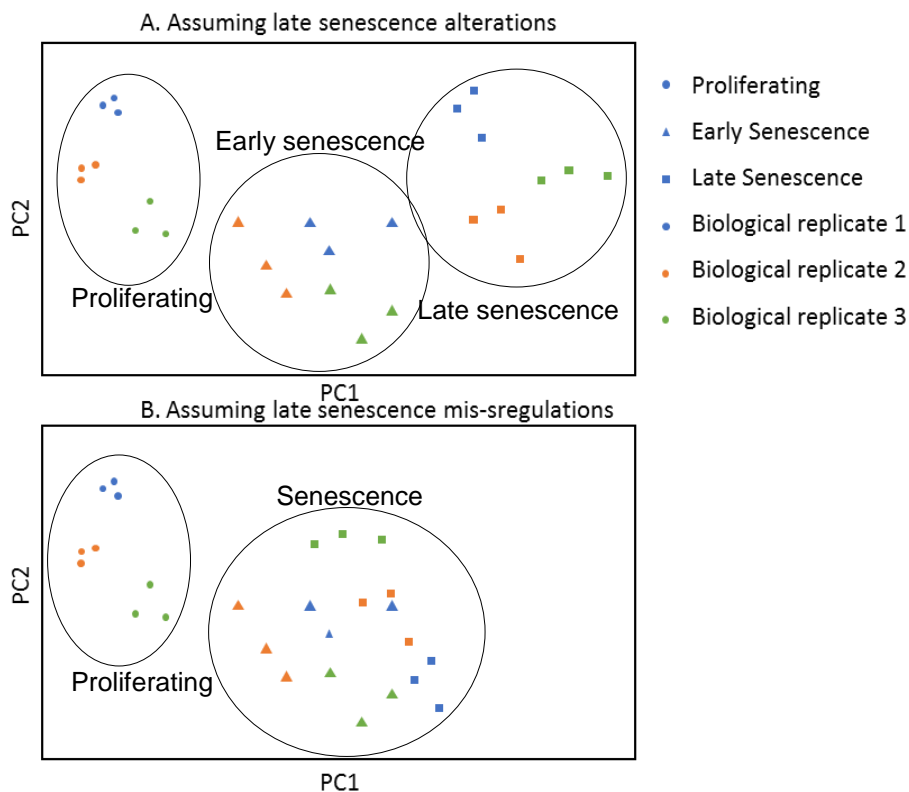


Figure 3-10. Hypothetical PCA plots of proliferating (circle) early senescence (triangle) and late senescence (square) for three biological replicates (red, orange and green). (A) if there was a distinct post senescence change. (B) if there was no post-senescent changes and the observed effects were due to mis-regulation.

used to improve a senescent tissues health? The answers to these questions can only be determined through new research. Replicating the study with an increased number of parallel biological replicates would show if there is a predetermined post senescence program or if the changes were due to missregulation. The study should also be replicated across different cells types to determine the prevalence of post senescence plasticity. Replication across different organisms would determine if this effect is common across living eukaryotes. This study has raised far more questions than it has answered.

WI38 show strong signs of senescence like the effects seen in the mice experiments. These include cell cycle arrest, immune system genes (SASPS), and splicing changes. There were strong splicing alteration with senescence induction

Non-splicing based analysis of WI38 cells (DESeq, Reactome, and network analysis) revealed that genetic expression patterns during the onset of senescence were not maintained in late senescence. Similarly, splicing analysis found splicing changes were not maintained during early and late senescence. Although immune system and energy metabolism genes remained spliced, there were many spliced pathways (NGF, DAP12, PDGF, Gastring-CREB, VEGF, and EGFR) which were only spliced during early senescence and not in late senescence.

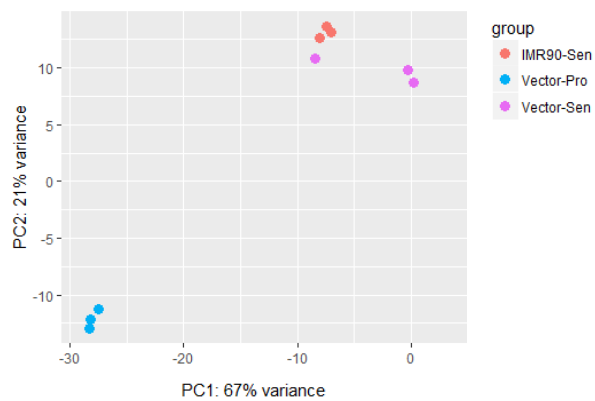


Figure 3-11. PCA analysis show that senescent cells with empty vector co-cluster with senescent without a vector.

IMR90 IONIZING RADIATION INDUCED SENESCENCE

INTRODUCTION

In a study conducted by Peter Adam's lab, aimed at elucidating the epigenetic role of the HIRA gene (HIR Histone cell cycle regulation defective homologue), IMR90 cells (human myofibroblasts foetal lung cells) and senescent HIRA mutant cells, were cultured and sequenced after undergoing ionizing radiation (IR) induced senescence. Peter Adam's lab cultured and sequenced three cohorts with three replicates each (Gene Expression Omnibus GSE53356). The samples we are interested are: senescent IMR90 without vector for control, senescent IMR90 with empty vector, proliferating IMR90 with empty vector.

In this section we study the effects of senescence on the transcriptome, and splicing, by interrogating IMR90 cells that were driven to irradiation induced senescence.

METHODS

The IMR90 data were aligned to the genome by STAR ¹⁶⁵ (page **Error! Bookmark not defined.**). Gene expression was counted using HTSeq ¹⁰⁸. Both senescent cohorts show co-cluster in principal component analysis (Figure 3-11). DESeq analysis of IMR90 (file [7]), showed little intrasample variation and large intersample variation between samples. 92% of the variance between proliferating and senescent samples could be attributed to senescence induced changes.

RESULTS

Reactome enrichment of IMR90 for up/down genes (Figure 3-12; file [8]), showed a very strong downregulation of mitosis Reactome annotations (M phase (R-HSA-68886), Mitotic Prophase (R-HSA-68875)), genome organization (Nucleosome assembly (R-HSA-774815), HATs acetylate histones (R-HSA-3214847), PRC2 methylates histones and DNA (R-HSA-212300), Telomere Maintenance (R-HSA-157579), and DNA methylation (R-HSA-5334118)). Upregulated genes were found to be enriched for cell death (TP53 Regulates transcription of cell death genes (R-HSA-5633008)), immune genes (Neutrophil degranulation (R-HSA-6798695), and Interferon alpha/beta signalling (R-HSA-909733)).

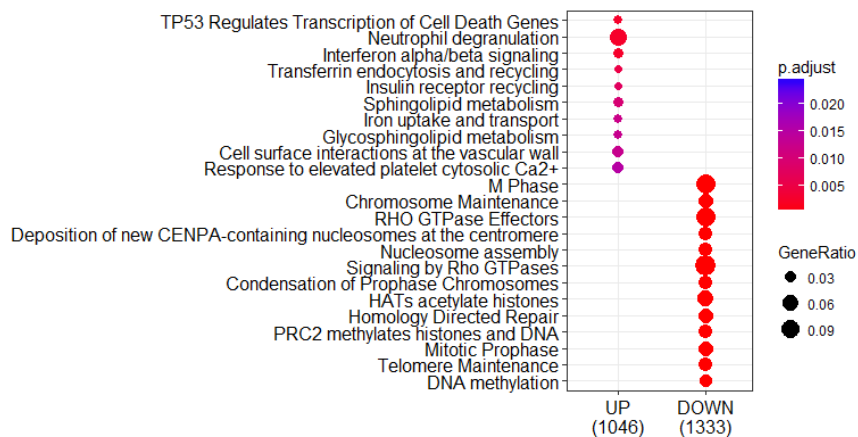


Figure 3-12. Reactome enrichment of upregulated and downregulated of IMR90 experiments.

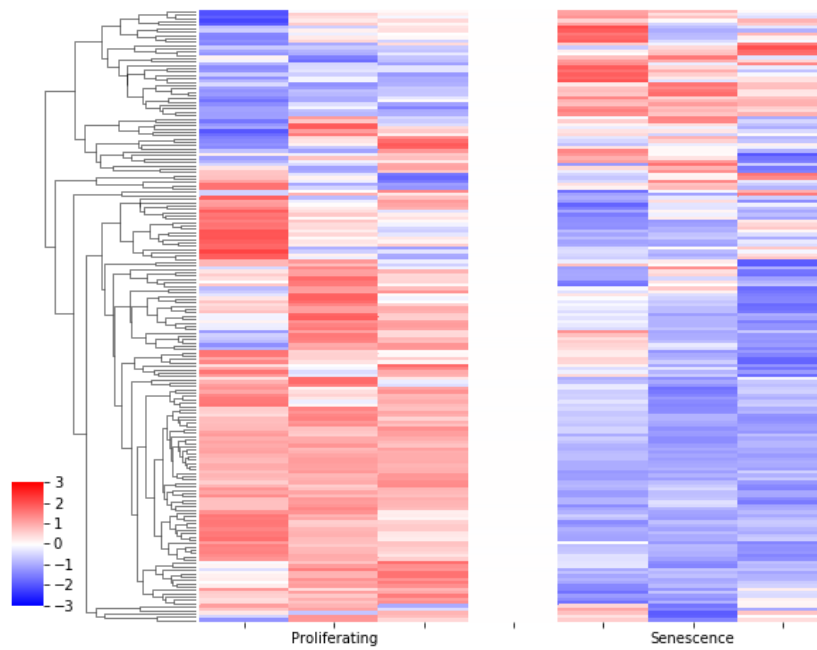


Figure 3-13. Z_scaled clustered heatmap of splicing genes of IMR90 (R-HSA-72203).

When looking at expression of splicing factor genes of the IMR90 experiment (Figure 3-13 ;R-HSA-72203), We see a large downregulation of splicing factor genes with a minority of them being upregulated, suggesting perhaps a decrease in splicing control.

SPLICING

After seeing evidence of differential expression of splicing factors and spliceosome (Figure 3-13), we focused on splicing analysis of the IMR90 datasets.

rMATS analysis of IMR90 senescence cells found 225 significantly spliced genes (filtered by P-value of at least 0.05 and inclusion level difference of at least 20%) between proliferating and senescent cells (Figure 3-14; file [9]). Energy metabolism and biological oxidation were highly enriched in the spliced events.

CONCLUSION

IMR90 show strong signs of senescence like the effects seen in the mice experiments. These include cell cycle arrest, immune system genes (SASPS), and splicing changes. There were strong splicing alteration with senescence induction.

Differential gene expression showed decreased mitotic and increased immune system gene transcription in IMR90. Splicing analysis revealed significant changes to metabolic pathways and an important change to an immune cell receptor CD44 increasing perhaps its activity in a compensatory fashion.

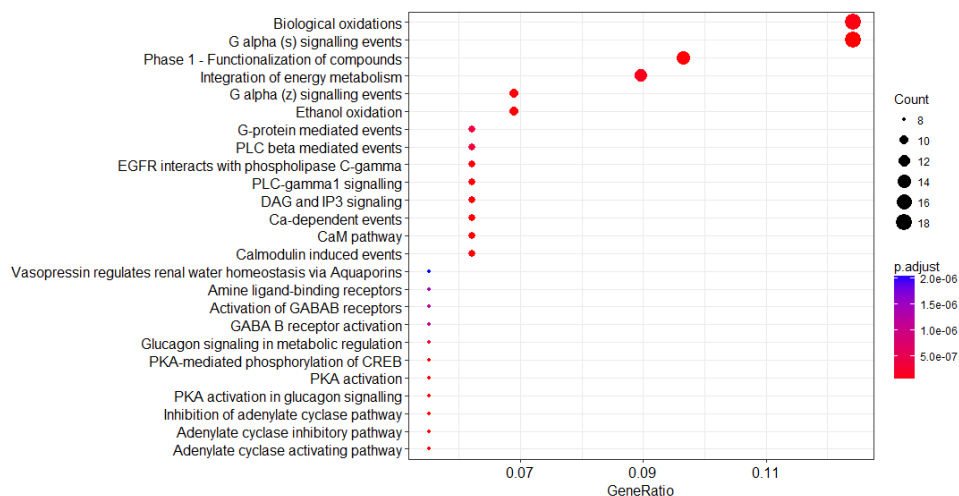


Figure 3-14. Reactome enrichment of significantly (P-value \leq 0.05 and inclusion level difference \geq 0.2) spliced genes of IMR90 during irradiation induced senescence.

CHAPTER 4 MALTESERS

Chapter objectives:

- To introduce a novel algorithm created to predict the effect of removal or addition of a functional motif on a protein after an alternative splicing event.
- Present MAItESERS, a novel tool for determining functional motif changes in alternative splicing.
- To review the application of MAItESERS and to study alternative splicing in a variety of senescence experiments.
 - In ageing mice and diet
 - Between proliferating, early senescence, and late senescence cells
 - In irradiation induced senescence cells

INTRODUCTION

Splicing changes often involve the inclusion or exclusion of a large domain. It would be beneficial to know how changes in alternative splicing affect a protein's function. Currently, each splice event is treated as an individual case. After a splicing event is identified, manual identification of the specific isoform is necessary. Then it is necessary to check if the isoforms mRNA processing is known in the Ensemble database ¹⁶⁶ (if the mRNA is processed into protein or not), or manually assert domain changes in the isoform changes, or do a literature search on the different phenotypes associated with each isoform. While the methods listed above can be productive, a single experimental treatment often results in hundreds of splice events. Researching each splice event individually becomes an impossible and heavily monotonous task. In addition to the large number of isoforms requiring manual identification, literature searches fail to provide functional results for novel splicing events.

There are some solutions available for getting functional changes of alternative splicing. MAISTAS developed in the instituto Pasteur Fondazione Cenci Bolognetti in the Sapienza University of Rome, it uses machine learning in human data to predict structural changes between different isoforms ¹⁶⁷. A similar tool SpliceDisease¹⁶⁸ uses a curated database to relate splicing changes with diseases, however.

A more practical way to summarize and study the functional consequences of splicing events. For this reason, we created MAItESERS. MAItESERS is a tool to annotate and analyse all splicing changes at once using the Prosite ps-scan ¹⁶⁹, and relate these changes to functional domain changes.

ALGORITHM

MALTESERS (Motif ALternative Exons Scanner Enrichment of RNA-Seq; Git repository available in github.com/aLahat/maltese; Figure 4-1), is a tool written in Python2.7 that systematically extracts the sequences of all exons enriched (through other tools such as rMATS ¹³⁷ or DEXSeq ¹⁰⁹) and their respective background genes

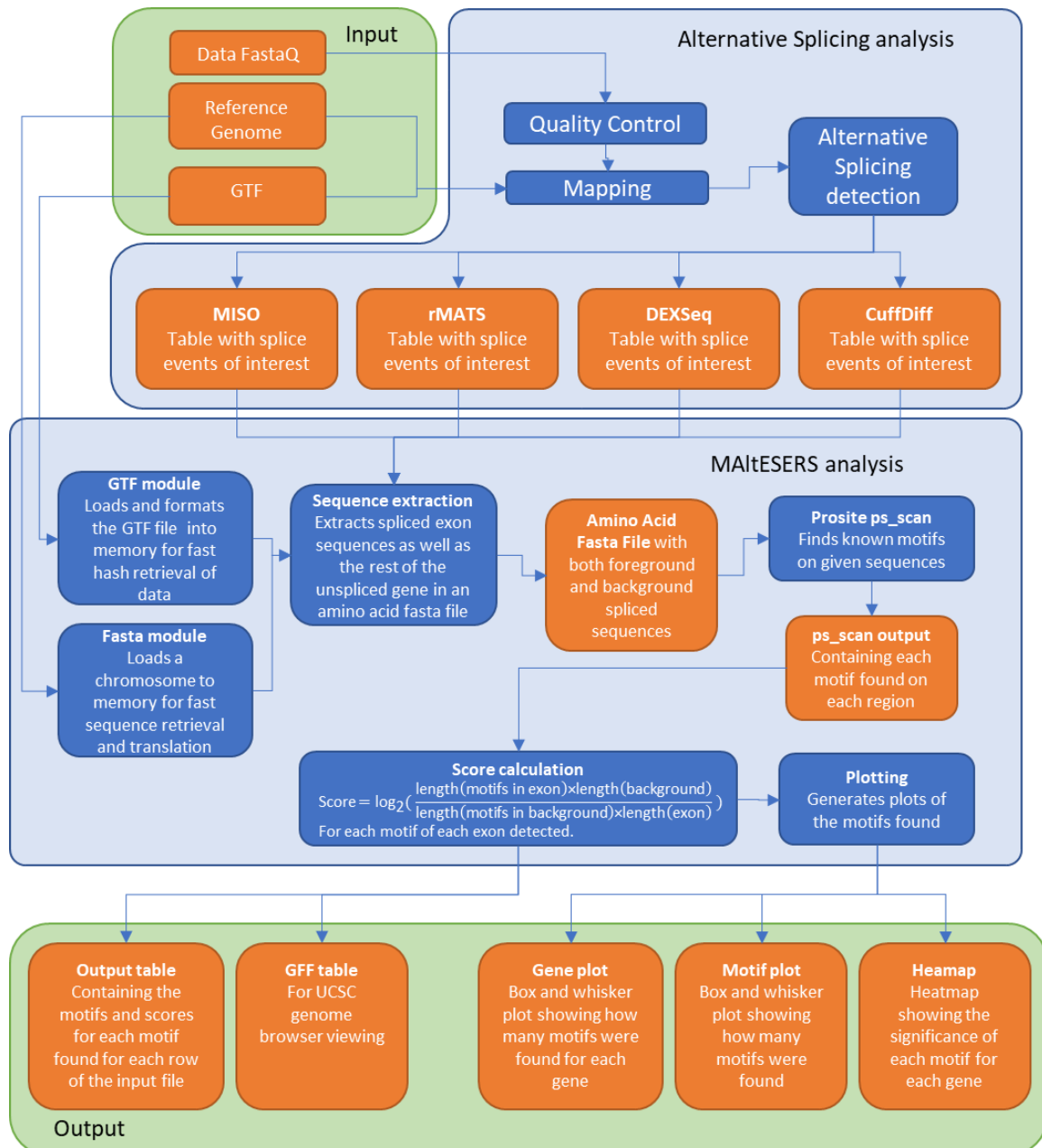


Figure 4-1. Diagrammatic representation of the MALTESERS algorithm. Each row of DEXSeq (a) output contains information about the coordinates of the exon (b). Using exon coordinates the sequence of the exon is retrieved, the sequence of the rest of the exons (sans the exon in question) are retrieved as well (c). ps_scan (d) is a tool that scans the sequences for known domains. The ps_scan output (e) is used to calculate the score for domains. The score is calculated by the fraction of the exon containing the motif divided by the fraction of background containing the motif (simplified in f). The output (g) is given for each row of the DEXSeq output. In the case where a motif is found in an exon and not in the background, it is returned with an 'N' (number) suffixing the number of motifs found in the exon.

and transcodes them from nucleotide sequences to amino acids. Once the genes and exons amino acid sequence are extracted into a FASTA format, MAItESERS finds known domains using the Prosite ps_scan tool ¹⁷⁰. Prosite_scan is an open source tool that matches known domain types with biological significance from the Prosite database ¹⁷¹ to a source FASTA file.

Once motifs have been found for each exon and background gene, MAItESERS counts the motif types for each exon and background and calculates the fold motif density enrichment:

$$Score = \log_2\left(\frac{\text{density motif in exon}}{\text{density motif in background}}\right)$$

$$Score = \log_2\left(\frac{\frac{\text{length(motifs in exon)}}{\text{length(exon)}}}{\frac{\text{length(motifs in background)}}{\text{length(background)}}}\right)$$

$$Score = \log_2\left(\frac{\text{length(motifs in exon)} \times \text{length(background)}}{\text{length(motifs in background)} \times \text{length(exon)}}\right)$$

If a motif in an exon has no representative in the background, then the score consists of the number of times the motif appears in the exon with an 'N' flag. This is important because if no motifs are found in the background (but are present in the exon), the result must not be discarded due to a simple division by zero error.

The output consists of the input table with rows prepended with motifs enrichments. In addition to the main output table, three plots are produced; a box and whisker plot for enriched motifs, a box and whisker plot for enriched genes, and a heatmap for enriched motifs of genes (Figure 4-2 B, C and D).

TESTING WITH PUBLICALLY AVAILABLE DATA

Public datasets were used to test MAItESERS. The dataset GSE59335¹⁷² was chosen from the gene expression omnibus¹⁷³. This study compared MDA-MB-231 cells treated with TRA2A/B siRNA to inhibit splicing versus the same cells without siRNA treatment. The raw FASTQ was trimmed and filtered using Kraken¹⁰¹, trimmed to 95pb, aligned using STAR ¹⁷⁴, and tested for splice events using rMAT¹³⁷ (Appendix 11).

MAItESERS was tested using only spliced exons with a significance of at least 5% and an absolute inclusion level of at least 85% (kept high to keep the results smaller for presentation).

MAItESERS outputs a table (Figure 4-2 A) as well as three plots (Figure 4-2 B, C, D). The main table output is based on the splicing table provided. For each motif

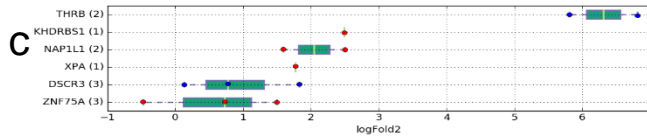
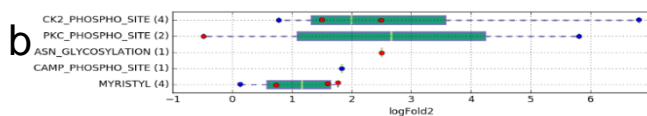
found on each row of the splice sites provided, there is a row in the output table. The output table contains a motif column with the motif name, log2fold density scores of the domain against the rest of the gene, P value calculated using a Fisher's exact test, the number of motifs found in the exon ('Motif Exon Count'), the length of the exon ('Exon Len'), the number of motifs found in the background gene ('Motif Gene Count'), the background gene length ('Gene Len'), and the rest of the columns are the original columns of the input file provided.

There are three plots outputted by MAItESERS. There is a table called [input file]_motif.[file format] (Figure 4-2 b) with an associated csv file which has the horizontal box and whisker plot for each motif found and the log2score and each dot is coloured by the inclusion level difference as well as a box and whisker representing the median, and the quartiles of the data distribution. This plot shows if some motifs are overrepresented regardless of the gene they were spliced from or into. Imagine as an example, that a specific domain is highly repressed or included in many genes. It would be difficult to notice if only looking at individual genes.

A similar plot [input file]_exons.[file format] (Figure 4-2 c) and its associated csv file represents the enrichment of motifs (regardless of motifs) for each exon spliced. This shows what genes are being affected the most by splicing. A third plot [input

a

motif	logFold2	Pvalue	Motif Exon Count	Exon Len	Motif Gene Count	Gene Len	Gene Symbol	chr	Strand	Exon Start_Obase	Exon End	PValue	Inc Level Difference
CK2_PHOSPHO_SITE	6.81	9.30E-08	4	32	140	126040	THRB	3	-	24297225	24297323	1.77636E-15	-0.933
PKC_PHOSPHO_SITE	5.81	2.98E-05	3	32	210	126040	THRB	3	-	24297225	24297323	1.77636E-15	-0.933
CK2_PHOSPHO_SITE	1.5	5.69E-04	16	174	146	4482	ZNF75A	16	+	3308312	3308836	8.02691E-14	0.865
ASN_GLYCOSYLATION	2.5	7.57E-03	4	34	337	16148	NAP1L1	12	-	76059776	76059878	1.12244E-13	0.909
CK2_PHOSPHO_SITE	2.49	8.36E-03	4	22	506	15673	KHDRBS1	1	+	32038551	32038619	3.7701E-12	1
MYRISTYL	1.77	1.33E-02	6	35	377	7482	XPA	9	-	97689533	97689639	0	0.907
MYRISTYL	1.59	2.22E-02	6	34	948	16148	NAP1L1	12	-	76059776	76059878	1.12244E-13	0.909
CAMP_PHOSPHO_SITE	1.83	3.07E-02	4	161	104	14833	DSCR3	21	-	37231492	37231975	5.22572E-09	-0.889
MYRISTYL	0.73	6.02E-02	18	174	279	4482	ZNF75A	16	+	3308312	3308836	8.02691E-14	0.865
CK2_PHOSPHO_SITE	0.78	1.53E-01	8	161	428	14833	DSCR3	21	-	37231492	37231975	5.22572E-09	-0.889
PKC_PHOSPHO_SITE	-0.48	5.83E-01	6	174	216	4482	ZNF75A	16	+	3308312	3308836	8.02691E-14	0.865
MYRISTYL	0.13	7.53E-01	12	161	1007	14833	DSCR3	21	-	37231492	37231975	5.22572E-09	-0.889



d



Figure 4-2. MAItESERS output. (a) original output lines are prepended with MAItESERS domains (blue), the log fold 2 significance, and the numbers of domains found in the exon and the rest of the gene. Each output line is repeated for each domain that was found but input lines with no motifs found are not present in the output (some columns were removed for display). (b) a box and whiskers plot showing log fold 2 scores of motifs found. The motifs found are also coloured by the change (Inc Level Difference). (c) a box and whiskers plot showing the log fold 2 scores of motifs by genes, also coloured by the change. (d) clustered heatmap showing the log fold 2 scores of motifs and their gene exon.

file]_motifExon.[file format] (Figure 4-2 d) and an associated csv file is a clustered heatmap of domain enrichment showing the most significantly enriched motifs as dark spots and enables to visually understand which domains are affected and by how much on the gene in question.

MAItESERS found some motifs that were significantly represented in the exons provided (Figure 4-2). This includes; cAMP and cGMP dependent protein kinase phosphorylation site (CAMP_PHOSPHO_SITE), Casein kinase II phosphorylation site (CK2_PHOSPHO_SITE), Protein kinase C phosphorylation site (PKC_PHOSPHO_SITE), N-glycosylation site (ASN_GLYCOSYLATION), and an N-myristoylation site (MYRISTYL).

MAItESERS found two highly enriched and significant phosphorylation sites (CK2 phosphorylation with $P=9.3E-8$ and PKC phosphorylation with $P=3E-5$) in the thyroid hormone receptor B (THRB). This hormone receptor needs to be phosphorylated in order to function ¹⁷⁵ and so this splicing event might be important in the cell's response to thyroid hormone. Even though this motif was enriched by two types of phosphorylation sites, it is important to note that the motifs for Casein kinase II and Protein Kinase C phosphorylation motifs ([ST]-x(2)-[DE] and [ST]-x-[RK] respectively) cannot overlap and are mutually incompatible. Therefore, both domains are independent of each other and do not represent the same domain mischaracterized as two plausible domains.

MAItESERS also found splicing changes at a significant CK2 phosphorylation site in p68 (KHDRBS1; $P=0.008$) which is activated by phosphorylation ^{176,177}.

TESTING WITH MICE AGEING DATA

Following the splicing analysis of the mice ageing experiments we use MAItESERS for further analysis of the rMATS data (Figure 4-3; looking for rMATS events with a significance of at least 0.001 and a differential inclusion level of 50%; splicing example in (Appendix 6)). MAItESERS identified 395 significant motifs in various splicing events. The motifs returned by MAItESERS were significantly differentially spliced, in a treatment dependant manner. For example, a bipartite nuclear localization signal (NLS_BP; domain needed for a protein to be translocated into the nuclei) was highly enriched in an exon of Bicc1 (RNA binding protein), that was spliced out significantly at AL15 but not in AL24, Young, or DR15. According to COMPARTMENTS subcellular location database¹⁷⁸, Bicc1 is present in the nucleus, therefore losing its nuclear localization signal might change its functionality. Another interesting example was Dennd1b a T-cell receptor that is alternatively spliced out

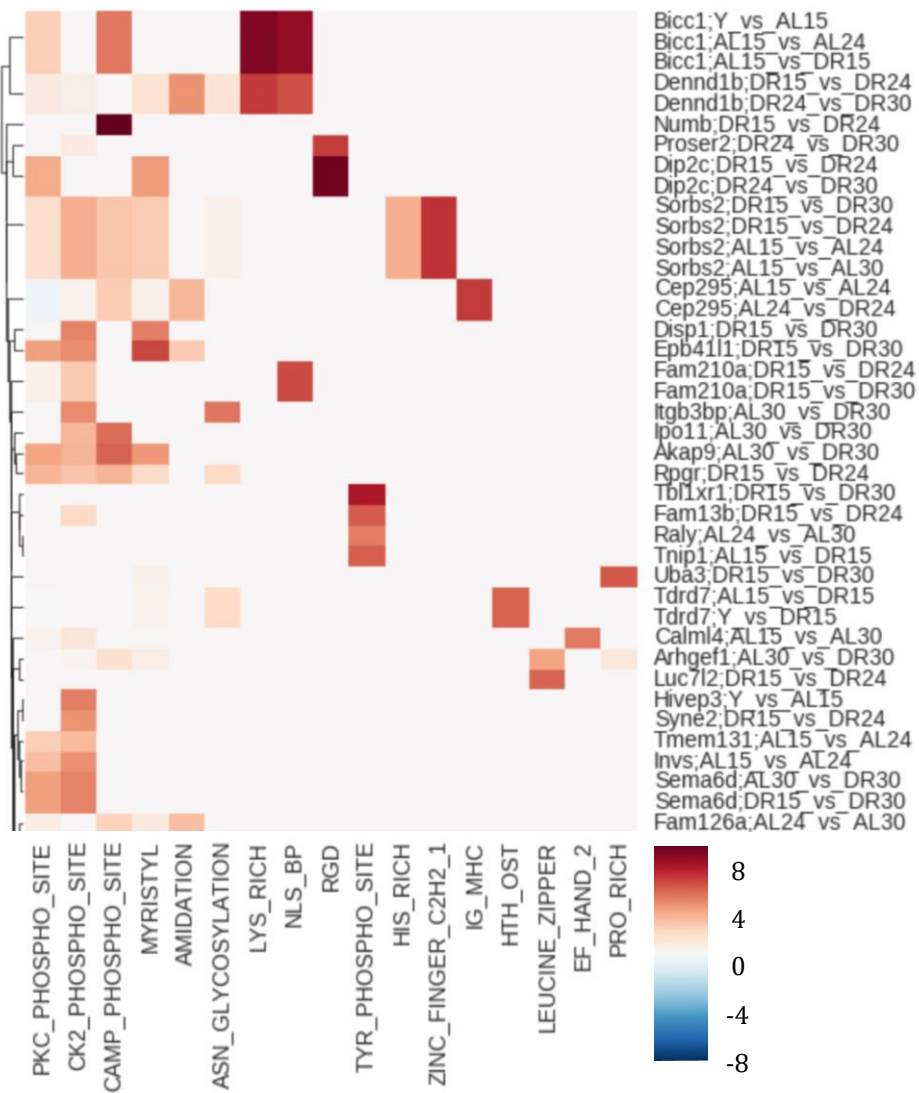


Figure 4-3. Top 40 MAItESERS results heatmap showing the log₂ (density difference of motif between spliced exon and rest of gene; “Y” samples are young 3-month AL).

in DR24 (compared with DR15 and DR30). The alternative spliced exon contains a phosphorylation site. Removal of a phosphorylation site is likely to be relevant for the protein's function, since *Denn1b* is activated by phosphorylation, DR24 mice might lack *Denn1b* function. This would lead to decreased immune sensitivity in DR. *Dip2c* (Disco Interacting Protein 2 Homolog C) has an exon containing an RGD tripeptide which is crucial for cell surface receptors. MAltESERS found the *Dip2c* protein to have increased RGD tripeptide splicing in DR24. *Dip2c* is located in the plasma membrane¹⁷⁸ and so this exonic loss might impair its function. Another interesting spliced gene found in MAltESERS was *Sorbs2*, an adapter protein that links ABL kinases to the cytoskeleton. MAltESERS found increased excision of the *Sorb2* zinc finger domain with age regardless of diet.

Differential gene expression changes (page 43) showed few genes significantly altered past 24 months of age. However, rMATS returned many changes in alternative splicing between both age and diet comparisons. Lipid transport genes were spliced when comparing 15 months DR to 30 months DR and in AL at 30 months. This means that lipid splicing might be affected by diet.

rMATS analysis also showed a splicing of spliceosomal genes between 24 to 30 months, this might be a cause of the continual splicing effects past senescence. Splicing changes to splicing genes also occurred between AL to DR (could this compensate for ageing effects?).

In conclusion, rMATS showed significant changes to lipid transport and splicing genes in both ageing and diet.

TESTING WITH IMR90 DATA

MAItESERS analysis of IMR90 splicing data () revealed many significant domains altered by irradiation induced senescence (Figure 4-4). ZNF207 is a kinetochore and microtubule binding protein with DNA binding functions. Alternative splicing and excision of a DNA binding leucine zipper in ZNF207 was found with senescence. C5orf42 is a ciliary protein whose activation is dependent on phosphorylation¹⁷⁹. C5orf42 gained a phosphorylation site with senescence. USP21 is a deubiquitinase, that stabilizes Nanog (an important protein that maintains pluripotency of stem cells) when phosphorylated¹⁸⁰. During senescence, USP21 loses a phosphorylation site which may affect stem cell pluripotency. CD44 is an important cell receptor that allows leukocyte migration during inflammation when phosphorylated¹⁸¹. During senescence MAItESERS showed that CD44 phosphorylation sites were increased in comparison with proliferating cells. This might facilitate inflammation in tissues.

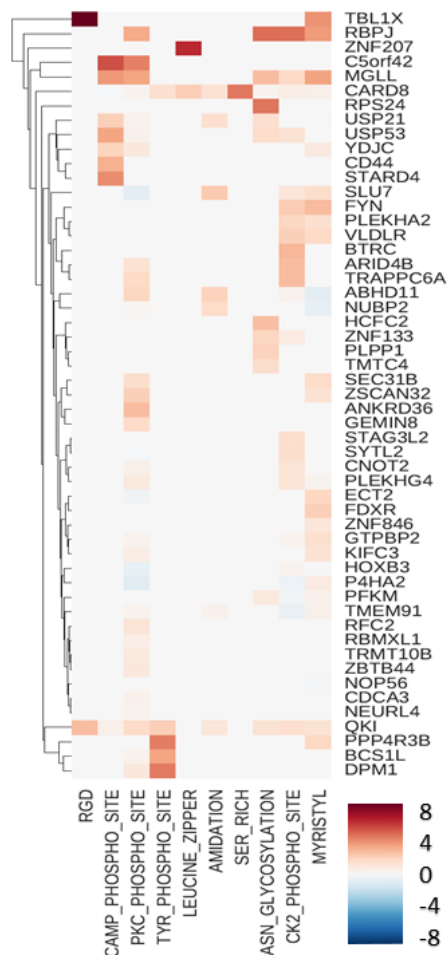


Figure 4-4. MAItESERS results heatmap showing the log2(density difference of motif between spliced exon and rest of gene).

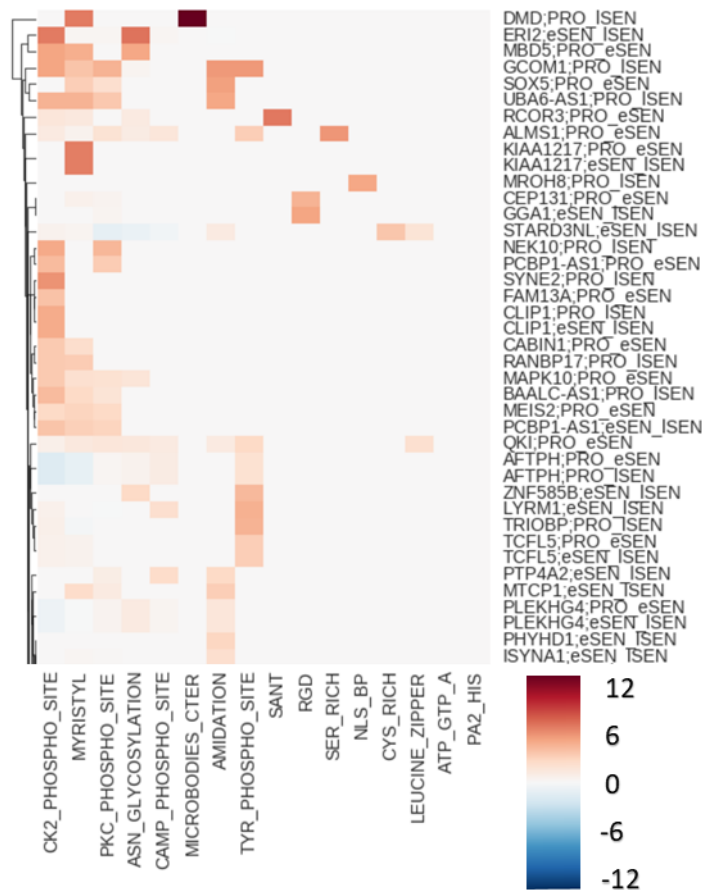


Figure 4-5. Top 40 MAltESERS results heatmap showing the log2(density difference of motifs between spliced exons and the rest of the gene).

TESTING WITH WI38 DATA

Following the splicing analysis conducted on the WI38 datasets (page 67), we run MAltESERS (Figure 4-5; file [10]) which returned 303 enriched spliced domains. Only enriched spliced domains found to be maintained in late senescence were analysed further. Two of the most significant domains included pertained to Clip1 and AFTPH. CLIP1 regulates microtubule dynamics by binding to the microtubule plus end. CLIP1 is activated by mTOR dependant phosphorylation. MAltESERS found the CLIP1 isoform with a phosphorylation site to be more prominent in senescence and maintained in late senescence. Alternative splicing of the CLIP1 phosphorylation site may regulate the mTOR downstream effects on the cytoskeleton. AFTPH (Aftiphilin) is a vesicle binding and transport protein, which gains a tyrosine phosphorylation site in senescence. It is not known if this AFTPH is phosphorylated or not, and the role of AFTPH in senescence is currently unknown.

In the WI38 senescence and late senescence study, Maltersers found an interesting target of regulatory splicing, CLIP1. MAltESERS found phosphorylation sites of

CLIP1 to be more prevalent in late senescence. CLIP1 binds vesicles to microtubules and needs to be phosphorylated by mTOR to be active. This gain in phosphorylation site containing isoforms might increase CLIP1 dependant vesicle transport during senescence for the senescent associated secretory pathway.

CONCLUSION

MAItESERS systematically finds possible motifs affected by splicing. Some of these motifs are already present in the literature but many putative motifs are not annotated yet. MAItESERS returns putative motifs that are lost or gained during a splice event. Therefore, MAItESERS can be used to predict the effects of splicing on a protein. However, MAItESERS likely outputs false positives which are impossible to disprove unless experimental techniques are utilized. For example, we would need to prove that a putative phosphorylation motif is in fact a developed phosphorylation region. However, MAItESERS is a useful tool for initial assessment regarding the functional effects of alternative splicing on a gene in a more practical fashion.

A challenge in bioinformatic data analysis is to find important factors in gigabytes of data. This is extremely hard when focusing on one datum at a time, as it would take weeks to scan through all significant events. In addition, scanning data manually is open to a large number of biases, such as cherry picking, or simply disregarding relevant results out of sheer exhaustion. MAItESERS significantly decreases the number of significant events worth investigating manually.

CHAPTER 5 CONCLUSIONS

Senescence Trifecta

Transcriptomic changes across all experimental data sets showed three major senescent effects; immune system (including Senescence-Associated Secretory Pathway), chromatin structural changes, and metabolic changes. These changes were observed in the WI38 post senescence plasticity experiment (immune system mostly), and the IMR90 irradiation induced senescence experiment (chromatin structure). Splicing analyses also showed significant isoform expression in genes for chromatin structure and metabolic changes in ageing mice. In addition to this, immune system genes were increasingly spliced in WI38 experiments, and biological respiration gene regulators were found spliced in IMR90 senescence.

As argued previously (page **Error! Bookmark not defined.**) each component of the biological trifecta is affected by the other two components. These three senescent factors may work in concert to reinforce each other and ensure the irreversibility of senescence. However, a proper modelling study needs to be conducted to properly ascertain this theory.

Senescence is important for inactivating cell division of cells with DNA damage which might become tumorigenic¹⁸². In order to stop mitotic activity in oncogenic cells, senescence has to be irreversible (although experimental exceptions do exist¹⁸³). For a certain cellular (or any system) state to be irreversible, there must be redundant mechanisms to inaction (akin to using more than one lock to secure a door). If there is DNA damage which might compromise one of the trifecta pathways, senescence could still be established by the other two reinforcing nearly independent mechanisms. After all, if senescence was only maintained by one pathway, a single well-placed mutation might reverse it and lead to a pathogenic tumorigenic phenotype.

It is important to consider the complexity of all the pathways involved indirectly in senescence. It is possible that these observations are based on Reactome biases. As both metabolism (contains 2116 genes) and immune system (contains 2229 genes) are amongst the largest Reactome classes. However, there are numerous experimental studies that support our thoughts. Systemic inflammation is common with ageing^{69,184,185}, as is metabolic changes¹⁸⁶.

Ageing is an Event or a Process?

The results of the mice ageing experiments propose an interesting question. Senescence induction was found to appear in a reasonably short period of time. The abrupt transcriptomic changes corresponded to a decrease in mouse survival rates.

These results were consistent with a sudden (possibly organism wide) onset of senescence. Considering this finding, it may be more accurate to view the cause of ageing as a sudden shift in cellular transcriptomic activity, instead of a gradual degradation in health over time.

One likely mechanism for sudden senescence induction is the senescence bystander effect ⁵⁷. The bystander effect occurs as the senescent secretory associated pathway initiates the release of inflammatory factors which can induce senescence in nearby cells. This effect might induce a rapid organ senescent blooming in tissues. This is because when cells become senescent in a tissue, the non-senescent cells are more likely to follow suit. Previous experiments which induced apoptosis in early senescent cells, found an increase in fitness and health in the mice ¹⁸⁷.

Currently, the dynamic of continual ageing in tissues past senescent induction is unknown. For example, if a liver becomes mostly senescent at an age of 50 years, will the liver tissues remain transcriptomically similar at an age of 80 years (besides accumulation of metabolic waste)? could a sudden senescence onset produce a gradual phenotype onset?

DESeq analysis was conducted on replication induced senescent cells to gain insight into the amount of transcriptomic plasticity post senescence. The mice ageing and WI38 trans-senescent experiments showed that the cells had changes between early and late senescence. Both WI38 and mice hepatic cells revealed that past senescent induction, the splicing landscape was still being altered (seen in both WI38 and mice liver) as well as the transcriptome landscape (seen only in WI38). It is possible that transcriptomic plasticity post senescence is universal in multicellular organisms (although very few late senescence experiments have been reported).

It was not evident if the late senescence pathway was predefined or resulted from missregulation. If it is based on missregulation, then senescent tissues will gradually deviate from the senescent phenotype. This could lead to an eventual tissue/organ/systemic failure leading to age-related mortality.

A plausible explanation for this event would be a survivor effect. As the study did not follow individual mice, as the mice aged, the probability of obtaining samples from exceptionally healthy mice increased. Meaning that the longest-lived mice were disproportionately healthy compared with the young mice. Therefore, it could be possible that a survivor effect would alter the results. It is however important to consider that all the mice were genetically identical and were kept in the same conditions in order to minimize any individual sample difference.

As ageing is perhaps the most common inducer of human morbidity and mortality, there is a strong movement to treat it as a disease. Treatments to ageing could possibly lead to biological immortality in humans¹⁸⁸

. However, with consideration for the senescent blooming hypothesis, speculative antiaging senescence targeting therapies would have little effect on patients who have already gone through the senescent blooming. Therefore, even if antiaging senescent targeting therapies appear, the aged demographic will be unable to receive it. As expressed earlier (page 85) senescence appears to be maintained by three nearly independent pathways. Therefore, reversal of senescence would involve navigating a complicated and multiplexed network of cellular interactions.

Future Research Questions

Post Senescence Plasticity

The observation of post senescence plasticity, raises three questions; how prevalent this effect in different tissues or taxa is, whether these changes are due to missregulation or a defined late senescence differentiation pathway, and can these changes lead to a pathogenic phenotype leading to tissue/organ/system failure?

The prevalence of late senescence plasticity can be addressed by screening transcription changes in proliferating, senescent, and late senescent cell lines derived from as many organisms and tissues as possible. The falling cost of sequencing would enable the analysis of more cell lines within budgetary constraints. The result would be apparent in a simple PCA plot (Figure5-1) where tissues or organisms with post senescence plasticity would have non-overlapping early and late senescence clusters, while cells with no late senescence plasticity would have overlapping early and late senescence clusters.

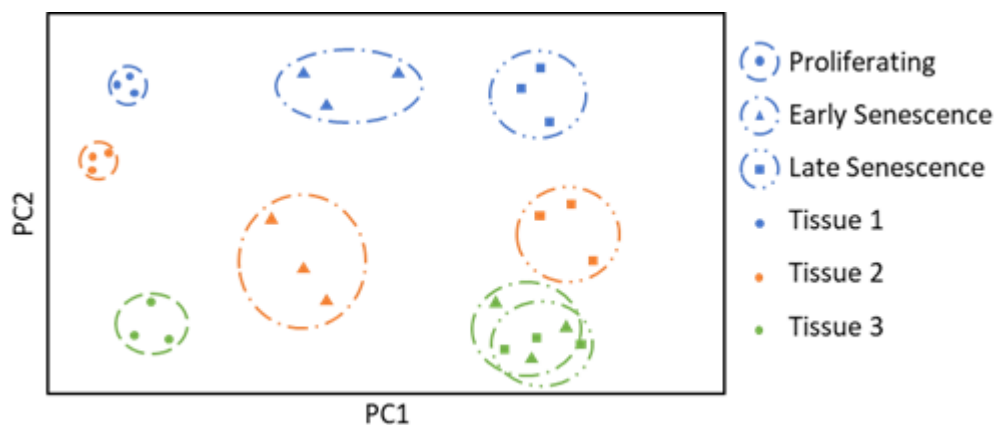


Figure 5-1. PCA showing hypothetical results for postsenescence plasticity study. Here the ageing effect is seen in the principal component 1. Both tissues 1 and 2 show post senescence plasticity, (non-overlapping early and late senescence). However, tissue 3 shows no senescence plasticity as early and late senescence overlap.

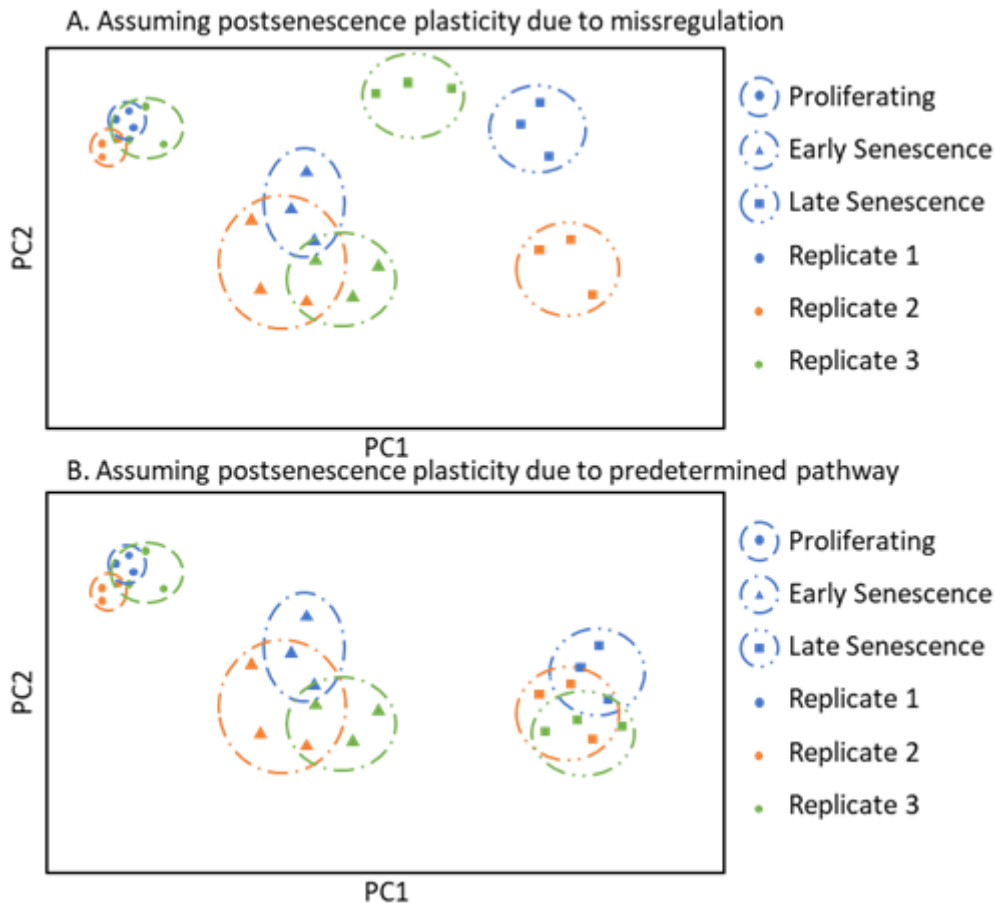


Figure 5-2. Hypothetical PCA plots to study if late senescence plasticity is due to missregulation (A) or a biological predefined developmental pathway (B)

In order to determine if late senescence plasticity is a predetermined differentiation pathway or due to cumulative missregulation, a similar experiment can be conducted where clonal tissues are grown in the same conditions independently (Figure 5-2). We would expect to see late senescence expression in a PCA plot diverge if the plasticity is due to missregulation (Figure 5-2A), or cluster together (albeit with some batch effects) if the changes were predetermined in the Waddington developmental landscape (Figure 5-2B).

Perhaps a more difficult and important question is whether late senescence plasticity is pathologically relevant. This might be difficult to study as we would need to know in detail the cause of a specific age-related organ failure and whether it was triggered by a gene affected during senescent plasticity.

Senescence Trifecta Modelling

To corroborate this triple senescent lock model in which the senescence trifecta acts as a way to impede senescent reversal, a very careful and complex model needs to be created using system biology techniques. As some level of transcriptomic data is available from the mice experiments and interactions between different pathways

are available in Reactome ¹¹⁴, a top-down approach can be used for modelling. If the model is successful, the ultimate test is to find targets on each of the three pathways to create a cell line with reversible senescence.

This is undoubtedly a difficult project, but it might open the door for development of senescence reversal clinical treatments which could significantly extend lifespan (especially healthy lifespan) on patients, while minimising increased risk of cancer.

APPENDIX

Appendix 1. Age Alteration of Metal Landscape	97
Appendix 2. Microscopy analysis of mice tissues	99
Appendix 3. GO enrichment for cluster fished genes for crossover mice senescence	100
Appendix 4. Regulation of Mitochondrial Gene Expression by tRNA Availability .	101
Appendix 5. Long non-coding RNA study on mice aging data.	103
Appendix 6. Sashimi plot showing detected splicing example of mice ageing data.	105
Appendix 7. Sashimi plot example of MAltESERS analysis of ageing mice.....	106
Appendix 8. WI38 differential expression without bath correction	107
Appendix 9. WI38 differential expression with sva BATCH correction	108
Appendix 10. WI38 differential expression with Combat correction.....	109
Appendix 11. Makefile used for testing MAltESERS	110

APPENDIX 1. AGE ALTERATION OF METAL LANDSCAPE

During analyses of the mice dietary longitudinal study, changes in metalloprotein expression were found (Figure 0-1). This might lead to alterations in metal regulation both intra or intercellular. Specifically, iron genes were found to be altered with age (except Aconitase2 which was upregulated in dietary restriction). Magnesium was also strongly affected by age and diet. It is known that iron ^{189,190}, magnesium ¹⁹¹ missregulation and accumulation happens in senescent cells.

In order to investigate the change in the metal landscape with ageing, biological samples extracted from the mice would need to be analysed for metallic ion concentrations. However, as the mice longitudinal study had already finished, there were few samples left from those mice to perform additional tests.

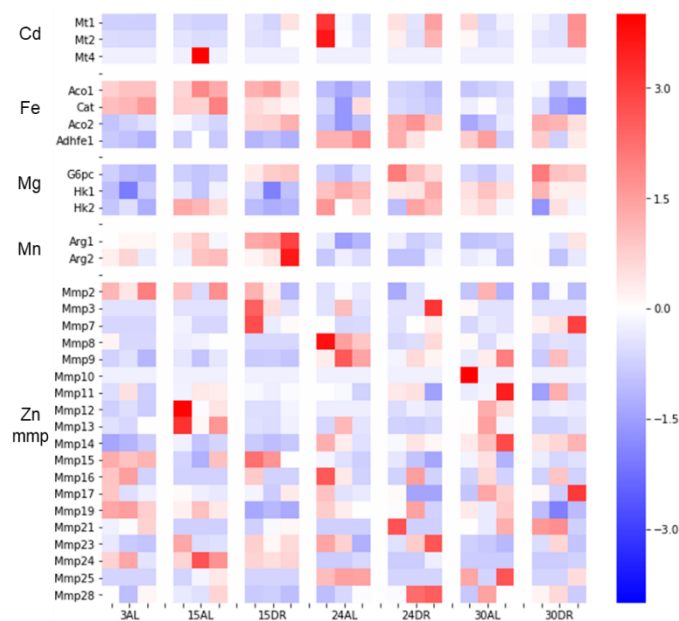


Figure 0-1. Unclustered z scored heatmap showing metalloprotein genes (sorted by their metal).

APPENDIX 2. MICROSCOPY ANALYSIS OF MICE TISSUES

Imaging was conducted by Mikolaj Ogrodnik at the Newcastle University Institute for Ageing, Institute for Cell and Molecular Biosciences¹⁴¹. Samples were prepared according to Wang, Jurk, Maddick, Nelson, Martin-Ruiz, and Von Zglinicki¹³¹. The OCT-embedded hepatic samples were washed three times with PBS and mounted with DAPI and imaged. Karyomegaly frequency was imaged by maximum Z projections with a nucleus being considered enlarged if the nuclear area was 127 μm^2 .

For telomere-associated foci counting, samples were first stained with γ - *H2A.X* (S139 no. 9718, rabbit monoclonal), then washed in PBS three times for 5 minutes each. Sections were then cross-linked with 4% paraformaldehyde for 20 minutes, rewashed with PBS three time for 5 minutes each and dehydrated and dried by graded ethanol for 3 minutes each. Sections were denatured at 82°C for 10 minutes in hybridization buffer (70% formamide, 25mM *MgCl*₂, 0.1 Tris (pH 7.2), 5% blocking reagent) containing 2.5 μgml^{-1} telomere-specific Cy-3-labelled (CCCTAA). Slides were then washed with 70% formamide (10 min), then twice with ssc (15 min), then twice with scc ad PBS (10 min), sections were then incubated with DAPI mounting media and imaged. Number of TAFs per cells were assessed by quantification of overlapping telomere probe and γ - *H2A.X*.

APPENDIX 3. GO ENRICHMENT FOR CLUSTER FISHED GENES FOR CROSSOVER MICE SENESCENCE

5.5	Fold Enrichment	P-value
immune system process (GO:0002376)	2.51	6.48E-08
positive regulation of biological process (GO:0048518)	1.66	7.55E-07
biological adhesion (GO:0022610)	2.96	1.64E-06
regulation of cell motility (GO:2000145)	3.36	1.82E-06
regulation of cell migration (GO:0030334)	3.43	1.83E-06
regulation of locomotion (GO:0040012)	3.19	2.44E-06
cell adhesion (GO:0007155)	2.93	3.91E-06
regulation of localization (GO:0032879)	2.05	5.99E-06
regulation of multicellular organismal process (GO:0051239)	1.99	7.26E-06
regulation of cellular component movement (GO:0051270)	3.09	9.81E-06
cell activation (GO:0001775)	3.59	1.19E-05
locomotion (GO:0040011)	2.7	4.05E-05
regulation of immune system process (GO:0002682)	2.49	1.90E-04
positive regulation of multicellular organismal process (GO:0051240)	2.25	2.03E-04
regulation of molecular function (GO:0065009)	1.91	4.61E-04
regulation of response to stimulus (GO:0048583)	1.76	5.01E-04
regulation of developmental process (GO:0050793)	1.93	6.22E-04
leukocyte activation (GO:0045321)	3.49	7.99E-04
single organismal cell-cell adhesion (GO:0016337)	3.47	9.11E-04
cell migration (GO:0016477)	2.87	1.09E-03
positive regulation of hydrolase activity (GO:0051345)	3.06	1.20E-03
positive regulation of cell adhesion (GO:0045785)	3.86	1.69E-03
regulation of cell adhesion (GO:0030155)	2.99	1.84E-03
negative regulation of cellular process (GO:0048523)	1.61	2.10E-03
positive regulation of cellular process (GO:0048522)	1.56	2.27E-03
cell-cell adhesion (GO:0098609)	3.11	2.29E-03
single organism cell adhesion (GO:0098602)	3.27	2.69E-03
single-organism process (GO:0044699)	1.23	3.09E-03
actin cytoskeleton organization (GO:0030036)	3.56	3.39E-03
localization of cell (GO:0051674)	2.67	3.60E-03
cell motility (GO:0048870)	2.67	3.60E-03
regulation of hydrolase activity (GO:0051336)	2.4	4.39E-03
regulation of catalytic activity (GO:0050790)	1.96	4.73E-03
negative regulation of biological process (GO:0048519)	1.56	5.27E-03
positive regulation of molecular function (GO:0044093)	2.14	5.67E-03
regulation of biological quality (GO:0065008)	1.74	5.69E-03
regulation of multicellular organismal development (GO:2000026)	2	6.15E-03
response to external stimulus (GO:0009605)	2.11	6.16E-03
defense response (GO:0006952)	2.46	6.70E-03
positive regulation of developmental process (GO:0051094)	2.21	7.15E-03
positive regulation of catalytic activity (GO:0043085)	2.29	7.28E-03
response to stress (GO:0006950)	1.75	7.72E-03
cellular component organization (GO:0016043)	1.55	8.26E-03
actin filament-based process (GO:0030029)	3.34	9.55E-03
localization (GO:0051179)	1.54	1.16E-02
leukocyte differentiation (GO:0002521)	3.69	1.23E-02
regulation of signal transduction (GO:0009966)	1.81	1.32E-02
response to stimulus (GO:0050896)	1.36	1.51E-02
lymphocyte differentiation (GO:0030098)	4.22	1.56E-02
regulation of anatomical structure morphogenesis (GO:0022603)	2.4	1.65E-02
negative regulation of multicellular organismal process (GO:0051241)	2.3	1.71E-02
regulation of leukocyte differentiation (GO:1902105)	4.14	1.95E-02
cell differentiation (GO:0030154)	1.63	2.14E-02
immune response (GO:0006955)	2.49	2.18E-02
regulation of cellular component organization (GO:0051128)	1.8	2.30E-02
phagocytosis (GO:0006909)	> 5	2.39E-02
lymphocyte activation (GO:0046649)	3.38	2.39E-02
lipid modification (GO:0030258)	> 5	2.76E-02
cytoskeleton organization (GO:0007010)	2.43	3.79E-02
regulation of vasculature development (GO:1901342)	4.16	3.83E-02
cellular component organization or biogenesis (GO:0071840)	1.5	3.95E-02
positive regulation of immune system process (GO:0002684)	2.61	3.96E-02
positive regulation of metabolic process (GO:0009893)	1.6	4.21E-02
regulation of actin filament-based process (GO:0032970)	3.52	4.39E-02
negative regulation of blood vessel morphogenesis (GO:2000181)	> 5	4.73E-02

Table showing all significant GO annotations for the cluster fished genes for crossover mice senescence. Some annotations have been coloured by type of annotation; immune system, cell-cell binding, and lipid metabolism.

APPENDIX 4. REGULATION OF MITOCHONDRIAL GENE EXPRESSION BY TRNA AVAILABILITY

$$\Delta Expression = \Delta RNA \times \frac{\Delta mtT_a \times Fraction(a) + \dots + \Delta mtT_v \times Fraction(v)}{20}$$

Equation 1. Simplistic model on how gene expression was affected both by changes in RNA and by the changes in tRNA landscape ($\Delta mtTX$) and the fraction of the gene composed of a specific aminoacid ($Fraction(X)$).

Differential expression analysis showed mitochondrial tRNAs to be significantly altered between different treatments (Figure 0-2 A). For example, the tRNA for alanine is more abundant after feeding than after fasting (PFAL and PFDR), and the tRNA for serine is highly present during youth (3 to 15 months, 3AL, 15AL, and 15DR). Because the tRNA landscape is flexible, and mitochondrial genes are homogenous (Figure 0-2 B) in their amino acid content, we can easily assume that a different tRNA landscape will affect the expression of mitochondrial genes.

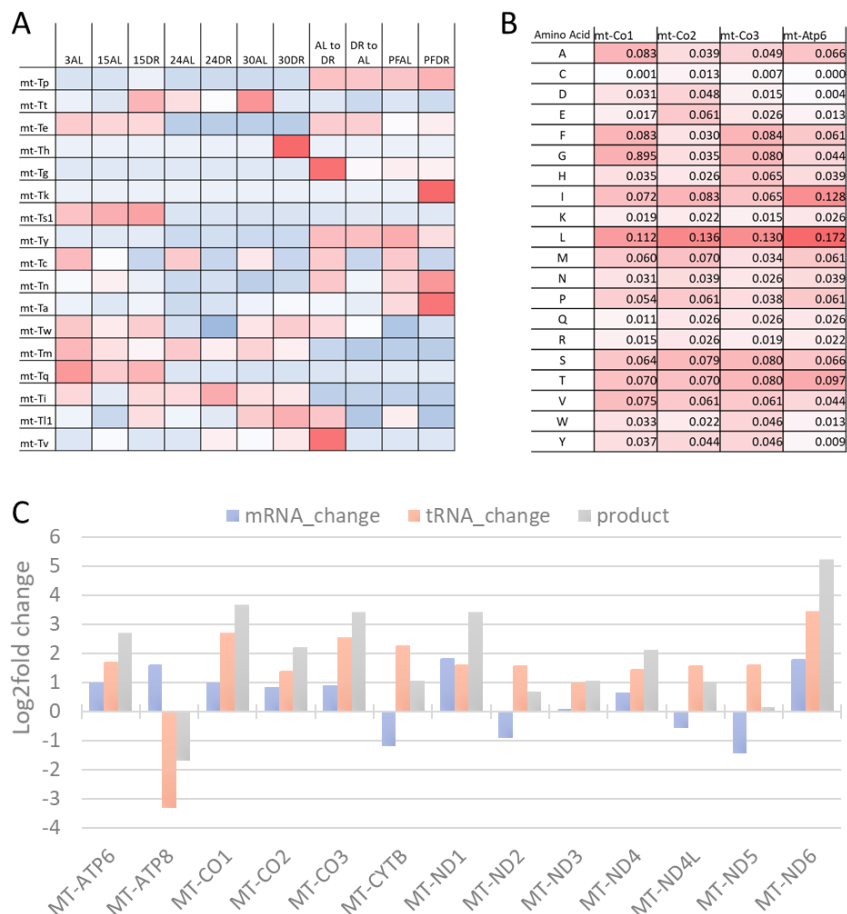


Figure 0-2. (A) Expression of mitochondrial tRNA for different amino acids at different treatments (z-scored). (B) Aminoacid fraction of four mitochondrial genes. (C) Plot representing the mRNA changes, the expected effect of different tRNA landscape, and the expected expression levels for mitochondrial genes.

A simplistic model can be designed where we assume that the more a tRNA is present the faster a gene will be translated into protein (depending on the fraction of the protein the amino acid in question is needed for (Equation 1)). When applying this model comparing expression and tRNA landscape among two arbitrary samples (Figure 0-2 C) some genes were found highly affected. According to RNA content alone, it seemed that mt-ATP8 was upregulated. However, the tRNA landscape seems very 'hostile' to mt-ATP8 so the model predicts less mt-ATP8 translation even with increased RNA transcription. Mt-ND5 had a lower expression but with a more 'favourable' tRNA landscape, so its translation levels seem unaltered.

This model might be too simplistic as it does not account for aminoacyl-tRNAs availability (observable in RNA-seq) and amino acid availability (not observable in RNA-seq).

This model and question were formulated on mitochondrial genes for two reasons. A practical reason being that during sequencing nuclear tRNAs were filtered out and so were not available for analysis. Another perhaps more theoretical reason is that there are very few genes in the mitochondria (yet all very important for energy metabolism) making it plausible to regulate gene expression by alterations of the tRNA landscape. However, there are circa 20,000 genes in the human genome and so its regulation by alteration of tRNA landscape seems much more difficult, as regulation of individual genes would not be possible.

A parallel study which focused on mitochondrial ageing using the same tissues from the sequenced mice utilized proteomic analysis by mass spectrometry. However, due to time constraints, further investigation of mitochondrial protein synthesis by tRNA regulation was not feasible.

APPENDIX 5. LONG NON-CODING RNA STUDY ON MICE AGING DATA.

Reactome enrichments on ageing and diet show changes in splicing and lncRNAs (Figure 2-11 and Figure 2-15). These changes largely occurred between 15 to 24 months of age, which corresponds to a drastic increase in mouse death rate (Figure 2-3). During these two timepoints, a clear demarcation between healthy young mice and feeble dying mice was present. Long non-coding RNAs followed a similar trend to differential expression analysis (Figure 0-3), with most changes occurring between

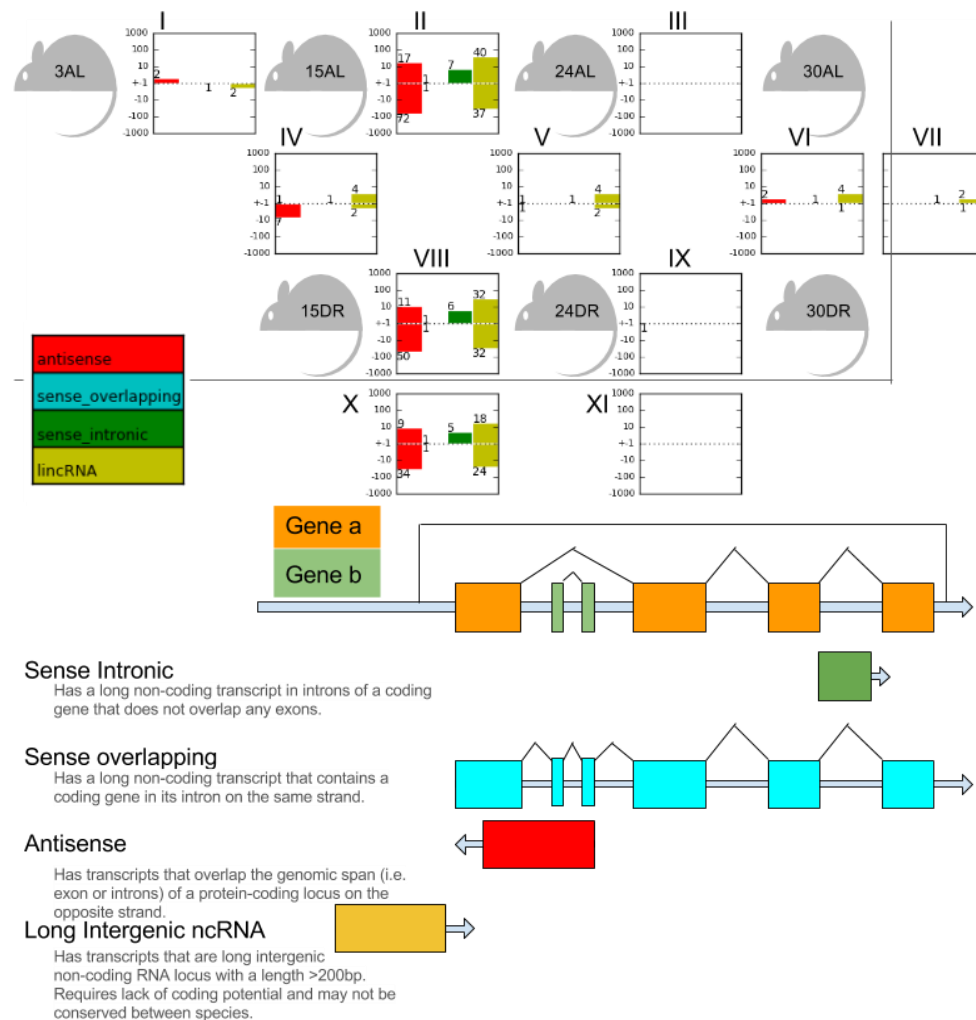


Figure 0-3. (top) Differential expression of long non-coding RNAs. Each plot represents the genes upregulated (on top) or downregulated (on the bottom) in a log10 scale between the treatments of the mice neighbouring the plot (ex. Plot I represents the changes between 3 months AL and 15 months AL). Plots VII, X, and XI represent the changes that are consistent between AL to DR (VII), 15 to 24 (X) and 24 to 30 (XI). (Bottom) diagrammatic representation of the varying types of lncRNAs studied.

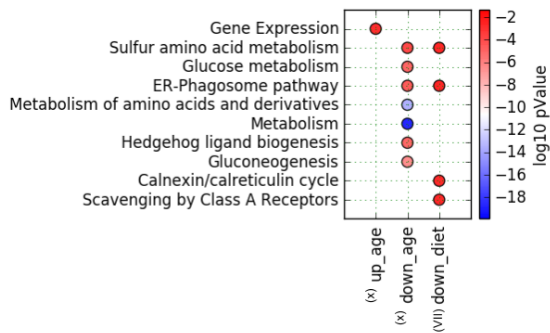
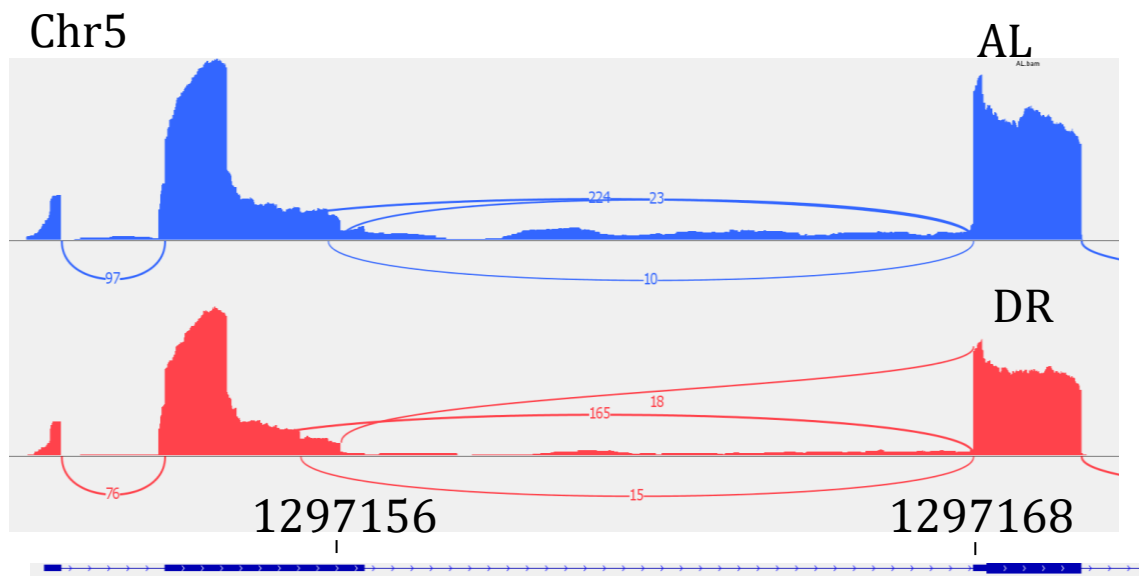


Figure 0-4. Reactome enrichment of differentially expressed long non-coding RNAs

15 to 24 months, and negligible changes thereafter. During the 15 to 24 month's time points, there was a decrease in antisense lncRNAs (compared to upregulated antisense lncRNA). Long non-coding RNAs were extracted using their respective GTF file annotation (GRCm38 release 91).

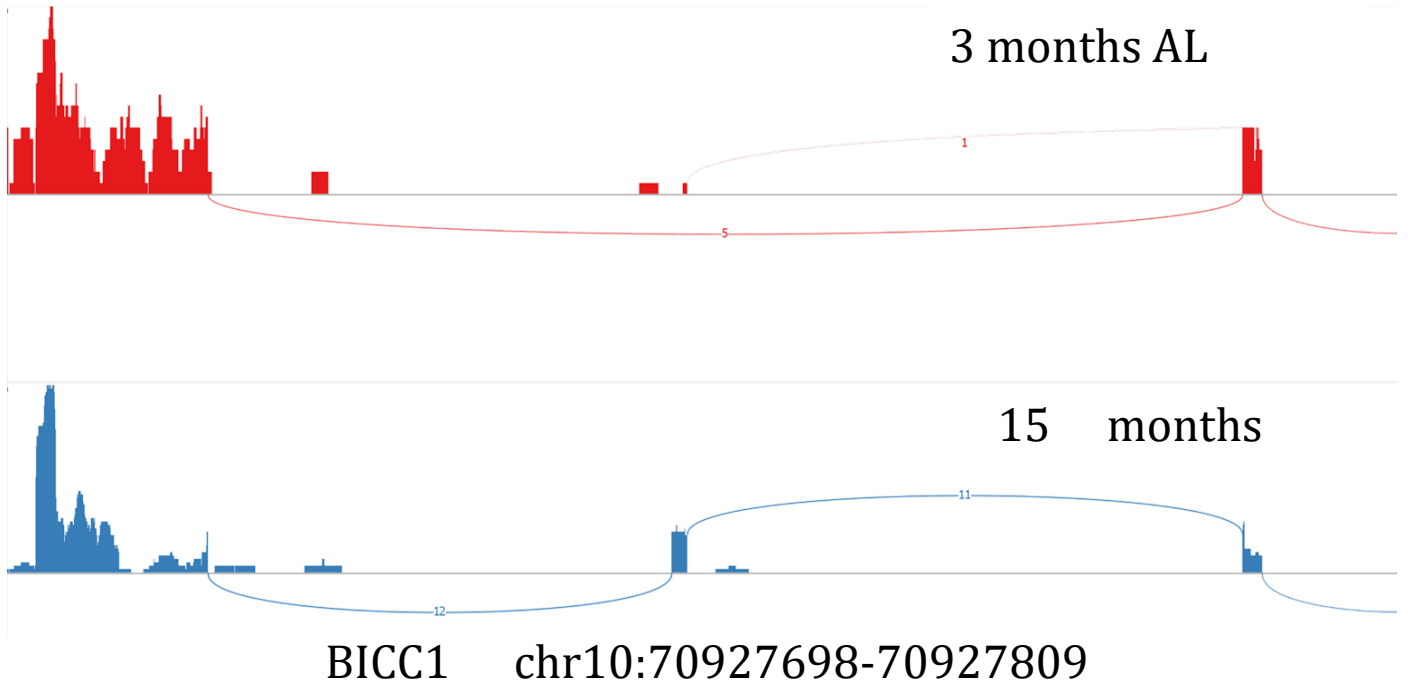
Reactome enrichment for long non-coding RNAs (Figure 0-4) showed that metabolism (R-MMU-1430728; including amino acid synthesis (R-MMU-71291) and gluconeogenesis (R-MMU-70263)) was downregulated at an increased age.

APPENDIX 6. SASHIMI PLOT SHOWING DETECTED SPLICING EXAMPLE OF MICE AGEING DATA.



A sashimi plot showing a retained intron detected on chromosome 5 showing a retained intron in AL for MRPS17, a mitochondrial riboprotein.

APPENDIX 7. SASHIMI PLOT EXAMPLE OF MALTESERS ANALYSIS OF AGEING MICE



APPENDIX 8. WI38 DIFFERENTIAL EXPRESSION WITHOUT BATH CORRECTION

```
library("sva")
library(DESeq2)
countdata = read.table('counts/merged.csv',header =TRUE,row.names=1,sep=',')
countdata = round(countdata) countdata = as.matrix(countdata)
coldata = data.frame( experiment= c( 'lSen_NGSs', 'eSen_NGSs', 'eSen_GSs', 'eSen_GSs',
                                     'Pro_NGSs', 'Pro_GSs', 'eSen_NGSs', 'eSen_NGSs',
                                     'lSen_NGSs', 'Pro_NGSs', 'eSen_GSs', 'Pro_GSs',
                                     'Pro_NGSs', 'Pro_GSs', 'lSen_NGSs'),
                      batch = c( '3','2','2','3','2','1','3','1','1','1','3','3','2','2'),
                      batchEffect=c( 'B','B','B','B','B','A','B','A','A','A','A','B','B','B','B'),
                      row.names=colnames(countdata) )
dds <- DESeqDataSetFromMatrix(countData = countdata, colData = coldata,
                              design =~batchEffect+experiment)

dds <- DESeq(dds)
rld <- rlogTransformation(dds)
pdf('DESeq/deseq.pdf')
plotPCA (rld, intgroup = c('experiment'))
plotPCA (rld, intgroup = c('batchEffect'))
sampleDists <- as.matrix(dist(t(assay(rld))))
heatmap(sampleDists)
plotDispEsts(dds)
dev.off()
norm <- counts(dds,normalize = T)
head(norm)
write.table(norm,'DESeq/normCounts.counts')
save.image(file = 'DESeq/Ranalysis.r')
```

APPENDIX 9. WI38 DIFFERENTIAL EXPRESSION WITH SVA BATCH CORRECTION

```
library("sva")
library(DESeq2)
countdata = read.table('counts/merged.csv',header =TRUE,row.names=1,sep=',')
countdata = round(countdata) countdata = as.matrix(countdata)
coldata = data.frame( experiment= c( 'lSen_NGSs', 'eSen_NGSs', 'eSen_GSs', 'eSen_GSs',
                                     'Pro_NGSs', 'Pro_GSs', 'eSen_NGSs', 'eSen_NGSs',
                                     'lSen_NGSs', 'Pro_NGSs', 'eSen_GSs', 'Pro_GSs',
                                     'Pro_NGSs', 'Pro_GSs', 'lSen_NGSs'),
                      batch = c( '3', '2', '2', '3', '2', '1', '3', '1', '1', '1', '1', '3', '3', '2', '2'),
                      batchEffect=c( 'B', 'B', 'B', 'B', 'B', 'A', 'B', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
                      row.names=colnames(countdata) )
mod = model.matrix(~as.factor(experiment) + as.factor( batchEffect),data=coldata)
fit = lm.fit(mod,t(countdata))
countdata=(t(combat_edata$fitted.values))
countdata[countdata<0]=0
storage.mode(countdata) = "integer"
dds <- DESeqDataSetFromMatrix(countData = countdata, colData = coldata,
                              design =~batchEffect+experiment)

dds <- DESeq(dds)
rld <- rlogTransformation(dds)
pdf('DESeq/deseq.pdf')
plotPCA (rld, intgroup = c('experiment'))
plotPCA (rld, intgroup = c('batchEffect'))
sampleDists <- as.matrix(dist(t(assay(rld))))
heatmap(sampleDists)
plotDispEsts(dds)
dev.off()
norm <- counts(dds,normalize = T)
head(norm)
write.table(norm,'DESeq/normCounts.counts')
save.image(file = 'DESeq/Ranalysis.r')
```

APPENDIX 10. WI38 DIFFERENTIAL EXPRESSION WITH COMBAT CORRECTION

```
library("sva")
library(DESeq2)
countdata = read.table('counts/merged.csv',header =TRUE,row.names=1,sep=',')
countdata = round(countdata)
countdata = as.matrix(countdata)
coldata = data.frame( experiment= c( 'lSen_NGSs', 'eSen_NGSs', 'eSen_GSs', 'eSen_GSs',
                                   'Pro_NGSs', 'Pro_GSs', 'eSen_NGSs', 'eSen_NGSs',
                                   'lSen_NGSs', 'Pro_NGSs', 'eSen_GSs', 'Pro_GSs',
                                   'Pro_NGSs', 'Pro_GSs', 'lSen_NGSs'),
                    batch = c( '3', '2', '2', '3', '2', '1', '3', '1', '1', '1', '1', '3', '3', '2', '2'),
                    batchEffect=c( 'B', 'B', 'B', 'B', 'B', 'A', 'B', 'A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
                    row.names=colnames(countdata) )

batchEffect = coldata$batchEffect
countdata = (countdata[(rowVars(countdata)) > 0,])
modcombat = model.matrix(~1, data=data.frame(countdata))
combat_edata = ComBat(dat=(countdata), batch=batchEffect, mod=NULL, par.prior=TRUE, prior.plots=FALSE)
countdata=(t(combat_edata$fitted.values))
countdata[countdata<0]=0
storage.mode(countdata) = "integer"
dds <- DESeqDataSetFromMatrix(countData = countdata, colData = coldata,
                             design =~batchEffect+experiment)

dds <- DESeq(dds)
rld <- rlogTransformation(dds)
pdf('DESeq/deseq.pdf')
plotPCA(rld, intgroup = c('experiment'))
plotPCA(rld, intgroup = c('batchEffect'))
sampleDists <- as.matrix(dist(t(assay(rld))))
heatmap(sampleDists)
plotDispEsts(dds)
dev.off()
norm <- counts(dds,normalize = T)
head(norm)
write.table(norm,'DESeq/normCounts.counts')
save.image(file = 'DESeq/Ranalysis.r')
```

APPENDIX 11. MAKEFILE USED FOR TESTING MALTESERS

```

.PRECIOUS:
.SECONDARY:
trimm_len = 101
cores = 4
rawFa = $(wildcard rawFasta/*.fastq.gz)
MATS = ../tools/rMATS.3.2.2.beta/RNASeq-MATS.py
GENOME = ../tamir/genome
GTF = $(GENOME)/Homo_sapiens.GRCh38.84.gtf
SJoutTAB = $(shell python -c 'print "
.join(set([i.replace("rawFasta/", "aligned/out1/").replace(".fastq.gz", "")[:-2]+"SJ.out.tab" for i in
"${rawFa}".split()])))' )
BAMS = $(shell python -c 'print "
.join(set([i.replace("rawFasta/", "aligned/out2/").replace(".fastq.gz", "")[:-2]+"Aligned.sortedByCoord.out.bam"
for i in "${rawFa}".split()])))' )
BAMS_TRIMM= $(shell python -c 'print "
.join(set([i.replace("rawFasta/", "aligned/${trimm_len}/").replace(".fastq.gz", "")[:-
2]+"Aligned.sortedByCoord.out.bam" for i in "${rawFa}".split()])))' )
COUNTS = $(shell python -c 'print " .join(set([i.replace("rawFasta/", "counts/").replace(".fastq.gz", "")[:-
2]+"count" for i in "${rawFa}".split()])))' )

pre: rawFasta/readCount.txt

rawFasta/readCount.txt:
echo making $@
for i in rawFasta/*.fastq.gz; do echo $$i $$$( expr $$$( zcat $$i | wc -l ) / 4 ) >>
rawFasta/readCount.txt ; done

test:
echo $(BAMS)

monitor:
watch "ls -lth aligned/*/*bam rMATS_Kidneys_Pancreas/* 2>/dev/null "

align: $(BAMS)

aligned/${trimm_len}:
mkdir -p $@

aligned/out1/SJ.out.tab: rawFasta/%_1.fastq.gz rawFasta/%_2.fastq.gz
mkdir -p aligned/out1 ;\
STAR --alignEndsType EndToEnd \
--alignSJOverhangMin 6 \
--alignIntronMax 300000 \
--chimSegmentMin 2 \
--readFilesIn $^ \
--genomeDir $(GENOME)/index \
--runThreadN $(cores) \
--readFilesCommand zcat \
--twopassMode None \
--sjdbGTFfile $(GTF) \
--sjdbOverhang 84 \
--outSAMstrandField intronMotif \
--outFilterMismatchNmax 3 \
--outFileNamePrefix aligned/out1/* \
--outSAMmultNmax 1 \
--outSAMtype BAM SortedByCoordinate \
--outFilterMultimapNmax 1 \
--outSAMprimaryFlag OneBestScore

aligned/out2/Aligned.sortedByCoord.out.bam: $(SJoutTAB) | rawFasta/%_1.fastq.gz rawFasta/%_2.fastq.gz
mkdir -p aligned/out2 ;\
STAR --alignEndsType EndToEnd \
--alignSJOverhangMin 6 \
--alignIntronMax 300000 \
--chimSegmentMin 2 \
--readFilesIn $| \
--genomeDir $(GENOME)/index \
--runThreadN $(cores) \
--readFilesCommand zcat \
--twopassMode None \
--sjdbGTFfile $(GTF) \
--sjdbOverhang 84 \
--outSAMstrandField intronMotif \
--outFilterMismatchNmax 3 \
--outFileNamePrefix aligned/out2/* \
--outSAMmultNmax 1 \
--outSAMtype BAM SortedByCoordinate \
--outFilterMultimapNmax 1 \
--outSAMprimaryFlag OneBestScore \
--sjdbFileChrStartEnd $(SJoutTAB)

trimmLen.txt: $(BAMS)
samtools view $^ | head | python optimalTrimm.py > $@
cat $@

aligned/${trimm_len}/%.bam: aligned/out2/%.bam aligned/${trimm_len} trimmLen.txt
samtools view -h $< | python trimmBam.py ${trimm_len} | samtools view -hb - > $@

rMATS: $(BAMS_TRIMM)
mkdir -p rMATS_Kidneys_Pancreas/
python $(MATS) \

```

```

-b1 $(shell python -c 'print ",".join(filter(lambda x: "kidney" in
x,"$(BAMS_TRIMM)".split()))') \
-b2 $(shell python -c 'print ",".join(filter(lambda x: "pancreas" in
x,"$(BAMS_TRIMM)".split()))') \
-gtf $(GTF) \
-o rMATS_Kidneys_Pancreas/ \
-t paired \
-len $(trimm_len)

counts:
mkdir -p counts

counts/%.count: aligned/${(trimm_len)}/%Aligned.sortedByCoord.out.bam counts
echo counting $*
python -m HTSeq.scripts.count \
-f bam \
< \
$(GTF) > $@ && \
chmod 555 $@ \

counts/merged.csv: $(COUNTS)
python -c 'import glob,re,pandas; \
files = glob.glob("counts/*.count"); \
df=pandas.concat([pandas.Series.from_csv(x,sep="\t") for x in files],axis=1); \
df.columns = [re.findall("./(.)\.",x)[0] for x in files]; \
print df.head();\
df.to_csv("counts/merged.csv")'

```


REFERENCES

1. Yarian, C. S. C. C. S., Toroser, D. & Sohal, R. S. R. R. S. Aconitase is the main functional target of aging in the citric acid cycle of kidney mitochondria from mice. *Mech. Ageing Dev.* **127**, 79–84 (2006).
2. Martínez, D. E. Mortality Patterns Suggest Lack of Senescence in Hydra. *Exp. Gerontol.* **33**, 217–225 (1998).
3. Miglietta, M. P., Piraino, S., Kubota, S. & Schuchert, P. Species in the genus *Turritopsis* (Cnidaria, Hydrozoa): a molecular evaluation. *J. Zool. Syst. Evol. Res.* **45**, 11–19 (2007).
4. Klapper, W. *et al.* Longevity of lobsters is linked to ubiquitous telomerase expression. *FEBS Lett.* **439**, 143–146 (1998).
5. Tan, T. C. J. *et al.* Telomere maintenance and telomerase activity are differentially regulated in asexual and sexual worms. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 4209–14 (2012).
6. Bishop, N. A. & Guarente, L. Two neurons mediate diet-restriction-induced longevity in *C. elegans*. *Nature* **447**, 545–549 (2007).
7. Wolff, S. & Dillin, A. The trifecta of aging in *Caenorhabditis elegans*. *Exp. Gerontol.* **41**, 894–903 (2006).
8. Bartke, A. Insulin and aging. *Cell Cycle* **7**, 3338–3343 (2008).
9. Feng, J., Bussière, F. & Hekimi, S. Mitochondrial Electron Transport Is a Key Determinant of Life Span in *Caenorhabditis elegans*. *Dev. Cell* **1**, 633–644 (2001).
10. Dillin, A. *et al.* Rates of Behavior and Aging Specified by Mitochondrial Function During Development. *Science (80-.)*. **298**, (2002).
11. Lee, S. S. *et al.* A systematic RNAi screen identifies a critical role for mitochondria in *C. elegans* longevity. *Nat. Genet.* **33**, 40–48 (2002).
12. Liu, X. *et al.* Evolutionary conservation of the *clk-1*-dependent mechanism of longevity: loss of *mclk1* increases cellular fitness and lifespan in mice. *Genes Dev.* **19**, 2424–34 (2005).
13. Copeland, J. M. *et al.* Extension of *Drosophila* Life Span by RNAi of the Mitochondrial Respiratory Chain. *Curr. Biol.* **19**, 1591–1598 (2009).

14. Taylor, R. C. & Dillin, A. Aging as an event of proteostasis collapse. *Cold Spring Harb. Perspect. Biol.* **3**, a004440 (2011).
15. Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res.* **25**, 585–621 (1961).
16. Strehler, B. L. Genetic instability as the primary cause of human aging. *Exp. Gerontol.* **21**, 283–319 (1986).
17. Carroll, B., Hewitt, G. & Korolchuk, V. I. Autophagy and ageing: implications for age-related neurodegenerative diseases. *Essays Biochem.* **55**, (2013).
18. Zhang, Y. *et al.* Protein biogenesis machinery is a driver of replicative aging in yeast. *PLoS One* **7**, e48275 (2015).
19. Jin, K. Modern Biological Theories of Aging. *Aging Dis.* **1**, 72–74 (2010).
20. Okuda, K. *et al.* Telomere Length in the Newborn. *Pediatr. Res.* **52**, 377–381 (2002).
21. Arai, Y. *et al.* Inflammation, But Not Telomere Length, Predicts Successful Ageing at Extreme Old Age: A Longitudinal Study of Semi-supercentenarians. *EBioMedicine* **2**, 1549–58 (2015).
22. Laura J, F. *et al.* Telomere Length Correlates with Life Span of Dog Breeds. *Cell Rep.* **2**, 1530–1536 (2012).
23. Campisi, J. Senescent Cells, Tumor Suppression, and Organismal Aging: Good Citizens, Bad Neighbors. *Cell* **120**, 513–522 (2005).
24. Grube, K. & Bürkle, A. Poly(ADP-ribose) polymerase activity in mononuclear leukocytes of 13 mammalian species correlates with species-specific life span. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 11759–63 (1992).
25. Smith, G. C. & Jackson, S. P. The DNA-dependent protein kinase. *Genes Dev.* **13**, 916–34 (1999).
26. Henshaw, P. S., Riley, E. F. & Stapleton, G. E. The Biologic Effects of Pile Radiations. *Radiology* **49**, 349–360 (1947).
27. Delp, M. D., Charvat, J. M., Limoli, C. L., Globus, R. K. & Ghosh, P. Apollo Lunar Astronauts Show Higher Cardiovascular Disease Mortality: Possible Deep Space Radiation Effects on the Vascular Endothelium. *Sci. Rep.* **6**, 29901 (2016).
28. Bohr, V. a & Anson, R. M. DNA damage , mutation and fine structure DNA repair in aging. *Mutat. Res.* **338**, 25–34 (1995).

29. Stiff, T. *et al.* ATM and DNA-PK Function Redundantly to Phosphorylate H2AX after Exposure to Ionizing Radiation. *Cancer Res.* **64**, 2390–2396 (2004).
30. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase chromatin domains involved in DNA double-strand breaks in vivo. *J. Cell Biol.* **146**, 905–16 (1999).
31. Panier, S. & Boulton, S. J. Double-strand break repair: 53BP1 comes into focus. *Nat. Rev. Mol. Cell Biol.* **15**, 7–18 (2013).
32. Lawless, C. *et al.* Quantitative assessment of markers for cell senescence. *Exp. Gerontol.* **45**, 772–778 (2010).
33. Hamilton, C. A., Brosnan, M. J., McIntyre, M., Graham, D. & Dominiczak, A. F. Superoxide excess in hypertension and aging: a common cause of endothelial dysfunction. *Hypertens. (Dallas, Tex. 1979)* **37**, 529–34 (2001).
34. Moon, S. K. *et al.* Aging, oxidative responses, and proliferative capacity in cultured mouse aortic smooth muscle cells. *Am. J. Physiol. Heart Circ. Physiol.* **280**, H2779–88 (2001).
35. Chen, H. *et al.* Age-related increase in mitochondrial superoxide generation in the testosterone-producing cells of Brown Norway rat testes: relationship to reduced steroidogenic function? *Exp. Gerontol.* **36**, 1361–73 (2001).
36. Donato, A. J. *et al.* Direct Evidence of Endothelial Oxidative Stress With Aging in Humans: Relation to Impaired Endothelium-Dependent Dilation and Upregulation of Nuclear Factor- κ B. *Circ. Res.* **100**, 1659–1666 (2007).
37. Jacobson, A. *et al.* Aging enhances pressure-induced arterial superoxide formation. *AJP Hear. Circ. Physiol.* **293**, H1344–H1350 (2007).
38. Lener, B. *et al.* The NADPH oxidase Nox4 restricts the replicative lifespan of human endothelial cells. *Biochem. J.* **423**, 363–374 (2009).
39. Mendoza-Núñez, V. M., Ruiz-Ramos, M., Sánchez-Rodríguez, M. A., Retana-Ugalde, R. & Muñoz-Sánchez, J. L. Aging-related oxidative stress in healthy humans. *Tohoku J. Exp. Med.* **213**, 261–8 (2007).
40. Lund, D. D., Chu, Y., Miller, J. D. & Heistad, D. D. Protective effect of extracellular superoxide dismutase on endothelial function during aging. *AJP Hear. Circ. Physiol.* **296**, H1920–H1925 (2009).
41. Miyazawa, M. *et al.* The role of mitochondrial superoxide anion ($O_2^{\cdot-}$) on physiological aging in C57BL/6J mice. *J. Radiat. Res.* **50**, 73–83 (2009).

42. Loeb, L. A., Wallace, D. C. & Martin, G. M. The mitochondrial theory of aging and its relationship to reactive oxygen species damage and somatic mtDNA mutations. *Proc. Natl. Acad. Sci.* **102**, 18769–18770 (2005).
43. Sell, D. R. & Monnier, V. M. Molecular Basis of Arterial Stiffening: Role of Glycation – A Mini-Review. *Gerontology* **58**, 227–237 (2012).
44. Kerr, F. *et al.* Dietary restriction delays aging, but not neuronal dysfunction, in *Drosophila* models of Alzheimer’s disease. *Neurobiol. Aging* **32**, 1977–89 (2011).
45. Guo, Z. *et al.* Dietary restriction reduces atherosclerosis and oxidative stress in the aorta of apolipoprotein E-deficient mice. *Mech. Ageing Dev.* **123**, 1121–31 (2002).
46. Jia, K. & Levine, B. Autophagy is Required for Dietary Restriction-Mediated Life Span Extension in *C. elegans*. *Autophagy* **3**, 597–599 (2007).
47. Jeyapalan, J. C. & Sedivy, J. M. Cellular senescence and organismal aging. *Mech. Ageing Dev.* **129**, 467–74 (2008).
48. Wagner, W. *et al.* Aging and replicative senescence have related effects on human stem and progenitor cells. *PLoS One* **4**, e5846 (2009).
49. Sager, R. Senescence as a mode of tumor suppression. in *Environmental Health Perspectives* **93**, 59–62 (1991).
50. Sharpless, N. E. *et al.* Loss of p16Ink4a with retention of p19Arf predisposes mice to tumorigenesis. *Nature* **413**, 86–91 (2001).
51. Price, J. S. *et al.* The role of chondrocyte senescence in osteoarthritis. *Aging Cell* **1**, 57–65 (2002).
52. Stehens, W. E., Wakefield, S. J., Gilbert-Barness, E., Olson, R. E. & Ackerman, J. Histological and Ultrastructural Features of Atherosclerosis in Progeria. *Cardiovasc. Pathol.* **8**, 29–39 (1999).
53. Saito, H. & Papaconstantinou, J. Age-associated differences in cardiovascular inflammatory gene induction during endotoxic stress. *J. Biol. Chem.* **276**, 29307–12 (2001).
54. Vasile, E., Tomita, Y., Brown, L. F., Kocher, O. & Dvorak, H. F. Differential expression of thymosin beta-10 by early passage and senescent vascular endothelium is modulated by VPF/VEGF: evidence for senescent endothelial cells in vivo at sites of atherosclerosis. *FASEB J.* **15**, 458–66 (2001).

55. Castro, P., Giri, D., Lamb, D. & Ittmann, M. Cellular senescence in the pathogenesis of benign prostatic hyperplasia. *Prostate* **55**, 30–8 (2003).
56. Choi, J. *et al.* Expression of senescence-associated beta-galactosidase in enlarged prostates from men with benign prostatic hyperplasia. *Urology* **56**, 160–166 (2000).
57. Nelson, G. *et al.* A senescent cell bystander effect: senescence-induced senescence. *Aging Cell* **11**, 345–349 (2012).
58. Hewitt, G. *et al.* Telomeres are favoured targets of a persistent DNA damage response in ageing and stress-induced senescence. *Nat. Commun.* **3**, 708 (2012).
59. White, R. R. *et al.* Controlled induction of DNA double-strand breaks in the mouse liver induces features of tissue ageing. *Nat. Commun.* **6**, 6790 (2015).
60. Duncan, A. W. *et al.* The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature* **467**, 707–710 (2010).
61. McCay, Carl M and Crowell, Mary F and Maynard, L. A. *The effect of retarded growth upon the length of life span and upon the ultimate body size. The Journal of Nutrition* **10**, (American Society for Nutrition, 1935).
62. Schleit, J., Wasko, B. M. & Kaeberlein, M. Yeast as a model to understand the interaction between genotype and the response to calorie restriction. *FEBS Lett.* **586**, 2868–73 (2012).
63. Schulz, T. J. *et al.* Glucose restriction extends *Caenorhabditis elegans* life span by inducing mitochondrial respiration and increasing oxidative stress. *Cell Metab.* **6**, 280–93 (2007).
64. Harrison, D. E. *et al.* Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* **460**, 392–395 (2009).
65. Weindruch, R., Walford, R. L., Fligiel, S. & Guthrie, D. The retardation of aging in mice by dietary restriction: longevity, cancer, immunity and lifetime energy intake. *J. Nutr.* **116**, 641–654 (1986).
66. Colman, R. J. & Anderson, R. M. Nonhuman Primate Calorie Restriction. *Antioxid. Redox Signal.* **14**, 229–239 (2011).
67. Mattison, J. A. *et al.* Caloric restriction improves health and survival of rhesus monkeys. *Nat. Commun.* **8**, 14063 (2017).
68. Rhoads, T. W. *et al.* Caloric Restriction Engages Hepatic RNA Processing

- Mechanisms in Rhesus Monkeys. *Cell Metab.* **27**, 677–688.e5 (2018).
69. Johnson, T. E. Recent results: Biomarkers of aging. *Exp. Gerontol.* **41**, 1243–1246 (2006).
 70. Fontana, L. & Klein, S. Aging, adiposity, and calorie restriction. *JAMA* **297**, 986–994 (2007).
 71. Powers, R. W., Kaeberlein, M., Caldwell, S. D., Kennedy, B. K. & Fields, S. Extension of chronological life span in yeast by decreased TOR pathway signaling. *Genes Dev.* **20**, 174–184 (2006).
 72. Bjedov, I. *et al.* Mechanisms of Life Span Extension by Rapamycin in the Fruit Fly *Drosophila melanogaster*. *Cell Metab.* **11**, 35–46 (2010).
 73. Majumder, S. *et al.* Lifelong rapamycin administration ameliorates age-dependent cognitive deficits by reducing IL-1 β and enhancing NMDA signaling. *Aging Cell* **11**, 326–335 (2012).
 74. Lamming, D. W. *et al.* Rapamycin-Induced Insulin Resistance Is Mediated by mTORC2 Loss and Uncoupled from Longevity. *Science (80-)*. **335**, 1638–1643 (2012).
 75. Hay, N. & Sonenberg, N. Upstream and downstream of mTOR. *Genes Dev.* **18**, 1926–45 (2004).
 76. Dos D. Sarbassov, D. D. *et al.* Rictor, a Novel Binding Partner of mTOR, Defines a Rapamycin-Insensitive and Raptor-Independent Pathway that Regulates the Cytoskeleton. *Curr. Biol.* **14**, 1296–1302 (2004).
 77. Chang, A. M. & Halter, J. B. Aging and insulin secretion. *Am. J. Physiol. Endocrinol. Metab.* **284**, E7-12 (2003).
 78. KLASS, M. & HIRSH, D. Non-ageing developmental variant of *Caenorhabditis elegans*. *Nature* **260**, 523–525 (1976).
 79. Kimura, K. D., Tissenbaum, H. A., Liu, Y. & Ruvkun, G. *daf-2*, an insulin receptor-like gene that regulates longevity and diapause in *Caenorhabditis elegans*. *Science* **277**, 942–6 (1997).
 80. Lin, K., Dorman, J. B., Rodan, A. & Kenyon, C. *daf-16*: An HNF-3/forkhead family member that can function to double the life-span of *Caenorhabditis elegans*. *Science* **278**, 1319–22 (1997).
 81. Ruvkun, G. *et al.* The Fork head transcription factor DAF-16 transduces insulin-

- like metabolic and longevity signals in *C. elegans*. *Nature* **389**, 994–999 (1997).
82. Blüher, M. *et al.* Adipose tissue selective insulin receptor knockout protects against obesity and obesity-related glucose intolerance. *Dev. Cell* **3**, 25–38 (2002).
 83. Kurosu, H. *et al.* Suppression of Aging in Mice by the Hormone Klotho. *Science (80-.)*. **309**, 1829–1833 (2005).
 84. Shevah, O. & Laron, Z. Patients with congenital deficiency of IGF-I seem protected from the development of malignancies: a preliminary report. *Growth Horm. IGF Res.* **17**, 54–7 (2007).
 85. Bonafè, M. *et al.* 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. *J. Clin. Endocrinol. Metab.* **88**, 3299–3304 (2003).
 86. Van Heemst, D. *et al.* Reduced insulin/IGF-1 signalling and human longevity. *Aging Cell* **4**, 79–85 (2005).
 87. Euser, S. M., van Heemst, D., van Vliet, P., Breteler, M. M. B. & Westendorp, R. G. J. Insulin/Insulin-like growth factor-1 signaling and cognitive function in humans. *J. Gerontol. A. Biol. Sci. Med. Sci.* **63**, 907–10 (2008).
 88. Suh, Y. *et al.* Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proc. Natl. Acad. Sci.* **105**, 3438–3442 (2008).
 89. Pawlikowska, L. *et al.* Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. *Aging Cell* **8**, 460–472 (2009).
 90. Anselmi, C. V. *et al.* Association of the *FOXO3A* Locus with Extreme Longevity in a Southern Italian Centenarian Study. *Rejuvenation Res.* **12**, 95–104 (2009).
 91. Flachsbart, F. *et al.* Association of *FOXO3A* variation with human longevity confirmed in German centenarians. *Proc. Natl. Acad. Sci.* **106**, 2700–2705 (2009).
 92. Li, Y. *et al.* Genetic association of *FOXO1A* and *FOXO3A* with longevity trait in Han Chinese populations. *Hum. Mol. Genet.* **18**, 4897–4904 (2009).
 93. Willcox, B. J. *et al.* *FOXO3A* genotype is strongly associated with human longevity. *Proc. Natl. Acad. Sci.* **105**, 13987–13992 (2008).

94. Rhoads, T. W. *et al.* Caloric Restriction Engages Hepatic RNA Processing Mechanisms in Rhesus Monkeys. *Cell Metab.* **27**, 677–688.e5 (2018).
95. Lee, B. P. *et al.* Changes in the expression of splicing factor transcripts and variations in alternative splicing are associated with lifespan in mice and humans. *Aging Cell* (2016). doi:10.1111/acel.12499
96. Gilbert, W. Origin of life: The RNA world. *Nature* **319**, 618–618 (1986).
97. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–5 (2008).
98. Ward, A. J. & Cooper, T. A. The pathobiology of splicing. *J. Pathol.* **220**, 152–63 (2010).
99. Lynch, K. W. & Maniatis, T. Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer. *Genes Dev.* **10**, 2089–2101 (1996).
100. Eriksson, M. *et al.* Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293–298 (2003).
101. Davis, M. P. A. A., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: {A} set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41–49 (2013).
102. Ostell, J. & McEntyre, J. The NCBI Handbook. *NCBI Bookshelf* 1–8 (2007). doi:10.4016/12837.01
103. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–11 (2009).
104. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, (2010).
105. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
106. DownloadMoreRAM.com - CloudRAM 2.0. Available at: <https://downloadmoreram.com/>. (Accessed: 22nd January 2018)
107. Wang, W.-A. *et al.* Comparisons and performance evaluations of RNA-seq alignment tools. in *2014 International Conference on Electrical Engineering and Computer Science (ICEECS)* 215–218 (IEEE, 2014).

doi:10.1109/ICEECS.2014.7045249

108. Anders, S., Pyl, P. T. & Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2014).
109. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
110. Love, M. I., Anders, S. & Huber, W. *Differential analysis of count data - the DESeq2 package.* *Genome Biology* **15**, (2014).
111. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
112. Theocharidis, A., van Dongen, S., Enright, A. J. & Freeman, T. C. Network visualization and analysis of gene expression data using BioLayout Express3D. *Nat. Protoc.* **4**, 1535–50 (2009).
113. van Dongen, S., Dongen, S. van & van Dongen, S. Graph Clustering by Flow Simulation. (University of Utrecht, 2000).
114. Vastrik, I. *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8**, R39 (2007).
115. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. (2016). doi:doi:10.1039/C5MB00663E
116. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
117. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
118. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–78 (2012).
119. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–52 (2011).
120. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–12 (2010).
121. Aschoff, M. *et al.* SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* **29**, 1141–8 (2013).

122. Britton, M. & Ph, D. RNA-Seq with the Tuxedo Suite : Differential Expression and Gene Construction Bioinformatics Analyst The Tuxedo Suite. (2013).
123. Katz, Y., Wang, E. T., Airoidi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–15 (2010).
124. Vitting-Seerup, K., Porse, B. T., Sandelin, A. & Waage, J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* **15**, 81 (2014).
125. GNU make. (1974).
126. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).
127. CMake. Available at: <https://cmake.org/>. (Accessed: 22nd January 2018)
128. SCons: A software construction tool - SCons. Available at: <http://www.scons.org/>. (Accessed: 22nd January 2018)
129. Dutta, S. & Sengupta, P. Men and mice: Relating their ages. *Life Sci.* **152**, 244–248 (2016).
130. Sedelnikova, O. A. *et al.* Senescing human cells and ageing mice accumulate DNA lesions with unrepairable double-strand breaks. *Nat. Cell Biol.* **6**, 168–170 (2004).
131. Wang, C. *et al.* DNA damage response and cellular senescence in tissues of aging mice. *Aging Cell* **8**, 311–323 (2009).
132. Cameron, K. M., Miwa, S., Walker, C. & von Zglinicki, T. Male mice retain a metabolic memory of improved glucose tolerance induced during adult onset, short-term dietary restriction. *Longev. Heal.* **1**, 3 (2012).
133. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
134. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
135. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
136. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* **12**, 477–479 (2015).

137. Shen, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 201419161 (2014).
138. Vermeij, W. P. *et al.* Restricted diet delays accelerated ageing and genomic stress in DNA-repair-deficient mice. *Nature* **537**, 427–431 (2016).
139. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Research* **31**, 51–54 (2003).
140. Borras, C. *et al.* Human exceptional longevity: transcriptome from centenarians is distinct from septuagenarians and reveals a role of Bcl-xL in successful aging. *Aging (Albany, NY)*. **8**, 3185–3208 (2016).
141. Ogrodnik, M. *et al.* Cellular senescence drives age-dependent hepatic steatosis. (2017).
142. Enright, A. J. & Ouzounis, C. A. BioLayout--an automatic graph layout algorithm for similarity visualization. *Bioinformatics* **17**, 853–854 (2001).
143. Stokkan, K.-A., Yamazaki, S., Tei, H., Sakaki, Y. & Menaker, M. Entrainment of the Circadian Clock in the Liver by Feeding. *Science (80-.)*. **291**, 490–493 (2001).
144. Kohsaka, A. *et al.* High-Fat Diet Disrupts Behavioral and Molecular Circadian Rhythms in Mice. *Cell Metab.* **6**, 414–421 (2007).
145. Sutton, G. M. *et al.* Biological aging alters circadian mechanisms in murine adipose tissue depots. *Age (Dordr)*. **35**, 533–47 (2013).
146. Stebbing, A. Hormesis - the Stimulation of Growth By Low-Levels of Inhibitors. *Sci. Total Environ.* **22**, 213–234 (1982).
147. Houtkooper, R. H. *et al.* The metabolic footprint of aging in mice. *Sci. Rep.* **1**, (2011).
148. Fok, W. C. *et al.* Mice fed rapamycin have an increase in lifespan associated with major changes in the liver transcriptome. *PLoS One* **9**, (2014).
149. Kim, S. J. *et al.* Mitochondria-targeted Ogg1 and aconitase-2 prevent oxidant-induced mitochondrial DNA damage in alveolar epithelial cells. *J. Biol. Chem.* **289**, 6165–6176 (2014).
150. Shoelson, S. E. & Goldfine, A. B. Fanning the Flames of Obesity-Induced Inflammation. *Nat. Med.* **15**, 373–374 (2009).
151. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444**, 860–867

(2006).

152. Lumeng, C. N. & Saltiel, A. R. Inflammatory links between obesity and metabolic disease. *J. Clin. Invest.* **121**, 2111–2117 (2011).
153. Hüttemann, M. *et al.* Regulation of mitochondrial respiration and apoptosis through cell signaling: Cytochrome c oxidase and cytochrome c in ischemia/reperfusion injury and inflammation. *Biochim. Biophys. Acta - Bioenerg.* **1817**, 598–609 (2012).
154. Borutaite, V., Matthias, A., Harris, H., Moncada, S. & Brown, G. C. Reversible inhibition of cellular respiration by nitric oxide in vascular inflammation. *Am. J. Physiol. Circ. Physiol.* **281**, H2256–H2260 (2001).
155. Rahman, I. Oxidative stress, transcription factors and chromatin remodelling in lung inflammation. *Biochem. Pharmacol.* **64**, 935–942 (2002).
156. Foster, S. L., Hargreaves, D. C. & Medzhitov, R. Gene-specific control of inflammation by TLR-induced chromatin modifications. *Nature* **447**, 972 (2007).
157. Sagelius, H. *et al.* Reversible phenotype in a mouse model of Hutchinson-Gilford progeria syndrome. *J. Med. Genet.* **45**, 794–801 (2008).
158. Lin, S.-J. *et al.* Calorie restriction extends *Saccharomyces cerevisiae* lifespan by increasing respiration. *Nature* **418**, 344–348 (2002).
159. Baker, D. J. *et al.* Clearance of p16 Ink4a-positive senescent cells delays ageing-associated disorders. *Nature* **479**, 232–236 (2011).
160. Melis, J. P. M. *et al.* Aging on a different scale - chronological versus pathology-related aging. *Aging (Albany, NY)*. **5**, 782–788 (2013).
161. Lee, J. S. *et al.* Meta-analysis of gene expression in the mouse liver reveals biomarkers associated with inflammation increased early during aging. *Mech. Ageing Dev.* **133**, 467–478 (2012).
162. Chandra, T. *et al.* Global reorganization of the nuclear landscape in senescent cells. *Cell Rep.* **10**, 471–83 (2015).
163. Aarts, M. *et al.* Coupling shRNA screens with single-cell RNA-seq identifies a dual role for mTOR in reprogramming-induced senescence. *Genes Dev.* (2017). doi:10.1101/gad.297796.117
164. Kong, X. *et al.* Gene Expression Profiles Deciphering Leaf Senescence Variation

- between Early- and Late-Senescence Cotton Lines. *PLoS One* **8**, (2013).
165. Dobin, A. STAR manual 2.5.1a. 1–39 (2016).
 166. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx1098
 167. Floris, M. *et al.* MAISTAS: a tool for automatic structural evaluation of alternative splicing products. *Bioinformatics* **27**, 1625–1629 (2011).
 168. Wang, J., Zhang, J., Li, K., Zhao, W. & Cui, Q. SpliceDisease database: linking RNA splicing and disease. *Nucleic Acids Res.* **40**, D1055–D1059 (2012).
 169. de Castro, E. *et al.* ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* **34**, 362 (2006).
 170. Gattiker, A., Gasteiger, E. & Bairoch, A. M. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinformatics* **1**, 107–108 (2002).
 171. Sigrist, C. J. A. *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–74 (2002).
 172. Best, A. *et al.* Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. *Nat. Commun.* **5**, 4760 (2014).
 173. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–10 (2002).
 174. Dobin, A. *et al.* {STAR}: ultrafast universal {RNA}-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
 175. Katz, D., Reginato, M. J. & Lazar, M. A. Functional regulation of thyroid hormone receptor variant TR alpha 2 by phosphorylation. *Mol. Cell. Biol.* **15**, 2341–8 (1995).
 176. Feracci, M. *et al.* Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68. *Nat. Commun.* **7**, 10355 (2016).
 177. Farriol-Mathis, N. *et al.* Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* **4**, 1537–1550 (2004).
 178. Binder, J. X. *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* **2014**, bau012-bau012 (2014).

179. Zhou, H. *et al.* Toward a Comprehensive Characterization of a Human Cancer Cell Phosphoproteome. *J. Proteome Res.* **12**, 260–271 (2013).
180. Jin, J. *et al.* The deubiquitinase USP21 maintains the stemness of mouse embryonic stem cells via stabilization of Nanog. *Nat. Commun.* **7**, 13594 (2016).
181. Peck, D. & Isacke, C. M. CD44 phosphorylation regulates melanoma cell and fibroblast migration on, but not attachment to, a hyaluronan substratum. *Curr. Biol.* **6**, 884–890 (1996).
182. Pérez-Garijo, A. & Steller, H. The benefits of aging: Cellular senescence in normal development. *EMBO J.* **33**, 99–100 (2014).
183. Beauséjour, C. M. *et al.* Reversal of human cellular senescence: Roles of the p53 and p16 pathways. *EMBO J.* **22**, 4212–4222 (2003).
184. Pawelec, G., Goldeck, D. & Derhovanessian, E. Inflammation, ageing and chronic disease. *Current Opinion in Immunology* (2014). doi:10.1016/j.coi.2014.03.007
185. Vasto, S. *et al.* Inflammation, ageing and cancer. *Mech. Ageing Dev.* **130**, 40–45 (2009).
186. Finkel, T. & Holbrook, N. J. Oxidants, oxidative stress and the biology of ageing. *Nature* **408**, 239–247 (2000).
187. Baar, M. P. *et al.* Targeted Apoptosis of Senescent Cells Restores Tissue Homeostasis in Response to Chemotoxicity and Aging. *Cell* **169**, 132–147.e16 (2017).
188. De Grey, A. D. N. J. The foreseeability of real anti-aging medicine: Focusing the debate. *Experimental Gerontology* **38**, 927–934 (2003).
189. Masaldan, S. *et al.* Iron accumulation in senescent cells is coupled with impaired ferritinophagy and inhibition of ferroptosis. *Redox Biol.* (2018). doi:10.1016/j.redox.2017.08.015
190. Killilea, D. W., Wong, S. L., Cahaya, H. S., Atamna, H. & Ames, B. N. Iron accumulation during cellular senescence. in *Annals of the New York Academy of Sciences* (2004). doi:10.1196/annals.1297.063
191. Killilea, D. W. & Ames, B. N. Magnesium deficiency accelerates cellular senescence in cultured human fibroblasts. *Proc. Natl. Acad. Sci.* (2008). doi:10.1073/pnas.0712401105
192. Storvall, H., Ramsköld, D. & Sandberg, R. Efficient and comprehensive

representation of uniqueness for next-generation sequencing by minimum unique length analyses. *PLoS One* **8**, e53822 (2013).

193. Juan Carlos Oliveros. Venny. An interactive tool for comparing lists with Venn's diagrams. (2007).

EXTERNAL FILES

- [1] "Differential expression results for the ageing mice datasets.," [Online]. Available: http://lahat.io/thesis/MICE_DESEQ.zip.
- [2] "Reactome enrichment of mice ageing deseq data.," [Online]. Available: http://lahat.io/thesis/MICE_DESEQ_REACTOME.csv.
- [3] "Mice clustering data," [Online]. Available: http://lahat.io/thesis/MICE_CLUSTERING.zip.
- [4] "Reactome enrichment of mice clustering analysis," [Online]. Available: http://lahat.io/thesis/MICE_CLUSTERING_REACTOME.csv.
- [5] "Reactome enrichment of genes altered between 15 months and 15 months post fed.," [Online]. Available: http://lahat.io/thesis/MICEPF_DESEQ_REACTOME.csv.
- [6] "Mice splicing results from rMATS," [Online]. Available: http://lahat.io/thesis/MICE_RMATS.csv.
- [7] "Differential expression results of empty vector proliferating vs empty vector senescent," [Online]. Available: http://lahat.io/thesis/IMR90_DESEQ.csv.
- [8] "Reactome analysis results of the differentially expressed genes between proliferating and senescence," [Online]. Available: http://lahat.io/thesis/IMR90_REACTOME.csv.
- [9] "rMATS of IMR90 data," [Online]. Available: http://lahat.io/thesis/IMR90_RMATS.csv.
- [10] "Maltesers results from WI38 study," [Online]. Available: http://lahat.io/thesis/WI38_MALTESERS.zip.
- [11] "Differentially expressed filtered genes used for reactome enrichment of WI38 data," [Online]. Available: http://lahat.io/thesis/WI38_DESEQ_Reactome.txt.

- [12] "Genes extracted through network fishing of P21 from the WI38 dataset," [Online]. Available: http://lahat.io/thesis/WI38_NetworkFishing.txt.
- [13] "Unmerged mice rmats reactome," [Online]. Available: http://lahat.io/thesis/MICE_RMATS_REACTOME_UNMERGED.csv.
- [14] "Reactome enrichment of IMR90 P21 network fishing," [Online]. Available: http://lahat.io/thesis/IMR90_P21_FISHING_REACTOME.csv.
- [15] "Reactome analysis of mice rMATS," [Online]. Available: http://lahat.io/thesis/MICE_RMATS_REACTOME.csv.
- [16] "Normalized expression for WI38 dataset," [Online]. Available: http://lahat.io/thesis/WI38_EXPRESSION.csv.