Developing RNA diagnostics for studying healthy human ageing

A thesis submitted by

Ms. Sanjana Sood MRes

Submitted in partial fulfillment of the requirements

for the award of

DOCTOR OF PHILOSOPHY

In the subject of

BIOINFORMATICS

To

School of Sport, Exercise and Health Sciences

Loughborough University

[November,2016]

Abstract

Developing strategies to cope with increase in the ageing population and age-related chronic diseases is one of the societies biggest challenges. The characteristics of the ageing process shows significant inter-individual variation. Building genomic signatures that could account for variation in health outcomes with age may facilitate early prognosis of individual age-correlated diseases (e.g. cancer, coronary artery diseases and dementia) and help in developing better targeted treatments provided years in advance of acquiring disabling symptoms for these diseases. The aim of this thesis was to explore methods for diagnosing molecular features of human ageing. In particular, we utilise multi-platform transcriptomics, independent clinical data and classification methods to evaluate which human tissues demonstrate a reproducible molecular signature for age and which clinical phenotypes correlated with these new RNA biomarkers.

Using machine-learning approach (kNN), applied to RNA data derived from muscle tissues from *healthy* donors, we developed a novel and statistically robust *neuro-muscular* 150 probe-set RNA signature and demonstrated its potential as a health-status diagnostic. Validated using multiple independent human muscle cohorts and external validation methods, the RNA signature was effective at distinguishing between young and old human muscle, brain and skin. In muscle, the RNA signature was not correlated with lifestyle regulated phenotypes in muscle or life-style diseases in blood (coronary vascular disease and Type 2 diabetes). This 150 probe-set *neuro-muscular signature* was related to cognitive status in two independent studies confirming that our 'ageing genes' were consistently regulated in muscle, hippocampus and blood tissue in humans.

To establish how unique this 150 probe-set neuro-muscular ageing was, we contrasted a "random" sampling approach with published genomic signatures of human ageing. This involved 'transferring' DNA and DNA methylation signatures to their equivalent RNA signature, before considering their prognostic or diagnostic performance. It was observed that our 150 neuro-muscular gene-set was the only one related to hippocampus ageing and cognitive health, while 'stress' resistant (selected from DNA analysis) and 'epidemiologically' selected linear models (RNA derived) were related with vascular disease. We then attempted to develop an RNA vascular ageing gene-expression model to complement our neuromuscular ageing diagnostic. While statistically significant, the gene-set did not contribute to clinical variance in a sufficient manner over and above key clinical variables e.g. blood pressure and chronological age.

In summary, vascular ageing appears to be distinct from neuro-muscular ageing, at least from the stand point of RNA gene-sets. Overall, this research has resulted in identifying a predictive diagnostic for human neuro-muscular ageing that could be potentially useful in assisting research aimed at finding treatments for and/or management of Alzheimer's disease.

Acknowledgments

I would firstly like to thank my supervisor Professor James A. Timmons for all of his advice and guidance on this project; without his continued enthusiasm, help and support, this research would not have been possible. I could not have imagined having a better advisor and mentor for my doctorate study.

To my dad, mum and Neha dee– you have supported me through all these years, thanks for your constant encouragement and love and for taking pride in my work.

Many thanks to Amo and Krzystof for making me see the bright side of things when I was stressed and low. A special thank you to Dr Lewis James and Dr Iain Gallagher, thank you for all your help, support and positivity.

My PhD would not have been the same without my amazing friends back home in India, thanks for being there '*doston*'. You all are true friends, and thank you for keeping me sane throughout my PhD!

Finally, to Gagan, thank you for being there, for your love and constant support – I would have not been able to do this without you by my side!

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer Disease
AUC	Area under the curve
CAD	Coronary Artery Disease
CC	Correlation Coefficient
CDF	Cell Definition File
CVD	Cardio Vascular Disease
DNAm	DNA methylation
EV	External Validation
fRMA	Frozen Robust Multi-array Analysis
GEO	Gene Expression Omnibus
GFR	Glomerular filtration Rate
GWAS	Genome Wide Association Study
IPA	Ingenuity Pathway Analysis
kNN	k- Nearest Neighbour
LOOCV	Leave One Out Cross Validation
MAP	Mean Arterial Pressure
MAQC	MicroArray Quality Control
MCI	Mild Cognitive Impairment
MMSE	Mini–Mental State Examination
MRI	Magnetic Resonance Imaging
NGS	Next Generation Sequencing
PGE	Positional gene enrichment analysis
PWV	Pulse Wave Velocity
QTL	Quantitative Trait Loci
RMA	Robust Multi-array Analysis
ROC	Receiver Operating Characteristic
SD	Standard deviation
SNP	Single Nucleotide Polymorphism
ULSAM	Uppsala Longitudinal Study of Adult Men
UNFPA	United Nations Population Fund
UTR	Untranslated Region

1.1 Introduction to thesis topic

Advances in infection control, medical diagnosis, and treatment have led to an improved 'health' span as well as an increase in human longevity. The extended life span has presented new medical challenges such as greater number of people with cardiac disease, cancer, and in particular neurodegeneration. This consequently is placing huge demands on our medical services. Currently, treatment of these age-associated diseases is based on interventions aimed to yield clinical benefits in randomized clinical trials (Wiesweg et al. 2013). To revolutionize current approaches in health care it is essential to identify effective strategies to substantially improve 'health span' in humans. Personalized selection of treatment strategies, or health advice, has an increasing impact on the planning of modern medical practice (Goldberger & Buxton 2013).

Personalized approaches to cancer diagnosis and treatment have been substantially influenced by molecular diagnostics (Abd El-Rehim et al. 2005; Shedden et al. 2008; Sebastiani et al. 2012). For human ageing, global RNA profiling based on linear correlative analysis have been utilised to search for consistent molecular events correlating with age across tissues (Willemijn M Passtoors et al. 2012; Gheorghe et al. 2014; Phillips et al. 2013; Glass et al. 2013; Peters et al. 2015). But these attempts failed to find any common gene-sets that could characterise human ageing as most of them were based on cohorts that blended in ageing, disease and drug-treatment and thus had very low reproducibility. Therefore, there are numerous challenges to both the development of, and implementation of personalized strategies for most major age-related diseases (Patnaik et al. 2010) such as the time to measure a biological profile that can provide reliable long-term prognostic information and the technological platform to utilize. Nonetheless, it would be interesting to explore if it is possible to find a common gene set for human tissue ageing and if this gene set signature has prognostic abilities to predict different types of age related diseases i.e. neurodegenerative, vascular etc., and health outcomes. This personalized molecular diagnostic approach could potentially not only estimate an individual's true biological age but will also help in developing therapies that could positively impact ageing and postpone related diseases.

1.2 Thesis Objectives

Developing a diagnostic tool for healthy ageing and applying that knowledge to precision medicine can lead to better therapeutics at an individual level targeted specifically to the genotype. To develop such a tool, a very strict set of methodologies and benchmarks, distinguishing them from descriptive studies of differential RNA expression should be applied. Therefore, the aim of this thesis was to:

- Identify 'healthy ageing genes' (RNA, gene expression) that could distinguish 'healthy' old muscle tissue from young sedentary muscle with high accuracy.
- Use the very same RNA signature to distinguish hundreds of old from young tissue across independent cohorts and tissue-types which has never been achieved before.
- Evaluate whether the validated RNA signature would relate to different human diseases thought to be 'caused' by ageing such as neurological, vascular etc.
- To compare if neuromuscular ageing is similar or different from vascular ageing.

To achieve these objectives, we would implement machine-learning methods on transcriptomics data from carefully phenotyped clinical samples to develop the first accurate molecular diagnostic that could discriminate between healthy young and healthy older humans. To this end, we hypothesized that production of an accurate and sensitive diagnostic, built using muscle tissue from young subjects contrasted with humans reaching their older age in good health, would provide the platform to produce a prognostic molecular 'signature' that could be applied to longitudinal studies for the purpose of forecasting health outcomes. To delve into the uniqueness of the discovered gene set we would further conduct a comparative analysis of some of existing signatures of human ageing. This should potentially provide a well-rounded perspective about relevance of different signatures of human ageing and their limitations, if any.

1.3 Outline

The research work consists of six chapters. The first chapter is the introduction and review of the research that establishes the concept of understanding the importance of biomarkers in ageing and describes all the important aspect that is necessary for completing the research. The second chapter begins with general methods to process and understand microarray data and discusses method development part of research through which the results will be conducted.

The third chapter is the results section that explains the reader the result and discoveries of the research topic. It is dedicated to the validation and application of the transcriptomic signature obtained from second chapter and thus provides the outcome of the research. Here, we investigate ageing with reference to neurocognitive health with particular emphasis on Alzheimer and Dementia. Further to fulfil the research goal and provide a comprehensive evaluation, we compare different genomic and transcriptomic signatures of human ageing and longevity in fourth chapter of this thesis. This chapter explores the uniqueness of our 150 gene set signature for neuro-muscular ageing with respect to the other published genomic signatures of human aging. In the next chapter of this research work we develop a linear model for vascular ageing to explore if vascular ageing is distinct from neuro-muscular ageing, from the stand point of RNA gene-sets. The results drawn

from this work will highlight the importance of clinical (practical) significance of an effect and the pitfalls of formulating inferences solely based on statistical weight. The last chapter is based on discussion, and explains in detail the most important factors in the findings. The discussion section translates the outcomes in this area and point out any additional findings. The conclusion segment in the discussion relates the study discoveries and clarifies and derives the overall outcome of the research with recommendations and implications.

1.4 A review of human ageing

Ageing is described as the process that influences the human physiological system and its performance and increases the chances of death and chronic diseases. It is a genetically complex multi-causal biological process that is certain and leads to a decline in adaptive capacities (K. Christensen et al. 2009). It unavoidably leads to death as it is accompanied by the development of age-related pathologies (Zaidi 2008). The general perception of human ageing includes the reduced ability of surviving chronic diseases as well as mobility loss, decline in cognitive or sensory functions and increase in health costs.

Visual examinations, biochemical analyses, physiological, psychological, and functional tests are used to examine age-related changes through the methods of conventional evaluation. The phenotype of ageing is considered as a complicated interaction of stochastic factors along with the genetic, environment and epigenetic variables (Rattan 2006). These variables favour molecular fidelity loss and intensify the random damages in the tissues, cells or in the human being as a whole. Along with these, the chances of death and disorder also increase (Candore et al. 2006).

However, it is observed that the processes that influence biological ageing are not apparent. This insight can be provided by identifying biomarkers as it can explain the heterogenetic aspect of the functional decline related to ageing (Niccoli & Partridge 2012). Thus, biomarkers are in need of urgent evaluation for the assessment of health conditions of elderly individuals, which could aid in developing therapeutic interventions.

1.4.1 Population ageing

The term ‘Population ageing’ is used to define the shift in the distribution of age of a nation towards older individuals. It is a concept that challenges gaining lives with longevity and is an outcome of decreased fertility and increased expectancy of life (Dobriansky et al. 2007). Older individuals and ageing population are considered as the most important demographic and global trend of the 21st century. As per UNFPA, every ninth individual in the world is above 60 years (UNFPA and HelpAge 2012). Greying of the population is inevitably going to occur in the next decades and this transition in demographics is unprecedented in human history, which will have

major implications on all aspect of life. It will be a real threat to retirement systems and social security systems. Nevertheless, the extent to which it will occur is uncertain and will particularly depend on future trends in mortality.

Population ageing is considered as the rapidly progressing problem worldwide. The countries that faced decline in fertility and mortality in the beginning are now facing increased proportion of elderly people (Lutz et al. 2008). The global share of older people (aged 60 years or over) has swollen from 9.2 in 1990 to 11.7 per cent in 2013 and is projected to reach 21.1 per cent by 2050. The demographics of UK and Europe are also dynamically changing. The demographic analysis conducted recently predicts that Europe's average age in 2020 will be 42.2 years in comparison to 39.8 years in 2010. Life expectancy rate has also been predicted to increase in 2020 to 77.84 years compared to 2005-2010 that was 75.34 years (United Nations, Department of Economic and Social Affairs 2013). As shown in Figure 1.1 the global median age has moved from 24 years in 1950 to 29 years in 2010, and will continue to increase to 36 years in 2050. Furthermore, within the older population itself the proportion of those aged 80 years or over has doubled from 7 per cent in 1950 to 14 per cent in 2013. This rise is occurring at a faster pace in the less developed regions than in the more developed regions (United Nations, Department of Economic and Social Affairs 2013; UNFPA and HelpAge 2012).

Thus, world is going through a demographic shift which is a direct consequence of health transition occurring globally at different speeds. People around the world are living longer, but many are also living sick-lives for long. This transitional period is impacted by different interconnected influences which includes the change from high to low fertility rates, a steady rise in life expectancy at birth and at advanced ages and a move from mainly infectious diseases to non-transmitted diseases and chronic conditions (International Conference of Social Security Actuaries and Statisticians 2009).

1.4.2 Social and Economic Effects of Ageing

The major factor that is connected with the addition of older ageing population is the severe challenges it poses on the traditional state of social welfare and economy of a nation (Turner et al. 1998). In many developed countries, it is observed that ageing increases pressure on social security programs. It is now considered that the policies are in need to be addressed in order to maintain and manage the retirement programs (Lefebvre & Goomar 2005). The ageing population is now a global concept for definite reasons. It not only poses a challenge to the society's security system, but is a great struggle for the health care systems as well. As humans age, the incidence of illness, disabilities and likelihood of age-associated diseases such as alzheimer, dementia, or diabetes

increases noticeably (Niccoli & Partridge 2012). In addition to this a substantial increase in health care expenses occurs, not only because of higher proportion of elderly people in the population, but also as a result of increasing costs per person, among others due to new and more expensive medical technology. Furthermore, increasing trends in physical and mental functioning may lead to an increased demand for formal and informal care, while at the same time sources of support decline (Rechel et al. 2013). One approach of addressing this health and social care challenge is by maintaining health and reducing disability among the elderly and more importantly ensuring to extend the period that older people remain healthy, independent and contributing to society (Figure 1.2) (Gavrilov & Heuveline 2003).

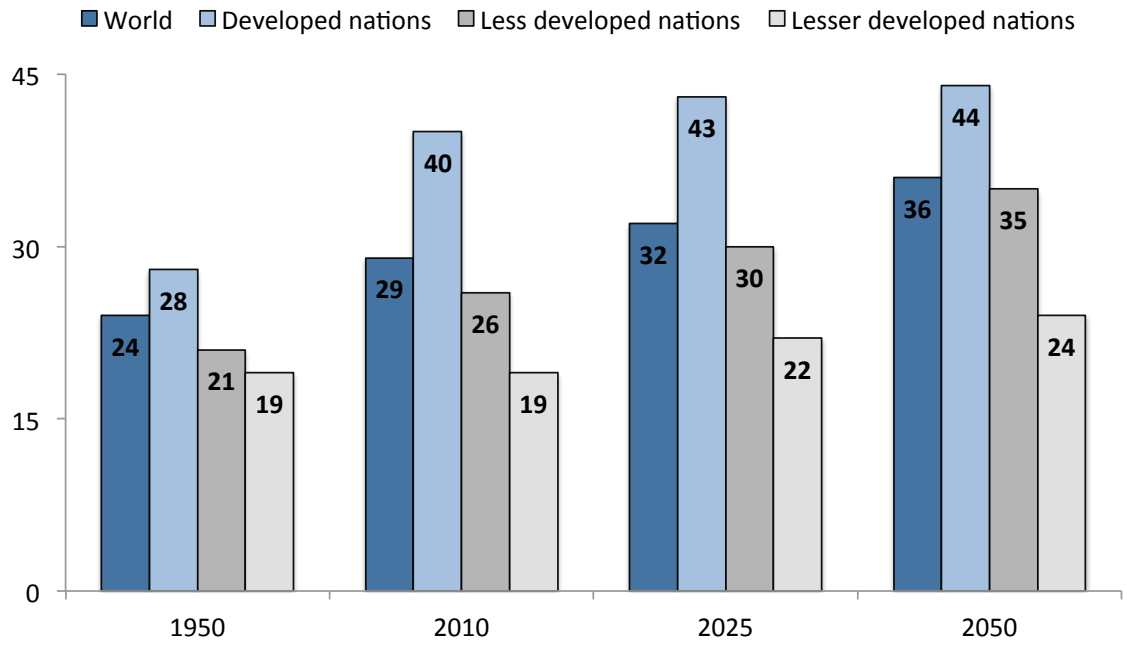


Figure 1.1 The median age of the population for the world and developed and developing nations. *The data by Department of Economic and Social Affairs Population Division, United nations across years shows that median age of the population is rising across the world in developed as well as developing nations thus establishing that population ageing is rapidly progressing problem.*

Economically, ageing has an impact on the labour market directly because of the influence of life expectancy and health on the way an individual behaves and make the decision of either working longer or retiring. The ages of 14-64 years old are considered to positively affect the economy and makes productive approaches, however ages 65 and above are considered dependent. Thus, in this regard population ageing can prove to be a challenge for the world economy (International Conference of Social Security Actuaries and Statisticians 2009). The problem due to ageing influences negatively in the growth of economy and in the participation rate of the labour

market which forces the researchers to analyse the changing directions through the trends of early retirements. In UK, the ratio of working individuals as compared to the ratio of people that are above 65 fell from 3.7 to 1 in the year 1999 and could fall from 2.1 to 1 in 2040 (Office for national statistics 2012). This estimation suggests that dependency ratio is about to be increased, which means more candidates of pension claiming individuals with less working individuals. This is a matter of concern for the government as the number of taxpayers will be reduced and dependency on government's funds will rise. The chances are that due to augmented dependency ratio, diseases, and human ageing, government will have to increase spending on pensions and healthcare as well. This will result in higher tax rates and less people paying it (World of Work 67 2009). Therefore, it is crucial to invest in discovering more successful ways of preventing or treating the major causes of illness and disability.

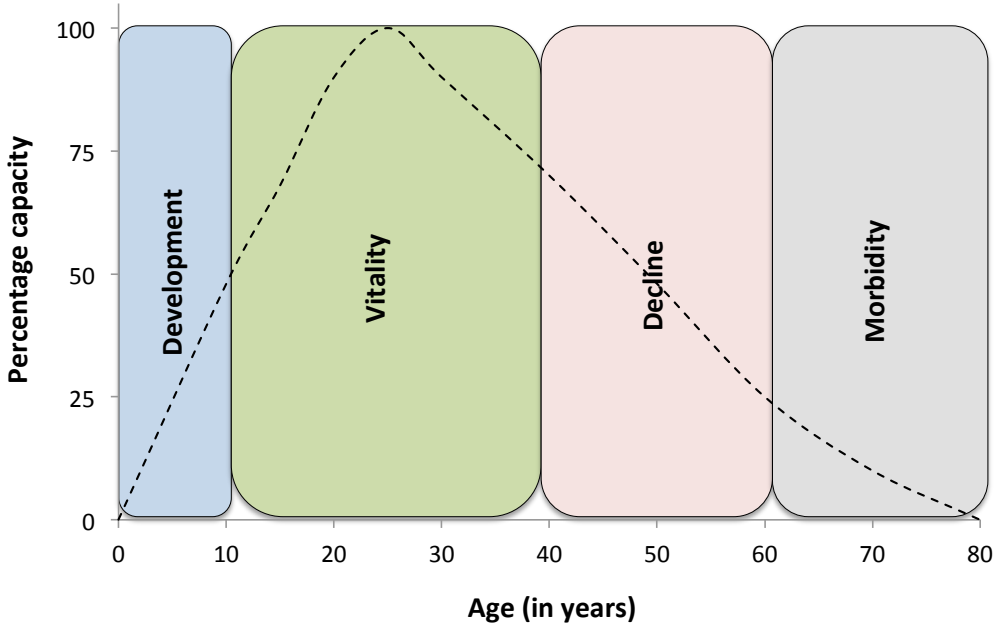


Figure 1.2: Different phases in lifespan of an individual. *An individual starts from an initial phase of development, followed by a peak in vitality of physical and physiological capacity, followed by a period of steady decline ending in morbidity. Instead of simply aspiring to extend the maximum lifespan it is important to have interventions that could possibly reduce the morbidity phase and move the curve towards right* (Larrick & Mendelsohn 2010).

1.5 Biological and molecular hallmarks of ageing

A progressive loss in the functional decline of an individual is considered as ageing, which leads to functional impairment and susceptibility to death. Throughout the history of humankind, the field of ageing has attracted a number of researchers (Zaidi 2008) as it is a basic risk-factor for many physiological and psychological human diseases such as diabetes, dementia, etc. (Niccoli & Partridge 2012). Many researchers have given their effort in the identification and categorisation of

cellular and molecular hallmarks of ageing. Here we discuss nine hallmarks of ageing, which have proved their contribution in ageing process, and mutually determine the phenotype of ageing (López-Otín et al. 2013).

1.5.1 Cellular Senescence

In 1961, Hayflick and Moorehead first described cellular senescence as a process that limited the proliferation of somatic human cells in culture. They suggested that this interruption of cell growth in culture reflects ageing in-vivo and also limits the lifespan of an organism (Hayflick & Moorhead 1961). It is understood that senescence is brought about by a variety of inherent and extrinsic stimuli including DNA damage, physiological stress, telomere shortening and stimulation of cancer-causing genes (Cech 2004). Senescent cells normally experience vivid structural and functional changes and are characterised by extremely distinguishable gene and protein expression profile. For example, increased adhesion to the extracellular framework and a levelled and highly augmented phenotype with a vacuolated morphology (Collado et al. 2007; Narita et al. 2003). Progression of ageing with senescence occurs in two ways that is by loss of the beneficial, replicative capacity of certain cell types and through the creation of proinflammatory cytokines which constitute the senescence-associated secretory phenotype (SASP) (Tchkonia et al. 2013).

This implies that aggregation of senescent cells with ageing is also associated with the production of proinflammatory factors. Chronic inflammation due to progressive accumulation of senescent cells is a defining characteristic of mammalian ageing, which promotes several age-related phenotypes, including neurodegenerative pathologies, such as loss of brain function, as well as proliferative diseases such as cancer (Newgard et al. 2013).

1.5.2 Telomere Attrition

Age-related DNA damage accumulation results in affecting the genome near-to-random, but some chromosomal region such as the telomeres are more prone to age-related deterioration (Blackburn et al. 2006). As cells divide repeatedly, small portion of telomeric DNA is lost with each cell division because of limitations of the DNA polymerases in completing the replication of the ends of the linear molecules, leading to telomere shortening with every replication. When telomere length reaches a critical limit, the cell undergoes senescence and/or cellular death. This restricted proliferative capacity due to telomere exhaustion is termed as Hayflick limit or replicative senescence (Blasco 2007). Thus, telomere length can serve as a sign of a cell's replicative activity

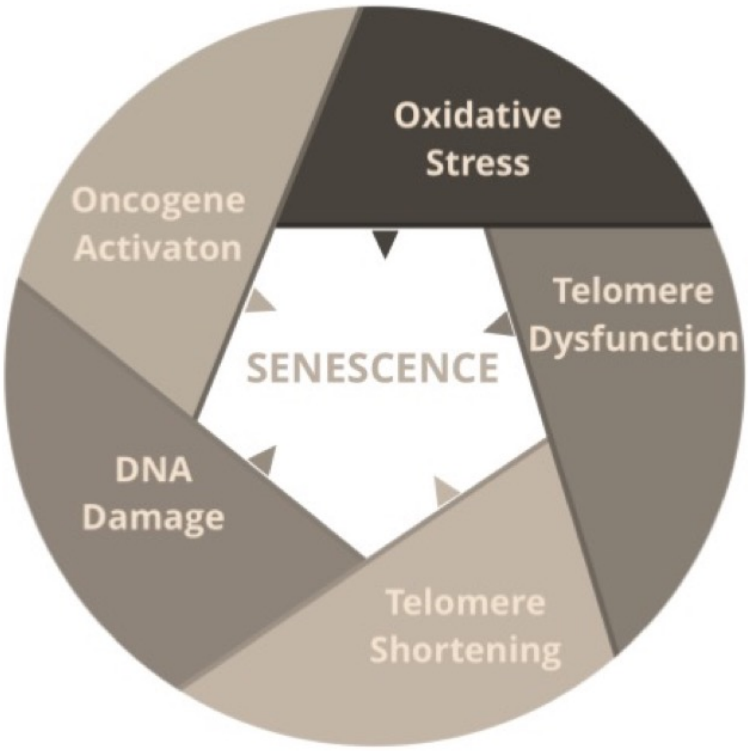


Figure 1.3: Proposed causes of cellular senescence. *Cellular senescence can be caused by a number of cellular stresses including oxidative stress, telomere dysfunction etc. Senescence prevents proliferation of potentially damaged cells and is initiated by inherent and extrinsic stimuli.*

rather than chronological age. Shorter telomeres have additionally been connected with the genomic instability and oncogenesis. Rate of telomere shortening has therefore been considered critical to an individual's health and pace of ageing (although more recently it has been accepted that telomere assays have proven insufficient to be prognostic for disease in the clinic). It has likewise been embroiled in pulmonary fibrosis, aplastic anemia and congenital dyskeratosis and all other premature developmental diseases that include the loss of regenerative capacity in varying tissues (Walne et al. 2008; Savage et al. 2008)

1.5.3 Genomic Instability

Genomic Instability affects a range of diseases and has been considered as one of the major causes behind ageing. Somatic cells are exposed to numerous sources of DNA damage such as environmental mutagens, UV radiation and exposure to reactive oxygen species(Nicholson et al. 2011). An intricate network of genome maintenance system acts to cope with millions of such attack on cell genome and restore the genomic base pair sequence that needs to be correct for normal functioning. Generally, the reason for epimutations and mutation is the flaw in this self-

repair phenomenon. However, occasional failures in correct replication of genome during the times of cell division is also a contributing factor (Boyette & Tuan 2014).

Genomic stability is an essential element in all eukaryotes' ageing but how it is related still remains unclear. The free radical theory of ageing hypothesizes that the oxidative damage to DNA and other cellular components is the key determinant of ageing (Harman 1955). Most exploratory proof for the free radical hypothesis of ageing originates from invertebrate models such as transgenic fruit flies. Recent alterations of this hypothesis states mitochondria as being responsible for oxygen species generation and oxidative damage (Vijg & Suh 2013).

1.5.4 Mitochondrial Dysfunction

Mitochondria are central regulators of various important cellular processes (Tait & Green 2012). The decrease in mitochondrial capacity compromises cellular integrity and has been proposed as one of the reasons for ageing (Sohal & Weindruch 1996). Mitochondrial DNA (mtDNA) mutations, dysfunction of the electron transport chain, oxidative stress, disparity in mitochondrial turnover, impaired trafficking and disruption of the fusion/ fission machinery are broadly involved in disease pathology and ageing (Wallace 2005).

Aged cells in post mitotic tissues, such as brain, heart and skeletal muscles, often have poor respiratory capacity because of mitochondrial dysfunction. Incidentally, these cells are regularly connected with aggregation of mtDNA mutations, surpassing the limit important for supporting mitochondrial capacity (Park & Larsson 2011). In spite of the fact that mtDNA just involve around 1% of the total DNA present in the cell, lots of evidence suggests that its role in cell physiology may well be far important than projected by its sum or size (Bratic et al. 2013). Because of the oxidative microenvironment of mitochondria and the absence of protecting histones, mtDNA becomes responsible for age-related mutations of somatic cells. Comprehending how mtDNA mutations proliferate and clonally expand in cells is critical in explaining the development of mitochondrial diseases along with the ageing process (Singh 2004).

1.5.5 Stem Cells Exhaustion

Adult stem cells dwell in most mammalian tissues, yet the degree to which they add to homeostasis and repair varies broadly (Wagers & Weissman 2004). These rare and specialised cells with the ability of self-renewal are required for tissue substitution all through the human lifespan. Tissues regenerative potential diminishes with age, hence a question emerges whether the attributes of an ageing tissue can be understood via declining utility of the adult stem cells that reside in it (Singh 2004). There are numerous reasons for stem and progenitors cell exhaustion with ageing, one of which is genomic instability as discussed before caused due to DNA damage which comes from a

variety of sources including the free radicals originating from normal metabolic respiration and failures amid DNA replication. Lack of an appropriate preventive response to DNA damage could lead to cancer initiation and progression and in order to avoid this regenerative potential is decreased. Thus, tumor suppressive response to DNA damage to prevent uncontrolled cell proliferation, could cause loss of stem cell function (Ruzankina & Brown 2007). Additionally researchers have also connected epigenetic modifications such as changes in DNA methylation, to the loss of regenerative capacity of stem cells with age (Vas et al. 2012).

1.5.6 Dynamic Alterations in gene expression and transcription

Among the three principle groups of biological macromolecules particularly connected with exchange and expression of genetic data i.e. DNA, RNA and proteins, the RNA stage is the least studied for conceivable age-related changes. Ageing process is connected with an increment in transcriptional noise, and an atypical generation and development of numerous mRNAs. A gradual loss of fine-tuning of gene regulatory pathways would presumably lead to a dysregulation of gene expression, which could have deleterious effects (Rivas et al. 2014). Some studies have tested the effects of ageing on gene expression using microarrays in model organisms and found variation in gene expression in some tissues of ageing mice (Weindruch et al. 2001). Microarray-based examinations of young and old tissues from a few animal types have recognised age-related transcriptional changes which are regulated by a small set of GATA transcription factors. The authors hypothesised that this network of GATA factors has evolved to regulate gene expression during development but may become unbalanced in old animals, thereby effecting the changes in gene expression observed with age (Budovskaya et al. 2008). Another study observed a shared gene expression signature for ageing in human, mouse and rat by comparing expression changes in seven microarray datasets from these organisms (Wennmalm et al. 2005). Further global RNA profiling using differential expression and regression models have been utilised to search for consistent molecular events correlating with age (Willemijn M Passtoors et al. 2012; Gheorghe et al. 2014; Phillips et al. 2013; Glass et al. 2013; Peters et al. 2015).

1.5.7 Proteostasis Dysfunction and Loss

All cells exploit a variety of mechanisms to safeguard the utility of their proteomes e.g. autophagy. Proteostasis includes systems for the adjustment of effectively collapsed proteins and proteotoxic stress, most distinctly the transcription factors including the heat shock group of proteins, forkhead factors and molecular chaperones (Taylor & Dillin 2011). Every one of these frameworks work in a facilitated manner to restore the structure of misfolded polypeptides or to evacuate and destroy them completely, consequently keeping the aggregation of harmed segments in control and

guaranteeing the consistent replenishment of intracellular proteins. Numerous studies have shown that this state of proteostasis is modified with ageing (Balch et al. 2008). In ageing cells over time, there is a functional decline in proteostasis machinery, resulting in continuous accumulation of damaged and mis-folded proteins, leading to reduced cellular viability and advancement of some age-related pathologies, for example, Alzheimer's disease, Parkinson's etc., (Rodier & Campisi 2011).

1.5.8 Altered Intercellular Communication

Intercellular communication is critical for coordination of physiology in multicellular organisms. Along with cellular level modifications, ageing also includes changes at the level of intercellular correspondence, impacting on endocrine and hormonal regulation (Downing & Miyan 2000). The functions of endocrine organs are linked in a such a way that reduced function in one alters the regulation of others. Another distinctive age-related change in intercellular communication relates to inflammation, which may come about because of numerous reasons such as aggregation of damage in pro-inflammatory tissue, decline in resistance framework's ability to successfully clear pathogens and the affinity of senescent cells towards pro-inflammatory cytokines. These modifications bring about an improved actuation of the NLRP3 inflammasome and other pro-inflammatory pathways which play a direct role in the pathogeneses of atherosclerosis, Type 2 diabetes and artery disease in the elderly (Baroja-Mazo et al. 2014). Overall, changes in metabolism and production of various hormones with age results in alterations in body composition characterized by decrease in lean body mass and bone mass and increase in fat mass. Furthermore, there is decline in functional status as well, such as reduced immune function, reduced capacity of the cardiovascular system, anaemia, insulin resistance and this leads to fatigue, depression and poor libido (Chahal & Drake 2007).

1.5.9 Immunosenescence or loss of immune function

As stated before, age correlated diseases such as rheumatoid arthritis (autoimmunity), cancer (e.g., prostate and lung), Type 2 diabetes(T2DM) and cardiovascular diseases (CVD) are major concerns for the elderly. The link between few of these diseases such as T2DM and CVD is immunity. With age, there is loss of immune functions known as immunosenescence which may explain the age-associated incidence of such diseases. Immunosenescence has been associated with an increased predisposition to diseases, malignancy, infections,poor response to treatments and impaired wound healing (Pawelec 2007).

The changes affecting the immune system often leads to global dysfunctions in both adaptive and innate immune system (Makinodan et al. 1991). An important contributor being impaired B and T

cells production in bone marrow and thymus and diminished function of mature lymphocytes because of which the elderly individuals do not respond to immune challenge as robustly as their younger counterparts (Montecino-Rodriguez et al. 2013). Immunosenescence also causes increased $CD8^+$ cytotoxic/suppressor cell numbers, and decreased $CD4^+$ T-cell and $CD19^+$ B-cell numbers which has been associated with increased morbidity and mortality (Ferguson et al. 1995). Further, in 'frail individuals' immunosenescence is often characterised by increased serum IL-6 and TNF-α levels causing chronic systemic inflammation, referred to as inflammaging (Franceschi et al. 2007; Franceschi & Campisi 2014)(Franceschi & Campisi 2014).

1.6 Age-Associated Diseases

Age associated diseases are conditions that generally manifest at advanced age causing disability or premature death (Partridge 2010). Physical and psychological disorders such as cancer, osteoporosis, cataract, arthritis, dementia, cardiovascular disease, diabetes, Alzheimer's disease, and hypertension are few examples of age-associated diseases (Boots et al. 2013; Wu et al. 2014; Barzilai et al. 2012). All these diseases and their incidence and severity increases with increase in age (World Health Organization 2011). Accumulation of damage and lack of repair mechanism at cellular, and molecular levels leading to the gradual decline of body function is considered as a starting point of several diseases and their pathogenesis. Some of these diseases result in change in hearing, muscular strength, vision, bone strength, nerve function and immunity (Niccoli & Partridge 2012). Ocular problems such as Glaucoma and cataract are also associated with ageing (Salvi et al. 2006). Ageing makes an individual more vulnerable to weaknesses and infections (Herbig et al. 2006). To comprehend this vulnerability of aged towards these diseases, it is crucial to understand the process of ageing and the underlying mechanism of these diseases.

1.6.1 Cardiovascular disease

Cardiovascular system pumps and supplies oxygenated blood to all parts of the body and is responsible for the health of every single tissue within an organism. Ageing being an inevitable part of life significantly affects the heart and arterial system and is the major risk factor for cardiovascular disease (CVD) including stroke, atherosclerosis (Coronary artery disease or CAD), myocardial infarction, and hypertension which are some of the leading causes of morbidity and death in the elderly population (North & Sinclair 2012). Various animal and human models have shown mechanisms such as oxidative stress and inflammation to play a central role in age-related cardiovascular dysfunction (Lakatta 2000; Judge et al. 2005). Ageing of the vasculature results in increased thickening and stiffness of the large blood vessels and also results in impaired endothelial function in the smaller blood vessels. Because changes in collagens are reported with ageing (Jani & Rajkumar 2006) and because large blood vessels contain high levels of collagens that determine

stiffness, it is possible large vessel status can be a biomarker for vascular ageing. Pulse wave velocity (PWV) is a widely used clinical measure of arterial stiffness. Alterations in the heart with age includes fibrosis, hypertrophy (increase in volume) and calcification. These physiological changes lead to increased systolic blood pressure, which increases the vulnerability for developing CVD and advances the risk of heart attacks.

In addition to non-modifiable risk factors such as gender and age, there are other risk factors for CVD, such as smoking, cholesterol, obesity, high blood pressure, lack of physical activity and diabetes (Anderson et al. 1991; Ambrose & Barua 2004; Sowers 2013) and these risk factors are thought to impact on small and large vessels to different extents. Different scoring systems have been developed to assess the risk of suffering from CVD, which work by allocating certain scores for each of the individual risk factors and then calculating a cumulative score, with higher score associated with higher risk profile. Amongst these, Framingham scoring system is the most popular approach that has been successfully applied to a wider population (Benjamin et al. 1994; Bhopal et al. 2005).

In terms of genomic knowledge of CVDs and stroke, enormous strides have been made over the past century, ranging from understanding cardiovascular physiology to molecular and cellular studies exploring the underlying risk factors (Feero et al. 2011). Methods such as whole genome sequencing and genome-wide association studies (GWASs) have recognized about twenty five loci associated with myocardial infarction and CAD by analyzing a large set of genetic variants by comparing case and control subjects from a population to determine which variants are associated with the disease in question (Anderson et al. 2010; Yasuno et al. 2010). Further, meta-analysis of these different GWAS based studies involving around ~90,000 CAD and control participants in addition to confirming the earlier observations have also identified 13 new loci associated with CAD (Smith et al. 2010; Feero et al. 2011; Schunkert et al. 2011). While this is a progressive step forward, it is important to know that CVD is a complex disease that occurs due to the sum of multiple polymorphisms, with each variant having a relatively small effect (<10%) on gene expression and disease. Transcriptomic signatures based on differential gene expression have been related to prognosis of CVD in humans and has indicated that quantitative differences in gene expression have the potential to define a person's phenotype (Heidecker et al. 2008). Thus combining genome-wide gene expression together with genetic variation could yield a far more promising approach that could disclose the link between transcriptional regulation and CVD (Dixon et al. 2007; Schnabel et al. 2012). Further age-related stiffening of large elastic arteries such as the aorta has also been an important predictor of future cardiovascular events. In chapter -5 we explore the possibility of building a vascular ageing signature using a skin gene expression data for subjects for whom clinical measure of arterial stiffness (Pulse wave velocity measure) were available. We

hypothesized that skin gene expression could potentially serve as a surrogate measure of elasticity and since with ageing there is a changes in collagen and loss of elastin, a model based on the transcriptome and clinical covariates could potentially provide an insight into a person's vascular (or extracellular matrix) ageing.

1.6.2 Sarcopenia

Sarcopenia is the progressive loss of skeletal muscle mass and function with age. It is often a result of or leads to decrease in physical activity which leads to functional impairment or disability and increases vulnerability towards other chronic ailments such as CVD and insulin resistance (Roubenoff & Hughes 2000). With age, there is a decline in mitochondrial biogenesis as well as reduction in the ability to promote muscle protein synthesis, which has substantial impact on the age-associated loss of muscle mass and strength. Both of which are the two most recognized risk phenotypes associated with sarcopenia. Previous work has reported enormous inter-individual variability in these phenotypes arising due to interaction of genetic and environmental factors (Cesari et al. 2006; Janssen 2006). Linkage and association studies have shown IGF1 (positive) and IL-6 (negative) group of genes to be statistically correlated with skeletal muscle strength and/or mass (Tan et al. 2012). Other factors that have been linked to the cause of sarcopenia include oxidative stress, poor nutrition and impaired regulation of growth hormones and sex steroids (McArdle et al. 2002; Rudman et al. 1990; Hickson 2006).

There is currently no fully accepted criteria or standardised technique that can diagnose and track sarcopenia related muscle decline. In clinical studies, it is commonly diagnosed using the appendicular lean mass calculation derived from the dual-energy X-ray absorptiometry (DXA) estimate. Though widely used, DXA tends to overestimate skeletal muscle mass as it is unable to discriminate muscle from water retention and muscle fat infiltration, thereby under estimating the extent of sarcopenia in an individual (Kim et al. 2002). This lack of an accurate and well established diagnostic criteria to identify patients with sarcopenia hinders the potential prevention and management options. Examining factors that determine skeletal muscle mass can possibly help in understanding and treating the age linked sarcopenia. Studies have indicated that resistance exercise training can stabilize the progress of sarcopenia as resistance training enhances muscle protein synthesis and improves muscle protein quality (Melov et al. 2007; Phillips et al. 2013).

1.6.3 Dementia and Alzheimer

Dementia is a loss of cognitive abilities in multiple domains that results in impairment in normal activities of daily living and loss of independence. Alzheimer's disease (AD) is a progressive neurodegenerative disease and is the most common form of dementia that accounts for 60–80% of all dementia cases (Jessen et al. 2011). The primary risk factor for AD is age. Around 7% of the

population above 65 years of age have dementia (at least 65% of these have AD), and with the shift in population demographics in the coming decades >150 million Europeans will be aged above 65 years (>1.2 billion, world-wide) (Harper 2014).

It is believed that the cellular and molecular alterations causing brain circuitry dysfunction in AD have a slow onset and full blown disease may take many years to develop. Alzheimer's disease is characterized by the accumulation of β-amyloid peptide (Aβ) within the brain and hyper-phosphorylated and cleaved forms of the microtubule-associated protein tau in elderly people (Kolarova et al. 2012). Also, unlike many other disorders and illnesses that can be associated to their fundamental cause, AD has been identified to be a result of combination of biological and environmental factors. These include mutations in the apo-lipoprotein, presenilin-1 and presenilin-2 genes, prior head injury, or having the ε4 allele of apo-lipoprotein (Hardy 1997; Kensinger & Corkin 2009; Mormino et al. 2014).

The early symptoms of AD are loss of episodic and working memory, which are due to network disconnections caused by oligomeric Aβ (Donohue et al. 2014). Alzheimer's disease causes severe suffering for patients and emotional distress to the family. The gradual disease progression is accompanied by cognitive decline related to memory impairment as well as decline in motor functions, attention, higher-order functions, personality as a whole and recognition of objects (DE Toledo-Morrell et al. 2000). Risk of AD is linked to ageing but also lifestyle diseases, such as T2DM and hence also related to physical activity. There has been preliminary evidence suggesting that physical activity such as walking and exercise may reduce some of the negative characteristics associated with cognitive impairment and reduce the risk of dementia (Ahlskog et al. 2011)

1.6.3.1 Diagnostics for Alzheimer's Disease

It is important to understand the early and asymptomatic states of the disease with the aim of proposing preventive therapeutic strategies. Around £26 billion is spent on health and social care activities for the 850,000 dementia patients in the UK alone and 250,000 new cases are expected each year (Dementia UK 2014). There is an urgent need to validate an AD diagnostic for primary physicians for a variety of social, medical and economic reasons – including a perceived under-diagnosis (e.g. Dementia Action Alliance report). Ultimately a treatment that will prevent or dramatically slow the progression of AD will be useful but this will not be possible without advances in population screening and robust diagnostics. MMSE scores (Folstein et al. 1983) is widely used to characterise cognitive decline in the elderly with scores below 24 commonly used to indicate a cognitive deficit. It serves as a neuropathological criterion for the diagnosis of AD.

However, the score is known to be affected by age and education and has a potential for misclassification or wrong diagnosis if not carefully reviewed by a trained clinician (Crum et al. 1993; Wind et al. 1997).

There are currently no drug-treatments for AD that halt or cure the disease (Salloway et al. 2014). Clinical view is that only the earliest possible intervention is likely to significantly impact on the structural features of neurodegeneration (e.g. such as anti-beta-amyloid compounds). However, currently available diagnostic techniques are neither scalable for mass population screening nor sufficiently cost-effective to be practical (Biasutti et al. 2012). For example, magnetic resonance imaging (MRI) combined with contrast agents has less than 5% utility in a 'screen and medicate' cost-effectiveness analysis. To date advances have been made in medical imaging and bioassays to confirm evidence of already extensive neurodegeneration (e.g. cerebral atrophy on MRI or CSF levels of beta-amyloid species) but these are expensive, invasive techniques requiring specialist medical centers and are technically restricted to specialists.

1.7 Human brain, ageing and cognition

Empirical studies of healthy ageing have indicated that older adults while still being mentally fit, become slower and gradually start developing cognitive problems evidenced by reduced performance on different kinds of memory tests (Folstein et al. 1983; Glisky 2007). The age-related decline in memory and performance is believed to be a result of failure to control the cluttering of irrelevant information. This loss of functional specialization could be a result of reduced neuronal integrity. It is believed that the cognitive functions that are functionally separated in young adults show reduced differentiation in older adults (Reuter-Lorenz 2002). Thus, cognitive impairment is one of the key determinants of advancing age and a major challenge for healthy ageing (Hanninen et al. 1996).

Cognitive processes are dependent upon the integrity of the brain, with age-related cognitive decline being widely associated with changes in the brain structure and function including neurochemical changes (Perry et al. 1982; Strong 1998), cerebral atrophy (Raz et al. 2005), reduction in brain volume (Walhovd, Fjell, Reinvang, Lundervold, et al. 2005) and reduced blood flow (Newberg et al. 2005). Research to understand the complexities of brain anatomy and its structure and function has been pursued for long and is still an on-going process. Though each of the brain regions have individual processes attributed to it such as cognition, problem solving ability, memory and response to sensory, spatial and visual stimuli (Goodale & Milner 1992; West 1996; Coutlee & Huettel 2012) brain functions as a single unit by interaction between the different sections (Spreng & Mar 2012). In spite of the functional interconnectivity, the advent of ageing

occurs at different rates and in different manners between these regions. These differences primarily can be observed in neurogenesis restricted to specific brain regions and the distinct deterioration in size with age (Raz et al. 2005). Biological changes associated with ageing eventually leads to dedifferentiation between cognitive and perceptual functions. Numerous genetic and environmental factors impact cognitive abilities (Winocur 1998; Mormino et al. 2014). To understand age-related decrease in cognitive functioning, it becomes imperative to study the changes in brain morphology and functioning. Deterioration in the brain regions was earlier considered to be the result of neuronal loss (Kemper 1994). However, with the advent of techniques such as positron emission tomography (PET) and MRI, researchers have found that neuronal loss is trivial, and atrophy could possibly be the result of reduced synaptic density, cell shrinkage and dendritic regression (Morrison & Hof 1997). Previous studies appear to present conflicting findings, especially when the brain is examined as a whole so as to assess the specific effects of ageing within each area it is important that each region of the brain is studied independently (Scahill et al. 2003).

1.7.1 Ageing and structural variability among the brain regions

Impact of the age-related effects are region dependent. Various cross-sectional studies have ascertained neuroanatomical age-related volume differences with different age trajectories for different brain regions. Some regions degenerate in a linear manner from early in life, whereas for some regions age-related volumetric changes are curvilinear where they continue to increase in volume, then stay constant (a plateau phase) and eventually begin to deteriorate (Walhovd, Fjell, Reinvang & Lundervold 2005). The greatest age-related change occurs in the striatum and frontal lobes with decrease in gray matter volume and an increase in white matter lesions. Volume losses within this region may contribute to age-related cognitive decline (Meguro et al. 2001; Abe et al. 2008). Deterioration within the gray matter structures has not been able to predict decline in cognitive functions. Presence of lesions in the white matter tracts that interconnects cortical to subcortical regions disrupts the neural transmission and might result in cognitive dysfunction (De Groot et al. 2000). Ageing is also accompanied with accelerated degeneration of the hippocampus and putamen (Jack et al. 1999; Laakso et al. 2000; Walhovd, Fjell, Reinvang & Lundervold 2005).

1.7.2 Ageing and brain transcriptome

Brain tissue has a high level of gene expression, with approximately ~45% known protein-coding genes expressed across all the different brain regions (Colantuoni et al. 2000; Myers et al. 2007). Ageing of the brain is characterized by varied complex events, and studies have shown existence of a robust relationship between gene expression levels and brain ageing (Berchtold et al 2008, Kumar et al 2013, Kang et al 2011), but only few of these age-related expression changes have been

consistently found across datasets. A comparison of global expression patterns across different human tissues have shown a distinct gene expression profile for brain than the rest (Saito-Hisaminato et al. 2002; Roth et al. 2006). Further, studies utilizing microarray and RNA-sequencing technology have shown higher expression and intricacy in the transcriptome of the human brain than other tissues (De La Grange et al. 2010; Ramskold et al. 2009), which could be because of extensive changes in the physiology and function of the human brain throughout the lifespan (Oldham et al. 2008). Thus, transcriptional profile of brain is always evolving with a brief duration of 10-15 years at ~30 to 45 years of age where it stays constant. Thereafter, with the advent of old age (>60 years) numerous changes in gene expression are evident impacting its mental and cognitive ability and brain plasticity (Somel et al. 2009; Colantuoni et al. 2011).

The enrichment of the brain transcriptome is accomplished by high level of alternative splicing events that enables the brain to express different isoforms of a gene (Wang et al. 2008; De La Grange et al. 2010). Aberrant splicing not only occurs in case of neurodegenerative disorders (Perez-Tur et al. 1995; Shehadeh et al. 2010), but also happens during normal ageing of the brain. Gene expression across brain regions also shows high variability due to anatomical and functional differences across regions (Roth et al. 2006). Over all cerebellum has the most distinct profile (Khaitovich et al. 2004) and this low concordance of gene expression is also observed within different cortical regions and neocortex (Strand et al. 2007). Therefore, it is safe to conclude that different brain regions exhibit changes with age, however with different rates and manifestations. Thus, to study the impact of ageing it is vital that every area of the brain is reviewed autonomously.

1.8 Biomarkers of ageing

It is critical to investigate healthy biomarkers of ageing to develop interventions that not only improves the healthy aspects of elderly but also stipulates approaches that monitor aspects of early or subclinical disease. It is also important to mention that our primary interest is in identifying biomarkers for ‘better’ or ‘worse’ ageing rather than a diagnostic of a disease. Diagnostics for diseases, such as Alzheimer’s are very challenging because of the low prevalence in the at risk population (e.g. 50-60yr old prevalence <5%) and the imprecise clinical criteria for defining Alzheimer’s disease in living individuals (e.g. ~90% correct). These combine, statistically, to make a high true-positive rate extremely challenging. Instead, if we could better define ageing, given it's the largest risk factor for Alzheimer’s, then we can identify an older population that has a much increased risk of Alzheimer’s disease in the future. In 1982, Reff and Schneider published a detailed set of criteria for the determination and measurement of a biomarker of ageing as follows (Reff & Schneider 1982):

a) **Highly reproducible in general and in cross-species comparison**
b) **Exhibits significant alterations over a relatively short time period**
c) **Critical for successful maintenance of health and disease prevention**
d) **Reveals a detectable parameter that can be predicted at a later age**
e) **Reflects some basic biological process of ageing and metabolism**

However, studies investigating into ageing biomarkers are restricted by several challenges. Several researchers study ageing with respect to morbidity and disease. However, age and disease are not biologically synonymous (Thompson & Voss 2009). These biomarker studies focus on age-related decline, which is not essentially a descriptor of healthy ageing. Inter-species differences in ageing also make this search for biomarkers more challenging. Living organisms are characterized by different lifespans, which implies that not all organisms age at the same rate (Piper et al. 2008). Genetic variation amongst different organisms have marked inter-species differences in the genes and proteins involved in the ageing processes thus implying that these processes are implemented and regulated differentially between organisms (Fontana et al. 2010). This restricted cross-species reproducibility becomes a greater challenge as most of the ageing studies are on model organisms. Many of the molecular mechanisms which extend the lifespan of laboratory animals have been reported to also positively impact on disease-free lifespan (Kenyon (2010) Nature 464: 504–512). Nevertheless, it has been difficult to establish if any of these are reliably modulated during human ageing (Phillips et al. 2013; Bell et al. 2012; Glass et al. 2013). Even if ageing-related molecular mechanisms are conserved across species, such molecules still may not represent reliable clinical biomarkers.

Validity of any biomarker with regard to ageing can be assessed by a well-accepted definition published by Baker and Sprott in 1988: “a biological parameter of an organism that either alone or in some multivariate composite way, in the absence of disease, better predict functional capacity at some later age than chronological age” (Baker & Sprott 1988). The molecules that naturally change with age are the only potential candidates for the signature of healthy ageing and in this sense a true biomarker of healthy ageing is unlike standard biomarkers that help in detecting or examining a disease. No single marker can give sufficiently high segregation of cases from controls as a diagnostic test for clinical applications. Thus, utilizing numerous markers consolidated in some kind of algorithm will be necessary to deliver requisite level of predictive ability. Figure 1.4 summarizes the characteristics of an ideal biomarker.

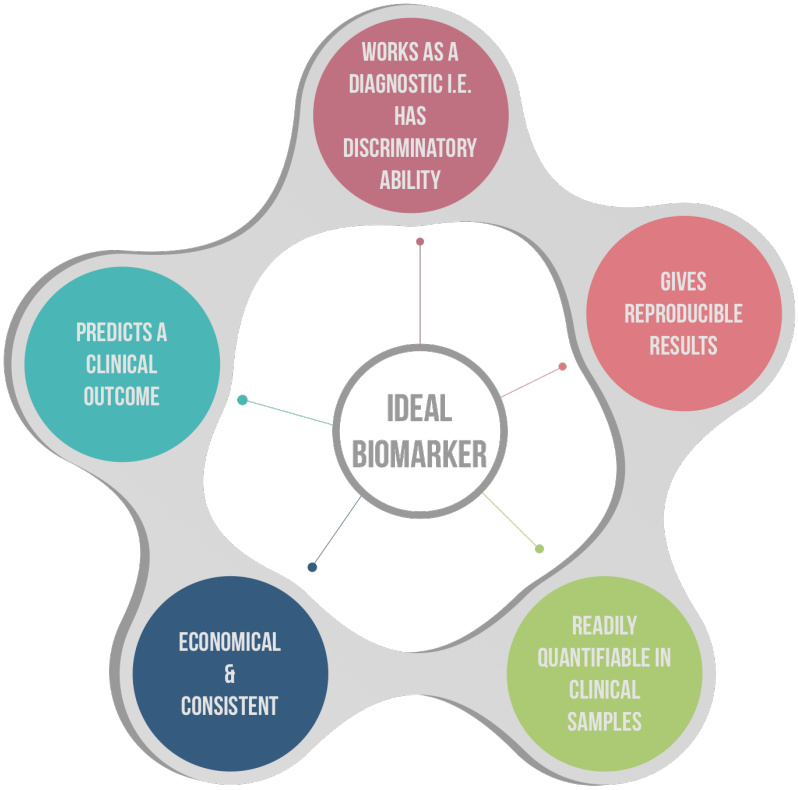


Figure 1.4 Characteristics of an ideal biomarker. *In addition to being an accurate reproducible diagnostic, a biomarker should have clinical abilities as well. Further, it should be quick, consistent, economical, and quantifiable in an accessible biological fluid or clinical sample.*

1.8.1 Approaches for Identifying Biomarkers

Different approaches are being explored to understand the distinctive physiological and pathophysiological processes that drive human ageing and longevity. These approaches can be broadly divided into genetic, epigenetic, transcriptomic and proteomic research (Deelen et al. 2013). In the past, genome-wide association (GWAS) methods, linear models of epigenetic regulation and differential gene expression have identified traits associated with ageing and exceptional longevity in humans and have attempted to explain factors driving age-associated disease risk (Sebastiani et al. 2012; Hannum et al. 2013; Horvath 2013).

1.8.1.1 Genomic (DNA) Approach

Identification of genetic variants associated with exceptional longevity in humans, using genome-wide association (GWAS) methods, is one approach that has been attempted to shed light on factors driving disease risk. For example, GWAS have reproducibly identified the APOE/APOC1 gene locus (Deelen et al. 2011; Sebastiani et al. 2012; Beekman et al. 2013), a locus associated with a rapid-ageing phenotype, Werner's syndrome (Yu et al. 1996) and a number of additional candidate

regions (Beekman et al. 2013). The largest study identified 281 single nucleotide polymorphisms (SNPs) that may explain up to 17% of exceptional longevity in humans (Sebastiani et al. 2012). However, long-lived humans share the common risk-alleles for coronary artery and other age-related diseases (Beekman et al. 2010) with people who have an average life-span, suggesting that long-lived humans have beneficial mutations that can compensate for these risk factors. The relationship between DNA variation and ageing has been proposed to involve many small-effect size variants (Yashin et al. 2010). This means that production of sensitive diagnostics from DNA samples alone, for the purpose of personalized medicine, may prove challenging. Indeed, establishment of a strong statistical association between a genomic variant does not establish if such a measure can be used to accurately diagnose risk, as the information available is unable to distinguish between two similar medical conditions with divergent treatment strategies.

1.8.1.2 Epigenomic Approach

Research examining animal models and twins has elucidated that individuals or organisms having highly similar genetic background can age at varying rates (Fraga et al. 2005). DNA undergoes several alterations as we age, some of these modifications occur without changing the genetic sequence or code (Holliday 1987). These modifications to DNA are epigenetic in nature and constitute DNA methylation, histone modifications and non-coding RNAs. Epigenetic changes at the key genomic regions such as transcription start site, promoter regions , etc., can switch on or off specific genes. Thus by controlling which genes are active in a particular cell, epigenome governs which proteins are transcribed locally within a cell type. The importance of epigenetic changes and its influence on longevity has already been established in model organisms such as yeast, worms, etc., (Greer et al. 2011; De Lencastre et al. 2010) as well as in humans (Fraga 2009). Epigenome is a primary location of gene-environment interactions and exposure to certain environmental stimuli can readily alter it (Aguilera et al. 2010). An age-related methylation drift has been observed and established, which is not uniform across the genome, and is quite variable between individuals of the same age (Rakyan et al. 2010; Hernandez et al. 2011). Thus this epigenetic landscape has been hypothesized as a 'biological marker' that reflects cell's identity, health and age.

Availability of affordable high-throughput techniques such as sequencing platforms and other genome wide technologies has helped to provide better insights of the epigenetic landscape specifically DNA methylation. Using an epigenome wide association approach scientists have investigated human longevity with methylation data on 172 females between an age-range of 32 y to 80 y (Bell et al. 2012). They noticed that the majority of age-related changes in DNA methylation were not related with phenotypic measures of healthy ageing such as telomere length, systolic blood pressure, etc., However, for small subset of genes they found that DNA methylation

mediated environmental and genetic effects on the age-related phenotypes. This implied that either DNA methylation has very small individual effect on measures of biological ageing or it may be associated with yet unknown ageing phenotypes or pathways.

Studies have examined the changes in DNA methylation as a potential marker for chronological age as well as biological age across species and cell types. These predictors of chronological age have been constructed across different tissues (Sinsheimer et al. 2011; Horvath et al. 2012; Hannum et al. 2013; Weidner et al. 2014). Hannum *et al.* built a multi-tissue linear model of DNA methylation in which age-related changes with DNA methylation state was closely related to chronological age but it could not distinguish between age and age-related disease. Horvath's epigenetic clock examined the relationship between DNA methylation and ageing by using ~8000 samples from brain, breast, skin, colon, kidney, liver, etc., with age range from newborns to 101 years. They developed a quasi-linear regression model of chronological age (~ 353 CpG sites) and transformed age in a unique manner for ages less than and greater than 20 y (log and linear transformation respectively). The divergence from chronological age (±3years only) was explained as the biological age of the sample. However, instead of being an actual disparity in ageing rates this slight deviation could imply a possible over-fitting of the specific model. Also, for most of the tissues the divergence from chronological age was minimal which raises the question on its utility to identify healthy ageing because a successful diagnostic of this type should show higher variability within a similarly aged population (chronological age).

For any epigenome-based markers it will be imperative to determine the effect of age-related changes in cell composition within tissues since methylation measures for these predictive markers are mostly taken on entire tissues (Zou et al. 2014). Even though there are many published epigenetic papers that have discovered markers of age and diseases such as cancer but very few of these relate the same markers in similar clinical specimens based on same assay technology. The consequence of employing diverse assays and varied markers discovered from them is that their real performance becomes incomparable. Further, these epigenetic assays are very poorly validated and need to be standardized if it have to be used as a diagnostic in clinical environment (Laird 2010).

1.8.1.3 Transcriptomic (RNA) Approach

Studies in model organisms and across different species have shown that ageing is characterised by molecular and physiological changes at cellular and tissue level. Identifying individual factors driving this multifaceted process is challenging as it is influenced not only by the genetic but also by the environmental factors such as diet, exercise, lifestyle, etc. Transcriptome has the ability to provide a better insight into age-related changes as the expression of RNA is under genetic

(Schwanhäusser et al. 2011; Westra & Franke 2014), epigenetic (Horvath et al. 2012) and environmental control (Keller et al. 2011; Larrouy et al. 2008). Transcriptomic study in skeletal muscle has successfully shown heterogeneity in gene expression profile for individuals from the age of 30 years, which implies that individuals have different ageing rate as they go through middle age and thus have diverse morbidity and mortality rates (Lu et al. 2004). The advent of global transcriptomic techniques such as next-generation sequencing and microarrays has helped researchers to unravel and understand the expression of whole transcriptome (Rodwell et al. 2004; Zahn et al. 2006). The potential of using RNA based classification approach for another major human phenotype i.e. adaptability of the aerobic capacity system in young and middle-aged adults has been previously established by our research group (Timmons et al. 2010). Machine learning methods applied to global transcriptomic profiles has already yielded sensitive and specific diagnostic and prognostic tools for cancer, using sets of gene-expression values of limited size (Shedden et al. 2008; Menden et al. 2013). However, in cancer studies gene expression changes are often of higher magnitude, to what can be expected from a study designed to identify healthy ageing, meaning it is unclear what size of gene-set would be required.

Ageing studies are usually carried out on cross-sectional datasets that cover individuals over a broad age-range. Several such transcriptomic studies have been designed and investigated across different tissue types such as skeletal muscle (Welle et al. 2004), kidney (Rodwell et al. 2004) , blood (Peters et al. 2015), brain (Erraji-Benchekroun et al. 2005) , etc. Even though there has been similarity in age affected pathways across these studies (partly reflecting inappropriate use of gene ontology analysis) but in terms of transcriptional changes with age there has been very limited overlap. This might imply that in each tissue different individual genes change their expression with age. However, a comparative analysis of transcriptomic data from different studies and different tissues such as muscle, kidney and brain found that along with tissue specific changes with age there is a possibility to find an underlying common ageing signature across tissues that might reflect the true biological age of the organism (Zahn et al. 2006). The Zahn study was limited by the fact that the old tissue samples originated largely from a very different type of muscle tissue than the young samples and was not possible to replicate in our more recent studies (Phillips et al 2013 PLos Genetics). A meta-analysis study performed on cross-sectional RNA data from healthy non-treated samples from adult mice, rat and human found that by integrating gene expression profiles from several studies it is possible to identify set of genes that are consistently regulated i.e., under or over expressed with age (De Magalhães et al. 2009) but without carrying out a formal robust tests as to the reliability of these genes to classify unknown samples.

Most of these earlier attempts of modelling the ageing phenotype have involved correlative

linear models that adjust for different age related covariates such as gender, blood pressure etc. This can be problematic as linear correlative approaches do not dissociate 'age' from age-related disease or drug treatment (unless the investigators strictly use healthy old subjects) specially when applied across a wide age range. For instance, a subject aged 40y can have the same measurement for an age related clinical covariate (blood pressure/ cholesterol) as a subject aged 60y. However, for latter the measurement could be an effect of a medication then an actual transcription profile reflective of its ageing. Therefore, both from a statistical and a clinical perspective linear correlative models for ageing are fraught with limitations without properly taking into account chronological age-range, concurrent drug-treatments etc. For this research work we have spent significant effort to build a good study design that could capture the trends of healthy ageing by taking into consideration strict set of benchmarks and established practices and have utilized classification statistical methods to evaluate which human tissues demonstrate a reproducible molecular signature for age (discussed in chapter-2 of the thesis). Thus, human ageing is characterised by focused changes in gene expression. If supported with good study designs and analysis, identifying such gene expression has the potential to yield robust biological ageing signature/biomarker.

1.8.1.4 Proteomics Approach

Proteome is defined as set of all expressed proteins that characterizes information flow within the cell or an organism and is considered as a dynamic reflection of both genes and environment. It is believed to hold a promise for biomarker discovery because proteins are ubiquitously affected in disease and disease response (Jain & Jain 2010). This is reflected in many protein disease biomarkers already available eg: CA-125 and alpha-fetoprotein (Bast Jr et al. 1997; Brock & Sutcliffe 1972). Thus, this approach is actively involved in the recognition of human physiology and its complexities. The techniques and procedures involved include mass spectrometry, western blot etc. While we gain much information from proteomic investigation it is complicated because of its domain size (>100000 proteins) and inability of the current technologies to detect low abundance proteins. Further, the quantity of data that is acquired with new techniques places new challenges on data processing and analysis (Chandramouli & Qian 2009).

Probing into the transcriptional range of a particular genome tells us more about the expression rather than its protein library (Hegde et al. 2003). This is because of the fact, that in eukaryotes there happens to be many regulatory RNAs which do not actively translated into proteins. Further there are long non-coding RNA which is mostly an enigma, till now (Kung et al. 2013; Cesana et al. 2011). Although they do not seem to be carrying anything out in particular, the sheer numbers by which they are transcribed indicate that they serve some purposes. Protein abundance regulation is much more convoluted then transcription regulation(Vogel & Marcotte 2012). In view of all this we believe that gene expression based biomarkers can potentially serve as a better proxy for biological activity associated with healthy ageing since transcriptomics are much cheaper and easier to do than proteomics.

1.9 Summary

Yielding strategies to cope with increase in the ageing population and age-related chronic diseases is important. Since the ageing process demonstrates large inter-individual variation, a sensitive diagnostic able to predict these characteristics would be useful. Personalized treatment strategies have high impact on modern medical practice and such strategies are essential if we are to achieve a greater degree of certainty that a particular treatment will benefit the individual patient. The challenge is to identify sets of molecular 'biomarkers' that provide sensitive and specific information, enabling long-term guidance for personalized treatment. However, there are numerous challenges to both the development of, and the implementation of personalized strategies for most major age-related diseases including economic constraints. Large experimental groups as well as a good study designs are required to enable reproducible conclusions to be made from studies of gene expression and ageing. Most studies for biomarkers of human ageing have been based on epidemiological cohorts that blend in ageing, disease and drug-treatment. However, a good study design and strategy to find candidate biomarkers for human ageing is to compare molecular traits in normative and healthy ageing in groups within the human population with no metabolic or chronic diseases. Further, there are multiple competing technological platforms that can yield plentiful data, but progress in integrating divergent data formats to yield robust and sensitive diagnostics for clinical decision making remains slow. Possibly, a pragmatic strategy will be to utilize a single technology platform with proven technical features that captures sufficient clinical variance, that it can provide a stand-alone and robust diagnostic for healthy human ageing.

2.1 Overview of the chapter

This chapter will describe the machine learning approaches used to develop an RNA expression signature of 'healthy ageing'. We begin by examining the strengths and limitations of technologies that provide insight into the human transcriptome including RNA-sequencing and microarrays, followed by a discussion of approaches we used to handle microarray data in order to generate the most robust signal possible. Following an overview of general methods relevant to this work, we describe in greater detail the development of our molecular diagnostic that was able to discriminate between healthy young and healthy older humans using a very strict set of methodologies and benchmarks. Thus, the principle goal of this chapter is to explain:

- Motivation for using the RNA classifier approach to produce the signature.
- Handling of the microarray data and extracting relevant information.
- Building the age-diagnostic.

2.2 Gene expression profiling

Nearly all individual cells within a multicellular organism contains of the same genome. However, within each cell, different genes are transcriptionally active, resulting in cells and tissue displaying different gene expression patterns. This results in a myriad of structural, biochemical, functional and phenotypic variations amongst cells and tissues that might play a role in the differences observed between health and morbidity. This complete set of transcribed genes expressed as mRNA within an individual is known as the transcriptome (Su et al. 2002). Gene expression profiles not only have the potential to explain cellular functions, regulation and biochemical pathways but when contrasted between cases and controls (e.g. normal vs healthy), the transcriptome may reveal insight into disease pathology and identify new therapeutic points of intervention, enhancing diagnosis and improving prognosis (Van't Veer et al. 2002; Xiong et al. 2013).

Transcriptomic changes are an important biological aspect of ageing (López-Otín et al. 2013; Glass et al. 2013). Indeed, variation in the regulation of gene expression, more-so than sequence variation, has been long postulated to be a more sensitive approach to studying ageing (King & Wilson 1975). The manifestation of profiling technologies and machine learning methods applied to global RNA profiles have already proven to yield sensitive and specific diagnostic and prognostic tools for cancer using sets of gene expression values of limited size (Patnaik et al. 2010; Shedden et al. 2008; Menden et al. 2013). While it is intuitive that a RNA profile obtained from a tumor demonstrates prognostic ability, the idea that a global RNA profile obtained from a non-diseased tissue sample can also produce an accurate and sensitive diagnostic that informs about

future disease has not been demonstrated.

2.2.1 Next-Generation sequencing and Microarrays

The development of transcriptome profiling technologies has allowed us unprecedented access to the world of RNA, with an ever-growing number of studies changing our view of its extent and complexity. Advances in molecular biology have brought utilization of microarrays and next-generation sequencing (NGS) technologies to the forefront of transcriptomics. Each of these technologies possesses a set of distinct features suitable for different applications and research goals. Current sequencing methods depend on the reconstruction of transcripts from sequenced fragments that generally do not exceed a few hundred nucleotides. These methods inevitably result in uneven coverage across the transcript (due to technical biases in the fragmentation and sequencing technologies), with the 5' and 3' ends often being the most problematic areas. With microarrays, RNA expression is measured through the amount of cDNA that hybridizes to pre-designed short DNA fragments, known as probes, immobilized on a chip. This limits the quantification of expression to areas in the genome that are matched by the probes. In addition to the need for having the correct type of probes, the distribution of probes must also be uniform (to an appropriate extent) across the transcripts' untranslated regions. Thus, arrays have a fundamental design bias i.e., one can only explore and analyze the transcriptomic regions for which probes have been designed. Also, arrays are highly dependent on reference databases from which they are designed. On the contrary, with NGS, reads are generated without any a priori knowledge of transcriptome, thus permitting analysis of novel transcripts, splice junctions and noncoding RNAs and defined based on current genome knowledge. Due to this potential for NGS technologies to provide a more detailed look at the transcriptome, researchers have been keen to use it for gene expression studies (Mutz et al. 2013).

Despite the methodological benefits of RNA-Seq, microarrays have several potential advantages over sequencing, particularly for detecting lower abundance transcripts. Hybridization in microarray typically uses higher concentrations of cDNA than RNA-seq assays, but the detection of each unique cDNA (or cRNA) is independent thereby avoiding the competitive detection scenario encountered with NGS data. With sequencing, the inability to detect a large proportion of lower abundant transcripts is caused by a few highly abundant RNA transcripts accounting for a very large proportion of a cDNA library (Lei et al. 2015). This inability to robustly detect low abundance transcripts leads to high variability in the quantitative measurement of transcript expression. Microarrays, on the other hand, provide coherent and accurate gene expression quantitation irrespective of transcript abundance. Use of microarrays for research remains prevalent, as the technology has been proven successful in consistently providing genomics insight

for the past two decades (Harrington et al. 2000; Trevino et al. 2006; Yan & Gu 2009). Also, microarrays are generally considered easier to use as protocols for sample labeling, array handling and data analysis are less intensive. Moreover, general agreement has emerged on the major methods for processing the data and a wealth of good tools exist to analyze them, while the same cannot be said yet for RNA-seq. Further, despite NGS advancements and a recent drop in the cost associated with NGS, expression arrays are still economical and easier when processing large numbers of samples (e.g., hundreds to thousands) and yield higher throughput.

There are pragmatic reasons for using microarray technology in a study such as ours as well. The primary research objective within this thesis was to find a biomarker or a diagnostic tool for healthy ageing that had prognostic abilities for a clinical outcome. There are many pre-existing datasets profiling a variety of tissues in young and old available on a variety of microarray platforms. Therefore, from a validation perspective microarray was a sensible choice (Figure 2.1). The different datasets used in our study were profiled on various microarray platforms including Affymetrix HGU133Plus2, Affymetrix HuEx-1.0 ST, HTA-2.0, Illumina HT-12 V3 beadchip and Illumina HT-12 V4 beadchip (for detail of the datasets see Appendix 1).

2.3 Handling microarray data by updating probe definition and annotation files

The most popular platform for genome-wide expression profiling is the Affymetrix GeneChip. However, the selection of probes to represent the totality of the transcriptome relies on genome and transcriptome annotation information available when a particular GeneChip was designed. Over time changes in the annotation of genome, leads to inaccuracies in the design time probe definition and this can affect the biological interpretation of the derived data (Sandberg & Larsson 2007). In this work, we tackled these critical concerns and implemented a solution for these design related drawbacks. A similar approach has been successfully implemented and employed in gene expression studies in the past (Dai et al. 2005; Greco et al. 2008).

A Chip Definition File (CDF) is an annotation file for Affymetrix chips that defines probes (cells in Affymetrix terminology) mapping to the genomic unit of interest. For instance, a CDF for gene expression will specify a sets of probes that maps to the same gene. Thus, different CDFs can be utilized to examine different genomic units (i.e. genes, transcripts, exons). Affymetrix provides CDFs based on design time annotations, collapsing a group of probes into an Affymetrix defined probeset. However, researchers have also developed 'custom CDFs' that are optimized for various genomic features. Custom CDFs reorganize the oligonucleotide probes on gene chip platforms based on the latest genome and transcriptome information allowing one to use the most updated annotation when analyzing the data (Dai et al. 2005). Custom CDFs can also be used define the

probes in alternative genomic contexts e.g. one could generate a custom CDF specifically targeting 5' and 3'UTR regions, noncoding regions etc. In addition, one can also resolve the polymorphism problem in the designed probes by removing the probes with SNPs' or indels (insertion/deletion of a nucleotide). Basically comprehensive polymorphism data are used to identify probes which cover regions with SNPs since polymorphism could affect signal integrity. These probes can then be removed from the CDF (Ramasamy et al. 2013). This can be useful as a genotype filter in case of unpaired analysis (see section 2.3.2).

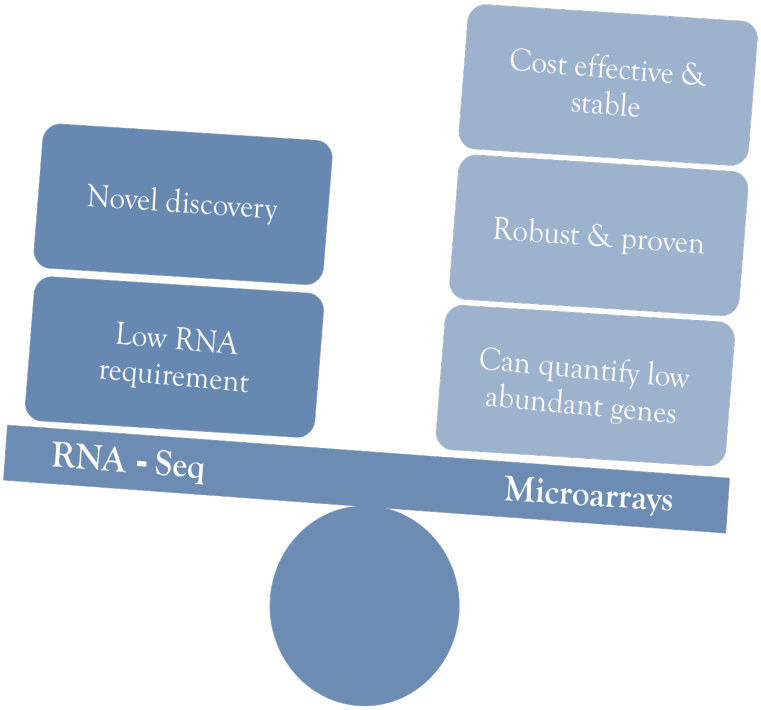


Figure 2.1: Comparison of gene expression profiling technologies. *Advantages of RNA-seq and Microarrays technology compared to each other. From personalized medicine and clinical perspective microarrays outweigh NGS.*

2.3.1 Generating custom CDF files

To generate these CDFs we start by aligning the probe sequences available in FASTA format (from the Affymetrix website) to the reference genome of interest using the bowtie alignment tool (Langmead & Salzberg 2012). We then extract the uniquely mapped probes from the alignment files i.e., probes that map only once to the reference genome, thus removing cross hybridising/multiple genomic loci probes. From the bowtie alignment file we create an annotation for each uniquely mapped probe comprising of the chromosome, strand, X and Y coordinate for the probe on the gene chip and also the probe start and stop position on reference genome. We additionally create a genome boundary file that defines the genomic region of interest for the CDF, which can range from a gene, transcript or an exon to 5' or 3' UTR. To create this file we first

download genome level data defining boundaries for different genomic units (genes, transcripts, exons) in General Transfer Format (GTF) from Ensembl. We extract the ensembl ID, chromosome, strand and start and stop position from the GTF file (Cunningham et al. 2015). It is important to note that we include only the exonic regions of a gene/transcript avoiding the regions spanning the introns, as these are unexpressed sections of a gene. So, for a gene level CDF, probes mapping to all exons from all transcripts are included in a single unit (gene in this case). Lastly using an in-house R script, we map the probe annotation file to the genome boundary file to produce a single tab-delimited file of all probes and their corresponding assignments to a genomic unit, with each row corresponding to a probe to be included in the custom CDF. The tab delimited file is converted into a custom CDF using a user defined function in R (Figure 2.2).

2.3.2 Resolving polymorphism in-probe problem

Microarray probes are typically designed to match one reference sequence only, based on reference genomes present in public databases at the design time. Sequences that depart from this reference, either due to the presence of SNPs' or due to the presence or absence of nucleotides (i.e. indels), often show a weaker binding affinity for the probe in question. Ramasamy et al proposes a solution to this problem which we adapted to generate custom CDFs by removing the probes with SNPs'/indels' (Ramasamy et al. 2013). These custom CDFs without SNPs will work as a genotype filter in case of unpaired analysis where we cross compare two samples from different subjects. The process comprises of three main steps as follows:

a) **Extracting the variant information by identifying and downloading the latest and most comprehensive polymorphism set.** There are different genetic variation databases available for reference in the public domain such as HapMap (The International HapMap Consortium 2010), exome variant server, 1000 genomes project (1000 Genomes Project Consortium 2010), differing in their completeness as well as diversity of the population profiled. For example: if our dataset comprises of subjects of European descent then the indels and SNPs information could be extracted from the European panel (n = 381) of 1000 genomes project. From this we would take into consideration only those SNPs and indels which have minor allele frequency (MAF) above a chosen threshold (> 1%) for polymorphism identification (threshold of 1% MAF in this case implies that the minor allele should exist in ~4 or more people out of the 381 profiled). We avoid being too stringent with the threshold because it can result in removing valid probes signals.

b) **Preparing the Probe file for the specific platform with genomic coordinates of each probe.** Unlike Ramaswamy et al. who used the design time annotations provided by the manufacturer to get the genomic coordinates of the probes on the reference genome

(GrCh37) we used the latest mappings, by aligning the probe sequences (in FASTA format) to reference genome GrCh37 using bowtie alignment tool.

c) **Comparing genomic coordinates of SNPs' and Indels' with probe coordinates.** Finally we used BedTools (Quinlan & Hall 2010) and files from step 1) and 2) to get the list of probes that contain these SNPs and Indels. Using the intersectBed functionality from BedTools we compare the probe coordinates and SNPs/indel coordinates and identify probes that overlap polymorphic sequence. We then remove these probes (as shown in Figure 2.2) before generating custom the CDF which collapses probes into gene, transcript or exon.

This ensures that our custom CDF's overcome the polymorphism problem in probes. One limitation of this approach is that it relies on variant databases for SNP/Indel information (currently completed for GrCh37 genome assembly) which are not updated along with reference genome (current version GrCh38) due to which one is unable to use the most updated genome information.

2.4 Feature selection from gene expression data

The goal of classification is to identify the features that can be used to predict class membership for new samples. Low reproducibility and the limited biological interpretability of candidate biomarker signatures identified from high-throughput data (microarrays, NGS etc.) is one of the key issues which impedes the use of discovered biomarker signatures into clinical applications. Under the circumstances, gene set analysis that investigates groups of genes instead of individual genes is becoming a trend in interpreting gene expression data.

2.4.1 Criticism of differential expression approach for biomarker discovery

With microarrays differential expression analysis of genes is key to classify features that relate to a phenotype and also helps in recognizing significant biological pathways. This feature selection based on differential expression analysis of gene expression data has been a widely used approach for identification of biomarkers. The differences in biological characteristics, e.g. genes, expressed across different species or conditions are generally investigated and those genes significantly changed are considered as differentially expressed. If a differentially expressed gene generally correlates very well with the phenotype of interest, then it is considered as a potential biomarker for that phenotype e.g. blood pressure, body mass index etc. The accumulation of wealth of the publically available gene expression data in databases such as Gene Expression Omnibus (GEO) has further helped in detecting genome-wide genes that are significantly differentially expressed between case and control samples or between different disease stages (Lewohl et al. 2000; De la Fuente 2010). Earlier approaches based on differential expression identified gene biomarkers by

setting arbitrary threshold/cut-offs for fold change and p values (Hibbs et al. 2004; Ginos et al. 2004). However, the technical noise inherited in the gene expression data and experimental variation in measured gene expression levels makes it challenging to detect significant gene expression differences that reproduce consistently across studies and contributes to false positive results. Indeed different threshold choices for differential expression can result in entirely different biological conclusions (Pan et al. 2005). Furthermore, there could potentially be many biologically important genes which are not considered because they are not significantly differentially expressed but are indeed related to the phenotype under consideration (Ben-Shaul et al. 2005; Goeman & Bühlmann 2007). Thus, a standard differential expression approach is unable to give a 'common' multi-tissue set of discriminatory RNA molecules that could function as a biomarker or a diagnostic (Glass et al. 2013) and can thus be inaccurate when used as a classifying tool. In this respect, machine learning approaches could alternatively provide a more useful and robust understanding of the large genomic datasets.

2.4.2 Machine learning approaches in genomics

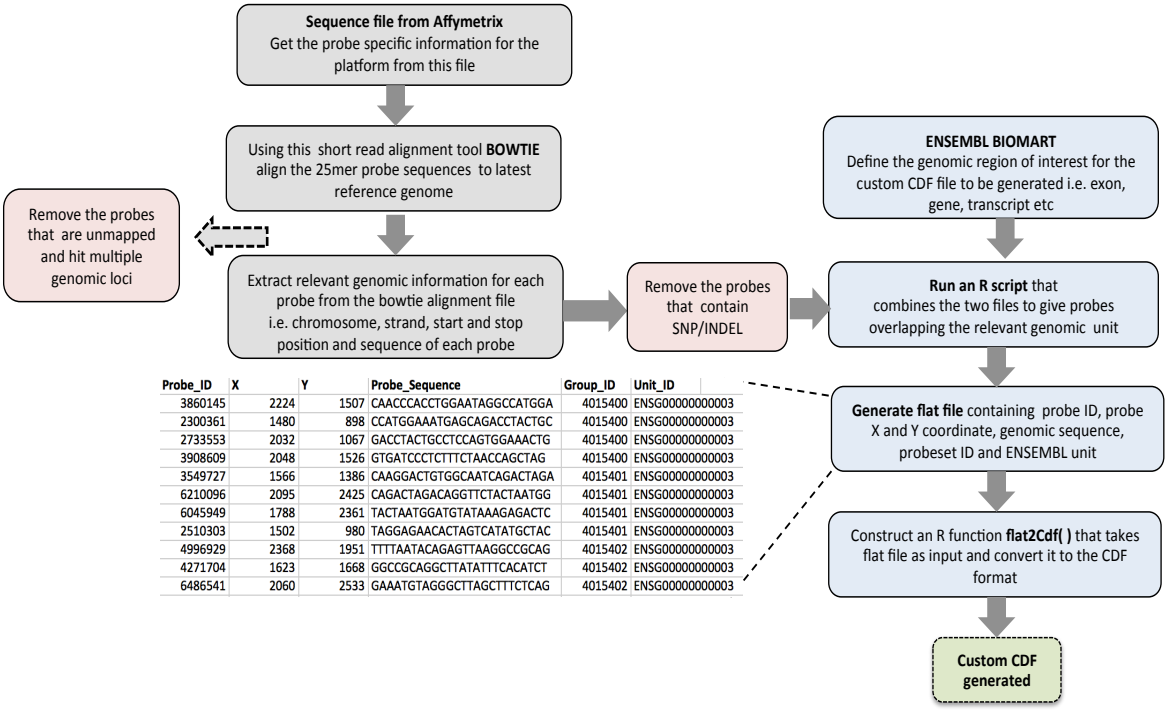
Machine learning is a term used to describe a broad range of automated algorithms that learns from data. By and large machine learning strategies have two applications i.e. prediction or interpretation (Libbrecht & Noble 2015). In genomics, specifically, machine learning has been utilized to predict the location and function of genes and regulatory elements, to identify non-coding RNA and to model and decipher gene expression data etc. (Aerts et al. 2004; Segal et al. 2003; Carter et al. 2001).

Supervised machine learning techniques are frequently used for classification purposes. These techniques train an algorithm to recognise features in the data which discriminate classes. The classification labels are 'seen' by the algorithm in the initial training hence the term supervised learning. After training the same algorithm is tested on unlabelled samples but using only the features of the data which were most useful for prediction in the initial training phase. From a genomics classifier or a diagnostic perspective, this machine learning approach distinguishes which features of the data are likely to be relevant on the basis of gene expression estimations. Thus supervised learning can be a method to deal with the task of feature selection (Baldi & Brunak 2001; Ding & Peng 2005) in classification problems.

Selecting an ensemble of features that can provide high discriminatory power between different biological groups or conditions has been successfully achieved before by using algorithms like support vector machines (SVM) (Guyon et al. 2002), k nearest neighbours (kNN) and the random forest approach (RF) (Diaz-Uriarte & De Andres 2006). A recent in depth review of

different supervised classifiers used for microarrays datasets (Statnikov et al. 2005; Saeys et al. 2007; Boulesteix et al. 2008) identified SVM and kNN as the most effective approaches for microarray data classification (Slonim 2002). The Microarray Quality Control Consortium in the second phase (MAQC-II) set out to evaluate different methods used for developing and validating microarray-based predictive models and reach consensus on the "best practices" for the use of these models in personalized medicine. After assessing the classifiers developed by thirty-six teams for thirteen different endpoints (breast cancer, liver toxicity etc.), they concluded that simple data analysis methods often perform as well as and sometimes even better then more complicated approaches (Shi et al. 2010). Further, the MAQC-II ranked kNN as one of the best performing supervised learning algorithms for microarray based predictive models. Briefly, kNN is a non-parametric method which assigns a label/class to an unknown sample on the basis of class membership of its k nearest neighbours, as determined by a Euclidean distance function. It is the simplest of all algorithms and has the power to perform well on non-linearly separable datasets, often giving better performance than more complex methods in many applications and is thus the approach we chose for our project. Therefore, for this research work we adopted the computationally inexpensive, relatively simple yet efficient kNN classifier for feature selection and classification.

The number of features(dimensionality) in a feature selection process can range from tens to thousands. Certain machine learning algorithms may perform poorly in high-dimensional data and this is referred as the curse of dimensionality (Donoho & others 2000; Van Der Laan & Bryan 2001). It's hard to know what true distance means when you have so many dimensions and the difficulty of searching through the space gets a lot harder. An easier, yet often exceptionally powerful, method for managing high-dimensional information is to diminish the feature space by eliminating some coordinates that seem irrelevant. With microarrays, this can often be done efficiently and simply, by excluding from consideration all those genes whose expression value doesn't vary across hybridization experiments (Hu et al. 2012). For our gene expression based classifier work we evaded this issue of dimensionality (~54K features) by reducing the feature space based on a modified t- statistic (from limma) and using a subset of top 200 features based on this statistic (Smyth 2004).



Probe_ID	X	Y	Probe_Sequence	Group_ID	Unit_ID
3860145	2224	1507	CAACCCACCTGGAATAGGCCATGGA	4015400	ENSG00000000003
2300361	1480	898	CCATGGAAATGAGCAGACCTACTGC	4015400	ENSG00000000003
2733553	2032	1067	GACCTACTGCCTCCAGTGGAAACTG	4015400	ENSG00000000003
3908609	2048	1526	GTGATCCCTCTTTCTAACCAGCTAG	4015400	ENSG00000000003
3549727	1566	1386	CAAGGACTGTGGCAATCAGACTAGA	4015401	ENSG00000000003
6210096	2095	2425	CAGACTAGACAGGTTCTACTAATGG	4015401	ENSG00000000003
6045949	1788	2361	TACTAATGGATGTATAAAGAGACTC	4015401	ENSG00000000003
2510303	1502	980	TAGGAGAACACTAGTCATATGCTAC	4015401	ENSG00000000003
4996929	2368	1951	TTTTAATACAGAGTTAAGGCCGCAG	4015402	ENSG00000000003
4271704	1623	1668	GGCCGCAGGCTTATATTTCACATCT	4015402	ENSG00000000003
6486541	2060	2533	GAAATGTAGGGCTTAGCTTTCTCAG	4015402	ENSG00000000003

Figure 2.2: Workflow for generating the custom CDF file. *The figure shows the workflow for updating the custom definition file for microarrays. This allows one to get the latest information from microarray probes in the biological context of interest (gene, transcript or exon level) using ENSEMBL annotation data.*

2.5 Building healthy ageing diagnostic

We took a hypothesis driven approach to study healthy muscle ageing. For this we utilized the kNN classification method, embedded within a nested loop for feature selection and classification to capture data-features that share non-linear interactions and have robust performance using methods consistent with the MAQC-II. We decided against taking the approach of building a simple linear model for ageing (Rodwell et al. 2004; Horvath et al. 2012; Hannum et al. 2013) as the validity of the linear approach to build a diagnostic of ageing status when applied to the entire adulthood chronological age-range is limited. Extensive molecular work has shown that abrupt changes in metabolism (i.e. a non-linear event) can occur in the 'early middle ages' of model organisms (López-Otín et al. 2013). Therefore, our focus was primarily on a binary predictor that could discriminate between healthy old and healthy young muscle.

2.5.1 Training Dataset

Our goal was to generate a valid molecular classifier of human age using tissue samples from healthy individuals, obtained across the decades during which chronic disease begins to emerge, i.e. the 3rd to 6th decades. Identification of a molecular pattern would then presumably reflect some form of adaptive program in healthy older subjects, since they were free from chronic diseases. Most ageing biomarkers or signatures are built on epidemiological cohorts that blend in ageing,

disease and drug-treatment and do not primarily reflect ageing. Our healthy-age prototype diagnostic was built using 15 young (19-28y, chronological age, VO_2max=2.52 L/min) and 15 older (59-77y chronological age, VO_2max=2.65 L/min) Scandinavian subjects free from metabolic and cardiovascular disease (Timmons et al. 2010; Keller et al. 2011). In humans, aerobic fitness has been found to be a powerful biomarker of all-cause mortality (Church et al. 2005; Wei et al. 1999; Blair et al. 1989; Myers et al. 2002), reflecting genetics (Timmons et al. 2010), and co-morbidities. Since the present aim was to develop a RNA diagnostic that when applied to *any* RNA tissue expression profile, would yield an accurate prediction of healthy physiological age and forecast long-term health, the younger and older samples, used in the prototype development, were matched for aerobic fitness to minimise the confounding effect of this aerobic fitness. We also constrained the gender effect in the study by including only the male subjects (Roth et al. 2002; Berchtold et al. 2008). The muscle biopsies from the samples were profiled on the Affymetrix HGU133Plus2 platform (GSE59880). Note that this dataset was solely used for feature selection and was discarded thereafter in the downstream steps to avoid over-fitting or bias.

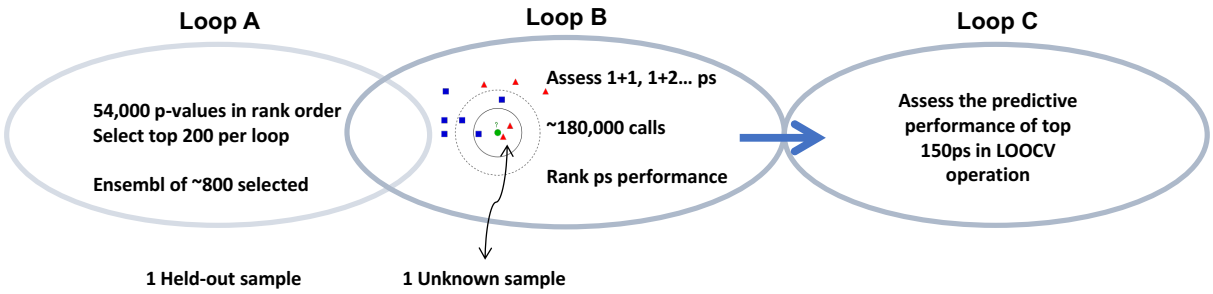


Figure 2.3: Building healthy ageing classifier. *A nested loop strategy along with kNN method was used to select features that could together discriminate between healthy young and healthy old samples.*

2.5.2 Array processing and classifier strategy

The probeset level intensities of the arrays were normalized using the Robust Multi-array Analysis method (RMA) (Irizarry et al. 2003) implemented within the R statistical software environment using the 'affy' package (Bioconductor project (Gentleman et al. 2004; Gautier et al. 2004)). The candidate probeset lists were created via a nested-loop, holding out two arrays at any one time to estimate two parameters from the data (Figure 2.3).

a. The first parameter was the conventional test set result i.e. is the array correctly classified Yes/No. We used this to derive classification success ratios of each of the top 200 probeset to classify each sample.
b. To calculate the second parameter (maximum appearance ratio), we first ranked the probesets based on number of times they correctly classified the array. We then estimated the maximum appearance ratio as the number of times a probeset appears in the top ranked list.

Two-hundred probesets were selected during each of the inner-most computational loops by ranking gene expression differences using an empirical Bayesian statistic (implemented as eBayes in the 'limma' package) (Smyth 2004). Following iterative assessment of all probesets on the gene-chip, involving ~180,000 permutations during which each one of the 30 samples was held-out of the ranking procedure, the best performing ~800 probesets were selected (based on the total number of correct sample classifications during the 180,00 iterations). We removed probesets that targeted multiple genomic loci (as discussed in section 2.3.1) and selected the top ranked 150 probesets (involved in >90% correct decisions) for further study. This reduced list was validated using multiple independent data sets using a kNN (n=3) classifier, implemented using the R 'class' package. To implement independent blind validation, we used both independent training and independent test muscle and brain data sets (chapter-3). The R code is included in Appendix 3 of the thesis.

2.6 Summary

Following a strict set of benchmarks we identified 150 RNA markers of muscle ageing using the following gene-chip profiles (GSE59880) from 15 young (19-28y) and 15 older subjects free from metabolic and cardiovascular disease (59-77y) (Keller et al. 2011; Gallagher et al. 2010). The RNA markers were selected using a nested-loop, holding out two arrays at any one time to estimate two parameters from the data. Following iterative assessment of all probesets and all samples, involving ~180,000 permutations, ~800 probesets were identified as having good performance (>70% correct classifications). After removing the probesets that targeted multiple genomic loci the top ranked 150 probesets for classification of healthy ageing were selected for further work. In the next chapter we test and validate the performance of this healthy ageing signature in independent datasets across different tissues and platforms and also explore its prognostic abilities (Figure 2.4).

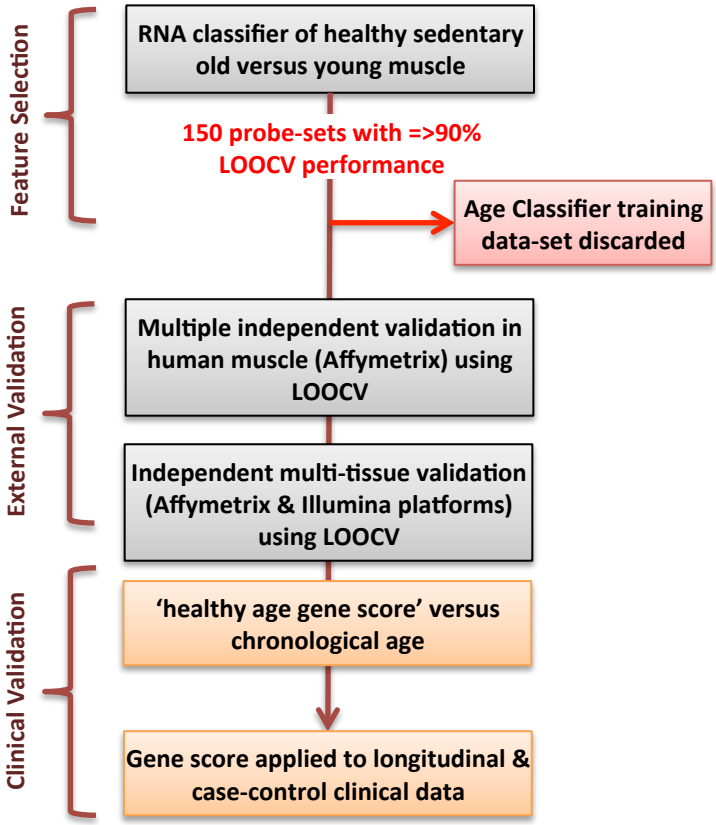


Figure 2.4: Molecular diagnostic for healthy ageing. *The figure summarizes the steps undertaken in this research work to build and validate the healthy ageing RNA signature. Firstly, we did feature selection using a machine learning kNN based approach. The signature was then independently validated to ensure it is not biased/over-fitted and its prognostic abilities were tested (explained in chapter-3).*

3.1 Overview of the chapter

The previous chapter described the transcriptomic approach used to build the RNA expression signature of 'healthy older tissue', by gene-chip profiling sedentary normal subjects who reached 65y in good health. The next step was to demonstrate its reproducibility and test the hypothesis that this gene expression pattern may provide reliable genomic predictors for healthy ageing and risk of age-related disease. Thus, the key objectives of this chapter are to:

- Perform independent validation of the healthy ageing signature on independent cohorts of human muscle, skin and brain tissue (n=594) to establish whether it is robust and reproducible.
- Examine the relationship between RNA classifier and confounding life-style factors and chronological age.
- Ascertain its clinical utility by examining its prognostic potential on a longitudinal study with 20y-follow up period.
- Explore the hypothesis that this gene expression pattern may provide reliable genomic predictors for risk of age-related disease.
- Investigate the biological narrative that governs this molecular signature for healthy ageing.

3.2 External validation across different tissues and technology platforms

Use of fully independent training and validation data sets allows for genuine external validation (EV) to be demonstrated. We implemented fully independent external validation (EV) of the 150-probeset healthy ageing classifier, a process that requires both independent 'known samples' and independent test gene-chips (Shao et al. 2013). When combined with LOOCV methods, this represents a 'gold standard' approach for validation of a classification model.

3.2.1 Independent validation cohorts and implementation

A new set of young and old muscle profiles (Selected from data-set 'Campbell', n=66 chips GSE9419) (Thalacker-Mercer et al. 2010) was used to represent the new 'expression space' of known samples. We then carried out evaluation of sets of independent gene-chip profiles from young and old human muscle (all Affymetrix U133+2) normalised using fRMA (McCall et al. 2010). The various fully independent samples were obtained from GEO or produced from our own clinical samples (Slentz et al. 2011). In each dataset the samples were selected to belong to either young (~25y) or old grouping (~65y) from a larger collection of samples. The sets of young and older samples were selected from 'Trappe' GSE28422(Raue et al. 2012) (*n*=48), 'Hoffman' GSE38718(Liu et al. 2013) (*n*=22), 'Derby' GSE47881 (Phillips et al. 2013) (*n*=26) and 'Kraus' GSE47969,[*n*=33). For all datasets, arrays were examined using hierarchical clustering and

Normalized Unscaled Standard Error (NUSE). In case we identified a small number of gene-chips (~2-3) that had evidence of technical defects and these were removed prior to any analysis. To assess if human brain and skin also demonstrated the same 150 age-related gene expression signature as healthy older muscle, we used young and old samples brain-bank array source (n=120, GSE11882) and the MuTHER cohort skin dataset (n=279, which includes subset of 3 replicates (n=131, n=124 and n=24)). The skin data was produced using the Illumina Human HT-12 V3 Bead chip (Arrayexpress: E-TABM-1140) and log-2 transformed signals were normalised using quantile normalisation. The 150 Affymetrix probesets were mapped to the Illumina platform (giving 129 probes). Due to differences in gene-chip technology, a leave-one-out cross validation (LOOCV) approach was used to classify age of each skin sample, using only the probes selected above. For skin, individuals aged < or = 45y were defined as young, and those > or = 70y as old to ensure balanced numbers of young and old samples existed to fairly assess the classifier performance. The R code for 'independent validation' is included in Appendix 3 of the thesis. The information about all the datasets used in this chapter is available in Appendix 1.

3.2.2 Reproducible RNA signature for age of human muscle, brain and skin

Using the 'Campbell' muscle data set (GSE9419) (Thalacker-Mercer et al. 2010) as the samples of known identity, we demonstrated that additional young and old muscle samples selected from 4 additional muscle data sets ('Trappe' GSE28422 (Raue et al. 2012), 'Hoffman' GSE38718(Liu et al. 2013) 'Kraus' GSE47969 and 'Derby' GSE47881 (Phillips et al. 2013)) could be classified with an average ~93% accuracy (70-100%) using only the 150 probesets selected at the start of the project. Substitution of Campbell with the other muscle data sets worked equally as well. These data shared a common microarray platform (Affymetrix HGU133Plus2) but as we demonstrate below, the classifier remains robust in the face of alternative platforms. Receiver operator curves (ROC) for kNN=5 demonstrating classifier performance for a number of tissue types are presented in Figure 3.1.

Using data from the HGU133Plus2 microarray platform for old and young samples of ectodermal origin (brain, n=120) (Gould et al. 1999) we confirmed that the 150 RNA 'healthy age' genes selected in muscle, could also distinguish the age of human brain one sample at a time, with a classification success rate up to 91% (Figure 3.1). Four brain regions were evaluated (Postcentral Gyrus, Entorhinal Cortex, Hippocampus and Superior Frontal Gyrus), GSE11882 and while they were confirmed disease-free by histopathology in the original study (Berchtold et al. 2008), unlike our muscle cohorts, their true functional status remains unknown. The Postcentral Gyrus samples were classified with 86% sensitivity and 89% specificity. Older hippocampal regions were often misclassified using the 150-genes (33% sensitivity) as 'young'. This higher misclassification rate

may relate to the substantial neurogenesis known to take place in the adult hippocampus or delays in tissue processing.

Tissue	Sample Size	Accuracy %	Sensitivity	Specificity
Muscle (Campbell)	66	96	0.94	0.97
Muscle (Derby)	26	100	1.00	1.00
Muscle (Trappe)	48	96	0.96	0.96
Muscle (Hoffman)	22	91	0.93	0.88
Muscle (Kraus)	33	70	1.00	0.60
Brain (SFG)	33	91	0.71	0.96
Brain (PCG)	31	88	0.86	0.89
Brain (Hippocampus)	31	85	0.33	1.00
Brain (EC)	25	72	0.43	0.94
Skin (MuTHER Cohort)	279	78	0.59	0.90

Table 3.1: Accuracy, sensitivity and specificity of the muscle-derived healthy age classifier when applied to multiple independent data sets. *The sensitivity and specificity of the top 150-probe-sets from the 672 probe-set derived from the STOCKHOLM U133+2 Affymetrix gene-chip data, was calculated for the human muscle data sets CAMPBELL, DERBY, HOFFMAN, TRAPPE AND KRAUS and the four brain regions derived from the Berchtold et al. study (Berchtold et al. 2008) and skin from the MuTHER cohort (Glass et al. 2013). The majority of data sets demonstrated both high sensitivity and high specificity using this un-optimised list. A young sample misclassified as 'old' (e.g. in HOFFMAN) is noted as a reduced sensitivity. If an old sample was misclassified as being young and healthy, as was the case for some of the Hippocampus regions, then this is defined as a reduction in specificity where young is a true-positive in this model. The likely contributing factors to these misclassifications include lack of standardisation of a single laboratory gene-chip protocol and variation in RNA quality and in some cases examples of older donors that have not induced the 'healthy ageing' signature to any measurable extent.*

Lastly, we evaluated whether the 150 genes could accurately classify tissue age of mesodermal origin (skin) using gene expression data in a total of 279 human skin samples of which there were up to three technical replicates per clinical sample (Glass et al. 2013). Notably these data originated from a different technology platform (Illumina Human HT-12 V3, Arrayexpress: E-TABM-1140) thus adding variability above that derived from a distinct tissue and potentially limiting the classification process. One hundred and twenty-nine genes were common to both gene-chip technologies, and we observed excellent classification of age of human skin (n=131,

AUC=0.85, Figure 3.1). The classification success was similar for all three replicates (71-78 raw classification success). Thus the technical performance of the 150-gene multi-tissue age classifier was excellent and robust, providing accurate classification despite inter-laboratory technical variation, different gene-chip platforms and ante-mortem tissues. We were therefore able to conclude that we have identified a reliable multi-tissue RNA signature of healthy tissue ageing in humans, something that has not been previously demonstrated (Glass et al. 2013; Phillips et al. 2013).

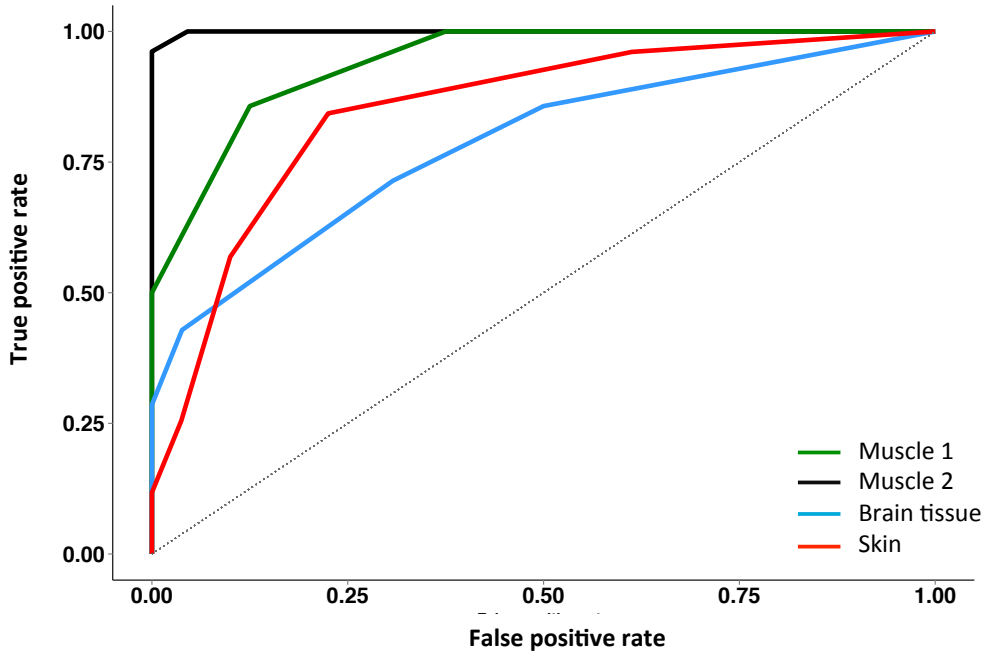


Figure 3.1. ROC curve showing predictive performance for tissue age classification using 'healthy age' biomarkers based on leave-one-out cross-validation (kNN=5) for muscle, brain and skin*. Using only the 150 probesets identified in the first stage of the project, this 'healthy age classifier' was able to correctly classify young and old samples across independent datasets with an accuracy of ~96%, 91%, 85% and 78%. We present two examples of independent muscle data (Raue et al. 2012) [50] and one example each for human brain (Berchtold et al. 2008) and skin datasets(Horvath 2013) with AUC of 0.99, 0.94, 0.78 and 0.85 respectively reflecting excellent separation of the age groups and hence accurate multi-tissue performance.*

3.3 Prognostic abilities of healthy ageing signature and relation with life-style related risk-factors

Ideally, a true diagnostic of 'healthy ageing' should not correlate with age associated phenotypes or risk factors for chronic disease (Baker & Sprott 1988). The specificity for 'healthy ageing' in our RNA signature was examined by assessing the relationship between the signature genes, chronological age and markers of life-style associated disease in a twenty-year longitudinal cohort

(ULSAM, Uppsala Longitudinal Study of Adult Men cohort). To achieve this, the RNA signature was transformed into a ranking metric by collapsing the expression pattern of each gene in our signature into a single score termed as 'healthy ageing score' which was then related to risk factors and health outcomes in the ULSAM study.

3.3.1 ULSAM longitudinal study and gene score calculation

We used a set of tissue samples from a birth cohort of men, such that the same chronological age (~70y) could be contrasted with the variation in 'healthy age gene score'. The ULSAM (Uppsala Longitudinal Study of Adult Men) is a cohort of men born in 1920-24 and living in Uppsala, Sweden to compare a constant chronological age (and similar environment) with the healthy muscle age gene score for each individual (Dunder et al. 2004). Dual-energy X-ray absorptiometry (DXA) scan measurements were performed during the last decade of the study and muscle mass status varied between -15% to +10% between age 70y to 88y and was unrelated to physical activity scores (recorded at 82y and 88y of age, with 80% being recorded as being moderately active). We had access to 129 skeletal muscle biopsies that were taken at age 70y (in 1992) and were processed in 2012 with the majority having excellent NUSE plot profiles. Total RNA was extracted from frozen muscle biopsy samples (*vastas lateralis*) using TRIzol reagent as previously described (Timmons et al. 2005). A total of 113 samples provided sufficient RNA and 50ng total RNA was amplified using Ambion's WT expression kit to produce cDNA. The cDNA was fragmented and labelled with GeneChip WT Terminal labelling kit (Affymetrix, Inc.). Unincorporated nucleotides from the IVT reaction were removed using the RNeasy column (QIAGEN Inc, USA). Hybridization, washing, staining and scanning of the arrays were performed according to the manufacturer's instructions (Affymetrix, Inc., USA).

One hundred and eight samples passed gene-chip quality control procedures. A cumulative gene-ranking based score was calculated using each of the 150 gene expression values for each of the 108 male subjects and the final score was compared in a linear fashion with a number of clinical parameters. For an RNA down regulated in the original training classification dataset (i.e. down regulated between 25y to 65y) the ULSAM subject with the highest expression was assigned a score of 1 and the subject with the lowest expression 108. For genes up regulated in the original age classification model, the opposite strategy was used. Thus both feature selection (genes) and direction of regulation were taken from the original model. The median sum of these rank scores (for all 150 genes) was calculated and that represented the 'healthy age gene score' for each 70y individual. Median rank ensured each gene provided equal weighting and regression analysis was used to study the variation in gene score in these men all of who had approximately the same chronological age. The R code for 'Gene score ranking' is included in Appendix 3 of the thesis.

3.3.2 Healthy ageing gene score is distinct from chronological age and is unrelated to life-style factors

The distribution of scores was examined for 70y old males and the scores were also correlated with markers of life-style associated disease (Figure 3.2). We ranked each subject for each of the 150 genes, taking the direction of gene expression change from the original classifier model into account (85% down-regulated). We then converted the individual gene scores into a summed median gene-score for each subject. We demonstrated that despite all subjects being ~70y of age at the time of the RNA sample, there was a very wide distribution in gene score (Figure 3.2A). Thus the 'healthy age gene-score' in muscle was very distinct from chronological age.

The 'healthy age gene-score' was regressed against a variety of continuous clinical variables (variables listed in Table A3.1 in Appendix 2). The gene-score at chronological age 70y was unrelated to conventional life-style regulated biomarkers at baseline e.g. renal function (estimated from cystatin-c, $r2<0.001$), systolic blood pressure (mmHg, $r2=0.0013$), 2hr glucose concentration following a standard oral glucose tolerance test (mmol, $r2=0.015$) or total cholesterol (mmol, $r2=0.002$). Gene score was also unrelated to resting heart rate or physical activity questionnaire. Infact the 'healthy-ageing' gene score was not correlated with any conventional risk factors (Figure 3.2B). This confirmed that the 150 gene expression markers were not reflecting a variety of life-style factors and diseases (e.g. exercise, diabetes).

3.3.3 Healthy ageing gene signature as prognostic of long-term health status in the ULSAM study

Our primary hypothesis was that a validated diagnostic of healthy physiological age could be used to predict health outcomes in a longitudinal study, where subjects were all the same chronological (calendar) age at the point of assessment. The relationship between the gene score at age 70y for subjects in ULSAM study and a number of clinical features was carried out using multi-factor models. At 70y three subjects had Cystatin C > 1.5 mg/l, while by 82y 36 of the subject studied in the present analysis had Cystatin C > 1.5 mg/l Cystatin C. A 1.5 mg/l Cystatin C corresponds to an estimated GFR of ~45 ml/min which is borderline for a moderately (30-45 ml/min) elevated risk for all-cause mortality. We estimated renal function using Cystatin C to calculate glomerular filtration rate (eGFR) as it is a robust marker for early renal impairment (Coll et al. 2000; Laterza et al. 2002) and demonstrated that the baseline healthy-age diagnostic ranking score was related to renal function 12 years later (age 82, $p=0.009$). While renal function is not sufficiently powerful to predict mortality in disease-free older subjects from the ULSAM cohort (Zethelius et al. 2008), we found that the healthy age diagnostic was able to strongly predict 20y survival in a cox-regression

model. Over the observation period mortality rate was 18% (19 events) and the relationship between mortality and gene-score was analysed as a continuous variable.

Remarkably, the 'healthy age gene-score' in muscle at 70y was independently related to 20-year survival (p=0.0295, Figure 3.3A) in a logistic regression model. While this observation should be interpreted cautiously, to illustrate the temporal relationship between the 'healthy age gene-score' and death, we divided gene-score into quartiles and applied a Cox-regression model (Figure 3.3B) and found a significant difference between the first versus the fourth quartile (p=0.04). In contrast to the 'healthy age gene-score', a median gene rank score based on inflammatory gene (GO:0006954) or mitochondrial gene (GO:0005739) expression in muscle demonstrated no relationship with health or mortality (Appendix 2 Figure A3.1, p=0.173 and p=0.337 respectively). For the cox-model we used the latest 'survival package' whereas the logistic regression model was estimated using the glm (generalized linear model) function and 'logit' model which models the log odds of the outcome as a linear combination of the predictor variables. For the Kaplan-Meier plots, gene-score was divided into quartiles and the plot was produced using the 'plot-survfit' function in the survival package. All three approaches yielded consistent results.

Thus, despite the limited sample size of the ULSAM cohort (n=108), we were able to establish that subjects with the highest muscle 'healthy ageing gene score' at age 70y had significantly better renal function 12 years later (at age 82 years) and a better survival rate 20 years later. The prediction of mortality in the ULSAM 20y follow-up study is of course preliminary, given the size of this part of our study, but it provides further support that *induction* of the age signature, by the 6th decade of life, represents a positive event since the directional shift in gene-expression and better 'health' was consistent for the renal and mortality analysis i.e. largest gene score in the ranking system was associated with better health in ULSAM.

3.4 Relation between 'healthy ageing gene signature' and cognitive health

Neurocognitive pathology (e.g. Alzheimer's disease – AD) becomes more pronounced with age and is often apparent in individuals who are otherwise healthy. Our analysis of the relationship between life-style factors and the 'healthy age gene score' in the ULSAM cohort suggested that the gene score was robust to confounding effects of life-style disease. We next examined whether the 'healthy ageing gene score' [median rank sum of the 150 RNA markers] was *selectively* useful in relation to identifying neurocognitive disease over life-style disease. To support this analysis, we utilised a large publically available gene-chip data-set derived from healthy human brain samples of various ages (Ramasamy et al. 2014). The BrainEac.org gene-chip resource (Ramasamy et al. 2014)(GSE60862) comprises 10 post-mortem brain samples from 134 subjects representing 1,231

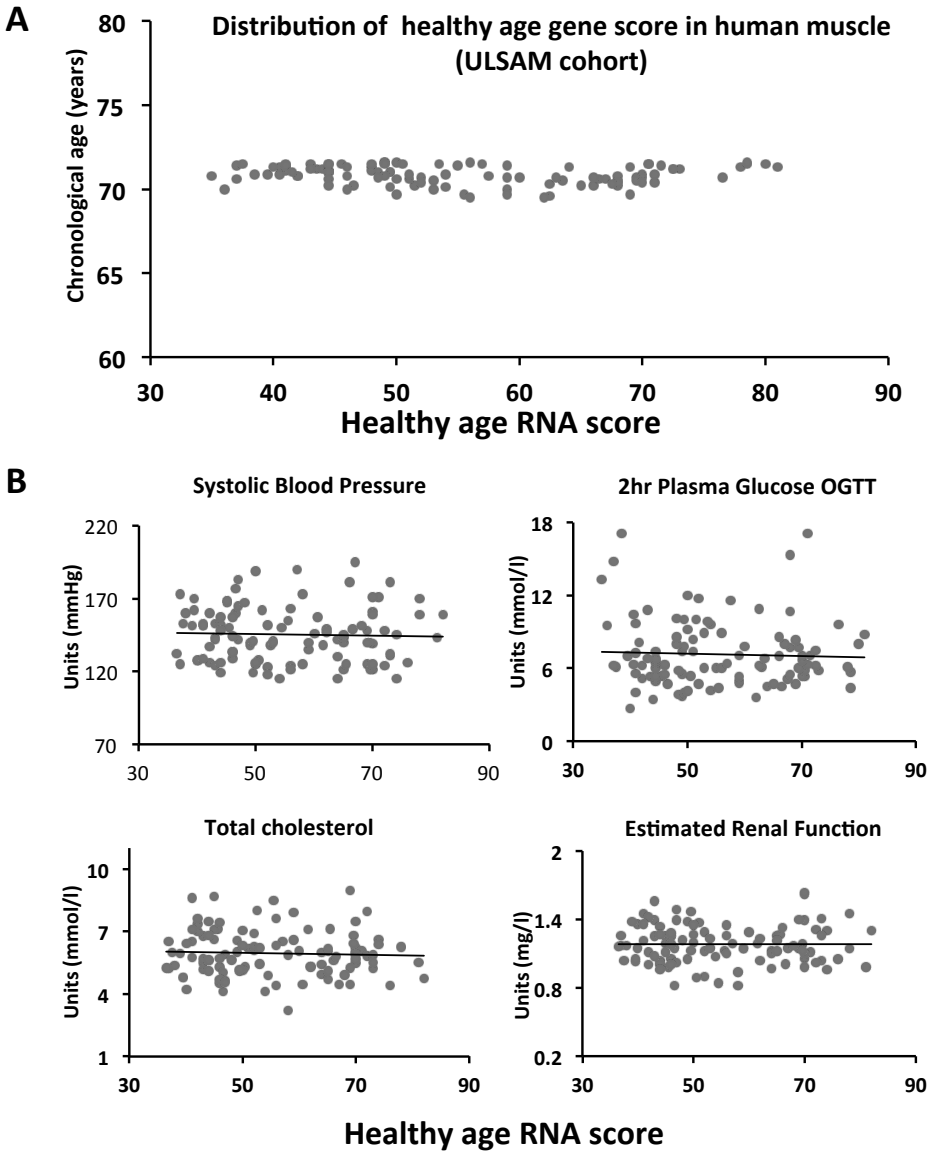


Figure 3.2. Distribution of healthy age gene score in ULSAM samples and its relation with clinical parameters. *At the date of assessment (1992), when the muscle biopsy was taken for subsequent gene-chip profiling, all subjects would be considered in reasonable health for their age and remained physically active. A) Distribution of Gene score based on the median rank for each of the 150 age genes. B) Clinical variables were determined as previously reported for ULSAM samples (chronological age=69-70y) (Huang et al. 2013)(Zethelius et al. 2008). Linear regression was used to examine the relationship between the healthy-ageing gene-score at ~70y and a variety of clinical parameters at age ~70y. No relationship between gene score and renal function (estimated from cystatin-c, r2<0.001), systolic blood pressure (mmHg, r2=0.0013), 2hr glucose concentration following a standard oral glucose tolerance test (mmol, r2=0.015) or total cholesterol (mmol, r2=0.002) was observed. Gene score was also unrelated to resting heart rate or physical activity questionnaire.*

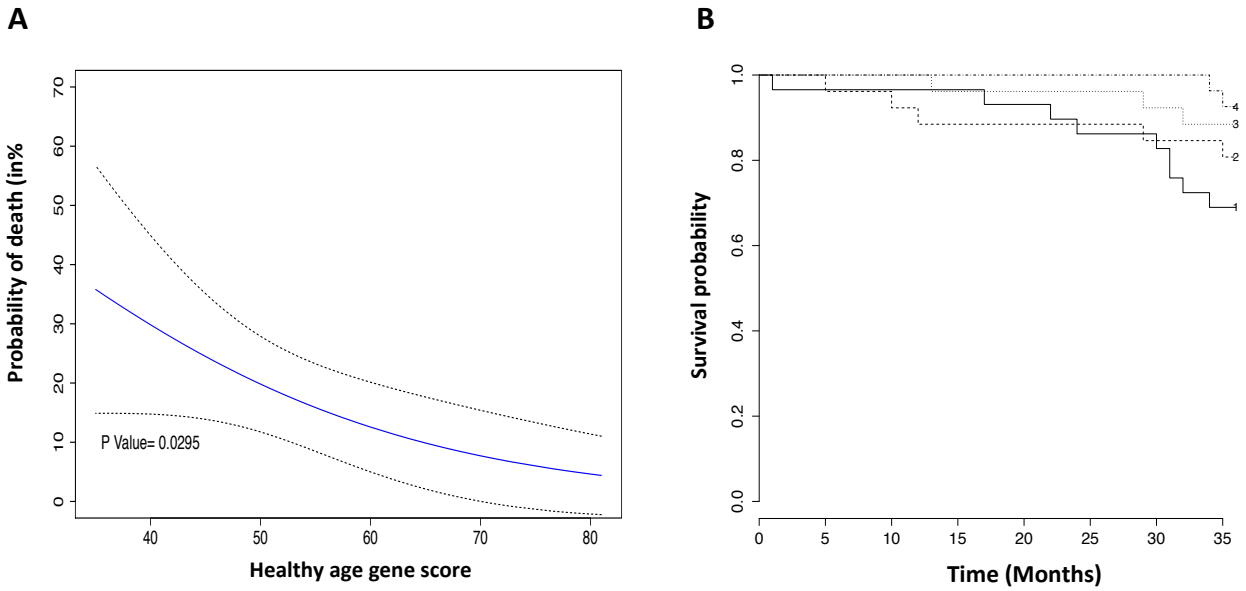


Figure 3.3 A cumulative ranking metric of the healthy ageing metric was prognostic for mortality over a 20-year follow-up period. *One-hundred and eight subjects provided a healthy tissue biopsy in 1992 that was suitable for RNA profiling and the fully annotated mortality data, covering 2009–2011, was retrieved from the Swedish national health registry. A) The rank score for healthy ageing gene expression was calculated from the top 150 genes of the healthy ageing prototype classifier (n = 108, male subjects all ~70 years of age). Logistic regression analysis performed using the cumulative ranking metric of the top 150 genes from original prototype was prognostic for mortality. It showed that those subjects with the lowest median healthy ageing gene score had a much higher probability of death during the 20-year follow-up period (p = 0.0295). B) The rank score for healthy ageing gene expression was calculated from the top 150 genes of the healthy ageing prototype classifier (n = 108, male subjects all ~70 years of age) and Kaplan–Meier plots were used to illustrate the temporal pattern of survival. Gene score was divided into quartiles and the plot was produced using the plot-survfit function in the R survival package. The plot allows us to compare overall survival rates between the four quartiles for gene score. The third and fourth quartiles differed from the first quartile (p < 0.04).*

samples. Using the same ranking approach as applied to the ULSAM cohort, the median sum of the rank score was calculated for each anatomical brain region (Figure 3.4). As before, in healthy older individuals the 'age' signature was 'switched on' (yielding a greater ranking score). Regulation of the healthy age gene score increased across individual healthy brain regions with chronological age, especially in the hippocampus ($p= 2 \times 10^{-8}$), as well as other regions such as putamen, thalamus, substantia nigra and the occipital, frontal and temporal cortex regions (all at least $p<0.002$ by Holm adjusted Mann-Whitney test). From this it was ascertained that the healthy ageing gene score was clearly evident in neuromuscular tissue, which suggested that it might relate to cognitive health.

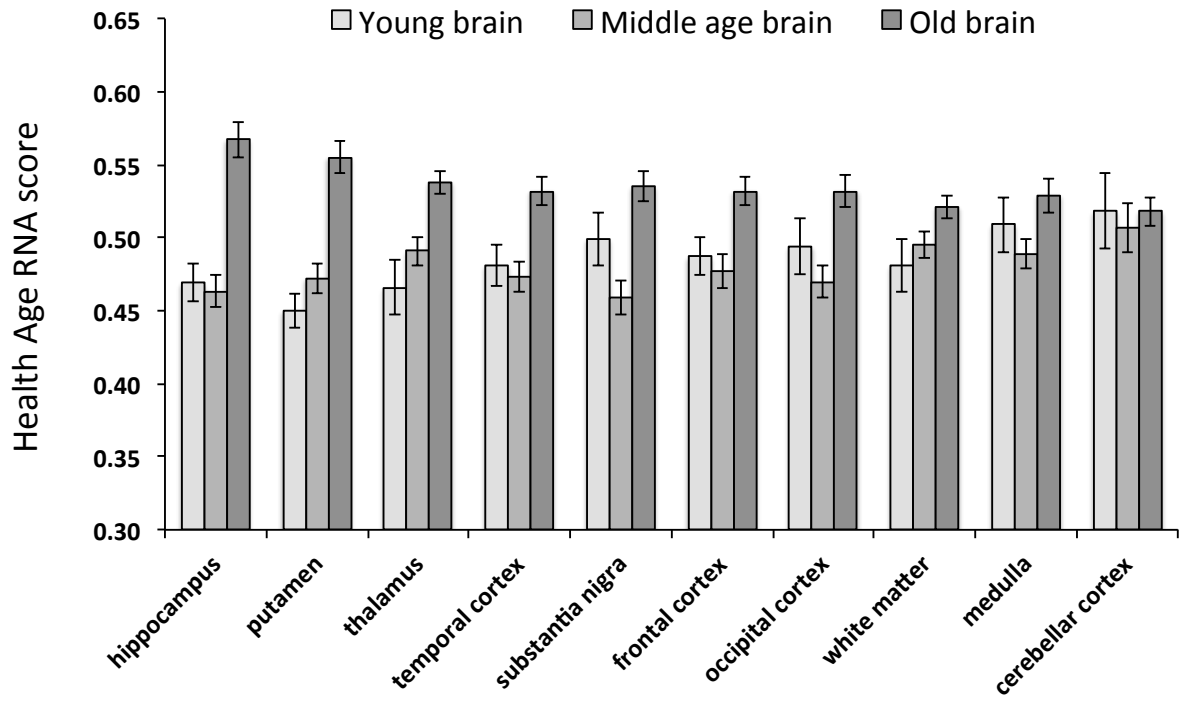


Figure 3.4 The 'healthy ageing' RNA signature was studied across diverse anatomical human brain regions in healthy individuals using BrainEac.org gene-chip resource. *The healthy Ten brain regions from 134 subjects representing 1231 samples were individually ranked and the median sum of the ranked scores calculated. Regulation of the healthy ageing genes differed across brain regions with age, as determined by a Kruskal Wallis Test (hippocampus $p = 2 \times 10^{-8}$, putamen $p = 4 \times 10^{-7}$, thalamus $p = 4 \times 10^{-5}$, temporal cortex $p = 0.0001$, substantia nigra $p = 0.0002$, frontal cortex $p = 0.001$, occipital cortex $p = 0.001$, white matter $p = 0.01$, medulla $p = 0.06$ and cerebellar cortex $p = 0.51$). Post hoc Mann–Whitney test, with correction for multiple comparisons (Holm), confirmed a striking 'increase' of the healthy ageing score in the healthy older samples (hippocampus, putamen, thalamus, substantia nigra, and the occipital, frontal, and temporal cortex regions; at least $p < 0.002$)*

3.4.1 Translating healthy ageing gene signature in Alzheimer/MCI cohorts

Based on the above observation our primary hypothesis was that, compared with control subjects of similar chronological age and gender, patients with AD would have a lower median healthy ageing gene score but the score would not distinguish diabetes or vascular disease patients from matched controls. To test this hypothesis, we used blood RNA profiles from subjects from the AddNeuroMed consortium, a large Cross-European AD biomarker study and a follow-on Dementia Case Register (DCR) cohort in London. Patient selection, design and clinical data have been reported previously (Lovestone et al. 2009; Lunnon et al. 2012). We used two independently produced gene-chip datasets from the consortia, one produced in a UK gene-chip facility and

another produced in the USA which have been deposited on GEO under GSE63060 and GSE63061. A summary of the cohort characteristics can be found in Table 3.2

Gender & Age matched cohorts	Age	Gender (F/M)	MMSE
Batch 1			
$Control_{MCI}$ (n=67)	69.6 (±4.2)	41/26 (61%F)	29.1 (±1.2)
MCI (n=39)	70.0 (±3.3)	24/15 (62%F)	27.5 (±1.6)
$Control_{AD}$ (n=64)	70.2 (±3.7)	41/23(64%F)	29.1(±1.2)
AD (n=49)	69.8 (±4.4)	34/15 (69%F)	21.8 (±4.5)
Batch 2			
$Control_{MCI}$ (n=71)	70.8 (±2.9)	44/27 (62%F)	28.9 (±1.9)
MCI (n=31)	69.5 (±4.5)	23/8 (74%F)	27.6 (±1.9)
$Control_{AD}$ (n=71)	70.8 (±2.9)	44/27(62%F)	28.9(±1.9)
AD (n=40)	69.9 (±4.3)	23/17 (58%F)	21.0 (±5.6)

Table 3.2: Clinical characteristics of batch 1 and batch 2 AD cohorts. *Case-control subjects that contributed to the blood gene chip profiles analysed and presented in Figure 3.5 and Figure 3.6*

Briefly, subjects were excluded from the study if they had neurological or psychiatric illness other than AD, unstable systematic illness or organ failure, or a geriatric depression rating scale score ≥ 4/5. AD was diagnosed using the National Institute of Neurological and Communicative Disease and Stroke and Alzheimer's disease (NINCDS-ADRDA) and Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria for possible or probable AD. All MCI subjects reported problems with memory, corroborated by an informant, but had normal activities of daily living as specified in the Petersen's criteria for amnestic MCI (Lovestone et al. 2009; Lunnon et al. 2012). All subjects underwent a structured interview and a battery of neuropsychological assessments including the Mini Mental State Examination (MMSE). Control and MCI subjects were further assessed using the CERAD battery and detailed information on subject recruitment and assessments can be found in other published studies describing the AddNeuroMed consortium (Lovestone et al. 2009; Snyder et al. 2014). RNA was obtained from whole venous blood and it was collected from the subjects who had fasted 2 hours prior to collection into a PAXgene™ Blood RNA tube (Becton & Dickenson, Qiagene Inc., Valencia, CA). The tubes were frozen at -20°C overnight prior to long-term storage at -80°C. RNA was extracted using PAXgene™ Blood RNA Kit (Qiagen) according to the manufacturer's instructions. Whole genome expression was produced using Illumina Human HT-12 v3 Expression BeadChips for the first case-control study (USA, 'Batch 1') and Illumina Human HT-12 v4 Expression BeadChips for the second case–control study (UK, 'Batch 2'). cDNA was synthesized from 200ng total RNA using TotalPrep™ RNA Amplification Kit (Ambion) which

was followed by amplification and biotinylation of cRNA and hybridization. The expression data was first transformed using variance-stabilization and then quantile normalized using the LUMI package in R.

For our primary analysis, control subjects were matched in a manner that created the largest possible group with the same chronological age and gender-balance as the AD or MCI groups. Thus our analysis was carried out on a sub-set of subjects deposited at GEO, with each case-control group having a similar median chronological age as the ULSAM cohort. A total number of 297 samples were utilised (Batch 1 CTR=67, MCI=39, AD=49 and Batch 2 CTR=72, MCI=30, AD=40). Retrospective inclusion of the entire cohort (n=717) did not alter the outcome of our analysis. The 150 Probesets were mapped from the Affymetrix platform to the Illumina platform yielding 128 genes from the original 150-gene list. For each case-control comparison the ranking metric was computed in the exact same manner as for the ULSAM subjects (see section 3.3.1). From Batch 1, 113 subjects were ranked for gene score, while 111 subjects were ranked in Batch 2 (Table 3.1). Wilcoxon rank sum test from the R stats package was used to test if the median gene score ranks between groups were significantly different or not. For data presentation, ranking scores were scaled to the total number of samples being ranked to ensure each data plot was on the same scale. The relative median rank score for AD patients was significantly lower that the age and gender matched controls (p=0.004, Figure 3.5), based on Wilcoxon rank sum test. Blood RNA from the second AD case-control cohort was profiled on the Illumina HT-12 V4 platform and in this case 122 genes were common to the 150-gene healthy ageing gene score.

As before, the median rank healthy ageing gene-score for AD patients in Batch 2 was significantly lower than in the control group (p=0.009, Figure 3.5). Furthermore, for both Batch 1 and Batch 2, the age-matched controls had a higher median gene score than subjects diagnosed with mild cognitive impairment (MCI, Figure 3.5 p=0.00005 and Figure 3.5 p=0.003).

It is important to note that the control samples used for comparison with MCI overlapped with those used for comparison with AD and that the MCI analysis cannot therefore be considered a fully independent observation. We also checked for overlap between the 150 healthy ageing gene markers and previous genomic and genetic disease markers of AD. Only three genes were in common (*SPN*, *NPEPL1* and *PDLIM7*) and none were from previously validated AD diagnostics. Their inclusion or exclusion did not impact our analysis.

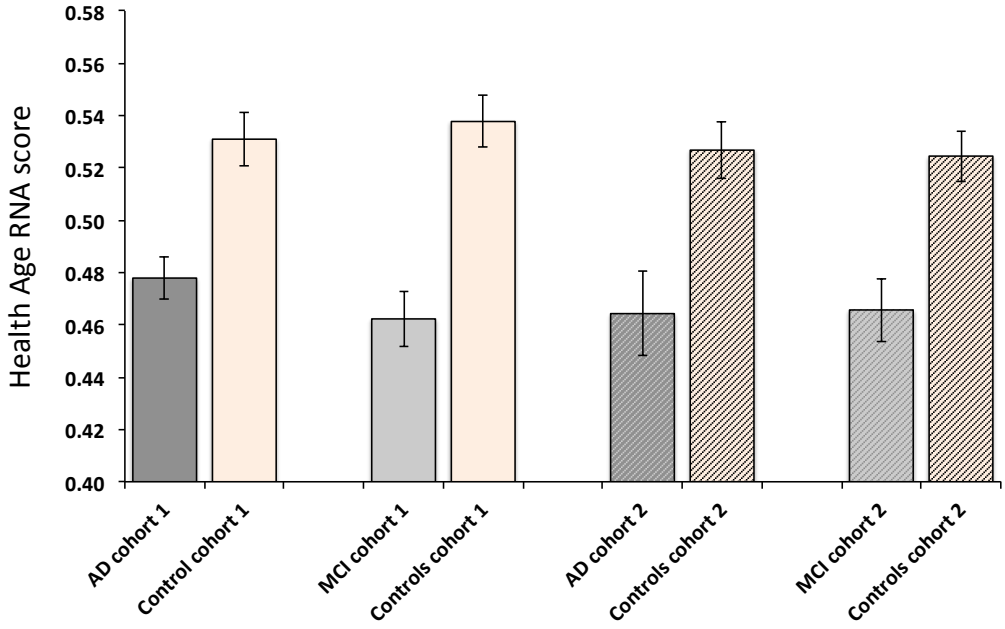


Figure 3.5 A cumulative ranking metric of the healthy-age metric could distinguish between control subjects with Alzheimer (AD) or Mild Cognitive impairment (MCI). *The healthy ageing RNA signature was studied in blood samples from two independently processed case–control studies of AD. In cohort 1 the control median gene score was greater ($p = 0.004$) than AD samples and greater ($p = 0.00005$) than that of the MCI samples (Wilcoxon rank sum test). In cohort 2 the median gene score of control samples was greater than that of AD samples ($p = 0.009$) and that of MCI samples ($p = 0.003$). Data are median gene score and standard error.*

3.4.2 Healthy ageing signature as AD diagnostic

We also formally evaluated whether the healthy ageing signature could act as a diagnostic for AD case-control cohorts using ROC analysis and found that it had robust independent performance on both cohorts that were used in the previous section (AUC=0.66-0.73, Figure 3.6). Our research group previously published a whole blood RNA based prototype AD diagnostic, consisting of 48 genes which was also identified using machine learning methods applied to Cohort 1 samples (Lunnon et al. 2013). We demonstrated that this prototype 'RNA disease signature' was independently validated in Cohort 2 using LOOCV.

Further, when we combined these two independently produced and validated gene expression classifiers (RNA age signature and RNA disease signature) we yielded an improved AD diagnostic (AUC=0.73-0.86, Figure 3.6), one which matches best in class (Snyder et al. 2014) for those blood-based AD diagnostics validated using independent data, but using a technology platform more suited to reproducible high-throughput diagnostics.

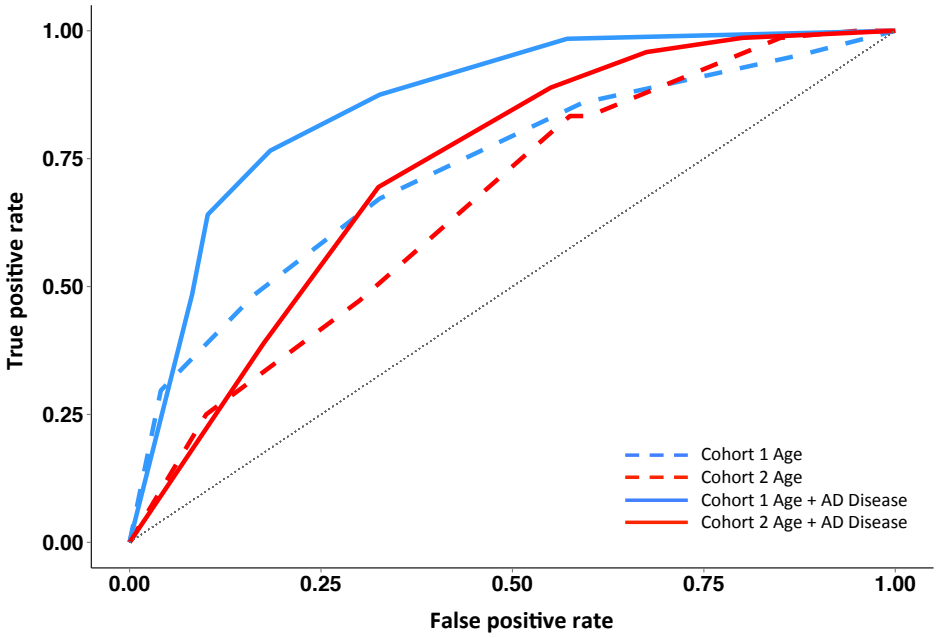


Figure 3.6 Validation of novel blood RNA classifiers as a diagnostic for Alzheimer's disease. *We used the independent batch 2 AD data set to test the predictive performance of our healthy ageing classifier and our previously published AD prototype diagnostic. The performance of each was evaluated using ROC curves. The healthy ageing gene classifier generated independent AUCs of 0.73 and 0.66 for AD in cohorts 1 and 2, respectively. For the combined 'healthy ageing' plus 'AD disease' RNA classifier (150 + 48 probesets) we obtained AUCs of 0.86 and 0.73 for AD without any attempt at optimization. The AD disease RNA classifier probesets were selected using cohort 1.*

3.4.3 Relationship between the healthy age gene score and chronic life-style diseases

Lastly, we utilised two additional large gene-chip clinical studies; one comparing blood RNA in Type II diabetes with control (Tabassum et al. 2014) and the other from our laboratory, comparing blood RNA in people with and without coronary artery disease (Sinnaeve et al. 2009). The main purpose of this analysis was to further establish that the 150 gene expression markers were not reflecting a variety of lifestyle-regulated diseases.

The diabetes data was profiled on Illumina Human HT.12.V4 arrays and comprised of 94 controls versus 50 cases (group mean age = 66 y) and the vascular disease data had 112 controls and 110 cases (group age = 53.3 y) on Affymetrix HG-U133A arrays. The case control analysis was done in same manner as for AD cohorts (section 3.3.1). Applying a Wilcoxon rank sum test, neither diabetes nor vascular disease ($p=0.588$ and $p=0.430$ respectively) was related to the healthy ageing gene score (Figure 3.7). This is consistent with our original hypothesis, and methods, that

the healthy ageing gene score is not related to lifestyle factors and it is also consistent with the results observed in the ULSAM cohort (Figure 3.2B).

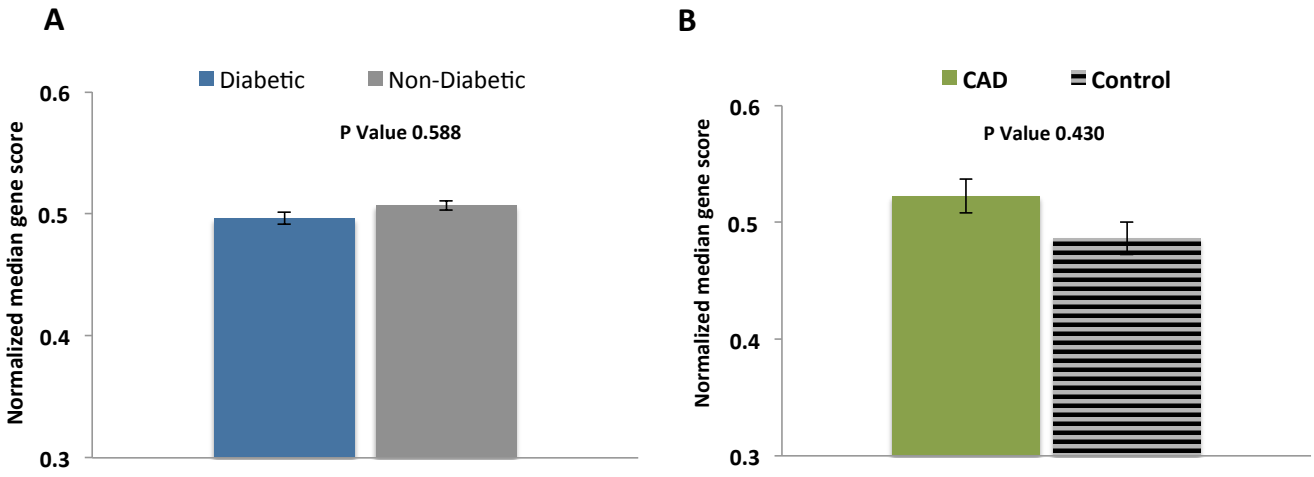


Figure 3.7 The healthy ageing signature activation was studied in blood samples from two independent large case–control studies of diabetes and vascular disease. *Applying a Wilcoxon rank sum test, neither diabetes nor vascular disease was related to the healthy ageing gene score. This is consistent with our original hypothesis, and methods, that the healthy ageing gene score is not related to lifestyle factors and it is also consistent with the results observed in the ULSAM cohort (Figure 3.2B). A) The diabetes data (94 controls versus 50 cases, group mean age = 66 y) originates from Tabassum et al. using Illumina Human HT.12.V4 arrays.* (Tabassum et al. 2014)*. B) The vascular disease data (112 controls and 110 cases, group age = 53.3 y) originates from Sinnaeve et al.*(Sinnaeve et al. 2009) *(using Affymetrix HG-U133A arrays).*

3.5 Biological features of the healthy age diagnostic

We were interested in whether the healthy-ageing diagnostic revealed any particular biological processes that might be open to therapeutic targeting. The bioinformatics tool, Ingenuity Pathway Analysis (IPA, http://www.ingenuity.com) was used to explore the biology of the age classifier genes. HUGO gene name identifiers were uploaded into IPA and queried against the verified IPA knowledge database. Out of the 150-gene list, a total of 127 genes were annotated in the database and revealed a few marginal functional associations (e.g. Nervous system development genes) but these did not remain significant following Benjamini and Hochberg correction. The top ranked database network (genes with published interactions) was defined as 'cell death and survival' and contained 31 molecules.

Then to establish the Gene ontology profile of the 150 genes (Appendix 1), we generated a null distribution of GO enrichment p-values by randomly sampling 10,000 lists of 150 probesets

from the HGU133Plus2 chip and testing each list for Molecular Function GO using the GOstats package in R. The entire population of probesets on the HGU133Plus2 microarray was used as the background population for these tests. We observed that the profile of ontological enrichment in the healthy-ageing diagnostic was not different from a random sample of 150 genes from the gene-chip, of which >99% of those 54,000 probesets had no ability to discriminate tissue age in our training model. In Figure 3.8 the density curves of p-values for each one of 10,000 hyper-geometric tests using randomly sampled gene-sets (n=150 in size) are plotted (black), along with the density curve of the p-values from the 150 healthy-ageing gene set (red).

Manual searching of PubMed and OMIM yielded some plausible connections with age-related and disease processes (Appendix 1), but such analysis is subjective. We did note that the 150 genes included some previously identified 'ageing' genes; *LMNA* (linked with Hutchinson-Gilford Progeria Syndrome), Unc-13 homolog (*UNC13C*) which is linked with beta-amyloid biology, as well as *COL1A1* (thought to change in skin-ageing). Finally, positional gene enrichment analysis (PGE) was used to identify whether the classification genes (or the classifier network genes) were significantly enriched within given chromosomal regions (De Preter et al. 2008) as previously implemented (Phillips et al. 2013). When we examined if the 150 age-related genes were over represented at genomic loci we found no significant associations. However, on using the top 670 genes from the first stage of the project (>70% success in training model) there were a number of significant findings with 3 genes originating from the top 150. In this analysis, 11q made a significantly greater contribution (adjusted p-value=0.005-0.007) to the enlarged prototype classifier than would be expected by chance (Figure 3.8 B), and there was a total of 15 genes from the 11q13 and 11q23 over-represented genomic locations (11q13 (*ALDH3B1, CAPN1, CDC42EP2, CORO1B, LTBP3, NRXN2, PPP1R14B, RCE1, RCOR2, SART1, SYT12* and *ZDHHC24,* P=0.0005) and 11q23 (*FXYD2, SCN2B* and *TMPRSS13*, P=0.0009)). Interestingly, 11q23 is the location for age-related genetic interactions, namely the apolipoprotein A family (Garasto et al. 2003; Feitosa et al. 2014) as well as a region containing genetic association single nucleotide variants (SNP) which modify the age of onset of colorectal cancer (Talseth-Palmer et al. 2013; Lubbe et al. 2012). Further, 11q13 harbours SNP's associated with age of onset of renal cell carcinoma and prostate cancer and modulating age-related disease emergence by 5 years (Audenet et al. 2014; Lange et al. 2012; Jin et al. 2012).

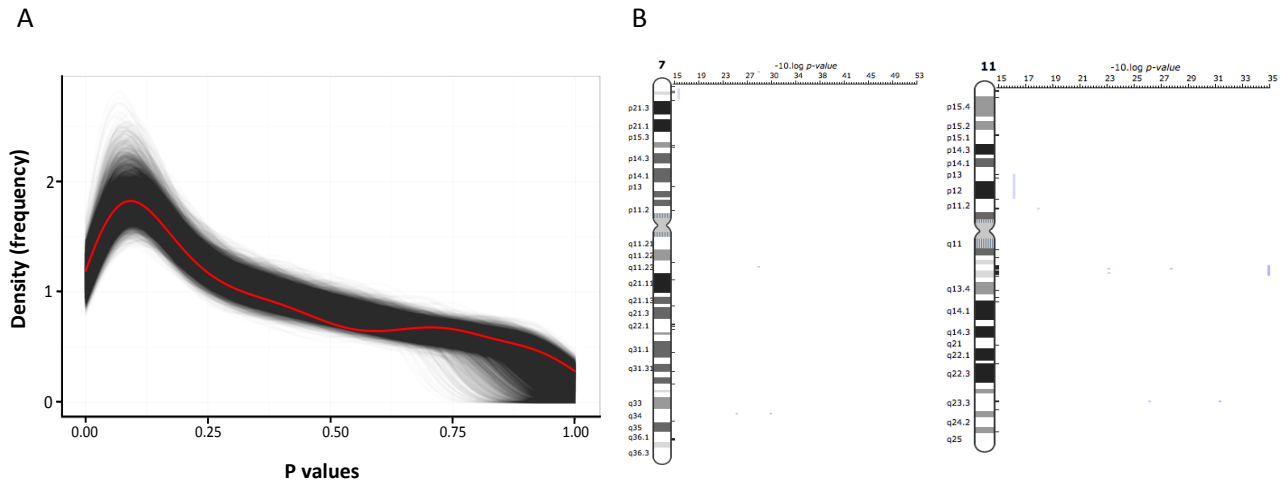


Figure 3.8. GO profile and chromosomal positional enrichment analysis for the healthy ageing RNA signature. *Pathway analysis and GO analysis indicate that the 150 healthy ageing genes are not related to a few specific biological processes but rather originate from across many biological processes. A) Density curves of raw p values for each of the 10,000 hypergeometric tests using randomly sampled probesets from the U133Plus2 gene-chip (n = 150 each time; black) and the density curve of the raw p values from a hypergeometric test using the 150 healthy ageing gene classifier probesets (red). B) Positional gene enrichment analysis for found over-representation at 7q22, 11q13 and 11q23. Those for 11q13 and 11q23 in particular were most significant, and contained genetic variants that influence the age of onset of various cancers.*

3.6 Discussion

Use of fully independent training and validation data sets allows for genuine external validation to be demonstrated and overcomes the over-fitting/bias caveat. The 'healthy ageing' signature fulfilled the first main criteria by providing independent and accurate tissue classification despite inter-laboratory technical variation and different gene-chip platforms. For being a novel diagnostic of 'healthy' ageing it was also important to consider whether the 150 RNAs were related to any likely confounding factors (e.g. life-style or metabolic disease). To test this, we profiled RNA from healthy members of the ULSAM cohort at age 70 years and analysed follow-up data over two decades. In 1992, these 70y old Swedish men had normal levels of physical activity "for their age" and most demonstrated longevity to 90y which is not exceptional in the Swedish population (Danielsson & Talbäck 2012). The healthy age score demonstrated a *four-fold* range (Figure 3.2 A) while chronological age varied by no more than one year across the group. Further the score did not correlate with any life-style related risk factors. We also illustrated the ageing signature's potential clinical utility in three different studies including ULSAM and two AD/MCI cohorts. In ULSAM greater induction of the RNA signature at baseline (~70y) was associated with improved survival over the ensuing 20y period and better renal function at 82y. Similarly, in AD cohorts higher gene

score was indicative of a better cognitive function. Both, renal function and cognition are important determinants of all-cause mortality (Zethelius et al. 2008; Raichlen & Alexander 2014). This concurrent reduction in clinically observed cognitive and renal function suggests both are subject to a general age-related decline in organ function (Romijn et al. 2014).

Neurological decline is predicted to contribute substantially to the economic burden of healthcare in the coming decades. AD is a multi-factorial disease (Hampel et al. 2014) with around 22 genetic loci potentially associated with disease risk or progression of symptoms. The strongest and most reproducible genomic association, APOE-ε4, is a modifier of risk, contributing to the age of onset of the disease by 3.7% (Naj et al. 2014). The remaining ~9 reproducible risk loci for late-onset AD (the most common form) contribute a further 2.2% of the variance in age of onset (Naj et al. 2014). In short, these DNA sequence variants will not be clinically useful for diagnosing or managing AD or even assessing risk, in the majority of people. Differential gene expression analysis and molecular classification have found disease related RNA markers of AD, using patient materials to build the model (Fehlbaum-Beurdeley et al. 2012). However, unknown features of the training dataset can bias such diagnostics. In contrast our 'healthy age genes' were selected via a hypothesis driven strategy that then relied on a validation process that included seven independent tissue cohorts involving multiple RNA detection technologies (so ruling out some unknown technical bias). Thus our healthy age gene expression signature has the key advantage of being a signature built using a paradigm and samples entirely distinct from Alzheimer's case-control samples. The healthy age gene score allowed us to demonstrate that patients diagnosed with AD have an altered healthy ageing RNA expression signature in blood that demonstrates significant association with disease.

Further, the muscle or blood gene score was unrelated to life-style diseases such as Type II diabetes and thus may be more clinically specific than earlier AD biomarkers (Laske et al. 2014; Lotz et al. 2013; Ray et al. 2007; Hye et al. 2014; O'Bryant et al. 2011; Hu et al. 2012; Sattlecker et al. 2014; Lunnon et al. 2013), most of which have already failed to replicate in independent clinical studies. We were able to provide independent validation for our earlier AD related 'disease' diagnostic (Lunnon et al. 2013), however, like many AD disease biomarkers (Fehlbaum-Beurdeley et al. 2012), it includes pro-inflammatory markers and oxidative stress, features that can be common to several diseases and thus it may not be specific in clinical practice. Nevertheless, when we combined the Lunnon *et al* AD biomarker (even after removing the 8 genes we found to be regulated in blood by diabetes or vascular disease) with the 'healthy age genes' we yielded an improved diagnostic for AD over and above either diagnostic alone (Figure 3.6). Ultimately, formal diagnosis of AD will continue to rely on a combination of diagnostics including invasive CSF

sampling, PET imaging and MRI. However, given the scale of screening required (e.g. > 1 million people in 2015/16) to deliver sufficient numbers of at risk subjects for AD clinical trials (www.iadrp.nia.nih.gov) a blood-based diagnostic will be extremely useful for pre-screening ahead of invasive and costly follow-up analysis. Enrichment of prevention trials with asymptomatic people most at risk for AD is required to ensure that event rates are sufficiently high to evaluate the multitude of drug-trials being considered for AD (Laske et al. 2014). Finally, while the lack of an apparent specific biological dialogue may be considered disappointing, the extensive independent clinical results strongly support that the novel 150 gene healthy ageing 'signature' is an important marker of healthy ageing in humans. Therefore regulation of this gene expression programme may in time reveal itself to be an important mechanism for maintaining human health and thereby a new opportunity for target development.

3.7 Summary

Our approach to develop the healthy age RNA signature was novel because we first sought to define a set of genes associated with 'healthy ageing' in 'normal' 65y subjects rather than with disease or extreme longevity. Indeed, we were able to demonstrate that the 150 'healthy ageing' genes are consistently modulated in several tissue types, but to very differing degrees in people of the same chronological age. Including the ULSAM analysis (males only), we have demonstrated in three independent clinical cohorts that greater 'healthy age gene score' associates with better health in men and women, suggesting that promotion of this gene expression profile may be beneficial and/or an adaptive compensatory response. Thus, we have identified a novel and statistically robust multi-tissue RNA signature of human healthy ageing that can act as a diagnostic of future health, using only a peripheral blood sample. This RNA signature has great potential to assist research aimed at finding treatments for and/or management of AD and other ageing-related conditions. In the next chapters we will be exploring tissue ageing specifically with respect to neuro-muscular and vascular ageing by using our healthy ageing signature and other external models for ageing/age associated diseases in literature and inferring how and where they might be useful.

4.1 Overview of the chapter

Approaches such as genome-wide association methods, linear models of epigenetic regulation and differential gene expression (transcriptomics) have identified genomic associations with ageing and exceptional longevity and have attempted to explain factors driving age-associated disease risk. As discussed in the previous chapters, using transcriptomics and machine-learning methods we have developed a robust diagnostic tool that successfully discriminates between healthy young and older humans and effectively predicts clinical outcomes. In this chapter, we present the first comparative analysis of some of the existing signatures of human ageing and longevity. We first established a representative RNA signature for each genomic signature and then evaluated the ability of these RNA signatures to classify neuro-muscular tissue age. We also examined how these multiple signatures relate to tissue ageing and health, clinical outcomes and each other. Together these assorted genomic signatures account for ~2% of the genes available on the gene-chip technology we used. We examined if 'random sampling' of the remaining ~98% of genes on the gene-chip could create any n=150 gene-set could replicate or exceed the performance of our RNA signature in age classification (Sood et al 2015). Thus, the principle goals of this chapter are:

- To establish if existing DNA, DNAm and/or non-muscle RNA 'age' signatures could be converted to a 'gene expression signature' that works as a binary classifier of healthy old versus healthy young human muscle and human brain tissue.
- To examine if these RNA versions of other 'age' signatures relate to cognitive health or age-correlated life-style diseases (diabetes and coronary artery disease).
- To examine whether there was any common biological context across different 'age' signatures.
- To investigate a random sampling approach to benchmark our ageing signature.

4.2 Different genomic signatures for ageing and longevity

We have generated a robust binary RNA diagnostic (150 genes) of healthy older human muscle tissue using transcriptomics and machine-learning methods in independent studies (chapter-2 and 3). Different studies have attempted to identify molecular associations with ageing/longevity, age-associated disease or survival by following different approaches. Healthy ageing per se has not yet been investigated. After generating and validating our RNA signature we set out to establish if the existing DNA, DNAm and/or non-muscle RNA 'age' signatures could be converted to a 'gene expression signature' that worked as a binary classifier for neuro-muscular age and Alzheimer's disease. Further, we were also interested in exploring a random sampling approach to examine the

robustness of our healthy ageing signature not only to establish the effectiveness of our machine learning approach but also to demonstrate that the performance of our particular set of 150 probesets was better than that of randomly sampled sets of 150 probesets from the same platform. In Table 4.1 we summarize the methods used to identify the different ageing and longevity signatures used in our comparison study.

Sood binary muscle Age RNA (Sood et al. 2015)	Muscle from healthy but sedentary 65y+ subjects was used to discover potential markers of healthy ageing
Peters linear blood Age RNA (Peters et al. 2015)	Blood expression and linear modelling used to determine a correlative profile of ageing in human blood
Wennmalm senescence RNA (Wennmalm et al. 2005)	Regression based meta analysis across several platforms to discover consistent in vitro senescence genes
Levine DNA SNP smoking-survival (Levine & Crimmins 2016)	Genome wide association study (GWAS) and network analysis used to discover DNA markers enriched in long-lived smokers (n=90)
Perl DNA SNP Longevity (Sebastiani et al. 2012)	Genome-wide association analysis linked with exceptional longevity
Hannum DNAmethylation (Hannum et al. 2013)	CpG sites correlated with age (penalized multivariate regression method)
Horvath DNAmethylation (Horvath 2013)	Quasilinear regression model using CpG sites across multiple human tissues and disease samples

Table 4.1: Ageing and longevity signatures. *Different Methods such as regression models and GWAS were used to identify the different ageing/longevity genomic signature.*

4.2.1 Producing representative RNA signature

A representative RNA signature for each ageing signature required the genomic features of the signature first to be mapped to equivalent gene symbols and then to the corresponding Affymetrix probeset ID. We used the BioMart tool and the Ensembl database to achieve the mapping. For all of the six ageing signatures used in this study (Table 4.1) the gene lists were provided by the authors in their respective publications. The authors of the smoking resistance DNA signature (Levine & Crimmins 2016) used PLINK, a whole genome analysis tool to map SNPs that fell within the designated GRCh37/hg19 coordinates of the gene instead of assigning upstream or downstream SNPs to a gene. Similarly, for the Horvath DNAm signature (Horvath 2013) the CpG sites located

in the promoter of a gene were assigned the corresponding gene symbol. Table 4.2 shows the number of genomic features in each individual signature and corresponding number of genes they mapped to respectively.

Existing genomic 'age' signature	N^{o} of Genes
Sood muscle Age RNA (150ps)	150
Peters blood Age RNA (1497 genes)	1497
Larsson cell senescence RNA (309 Genes)	309
Levine SNP-smoking (215 SNPs)	215
Perl Longevity DNA (281 SNPs)	130
Hannum DNAm (71 CpG sites)	83
Horvath DNAm (353 CpG sites)	353

Table 4.2: Mapping of genomic features identified in ageing studies to gene symbols. *Each of the different ageing and longevity signatures consisted of different genomic features which were transformed to a representative RNA signature for a comparative analysis.*

4.2.2 Overlap with the healthy ageing signature

After mapping genomic features to gene symbols we noted that the overlap in gene symbols between the different studies was very low. We also checked the overlap of the different ageing signatures with our RNA signature (both n=150 and the n=670 probeset lists with success rate >70%) in particular and tested if the overlap was significant or not based on Fisher's exact test. The only statistically significant results were for the study of DNA markers enriched in long-lived smokers (Levine & Crimmins 2016) and the correlative RNA signature profile of ageing in human blood (Peters et al. 2015). These had had overlaps of 11 and 48 genes respectively with our 670 probeset list (p-value<0.05).

4.3 Random sampling

Our original study took a hypothesis driven approach to study a physiological phenomenon, namely healthy muscle ageing and selected a single high performing gene-set to represent the hypothesis to be tested in thousands of independent samples. A consistent reliable signature had never been achieved before for healthy ageing. Recently a pre-print study based on our microarray data that stated that our healthy-ageing marker genes can be replaced by essentially any random set of 150 genes, with essentially equivalent performance (Jacob et al - http://biorxiv.org/content/biorxiv/early/2016/04/05/047050.full.pdf). However, the 'random sampling' strategy used had the major caveat in that it did not address our primary aim of finding a

single gene-set that works across all data-sets. Instead Jacob et al generated a separate gene-set each time and hence demonstrated low reliability. Further the idea that one can select a 'random' gene-set is, from a biological perspective, flawed as there are many genes regulated with human ageing (Sood et al. 2015; Peters et al. 2015; De Magalhães et al. 2009; Wennmalm et al. 2005) and none of these age-correlated gene-sets are random in a true sense. For a genuine test of a random sampling approach such age correlated genes should be excluded from the sampling. This was not done by Jacob et al.

To truly demonstrate the performance of random gene-sets as tissue age classifiers we first removed the genes we had originally identified with classification ability in ageing (n=670) (Sood et al. 2015) from the starting pool of genes to be sampled. We then evaluated the ability to classify, with statistical significance (Fisher's exact test), age or disease in 10,000 'random gene-sets' of n=150 genes in multiple tissues. Importantly, unlike the Jacob et al we used the same set of 10,000 different n=150 gene-set lists across all tissues examined. Classification performance was assessed using accepted methods (Speed 2003) with external validation and LOOCV so that each gene-set is judged in an independent data set (see section 4.5). In order to compare the performance of the random gene-sets across each muscle dataset we ranked the area under the curve (AUC) values generated from receiver-operator characteristic (ROC) curves of the random gene-sets along with AUC of our gene-set and six other literature derived ageing signatures (Table 4.1). We then calculated the cumulative median rank for each of the 10,000 random gene-sets and each of the age signature gene-sets (Figure 4.1). The performance of our published 150 probeset (red dot) exceeded the performance of all 10,000 random selections of 150 probesets (boxplot; median and quartiles) and previously published ageing/longevity signatures as shown (various coloured dots in Figure 4.1).

We then selected the top 764 performing (AUC >0.8) random gene-sets and inspected the genes in each of these. From the redundant pool of ~114,600 genes we found only a subset of 131 genes that occurred in more than one percent of the gene-sets. We created a new gene-set from these frequently appearing genes and tested classifier performance in muscle and brain age (Figure 2) and in the Alzheimer cohorts(Lovestone et al. 2009; Sood et al. 2015) . We also repeated the same analysis by using the random n=150 genes list that had a median rank order of ~5000 (AUC=0.73). These two random sets of n=131 genes and n=150 genes respectively were then compared to our RNA signature and other ageing gene-sets.

4.4 Neuro-muscular tissue age classification

We evaluated the ability of the different ageing signatures and each of the 10,000 probeset lists to distinguish neuro-muscular tissue age (young vs old) using fully independent external validation a

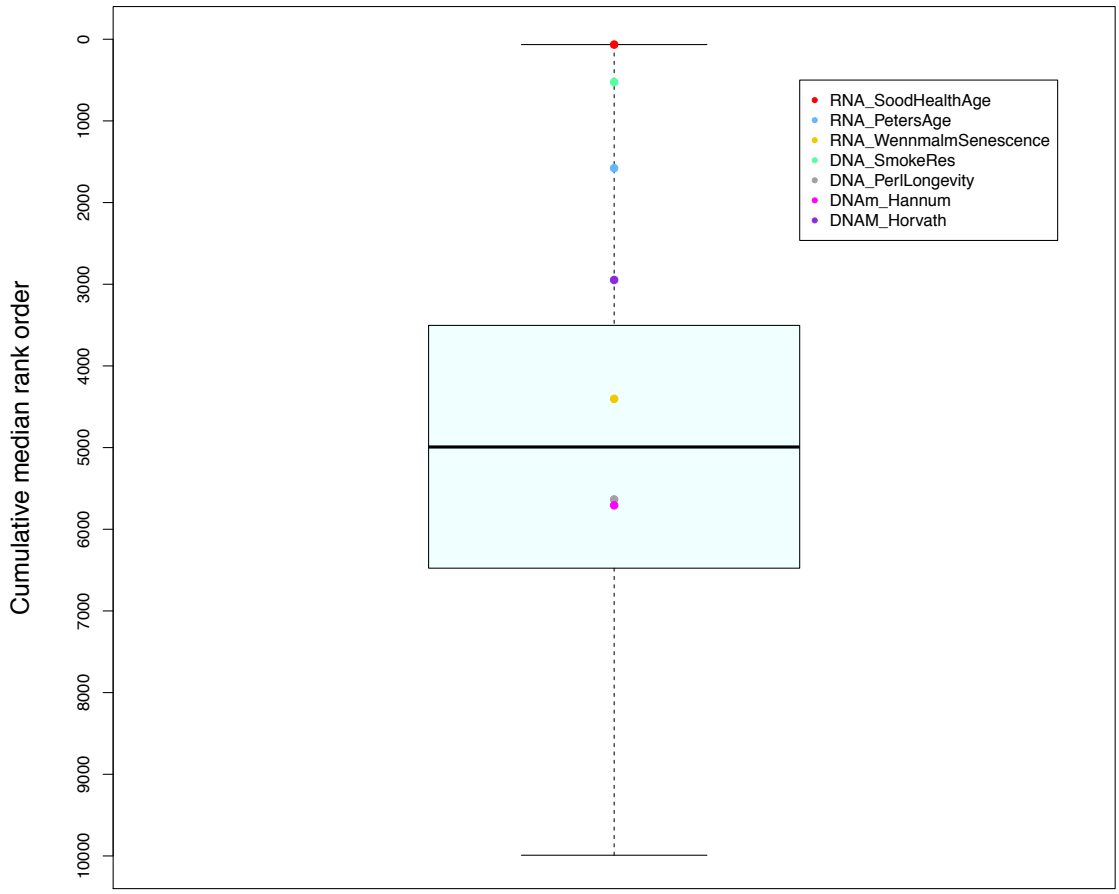


Figure 4.1. The rank order for area under curve for ROC analysis on 10,000 'random' samples of 150 probesets. *Following removal of our previously identified 'healthy age' genes (with AUC >0.7, 670 genes) we produced 10,000 random sets of n=150 probesets from the gene-chip ('Stockholm' healthy age samples). We assessed each set of 150 probesets for its ability to classify muscle tissue age, using gold-standard external validation methods and 5 independent gene-chip studies (5 nearest neighbor KNN classifier as implemented by Speed and Jacob). We compared the performance of these 10,000 probesets with our published 150 probeset and 6 additional published 'ageing' related gene-sets. The performance of our published 150 probeset (red dot) exceeded the performance of all 10,000 random selections of 150 probesets (boxplot; median and quartiles). Each previously published ageing/longevity is shown by a different colored dot. The median AUC from random sampling was 0.73*

process that requires both independent 'known samples' to define the expression space and independent test gene-chips (Shao et al. 2013). We used Frozen Robust Multi-array Analysis (fRMA) (McCall et al. 2010) for normalization and CoMBAT for batch adjustment (Johnson et al. 2007) to account for technical variance due to different laboratories and operators. Then, kNN was used on 4 independent muscle data-sets (the 5th was the CAMPBELL dataset for external validation) and on four distinct human brain regions (120 samples) from brain-bank array source (Berchtold et al. 2008). We tested for statistical significance in classifier performance using

Fisher's exact test. The resulting p values were then corrected for multiple testing using Benjamini-Hochberg correction method(Yoav Benjamini 1995).

The AUC for classification of muscle by these different signatures (row-wise Figure 4.2) was 0.92, 0.83, 0.76, 0.69, 0.85, 0.73, 0.76, 0.91 and 0.67 respectively. For brain the AUC were 0.67,0.79,0.61,0.64,0.69,0.59,0.75,0.6 and 0.58 respectively. The absolute $\log_{10}$ of the adjusted p-values are shown in the form of a heat map (Figure 4.2). Muscle tissue age was successfully determined with RNA signatures selected from various genomic signatures. However, in the case of brain tissue only our healthy ageing RNA signature, Peters RNA blood signature and Horvath DNAm signature performed with statistical significance ($p<0.05$). One of the interesting observation from our analysis was that relation between AUC and p-value was not necessarily 1:1 that is a higher AUC didn't correspond to a lower p value and vice versa. A possible explanation for this could be uneven class distribution (more samples in young and less samples in old or conversely) which could have impacted the AUC values (Daskalaki et al. 2006).

We also studied the different gene lists ability to classify tissue age across 3 brain regions, hippocampus, putamen and cerebellar cortex with healthy samples from BrainEac.org gene-chip resource study (Ramasamy et al. 2014). The hippocampus and putamen are both associated with neurodegeneration whereas cerebellar cortex is not subject to substantial age-related anatomical changes and thus serves as a control in this analysis (Ramasamy et al. 2014; Horvath et al. 2015). Using the cumulative gene score ranking approach (explained in the section 4.6) each brain region from each of the 134 subjects were individually ranked and the median sum of the ranked scores was calculated. The Wilcoxon rank sum test was used to evaluate if the expression of the different gene lists differed across the brain regions with age. As already mentioned in chapter 3, across the 3 human brain regions the RNA signature derived from healthy old muscle (Sood et al. 2015) was highly regulated in regions associated with neurodegeneration (hippocampus and putamen, Appendix 2 Figure A4.1A). The Peters blood RNA signature also tracked human brain age, albeit to much lesser extent (Appendix 2 Figure A4.1B). Consistent with multiple published observations (Ramasamy et al. 2014; Horvath et al. 2015), human cerebellar cortex did not appear to be subject to substantial age-related changes (Appendix 2 Figure A4.1).

4.5 Testing prognostic abilities of signatures in clinical studies

For clinical case-control analysis, each RNA signature was converted to a cumulative gene-ranking score based on individual RNA expression in the muscle classification dataset i.e. if the gene was

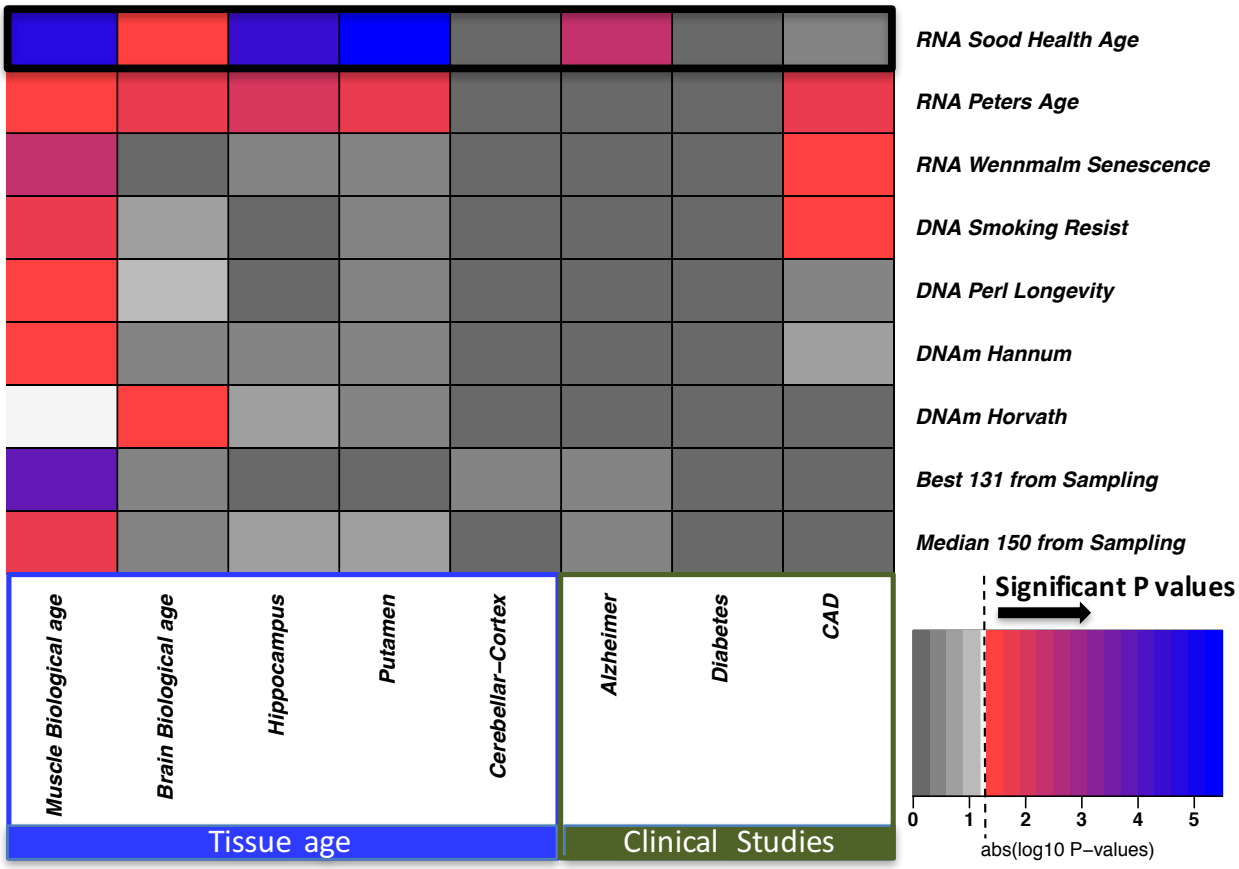


Figure 4.2: Heatmap representation of p-values for the application of each gene-set in multiple tissues and clinical disease samples (as per Sood et al 2015). *Each row of the heat map is one gene-set representing one RNA signature. Columns represent the human muscle, brain, brain regions and clinical data-sets. The statistical analysis from Sood et al 2015 is contrasted with the adjusted p-values for the 8 additional gene-sets (absolute log_{10}). Grey to dark grey is non-significant, near-white being marginally significant (p=0.05) with red towards blue representing an increasing order of significance. Interestingly, when the overlapping age genes (~48) from peters signature (section 4.2.2) were excluded it was no longer able to classify human brain age (second column and second row in the figure).*

down regulated in human muscle, from 25y to 65y, the sample with the highest expression was assigned a rank score of 1 and the subject with the lowest expression value was assigned the highest rank value. For genes up-regulated with age, the opposite ranking strategy was used. The median sum of these rank scores was calculated for each clinical sample (each gene provided equal weighting) (Sood et al. 2015). For the case-control analysis, feature selection (genes) was therefore independent of the clinical studies, while the direction of regulation reflected regulation of that gene in healthy old muscle versus healthy young muscle. To test if the cumulative gene score differed significantly between case and control we used Wilcoxon rank sum test.

As previously discussed, consistent with the substantial modulation of our RNA signature in human hippocampus (p=0.00005, Appendix 2 Figure A4.1A), expression of these RNA could also uniquely distinguish between Alzheimer's disease cases and controls. None of the other RNA representative signatures including Peters RNA signature (modulated in hippocampus to some extent) could predict cognitive health status in either of two independent Alzheimer cohorts (Figure 4.2) (Lovestone et al. 2009). We also used the clinical studies on blood RNA in type II diabetes (Tabassum et al. 2014) and blood RNA in people with and without coronary artery disease (CAD) (Sinnaeve et al. 2009) to test if any of the age related RNA signatures could capture some aspect of these age related diseases. None of the signatures could distinguish diabetics from controls. Two RNA signatures (Cell senescence (Wennmalm et al. 2005) & blood age (Peters et al. 2015)) and one DNA SNP (survival in smokers (Levine & Crimmins 2016)) derived RNA signature were diagnostic for coronary artery disease (CAD) versus control (n=222, adjusted p<0.05, Figure 4.2 and Figure 4.3).

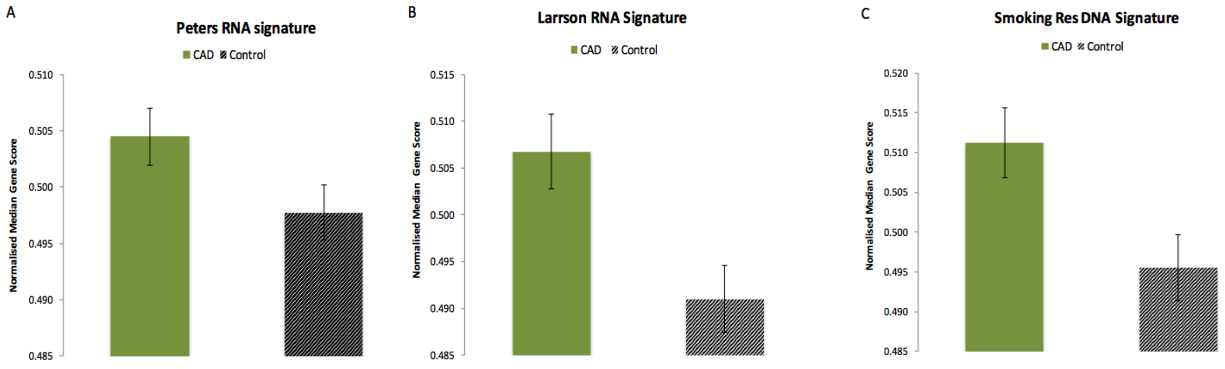


Figure 4.3 Vascular disease plot. *RNA rank-score in blood samples from a case-control study of middle-aged coronary vascular disease (112 controls and 110 cases, group age=53.3y, Affymetrix HG-U133A) was studied in the different ageing/longevity signatures. A Wilcoxon rank-sum test tested if cumulative gene ranking score for controls was significantly different from patients (CAD). Peters RNA signature included 48 of our 150 age signature genes, removing which didn't alter its ability to distingusih CAD from controls.*

Smoking is a major risk-factor for CAD and the Levine et al SNP results appear to translate to an RNA signature (survival to 85y despite smoking), while cellular senescence has been shown to play a role in vascular disease (Levine & Crimmins 2016; Wennmalm et al. 2005). The Peters RNA signature (Peters et al. 2015) was derived using samples with cardiovascular disease (e.g. hypertension) rather than age per se (Figure 4.2) and thus it comes as no surprise that it could discriminate CAD from controls. In all three signatures CAD had a higher gene score implying that

genes most regulated with age are more regulated in CAD then controls which could indicate response to damage. We lastly investigated if the three signatures were enriched for common biological process and/or molecular function gene ontologies. No overlapping ontologies were found across these three gene-lists.

4.6 Discussion

Population ageing and the shift from infectious to chronic diseases as major causes of death have together created an urgent need for the discovery of ageing and age related biomarkers/signatures. Targeting ageing is theoretically better than treating individual chronic diseases, however up to this point, translational routes to accomplish this objective have been purely speculative (Kaeberlein et al. 2015). While this subject has received considerable recent attention, there are numerous challenges to both the development and the implementation of diagnostics for ageing (Goldberger & Buxton 2013), including economic considerations.

Given the fact that the choice of method and cohorts used for developing biomarkers has a great influence on the subsequent statistical analysis and biological answer, the quality and credibility of these methods need to be assessed fairly. There are multiple competing technological platforms that yield plentiful data adding to the challenge for scientists to find a way to integrate information across different studies. So far progress in integrating different data formats to yield robust and sensitive diagnostics for clinical decision making remains slow (Goldberger & Buxton 2013). To this end, our comparison study in this chapter deals with seven representative RNA signatures (including our healthy ageing RNA signature) compared on healthy neuro-muscular ageing datasets and on clinical datasets representing various age related morbidities.

The utility of DNA sequence variation to guide treatment of cardiovascular disease or neurodegeneration is just being explored (Sawhney et al. 2012). However, this approach is severely limited by the total contribution that DNA variants make to the heterogeneity of these types of diseases. A study of exceptional longevity using Genome-wide association analysis linked 281 DNA variants with exceptional longevity (Sebastiani et al. 2012) and collectively explained only 17% of the variance in humans with an average AUC value of 0.65. However, long-lived humans appear to have a similar genetic burden for common DNA disease variants, suggesting the human exceptional longevity model may not be reflective of the processes that determine average longevity (Gierman et al. 2014).

In their work on the transcriptional profile of ageing in peripheral blood, Peters et al found 1497 genes to be associated with age. From this they identified only 163 genes (~11%) in cerebellar cortex to be significantly associated ($p<0.05$) with age in same direction as in whole blood. This

not only highlights the caveat of correlative models being highly tissue specific but is also primarily at odds with studies of the human cerebellar cortex (Horvath et al. 2015; Sood et al. 2015) as this brain region typically does not show a genomic signature of ageing nor demonstrate typical morphological changes with age (Ramasamy et al. 2014). This suggest that the Peters signature is not age, or not only age, but may reflect a number of other biological phenomena. For example, the Peters et al gene set is strongly correlated with hypertension but not with neuromuscular measures of ageing like MMSE scores (muscle function or cognitive status). A number of other published datasets seem to have this characteristic. Gene-sets like Peters et al are generated from epidemiological cohorts, validated in a single tissue type, and this type of linear covariate analysis can introduce statistical artefacts and requires validation in multiple cohorts. We produced a robust multi-tissue binary age classifier, using a single machine learning model (Chapter 2). Nonetheless it is always plausible that a number of other genes, in combination, could classify muscle age and indeed using literature 'age' signatures we have shown that other combinations were significant classifiers of muscle age (first column in Figure 4.2). Thus we demonstrate that human muscle tissue age can be determined with RNA signatures selected from diverse genomic signatures (e.g. derived from DNA sequence or methylation). As observed in the comparative analysis, only the Peters and Horvath gene-sets were able to classify both human muscle and brain but performance of these RNA signatures was modest. In clinical analysis, the RNA signature derived from genes selected to provide protection from environmental stress (survival to 85y despite smoking) was diagnostic for coronary artery disease (CAD) versus control (n=215, adjusted $p<0.05$) (Ambrose & Barua 2004). Cellular senescence has been shown to play a role in vascular ageing (Erusalimsky & Kurz 2005; Fyhrquist et al. 2013) and the senescence RNA signature could also distinguish between CAD versus control (n=309, adjusted $p<0.05$). Similarly, the Peters RNA signature derived in blood could distinguish between CAD versus control (number of genes=1497, adjusted $p=0.03$). In this case the cohorts used in the study contained older participants with age-associated diseases and thus the ability to classify CAD was not surprising.

Activation of different ageing signatures was studied in blood samples from two independent large case-control studies of Alzheimer's disease. Except for our RNA signature none demonstrated significant results for cognitive health status. This could be explained by increased regulation of the healthy age gene score with chronological age in the hippocampus ($p=0.00005$), as well as putamen ($p=0.000005$) both regions associated with neurodegeneration (Laakso et al. 2000; Erickson et al. 2011; Taupin 2006; Sekar et al. 2014; Lunnon et al. 2013). Additionally, we noticed that a small subset of genes could drive the overall performance of the literature gene-sets. For example, the large Peters et al gene-set (n=1497) has 48 genes in common with our healthy ageing signature and when our 48 genes were removed, the remaining Peters et al gene-set was no longer

able to classify human brain age (p-value =0.08), but its relationship to CAD remained unchanged. This implies that the ability of the age signature of Peters et al to classify brain age is driven to some extent by the 48 overlapping genes from our RNA signature.

4.7 Summary

This chapter provides an important perspective on both the utility and limitations of different genomic signatures of ageing when transformed to equivalent RNA (gene expression) signatures. In this analysis we observed that our muscle derived gene-set was the only one related to hippocampus ageing and cognitive health while 'stress' resistant and 'epidemiologically' selected linear models related with vascular disease (CAD) Analysis of the global transcriptome (RNA), using machine learning methods, has produced sensitive tools for cancer diagnosis and prognosis in the past (Abd El-Rehim et al. 2005; Shedden et al. 2008; Patnaik et al. 2010; Menden et al. 2013). Further, it has been possible to select features from a tumour global RNA profile that predicts drug sensitivity (Knudsen et al. 2014). Given the superior technical reproducibility and throughput of the Affymetrix gene-expression platform over DNA-methylation assays, the study shows that RNA signatures may represent an ideal approach for optimizing 'age' diagnostics as well. Also, in this analysis we have shown that a hypothesis driven approach based on machine learning methods is more reliable than the random sampling approach detailed by Jacob et al. The latter approach fails to produce a single gene-set capable of acting as a multi-tissue classifier of age and with discriminatory power in Alzheimer's disease. We originally identified one gene-set that works in all our data-sets, and in this chapter showed that after excluding our gene-set, 10,000 random samples cannot replicate the exceptional performance of our 'healthy ageing gene-set. This highlights the limitations of 'big data' analysis strategies when applied in the absence of clinical or biological insight.

5.1 Overview of the chapter

One of the earliest signs of vascular ageing is arterial stiffness, a process that is thought to be accelerated by arterial hypertension, resulting in an increased risk of cardiovascular morbidity and mortality. Arterial stiffness can be categorized by estimating vessel compliance, and decreased arterial compliance is one of the earliest indications of adverse structural and functional changes within the vessel wall. It is defined in the clinic by a parameter called pulse wave velocity (PWV), a measurement that is influenced by the end-user (technician) and the technology being used. In this chapter we investigate the possibility of developing a model for vascular ageing using gene expression data and clinical parameters (as dependent variables) known to co-vary with PWV measurements such as blood pressure and chronological age. Through this analysis we show the importance of understanding the difference between statistical and quantitative significance in diagnostics or biomarker development. Thus, the principle goals of this chapter are:

- To identify number of genes across the pre-existing skin tissue expression data, whose baseline expression correlate with PWV in vivo and transform them to a feature score metric.
- To evaluate a regression model of the resulting feature score combined with BP and/or age.
- To test the robustness of the model in validation datasets by comparing the predicted clinical response (PWV) with the actual observed PWV measure.
- Lastly, to test if our healthy neuro-muscular age signature of 150 genes (derived from a non-linear, binary approach) could be converted to a linear regression model for vascular stiffness (which we expected to be negative).

5.2 Over-view of vascular ageing and arterial stiffness

Ageing being an inevitable part of life significantly affects the heart and arterial system and is accompanied by decline in various physiological capacities – such as vasodilation and aerobic fitness. Vascular ageing is associated with changes in the structural and mechanical properties of the vascular wall that leads to the loss of arterial elasticity and impaired endothelial function (North & Sinclair 2012; Laurent 2012). Breakdown of elastic components mainly elastin and collagen in the aortic wall, results in its parallel stiffening and dilation (Jani & Rajkumar 2006). Thus, vascular changes contribute to the age dependent risk in developing vessel disease (e.g. tissue remodeling and atherosclerosis). In addition to non-modifiable risk factors such as chronological age and gender, there are other important life-style influenced risk factors such as hypertension and diabetes mellitus that also accelerates arterial stiffening increasing vulnerability for developing cardio vascular disease (Benetos et al. 2002).

Arterial stiffness has independent prognostic value for coronary and cardiovascular morbidity and mortality (Zieman et al. 2005; Nilsson et al. 2009). It can be measured by different non-invasive parameters like pulse wave velocity (PWV), augmentation index, pulse contour analysis etc. (Kelly et al. 1989). PWV measures the velocity of the propagation of the forward and backward pressure waves between two points of the artery and it is a reproducible and technically demanding parameter for estimating blood vessel function (Yamashina et al. 2002; Hansen et al. 2006; Laurent 2012), which is often considered as a clinical gold standard of measuring arterial stiffness. PWV varies greatly by blood pressure and is thus often adjusted for this variable (Ruitenbeek et al. 2008; Reference Values for Arterial Stiffness' Collaboration 2010) while our model was adjusted for chronological age as well to establish if there was also an underlying RNA signature that could contribute to accurately estimating vascular 'health'.

In terms of genomic knowledge of vascular age, methods such as whole genome sequencing and genome wide association studies (GWAS) have identified different genes and chromosomal regions potentially involved in arterial stiffness (Medley et al. 2002; Turner et al. 2006). SNP association studies have recognized various population specific variant that relate to age related vasculature stiffness (Lajemi et al. 2001; Ye 2006). While this is a progressive step forward, it is important to note that arterial stiffness is a polygenic condition that occurs due to the sum of multiple polymorphisms, with each variant having a relatively small effect (<5%) on the phenotype (Lacolley et al. 2009). Moreover, GWAS studies often lack the knowledge about which gene a SNP/variant impacts on, as it could be a gene far removed from the SNP location. This gets more complicated in polygenic conditions like arterial stiffness where each SNP could potentially be associated with more than one gene and modelling such data is still considerable very challenging. Therefore, rather than individual gene variants it will be more interesting to inspect whether vascular ageing phenotype may relate more closely to gene expression. In the past transcriptomic signatures based on differential gene expression have been studied as biomarker of arterial stiffness in humans and have indicated that quantitative differences in gene expression have the potential to define a person's phenotype (Durier et al. 2003; Heidecker et al. 2008). Thus, combining gene expression together with machine learning methods and clinical covariates we could yield a more promising approach that could also disclose the link between transcriptional regulation and vascular ageing.

5.3 Methods

5.3.1 Dataset and participants

Gene expression data from SKIN tissue from female Caucasian twins (~340 samples) from the Twins UK cohort was available on arrayExpress (Illumina Human HT-12 V3, E-TABM-1140), a cohort with characteristics similar to the general U.K. population (Andrew et al. 2001). The baseline age ranged from 39 to 85 years with a mean age of 59 years. All women underwent assessment of arterial stiffness by measurement of carotid-femoral pulse wave velocity (cfPWV). Gene expression levels were measured using Illumina Human HT-12 V3 BeadChip from skin tissue. From these 340 participants with gene expression data, 84 women had repeated vascular measures at a 4.3 ± 1.4 year follow-up.

5.3.2 Mean Arterial Blood pressure and PWV Measurements

Vascular measurements were performed in a temperature-controlled vascular laboratory (~ 24ºC). Brachial blood pressure was measured using a validated oscillometric device (Omron 705CP, Omron, Tokyo, Japan) after subjects had been supine for at least 10 minutes. SphygmoCor system (Atcor Medical, Sydney, Australia) was used to measure cfPWV by sequentially recording the carotid and femoral artery pulse by applanation tonometry (Nelson et al. 2010) with a high-fidelity transducer (Miller Instruments, Houston, Texas). Difference in time of pulse arrival from the R-wave of the electrocardiogram between the two sites was taken as the transit and difference in distance was estimated from the distance between the sternal notch and femoral artery at the point of applanation (Cecelja et al. 2009). Measurements were made in triplicate and mean values were used for analysis.

5.3.3 RNA extraction and expression profiling

For expression analysis, punch biopsies (8mm) were taken from relatively photo-protected infra-umbilical skin. RNA was extracted from homogenized tissue samples using Trizol Reagent (Invitrogen) according to manufacturer's protocol. RNA quality was assessed with the Agilent 2100 BioAnalyzer (Agilent technologies) and the concentrations were determined using NanoDrop ND-1000 (NanoDrop Technologies) and samples were stored in -80°C until ready to use. Expression profiling was performed using the Illumina Human HT – 12 V3 BeadChips (Illumina Inc) where 200 ng of total RNA was processed according to the protocol supplied by Illumina. For quality control, expression profiling was repeated two to three times on different beadchips. The expression data were first transformed using variance stabilization and then quantile normalized using the LUMI package in R as it is neither too strict nor too negligent while normalizing the data (Du et al. 2008; Ritchie et al. 2011).

.

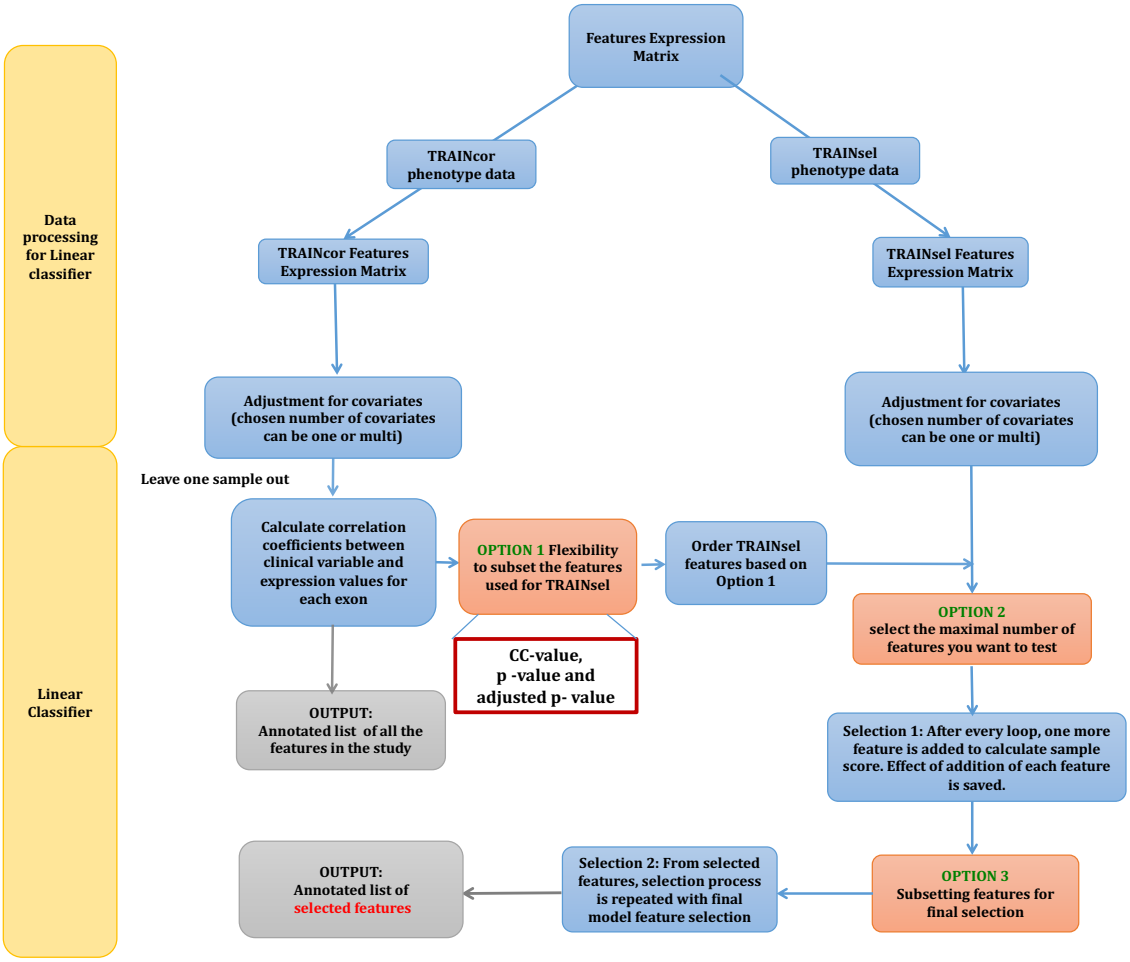


Figure 5.1 Workflow for developing a linear signature of vascular ageing using machine learning methodology. *We developed a pipeline for selecting the features from the transcriptomic data that linearly correlates with the clinical endpoint, PWV in this case. The workflow gives user the flexibility to subset the features based on different criterions (orange boxes shows the different options).*

5.3.4 Machine learning approach for predictor development

The aim of this analysis was to build a linear regression model that incorporates gene expression data along with clinical covariates such as age and blood pressure (variables that have been reported to relate to vascular health) to determine the PWV measure for a subject which in turn could provide insight into a subject's vasculature health or age. In order to determine the set of genomic features for this vascular age model, we developed a pipeline (Figure 5.1) that gives a list of features that correlates with the clinical endpoint (PWV in this case) but is able to select the features based on a variety of selection criteria when applied to a second data-set. When we explored the development of this prototype method, we considered both the direction of association with the clinical phenotype and the slope

5.3.4.1 Data pre-processing for linear classifier pipeline

To begin with we split the dataset into three groups, two splits/subsets were used as training datasets in the pipeline and one subset was used for validation (not part of the pipeline). For the training datasets, split 1 (~141 samples) was used for ***'feature selection'*** and split 2 (~116 samples) was used for ***'model selection'*** i.e., to combine the selected genomic features into a model that correlates with PWV. We ensured that distribution of the clinical variables is similar between the splits by plotting their distributions. While blind validation is far more robust, we did not have a validation set generated in a different laboratory and so relied on splitting the cohort with the third group being used for the statistical validation step.

We processed the RNA expression data using a standard deviation filter to remove probes from the study that had both a low and invariant expression signal. To select a suitable SD filter value, we plotted the distribution of SD values and the peak SD in the distribution was chosen as the threshold for detection (in this case features with SD >8 intensity values were considered as detected). For the detected features (~25,107 probes) expression measurements from LUMI were logit normalized i.e. log transformed and scaled to mean zero and SD 1 (Chen et al. 2012; Knudsen et al. 2014). The expression values ranged over several orders of magnitude and transformation ensured that when we got to the model selection phase each gene would be equally weighted when considering arithmetic combinations. Further, since PWV intrinsically varies with chronological age and mean arterial pressure (Figure 5.2), we adjusted the transformed gene expression data, by the residuals extracted from the linear model fitted through the response variable (PWV) as a function of these two clinical covariates (age and Mean arterial pressure). These steps correspond to the first main block (in yellow) of Figure 5.1.

5.3.4.2 Selection of features using feature selection dataset

To calculate the 'final' correlation between PWV and RNA expression for a given probe, in the 'feature selection' phase, one sample was left out of the calculation and Pearson correlation was calculated along with p-values, adjusted p-values and slope of the best fit line, and the median values for these parameters were retained. Features were taken forward if they had a median adjusted p-value <0.05 (option 1 in Figure 5.1) and these were ranked by the median correlation coefficient (CC) values. Next, from this rank ordered list the number of features to be carried forward downstream into the model selection run were selected (option 2 in Figure 5.1). The number of features selected for this prototype was arbitrary, however we tended to select a larger than optimal list as we wished to attempt to get a robust gene ontology profile to help interpret the underlying biology.

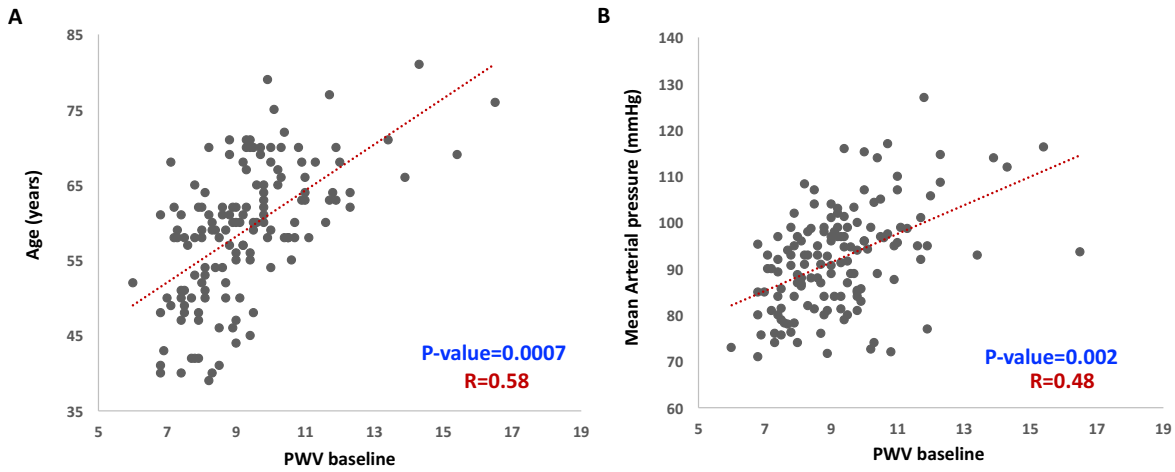


Figure 5.2 Relation of PWV with covariates*. Pulse wave velocity, the measure of arterial stiffness is known to intrinsically vary with chronological age and mean arterial pressure which holds true in our data as well.*

5.3.4.3 Selecting gene sets using model selection dataset

From the feature selection dataset, we obtained a number of interesting genes that may or may not combine together to produce a statistically significant linear model. Thus, the gene list obtained was not the end point of the analysis but the beginning of a multi-step process that includes arbitrary decisions according to the the pipeline shown in Figure 5.1. Our aim was to search for genes which when combined together could potentially work as predictor of vascular age and inform us about the underlying biology. The ***'model selection'*** dataset is used in a two-step selection process, where the first step, is an assessment of all statistically significant individual features (selected from the feature selection step) and all samples in model selection dataset. The impact of sequentially adding (i.e. 1,1+2,1+2+3, …, 1+2+…+ n) each feature at a time is calculated and the model plotted (Figure 5.3A).

In order to combine features together to form a model we use features scores where are determined by collapsing the expression values of the chosen genes into a single metric for each sample. Two different ways of estimating the feature score were investigated. Either as sum of the logit RNA expression or it could be computed as mean of logit expression values of positively correlated features minus the mean of logit expression values of negatively correlated features. A linear model of a resulting feature score versus the clinical phenotype (PWV) is created. In the second selection step the recorded effect (e.g. Figure 5.3B) of each feature being added to this model is used to decide which individual features combine together to generate the best model (in this case based on the final correlation coefficient). The potential effect of each feature on the

model are list in Table 5.1 which summarizes selection criterions used. In the pipeline this corresponds to option 3 of Figure 5.1.

	Description
Nosubset	no refined subset selected
Poscor	selects features that are positively correlated with the clinical variable in the 'feature-selection' dataset.
Negcor	selects features are negatively correlated with the clinical variable in the 'feature-selection' dataset
UpPlus	selects features that increase the correlation in an already positively correlated model
Up	selects features that drive either a positively or negatively correlated model in a positive direction.
DownMinus	selects features that drive a negative model correlation towards more negative values.
Down	selects features that drive either a positively or negatively correlated model in a negative direction.
UpPlus|DownMinus	selects features that drive the model in a negative direction if the model is already negative or in a positive direction if the model is already positive. Unlike the Up or Down options the direction and the strength of the association are considered.

Table 5.1: The different selection criterions for the 'model selection' dataset that takes into account the effect each feature has on the model.

5.3.5 Final regression model for vascular ageing and validation

The present linear classifier approach allowed us to explore the transcriptomic search space to find a set of features (genes) that together correlate with vascular age/stiffness. Once we have the knowledge about features relating with PWV, we built a final multiple regression model which was then used to establish if we could predict PWV values from gene expression (with or without covariates like chronological and blood pressure). The *lm* function in stats package from R(R Core Team 2015) was used for this purpose. The transformed feature scores (explained in section 5.3.4.3) were combined with mean arterial pressure and chronological age to predict the PWV for an unknown sample.

To validate each model 84 'new' subjects from the same cohort (the third split which was not used in training pipeline) along with baseline clinical and PWV values measured were used. For a subset of the aforementioned 84 subjects (n=75), PWV value and mean arterial pressures measures from a 4yr follow-up period (±4.3yr) were available and these were used to explore if there was a relationship between gene score and progression of aortic stiffness. The robustness of the model in these validation samples was tested by visualizing the results using Bland-Altman plots which

allows the comparison between the actual and predicted values. In these plots in general, if the plotted data clusters around the mean of the differences (called the bias), and is within ±1.95 standard deviations of the mean known as 'limit of agreement' then the observed and predicted values are considered to be in agreement with each other.

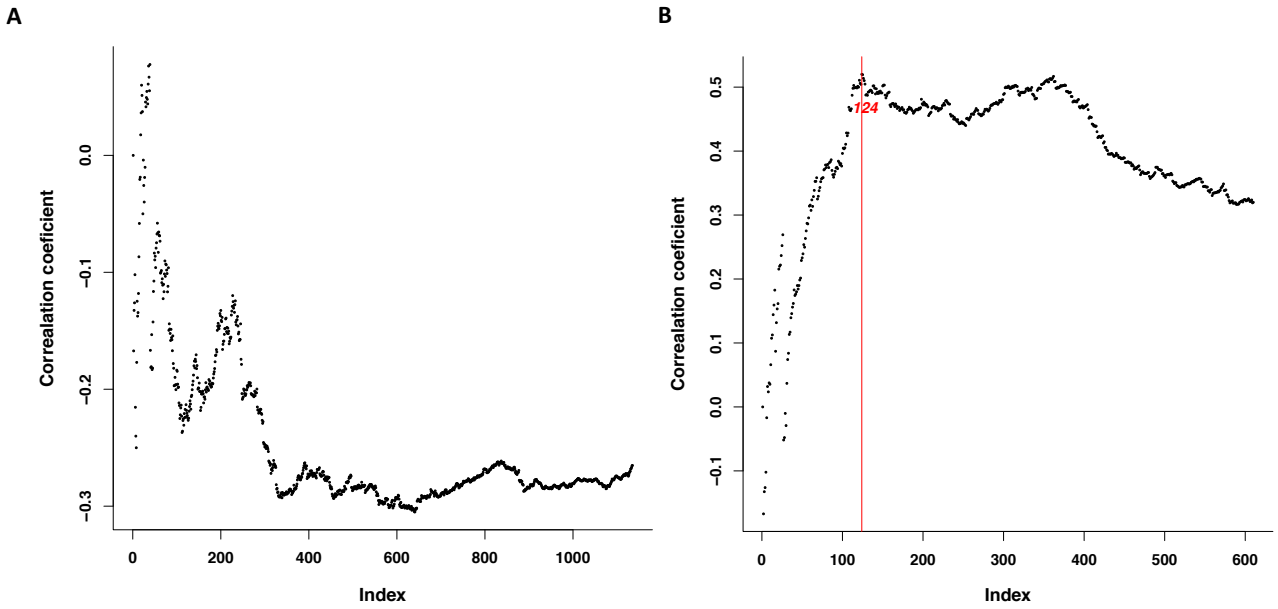


Figure 5.3 Selection criteria in 'model-selection' dataset that takes into account the effect each feature has on the model. *A) We iteratively add one feature at a time and compute the correlation coefficient of the gene set with PWV values. Then we record if adding the feature make the model better or worse and select one of the criteria from Table 5.2 B) Using the sub selection criteria we get the best model which in this case is a set of 124 features.*

5.3.6 Transforming healthy neuro-muscular age signature into a linear model

Using KNN method of binary classification we have found a healthy neuro-muscular signature effective at distinguishing between young and old human muscle and brain while in human blood, it was related to cognitive status in two independent studies (Chapter 3). We were interested in testing if this multi-tissue healthy age signature of 150 genes (~128 Illumina probes) could be potentially converted to a linear model for vascular stiffness. Since the age signature was developed as a non-linear model, it was necessary to find subset of genes from the 150 genes that followed a linear pattern of change with chronological age. Using correlative analysis and classification modeling we identified genes in the nonlinear model that had some linear features using a set of human muscle samples from the STRRIDE II cohort, a pre-clinical sedentary cohort of adults (age range of 25 to 68 years). From these linear genes we use only those that were expressed in both blood and skin as if we obtained a positive outcome we would have progressed the validation using blood-based

gene-expression cohorts. We combined the subset of expressed linear genes following the same approach as for model 1 and 2 described below i.e., adjusted the gene expression values for age and blood pressure and collapsed it into a feature score metric and subsequently analyse its potential as a model of vascular ageing.

5.4 Results and discussion

The linear classifier pipeline provided the flexibility of choosing different selection thresholds and criteria first for individual features and then for model selection as explained in the methods section above. Based on this pipeline we selected and validated two different PWV prediction models.

5.4.1 Predictor genes from machine learning approach

Using the detected Illumina probes the correlation between PWV and probe RNA expression were calculated. The features with a median adjusted p-value <0.05 were selected (~1135 features/probes) and ranked by the median correlation coefficient (CC). We then added each feature score metric (summation of logit expression values) in model selection dataset with 116 samples. In first instance the 'UP' selection criteria (Table 5.1, option 3 in Figure 5.1) gave us the best model of 124 features (Figure 5.3B) with CC value 0.51 (Figure 5.4A) and 0.82 (Figure 5.4B) in the model selection and feature selection datasets respectively (p-value < 10^{-9}). Using an alternative feature score calculation method (difference between the mean of logit expression values of positively correlated features and negatively correlated features) and UP+ criteria (Table 5.1) in 'model-selection' dataset we obtained a different set of 431 features (Appendix 2 Figure A5.1) that when combined together gave us a prototype model with CC of 0.56 (Figure 5.4 C) and 0.66 (Figure 5.4 D) in the model selection and feature selection datasets respectively (p-value < 10^{-11}). None of the gene sets were enriched for any significant gene ontologies when corrected for background bias.

5.4.2 Linear regression models for PWV prediction and validation

Thus, the linear feature selection pipeline gave two different gene sets that correlated with PWV and comprised of 124 and 431 features. By collapsing the gene sets into a feature score metric (explained in methods section) and combining it with age and blood pressure as additional explanatory variables two final regression models for PWV (the dependent variable) were attained. Model 1 had an adjusted R^2 =0.53 (p-value < 10^{-15}) and Model 2 had adjusted R^2 =0.46 (p-value < 10^{-12}).

$$PWV = 9.532 + 0.233 * feature\_Score + 0.072 * MeanArterialPressure + 0.104 * Age \quad [1]$$

$$PWV = -5.347 + 17.459 * feature\_Score + 0.063 * MeanArterialPressure + 0.109 * Age \quad [2]$$

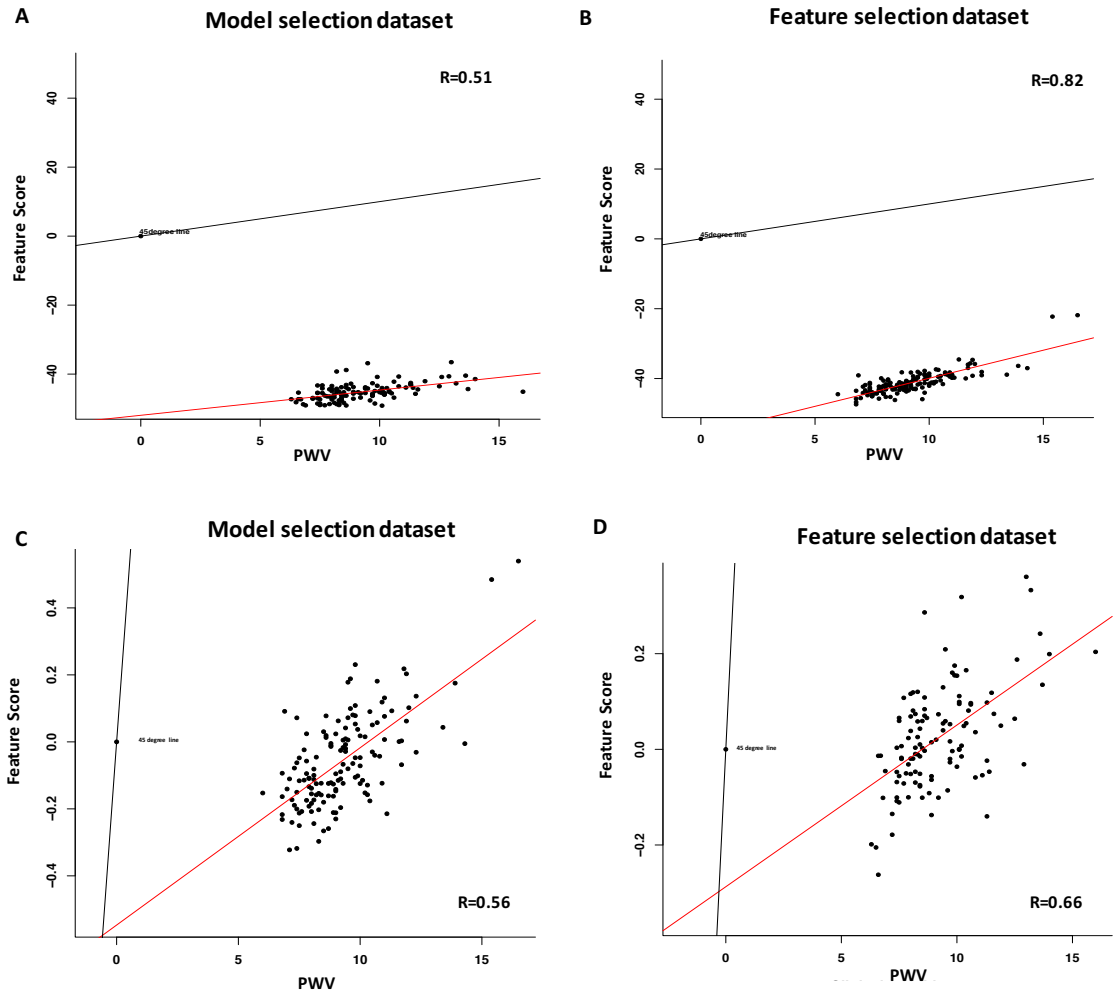


Figure 5.4 Relation between PWV values and feature score. *The figure shows the correlation between the the PWV values and feature score calculated by summation of expression estimates from the two different prototype models obtained from the machine learning linear classifier comprising of gene sets of 124 features (A and B) and 431 features (C and D) in model selection and feature selection datasets respectively.*

To verify that the models developed were able to predict the PWV values for vascular ageing, we predicted PWV measure based on the baseline gene expression and clinical data and compared it with the observed measures of PWV in the third data-set. Similarly, we also investigated if based on baseline gene expression data we could predict the PWV values ±4.3yr baseline.

For each model, the Pearson correlation between predicted and observed values were 0.6 and 0.65 (Figure 5.5) in the two validations datasets and with Model 2 they were 0.58 and 0.61 respectively. However, correlation coefficient is not the only parameter to judge the value of a model and we used Bland Altman plots to visually check the robustness of these models. Majority of observations for both model 1(Figure 5.5) and model 2 values (Appendix 2 Figure A5.2) were within the limit of agreement (±1.95 SD), thus implying that the difference between actual and

predicted PWV values were trivial. These were robust if not extremely strong models, capturing over 50% of the variance. However, inspection of the components of model 1 and 2 indicates that gene expression feature score (i.e. information from gene expression) was making a very modest contribution to the models. To establish this, we derived another regression model (Model 3) solely based on mean arterial pressure and age and without the gene expression data (adjusted R^2 =0.44, p-value<10^{-15}).

$$PWV = -3.123 + 0.065 * MeanArterialPressure + 0.106 * Age \quad [3]$$

On validation datasets 1 and 2, Model 3 had a CC value of 0.64 and 0.67 respectively (Appendix 2 Figure A5.3) and Bland-Altman plots revealed that its performance on validation datasets was at par with models that included the gene expression data thus implying that for a reasonable model for vascular "stiffness" can rely on chronological age and blood pressure.

5.4.3 Healthy neuro-muscular age signature as a model for vascular ageing

Examination of the 150 neuro-muscular gene signature (Chapter 3) we found a subset of ~20 genes expressed in skin and blood and showing a linear correlation with chronological age. We then tried to build a linear regression model by combining these linear features with the same covariates as above (age and mean arterial pressure) and observed that feature score (gene expression data from 20 linear age genes) didn't significantly added to a linear model (p value= 0.503) while age and MAP did.

Thus, a simple regression model using a subset of neuro-muscular healthy 'age' genes did not relate these features to vascular age. Indeed, we previously found neuro-muscular healthy 'age' signature did not relate to coronary vascular disease (Figure 3.7) and this additional analysis further ascertained the original interpretation that the 'healthy ageing gene' score was selectively useful in identifying neuro-muscular ageing and not vascular ageing (or that the vascular phenotypes are not directly related to biological ageing). Interestingly, in Chapter-4 of this thesis we had discussed Peters RNA signature and showed through our analysis that it captures the differences between CAD patients and controls (Figure 4.3A). Also in the original paper (Peters et al. 2015), the signature was strongly linked to blood pressure (p-value<10^{-5}). Since vascular ageing is coherently related to both blood pressure and CAD (Lakatta & Levy 2003; Laurent 2012), therefore there is a strong possibility that the Peters RNA signature represents vascular ageing instead of a general model of ageing as reported by the authors.

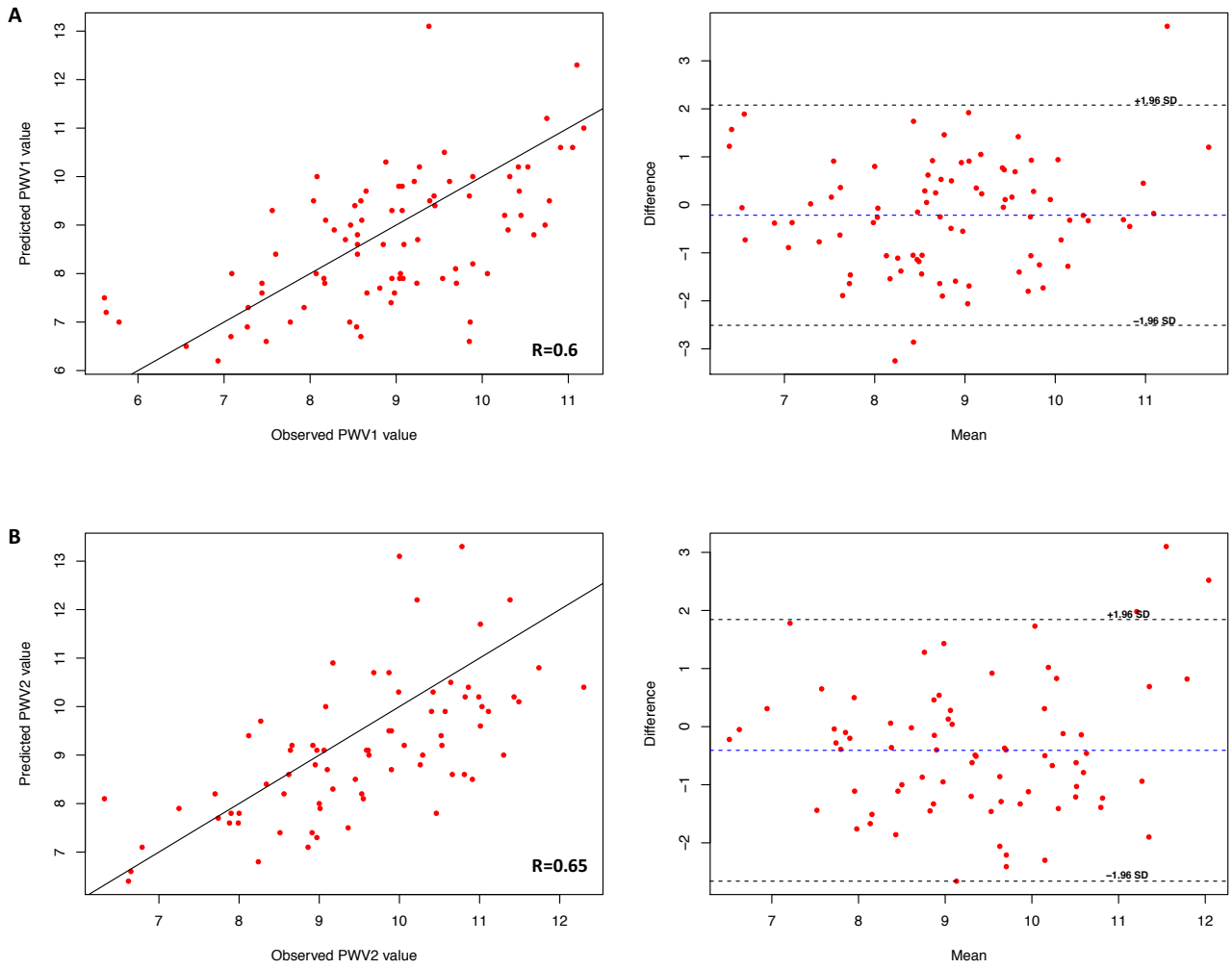


Figure 5.5 Validation of gene expression based vascular ageing signature. *Bland Altman plots (on the RHS) showing robustness of Model 1 in Validation datasets 1 (A) and 2 (B) respectively. Majority of observations for both model 1 and model 2 values were within the limit of agreement (dashed lines), thus implying that the difference between actual and predicted PWV values were small.*

5.5 Summary

We aimed to find an RNA signature for 'vascular' ageing (assuming 'stiffness' = ageing) using the gene expression data from skin from a TWIN cohort as skin is structurally similar to one of the large arteries (Nilsson et al. 2015). In order to do this, we developed a new machine learning linear classifier strategy which gives control over feature selection. We developed two different RNA models both of which statistically validated on independent set of 84 samples with baseline PWV measure and 75 samples with PWV measure, MAP and age measure at a different time point (~±4.3yr from baseline). On careful inspection we observed that feature score i.e. gene expression was making only a very modest contribution and it is possible to obtain a good model for vascular "ageing" that relies only on age and blood pressure. This is in accordance with the literature (Najjar et al. 2008; Kim et al. 2007) which shows that blood pressure is one of the strongest factors

influencing PWV followed by age and gender (Vermeersch et al. 2008). This also shows that having a statistically significant model (determined by p value) does not always translate to a model that has clinical significance as well.

Nonetheless, we cannot completely dismiss a genomic aspect to vascular ageing (as Peters RNA signature seems to grab some aspect of vascular ageing) or that changes in vascular function are causally related to life-style – which is also a covariate of chronological age and blood pressure (e.g. we become inactive and over-weight). It is also possible that our experimental design comprising of skin RNA does not best capture the effect of vascular age and blood RNA or vascular tissue might encompass more information. Another caveat of our approach could be that we built a linear model to explore the transcriptome search space purpose however it is possible that a non-linear model might be more suitable, nonetheless these are unlikely to surpass BP and chronological age but may add further information.

6.1 Overview of the chapter

This chapter aims to give a summary of the overall conclusions of this research in relation to the objectives set at the beginning of this thesis and how it contributes to existing knowledge. I further discuss potential future research directions related to this project with recommendations and implications. We had five key observations in our analysis which were as following:

1. We found a signature or gene-set of 150 genes >90% accurate in classifying 'healthy' old muscle tissue samples from young muscle samples.

2. Using hundreds of new muscle gene-chip profiles from independent human cohorts we found the same 150 signature was ~93% accurate in muscle. Further, this same signature could distinguish old from young human brain and skin tissue

3. We produced new muscle gene-chip profiles from a 20yr longitudinal tissue cohort (ULSAM) and we gained access two independently produced case-control gene-chip datasets for AD. We found that a greater gene score in a person with given birth year (e.g. 70yrs) correlated with better long term renal function, mortality or better cognitive status (using the same 150 genes).

4. In all three independent clinical cohorts there was a consistent directional pattern of gene expression in muscle and blood, associated with good health.

5. The existing 'stress' resistant and 'epidemiologically' selected linear models when transformed to equivalent RNA signatures related with vascular disease whereas our 150 healthy age signature worked specifically for neuromuscular ageing. Thus, we have revealed that vascular ageing had a distinct profile from neuro-muscular ageing (Zahn et al. 2007).

6.2 Discussion

As the number of people routinely living into their eighth decade and beyond rises, the prevalence of age-related diseases has significantly increased and at differing rates across the various countries. An example of age-related disease (i.e. very low prevalence in young and middle-aged adults) include skeletal muscle atrophy and dysfunction ('sarcopenia') and neurological disorders such as dementia. These age-related health problems have massive economic and social consequences (Janssen et al. 2004; Gustavsson et al. 2011). To maintain long term effective performance in any job role attainment of healthy ageing would be ideal. Furthermore, age is a routine parameter in most clinical decision making trees e.g. decision to screen or not for age related disease. Identifying the molecular processes governing healthy human ageing (and longevity) is of great medical importance, but there have been few human based discoveries, mainly due to the inability to effectively account for influential physiological and environmental factors and lack of large well-

funded multigenerational studies.

In the recent years, there has been a surge in the use of machine learning methods which has facilitated researchers to develop classifiers for identification or diagnosis of diseases. These computational methods accompanied by a good study design promises to aid clinicians in identifying patients at high-risk for poor outcomes, and in general improve patients' health while minimizing costs and improving overall patient management. Cancer diagnosis and treatment have been influenced by these machine learning approaches (Tokuda et al. 2009; Patnaik et al. 2010), and this arguably represents where the greatest progress has been made in terms of personalized medicine.

There is a generous amount of proof that gene expression changes with ageing in different tissue types and in the organism as a whole (Zahn et al. 2007; Glass et al. 2013). Global RNA (W M Passtoors et al. 2012; Phillips et al. 2013; Glass et al. 2013; Gheorghe et al. 2014) and DNA methylation profiling (B. C. Christensen et al. 2009; Horvath 2013; Bell et al. 2012) has been utilised to search for consistent molecular events correlating with age, where samples come from cross-sectional studies spanning 5-8 decades. Such correlation analyses yield highly significant linear associations, yet by design, such models must be influenced by disease as much as the ageing process *per se*. Further, each study identified a distinct list of genes or pathways. For example, Hannum *et al* built a multi-tissue linear model of DNA methylation age-related changes that correlated with chronological age over seven decades (Hannum et al. 2013). However, this type molecular profile is not, for example, very useful for distinguishing how successful a person was ageing among a group with the same birth-year (Horvath 2013; Hannum et al. 2013) as chronological age and methylation status tends to co-vary tightly and in epidemiological cohorts, these DNAm models only added ~6% to models examining rates of mortality (Marioni et al. 2015).

Further, studies exploring the genetics of human ageing most commonly consider exceptional longevity (e.g. 100 years or more) as phenotype of interest for human ageing. While longevity is driven by a strong genetic contribution (Sebastiani et al. 2012) being fit and healthy at age 65 year is a more common occurrence and likely to reflect complex molecular factors (Kenyon 2010; Sabia et al. 2012). Discovery of these molecular factors could help screen for drugs that help people age ‘better’. In the present body of work, a novel tool has been provided that should enable the future translation of basic science into clinical advances, namely a robust diagnostic of healthy neuromuscular ageing. For our work, human skeletal muscle provided the ideal starting tissue from which to generate a 'clean' ageing molecular classifier, as skeletal muscle RNA is easily accessible (with a relatively uniform cell content e.g. >90% myocytes) and its functional status can be studied

in great detail prior to tissue sampling in all age groups (Gallagher et al. 2010; Timmons et al. 2005). This lies in very distinct contrast to using post-mortem brain samples, hard to access myocardium tissue or any one of a number of other potential human tissue sources. In addition, it was possible to discover this robust set of marker genes for healthy physiological age as the research strategy involved tissue samples obtained from 65 year subjects who had demonstrated *successful* ageing i.e. they were selected to have good metabolic and cardiovascular health despite having behaviour that was sedentary (Gallagher et al. 2010; Keller et al. 2011). At this stage we do not know if other aspects of their life-style was unique (they were non-smokers), for example diet but the impact on selected nutrients on human disease and ageing has not proven to yield a plausible biological affect (Timmerman 2013). So the 'healthy ageing' component constitutes a novel aspect of our study that has not been used in previous studies.

The usefulness of supervised machine learning approach in developing clinically useful biomarkers and diagnostic is often limited by access to multiple data-sets as the methods are generally prone to over-fitting (single data set bias) which do not appear evident straightaway to many researchers (Ambroise & McLachlan 2002; Simon et al. 2003). Without independent validation, these computational methods give spurious associations by developing classifiers/predictors that perform impressively well on the original training study but then fail miserably when applied to new dataset.

In our work the discovered gene-set signature that was then extensively validated by using samples from different cohorts, generated in different laboratories and profiled on different gene detection technologies. In the absence of independent validation datasets validation methods such as bootstrapping and cross validation which combines training and validation of classifiers in one process are often used (Kohavi 1995; Steyerberg et al. 2003). Unfortunately, a large number of studies employing such computationally intensive approaches are not descriptive enough about their validation strategy which makes it difficult to assess the validity of their results. Thus, validating a diagnostic on subjects independent of the training set is of prime importance. However, researchers are often limited by the availability of publically accessible data sets that fits their study design and the relevant biological question.

Hence, it is useful to have the raw data from developmental prediction/diagnostic studies in public domain as well as the computational methods employed (including code). This would not only ensure the reproducibility of the original studies but would also potentially provide a bigger pool of datasets to scientists for external validation and meta-analysis (Simon 2005; Kattan 2004). All of the datasets used in this study were made freely accessible at GEO. Our signature was

consistently modulated in several tissue types (muscle, brain and skin), but to very differing degrees in people of the same chronological age (Figure 3.2A), clearly illustrating that biological age is indeed different from chronological age. By this it fulfilled the first main criteria for being a novel diagnostic of healthy (or biological) ageing. Hence had more likelihood for predicting of an individual having an age-related clinical adverse event or developing an ageing-related disease such as Alzheimer's, CAD etc.

Neurocognitive pathology (e.g. Alzheimer's disease) becomes more pronounced with age and is often apparent in individuals who are otherwise healthy. On examining the 'healthy ageing' signature in relation to identifying neurocognitive disease we found that it could distinguish AD samples from age-matched controls. Further the healthy ageing signature was regulated in a distinct manner across individual healthy brain regions with chronological age, especially in the hippocampus (Figure 3.4), a region associated with neurogenesis (Gould et al. 1999; Taupin 2006). Our analysis of the relationship between lifestyle factors and the 'healthy age gene score' (longitudinal study ULSAM) suggested that the gene score was robust to confounding effects of these factors. The lack of association with lifestyle modulated diseases such as diabetes, CAD further ascertained this.

Therefore, our 'healthy ageing' signature appears '*selectively'* useful in relation to identifying risk for neurocognitive disease over and above lifestyle or vascular diseases. This is not surprising since ageing is thought to be a continuous physiological process that could be expected to have a gene expression signature distinct from lifestyle related (e.g. Type II diabetes) or mutation driven (e.g. cancer) pathologies. Further, as discussed before, ageing is a multifaceted process that has different levels of complexity and variability across cells, organs, organ systems, organism and species (Cevenini et al. 2008; Cevenini et al. 2010) nevertheless some aspects of ageing do appear consistent across tissue types based on our analysis.

In mouse, based on the pattern of age-related transcriptional changes researchers have categorised tissues into three different ageing processes, that is a pattern common to neural tissues, a pattern for vascular tissues, and a pattern for glandular tissues (Zahn et al. 2007). Mouse studies are challenging to interpret because of the use of inbred strains and very controlled environments – a situation very different from humans. Another study replicated these findings and through tissue co-expression network analysis, claimed that the distinct gene expression changes with age are potentially synchronized at different levels with an individual (Fu et al. 2006; Huang et al. 2011). A similar study in humans observed tissues like heart, lung, and whole blood sharing a stronger co-ageing pattern in comparison to tissues like muscle. This inter tissue synchronization could be a

reflection of their functional connectivity in the early developmental stage which probably extends into the late stage of their lifespan (Yang et al. 2015). In contrast, you might expect that muscle use and cardiac use would share a strong link to the same environmental factor like exercise while lung tissue remodels in the face of pollution and factors like smoking. Thus, neuromuscular ageing may have a distinct gene expression profile than vascular or lung ageing, with latter being more susceptible to lifestyle or environmental related perturbations.

The endeavor to discover biomarkers for the ageing process has prompted the development of large ageing cohort studies and different ageing signatures. However, if the research question is ageing par se it is critical to have thoughtful use of study design, to avoid confounding the studies with subjects with age associated diseases and drugs. In chapter-4 we have effectively shown that our muscle derived gene-set was the only one related to hippocampus ageing and cognitive health while 'stress' resistant and 'epidemiologically' selected linear models related with vascular diseases. By transforming the different genomic signatures of ageing to representative RNA signatures we have shown that it is possible to utilize a single technology platform (Transcriptomic/RNA profiling) to capture sufficient clinical variance of different aspects of ageing such as neuromuscular, vascular etc. (Figure 4.2). We have also successfully exhibited that a hypothesis driven machine learning method is different and reliable then the random sampling approach which fails to produce a single ageing 'gene-set' which works as a multi-tissue age-classifier or a discriminator for Alzheimer's disease. This provides the first RNA risk-factor for Alzheimer's disease that was not built/selected using Alzheimer's disease clinical samples and thus is the most independent. However, because Alzheimer's disease has low prevalence until the 80th decade and is a complex clinical diagnosis, an RNA *diagnostic* is going to be very challenging to develop.

Vascular ageing is related to ageing of vessels or arteries that helps in circulation of blood. Age-related pathologies generally effect the large elastic arteries which are rich in elastin and collagen with latter providing the strength to vasculature at higher blood pressures (Jani & Rajkumar 2006). Loss of elasticity damages coronary flow and results in coronary artery disease or atherosclerosis (McCullagh & Balian 1975). In chapter-5, we investigated if using gene expression data from skin tissue, as skin structure is close to one of the large arteries (Nilsson et al. 2015), along with PWV measures to gauge vasculature health, we can build a model for vascular ageing. Since PWV is known to strongly co-vary with both blood pressure and chronological age (Ruitenbeek et al. 2008; Vermeersch et al. 2008; Elias et al. 2009) we decided to use a machine learning linear modeling/regression strategy instead of the binary approach used for our neuromuscular signature as in the former we could adjust for the two covariates i.e. blood pressure

and age. We built three distinct regression models based on different criteria from our feature selection pipeline. We observed that gene expression made only a very modest contribution to the model when compared to blood pressure and chronological age. It is crucial to understand that this particular analysis was based on the hypothesis that skin mRNA could capture the effect of vascular ageing which could have been one of the limitations of our approach and possibly blood mRNA encompasses more information about vascular health and age.

6.3 Conclusion

In conclusion, we have discovered a novel and statistically robust multi-tissue RNA signature of 'biological ageing' that has potential as a health diagnostic. In particular, it was prognostic for long-term health in older humans, remotely informing about organ function (including cognitive functioning) using only a peripheral blood sample. Thus, we believe that our diagnostic represents a reliable proxy of 'biological' age that could be used in clinical decision-making, currently reliant on calendar age. Notably, ours is the first genomic signature able to identify AD from controls based entirely on an independently developed research hypothesis that does not include feature selection using disease cohorts. We believe that when combined with clinical data, our healthy age diagnostic could aid identification of at-risk late middle-aged non-symptomatic people.

6.4 Future directions

Novel easy to administer diagnostics that accurately and sensitively predict future health risk or help guide preventative measures would enable the evaluation of tailored treatment strategies for the individual. The biomarkers discovered in this thesis provides a novel way to assess whether an individual has a higher or lower probability, or risk, of developing an ageing-related disease, depending on the expression levels of these marker genes. It is advantageous to be able to assess an individual's biological age accurately, so that if an individual is identified to have a high risk of developing an ageing-related disease they can act accordingly to reduce their risk, such as through lifestyle changes or prophylactic treatment. The link between induction of the signature, renal decline, mortality and cognitive function suggests our signature transcends tissue specificity and also that it may be possible to facilitate healthier ageing e.g. to evaluate anti-ageing treatments using cell-based screening or to predict long-term safety in drug development. The signature could potentially be also used in predicting the quality of an organ based on the biological age and thus estimating the likelihood from a person over > 50 years of age being successfully used for transplantation into a donor patient by estimating the biological age of the organ.

We also believe that it will be informative to replace age with our healthy ageing gene diagnostic for many conditions. For example: In diabetes patients, where age is by far the more

powerful predictor of future dementia rather than severity of the diabetes measured using glycosylated hemoglobin A1 (HbA1) (Exalto et al. 2013) and in these cases replacing chronological age by biological age would potentially provide a better prognosis. This highlights that, clinically, various decision trees exist and our healthy ageing score could be integrated to help decide which middle-aged subjects could be offered entry into a preventative clinical trial many years before the clinical expression of AD. However, like many genomic diagnostics, the full clinical utility of ours will only emerge when combined with additional data and clinical insight.

References

1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–1073.

Abd El-Rehim, D.M. et al., 2005. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *International journal of cancer*, 116(3), pp.340–350.

Abe, O. et al., 2008. Aging in the CNS: comparison of gray/white matter volume and diffusion tensor data. *Neurobiology of aging*, 29(1), pp.102–116.

Aerts, S. et al., 2004. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20(12), pp.1974–1976.

Aguilera, O. et al., 2010. Epigenetics and environment: a complex relationship. *Journal of applied physiology*, 109(1), pp.243–251.

Ahlskog, J.E. et al., 2011. Physical exercise as a preventive or disease-modifying treatment of dementia and brain aging. In *Mayo Clinic Proceedings*. pp. 876–884.

Ambroise, C. & McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10), pp.6562–6566.

Ambrose, J.A. & Barua, R.S., 2004. The pathophysiology of cigarette smoking and cardiovascular disease: an update. *Journal of the American College of Cardiology*, 43(10), pp.1731–1737.

Anderson, C.D. et al., 2010. Chromosome 9p21 in ischemic stroke population structure and meta-analysis. *Stroke*, 41(6), pp.1123–1131.

Anderson, K.M. et al., 1991. Cardiovascular disease risk profiles. *American heart journal*, 121(1), pp.293–298.

Andrew, T. et al., 2001. Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin research*, 4(6), pp.464–477.

Audenet, F. et al., 2014. Germline genetic variations at 11q13 and 12p11 locus modulate age at onset for renal cell carcinoma. *The Journal of urology*, 191(2), pp.487–92.

Baker, G.T. & Sprott, R.L., 1988. Biomarkers of aging. *Experimental gerontology*, 23(4), pp.223–239.

Balch, W.E. et al., 2008. Adapting proteostasis for disease intervention. *science*, 319(5865), pp.916–919.

Baldi, P. & Brunak, S., 2001. *Bioinformatics: the machine learning approach*, MIT press.

Baroja-Mazo, A. et al., 2014. The NLRP3 inflammasome is released as a particulate danger signal that amplifies the inflammatory response. *Nature immunology*.

Barzilai, N. et al., 2012. The critical role of metabolic pathways in aging. *Diabetes*, 61(6), pp.1315–1322.

Bast Jr, R.C. et al., 1997. CA 125: the past and the future. *The International journal of biological markers*, 13(4), pp.179–187.

Beekman, M. et al., 2010. Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(42), pp.18046–9.

Beekman, M. et al., 2013. Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging cell*, 12(2), pp.184–193.

Bell, J.T. et al., 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS genetics*, 8(4), p.e1002629.

Ben-Shaul, Y., Bergman, H. & Soreq, H., 2005. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21(7), pp.1129–1137.

Benetos, A. et al., 2002. Influence of age, risk factors, and cardiovascular and renal disease on arterial stiffness: clinical applications. *American journal of hypertension*, 15(12), pp.1101–1108.

Benjamin, E.J. et al., 1994. Independent risk factors for atrial fibrillation in a population-based cohort: the Framingham Heart Study. *Jama*, 271(11), pp.840–844.

Berchtold, N.C. et al., 2008. Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proc Natl Acad Sci U S A*, 105(40), pp.15605–15610.

Bhopal, R. et al., 2005. Predicted and observed cardiovascular disease in South Asians: application of FINRISK, Framingham and SCORE models to Newcastle Heart Project data. *Journal of Public Health*, 27(1), pp.93–100.

Biasutti, M. et al., 2012. Cost-effectiveness of magnetic resonance imaging with a new contrast agent for the early diagnosis of Alzheimer's disease. *PloS one*, 7(4), p.e35559.

Blackburn, E.H., Greider, C.W. & Szostak, J.W., 2006. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nature medicine*, 12(10), pp.1133–1138.

Blair, S.N. et al., 1989. Physical fitness and all-cause mortality. A prospective study of healthy men and women. *Jama*, 262(17), pp.2395–2401.

Blasco, M.A., 2007. Telomere length, stem cells and aging. *Nature chemical biology*, 3(10), pp.640–649.

Boots, A.M.H. et al., 2013. The influence of ageing on the development and management of rheumatoid arthritis. *Nature Reviews Rheumatology*, 9(10), pp.604–613.

Boulesteix, A.-L. et al., 2008. Evaluating microarray-based classifiers: an overview. *Cancer Informatics*, 6, p.77.

Boyette, L.B. & Tuan, R.S., 2014. Adult stem cells and diseases of aging. *Journal of clinical medicine*, 3(1), pp.88–134.

Bratic, A., Larsson, N.-G. & others, 2013. The role of mitochondria in aging. *The Journal of clinical investigation*, 123(3), pp.951–957.

Brock, D.J.H. & Sutcliffe, R.G., 1972. Alpha-fetoprotein in the antenatal diagnosis of anencephaly and spina bifida. *The Lancet*, 300(7770), pp.197–199.

Budovskaya, Y. V et al., 2008. An elt-3/elt-5/elt-6 GATA Transcription Circuit Guides Aging in C.

elegans. *Cell*, 134(2), pp.291–303.

Candore, G. et al., 2006. Immunogenetics, Gender, and Longevity. *Annals of the New York Academy of Sciences*, 1089(1), pp.516–537.

Carter, R.J., Dubchak, I. & Holbrook, S.R., 2001. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research*, 29(19), pp.3928–3938.

Cecelja, M. et al., 2009. Increased wave reflection rather than central arterial stiffness is the main determinant of raised pulse pressure in women and relates to mismatch in arterial dimensions: a twin study. *Journal of the American College of Cardiology*, 54(8), pp.695–703.

Cech, T.R., 2004. Beginning to understand the end of the chromosome. *Cell*, 116(2), pp.273–279.

Cesana, M. et al., 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2), pp.358–69.

Cesari, M. et al., 2006. Frailty syndrome and skeletal muscle: results from the Invecchiare in Chianti study. *The American journal of clinical nutrition*, 83(5), pp.1142–1148.

Cevenini, E. et al., 2008. Human models of aging and longevity. *Expert opinion on biological therapy*, 8(9), pp.1393–1405.

Cevenini, E. et al., 2010. Systems biology and longevity: an emerging approach to identify innovative anti-aging targets and strategies. *Current pharmaceutical design*, 16(7), pp.802–813.

Chahal, H.S. & Drake, W.M., 2007. The endocrine system and ageing. *The Journal of pathology*, 211(2), pp.173–180.

Chandramouli, K. & Qian, P.-Y., 2009. Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity. *Human Genomics and Proteomics : HGP*, 2009, p.239204.

Chen, J.J. et al., 2012. A 71-gene signature of TRAIL sensitivity in cancer cells. *Molecular cancer therapeutics*, 11(1), pp.34–44.

Christensen, B.C. et al., 2009. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS genetics*, 5(8), p.e1000602.

Christensen, K. et al., 2009. Ageing populations : the challenges ahead. *Lancet*, 374(9696), pp.1196–1208.

Church, T.S. et al., 2005. Cardiorespiratory fitness and body mass index as predictors of cardiovascular disease mortality among men with diabetes. *Arch Intern Med*, 165(18), pp.2114–2120.

Colantuoni, C. et al., 2000. High throughput analysis of gene expression in the human brain. *Journal of Neuroscience Research*, 59, pp.1–10.

Colantuoni, C. et al., 2011. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, 478(7370), pp.519–23.

Coll, E. et al., 2000. Serum cystatin C as a new marker for noninvasive estimation of glomerular filtration rate and as a marker for early renal impairment. *American journal of Kidney diseases*, 36(1), pp.29–34.

Collado, M., Blasco, M.A. & Serrano, M., 2007. Cellular senescence in cancer and aging. *Cell*, 130(2), pp.223–233.

Coutlee, C.G. & Huettel, S.A., 2012. The functional neuroanatomy of decision making: prefrontal control of thought and action. *Brain research*, 1428, pp.3–12.

Crum, R.M. et al., 1993. Population-based norms for the Mini-Mental State Examination by age and educational level. *Jama*, 269(18), pp.2386–2391.

Cunningham, F. et al., 2015. Ensembl 2015. *Nucleic Acids Research*, 43(D1), pp.D662–D669.

Dai, M. et al., 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, 33(20), p.e175.

Danielsson, M. & Talbäck, M., 2012. Public health: an overview: Health in Sweden: The National Public Health Report. *Scandinavian journal of public health*, 40, pp.6–22.

Daskalaki, S., Kopanas, I. & Avouris, N., 2006. Evaluation of classifiers for an uneven class distribution problem. *Applied artificial intelligence*, 20(5), pp.381–417.

Deelen, J. et al., 2011. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging cell*, 10(4), pp.686–98.

Deelen, J. et al., 2013. Identifying the genomic determinants of aging and longevity in human population studies: Progress and challenges. *BioEssays*, 35(4), pp.386–396.

Dementia UK, 2014. Dementia 2014 infographic.

Diaz-Uriarte, R. & De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), p.1.

Ding, C. & Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2), pp.185–205.

Dixon, A.L. et al., 2007. A genome-wide association study of global gene expression. *Nature genetics*, 39(10), pp.1202–1207.

Dobriansky, P.J., Suzman, R.M. & Hodes, R.J., 2007. Why Population Aging Matters - A Global Perspective. In *US Department of State*. pp. 1–32.

Donoho, D.L. & others, 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1, p.32.

Donohue, M.C. et al., 2014. The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA neurology*, 71(8), pp.961–70.

Downing, J.E.G. & Miyan, J.A., 2000. Neural immunoregulation: emerging roles for nerves in immune homeostasis and disease. *Immunology today*, 21(6), pp.281–289.

Du, P., Kibbe, W.A. & Lin, S.M., 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, 24(13), pp.1547–1548.

Dunder, K. et al., 2004. Evaluation of a scoring scheme, including proinsulin and the apolipoprotein B/apolipoprotein A1 ratio, for the risk of acute coronary events in middle-aged men: Uppsala Longitudinal Study of Adult Men (ULSAM). *American heart journal*, 148(4), pp.596–601.

Durier, S. et al., 2003. Physiological genomics of human arteries quantitative relationship between

gene expression and arterial stiffness. *Circulation*, 108(15), pp.1845–1851.

Elias, M.F. et al., 2009. Arterial pulse wave velocity and cognition with advancing age. *Hypertension*, 53(4), pp.668–673.

Erickson, K.I. et al., 2011. Exercise training increases size of hippocampus and improves memory. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7), pp.3017–22.

Erraji-Benchekroun, L. et al., 2005. Molecular aging in human prefrontal cortex is selective and continuous throughout adult life. *Biological psychiatry*, 57(5), pp.549–58.

Erusalimsky, J.D. & Kurz, D.J., 2005. Cellular senescence in vivo: Its relevance in ageing and cardiovascular disease. *Experimental Gerontology*, 40(8–9), pp.634–642.

Exalto, L.G. et al., 2013. Risk score for prediction of 10 year dementia risk in individuals with type 2 diabetes: a cohort study. *The lancet. Diabetes & endocrinology*, 1(3), pp.183–90.

Feero, W.G. et al., 2011. Genomics of cardiovascular disease. *New England Journal of Medicine*, 365(22), pp.2098–2109.

Fehlbaum-Beurdeley, P. et al., 2012. Validation of AclarusDx™, a blood-based transcriptomic signature for the diagnosis of Alzheimer's disease. *Journal of Alzheimer's disease : JAD*, 32(1), pp.169–81.

Feitosa, M.F. et al., 2014. Genetic analysis of long-lived families reveals novel variants influencing high density-lipoprotein cholesterol. *Frontiers in genetics*, 5, p.159.

Ferguson, F.G. et al., 1995. Immune parameters in a longitudinal study of a very old population of Swedish people: a comparison between survivors and nonsurvivors. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 50(6), pp.B378--B382.

Folstein, M.F., Robins, L.N. & Helzer, J.E., 1983. The mini-mental state examination. *Archives of general psychiatry*, 40(7), p.812.

Fontana, L., Partridge, L. & Longo, V.D., 2010. Extending healthy life span--from yeast to humans. *science*, 328(5976), pp.321–326.

Fraga, M.F. et al., 2005. Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), pp.10604–10609.

Fraga, M.F., 2009. Genetic and epigenetic regulation of aging. *Current opinion in immunology*, 21(4), pp.446–453.

Franceschi, C. et al., 2007. Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mechanisms of ageing and development*, 128(1), pp.92–105.

Franceschi, C. & Campisi, J., 2014. Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 69(Suppl 1), pp.S4--S9.

Fu, C. et al., 2006. Tissue specific and non-specific changes in gene expression by aging and by early stage CR. *Mechanisms of ageing and development*, 127(12), pp.905–916.

Fyhrquist, F., Saijonmaa, O. & Strandberg, T., 2013. The roles of senescence and telomere

shortening in cardiovascular disease. *Nat.Rev.Cardiol.*, 10(5), pp.274–283.

Gallagher, I.J. et al., 2010. Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin resistance in type 2 diabetes. *Genome medicine*, 2(2), p.9.

Garasto, S. et al., 2003. The study of APOA1, APOC3 and APOA4 variability in healthy ageing people reveals another paradox in the oldest old subjects. *Annals of human genetics*, 67(Pt 1), pp.54–62.

Gautier, L. et al., 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)*, 20(3), pp.307–15.

Gavrilov, L.A. & Heuveline, P., 2003. Aging of population. *The encyclopedia of population*, 1, pp.32–37.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10), p.R80.

Gheorghe, M. et al., 2014. Major aging-associated RNA expressions change at two distinct age-positions. *BMC Genomics*, 15(1), pp.1–12.

Gierman, H.J. et al., 2014. Whole-Genome Sequencing of the World's Oldest People P. Lewis, ed. *PLoS ONE*, 9(11), p.e112430.

Ginos, M.A. et al., 2004. Identification of a gene expression signature associated with recurrent disease in squamous cell carcinoma of the head and neck. *Cancer research*, 64(1), pp.55–63.

Glass, D. et al., 2013. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome biology*, 14(7), p.R75.

Glisky, E.L., 2007. Changes in cognitive function in human aging. *Brain aging: models, methods, and mechanisms*, pp.3–20.

Goeman, J.J. & Bühlmann, P., 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)*, 23(8), pp.980–7.

Goldberger, J.J. & Buxton, A.E., 2013. Personalized medicine vs guideline-based medicine. *Jama*, 309(24), pp.2559–60.

Goodale, M.A. & Milner, A.D., 1992. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1), pp.20–25.

Gould, E. et al., 1999. Neurogenesis in the neocortex of adult primates. *Science*, 286(5439), pp.548–552.

Greco, D. et al., 2008. Gene expression in human NAFLD. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 294(5), pp.G1281--G1287.

Greer, E.L. et al., 2011. Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. *Nature*, 479(7373), pp.365–371.

De Groot, J.C. et al., 2000. Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. *Annals of neurology*, 47(2), pp.145–151.

Gustavsson, A. et al., 2011. Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(10), pp.718–779.

Guyon, I. et al., 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1–3), pp.389–422.

Hampel, H. et al., 2014. Perspective on future role of biological markers in clinical therapy trials of Alzheimer's disease: a long-range point of view beyond 2020. *Biochemical pharmacology*, 88(4), pp.426–49.

Hanninen, T., Koivisto, K. & Reinikainen, K.J., 1996. Prevalence of ageing-associated cognitive decline in an elderly population. *Age and ageing*, 25(3), pp.201–205.

Hannum, G. et al., 2013. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2), pp.359–67.

Hansen, T.W. et al., 2006. Prognostic value of aortic pulse wave velocity as index of arterial stiffness in the general population. *Circulation*, 113(5), pp.664–670.

Hardy, J., 1997. Amyloid, the presenilins and Alzheimer's disease. *Trends in neurosciences*, 20(4), pp.154–159.

Harman, D., 1955. Aging: a theory based on free radical and radiation chemistry. , pp.298–300.

Harper, S., 2014. Economic and social implications of aging societies. *Science*, 346(6209), pp.587–591.

Harrington, C.A., Rosenow, C. & Retief, J., 2000. Monitoring gene expression using DNA microarrays. *Current opinion in Microbiology*, 3(3), pp.285–291.

Hayflick, L. & Moorhead, P.S., 1961. The serial cultivation of human diploid cell strains. *Experimental cell research*, 25(3), pp.585–621.

Hegde, P.S., White, I.R. & Debouck, C., 2003. Interplay of transcriptomics and proteomics. *Current opinion in biotechnology*, 14(6), pp.647–651.

Heidecker, B. et al., 2008. Transcriptomic biomarkers for individual risk assessment in new-onset heart failure. *Circulation*, 118(3), pp.238–46.

Herbig, U. et al., 2006. Cellular Senescence in Aging Primates. , p.16.

Hernandez, D.G. et al., 2011. Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Human molecular genetics*, 20(6), pp.1164–1172.

Hibbs, K. et al., 2004. Differential gene expression in ovarian carcinoma: identification of potential biomarkers. *The American journal of pathology*, 165(2), pp.397–414.

Hickson, M., 2006. Malnutrition and ageing. *Postgraduate Medical Journal*, 82(963), pp.2–8.

Holliday, R., 1987. The inheritance of epigenetic defects. *Science*, 238(4824), pp.163–170.

Horvath, S. et al., 2012. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome biology*, 13(10), p.R97.

Horvath, S., 2013. DNA methylation age of human tissues and cell types DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10), p.R115.

Horvath, S. et al., 2015. The cerebellum ages slowly according to the epigenetic clock. *Aging (Albany NY)*, 7(5), pp.1–13.

Hu, W.T. et al., 2012. Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease. *Neurology*, 79(9), pp.897–905.

Huang, T. et al., 2011. Crosstissue coexpression network of aging. *Omics: a journal of integrative biology*, 15(10), pp.665–671.

Huang, X. et al., 2013. Serum fatty acid patterns, insulin sensitivity and the metabolic syndrome in individuals with chronic kidney disease. *Journal of internal medicine*, 275(1), pp.71–83.

Hye, A. et al., 2014. Plasma proteins predict conversion to dementia from prodromal disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 44, pp.1–9.

International Conference of Social Security Actuaries and Statisticians, 2009. Improvements in life expectancy and sustainability of social security schemes. In pp. 16–18.

Irizarry, R.A. et al., 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2), pp.249–264.

Jack, C.R. et al., 1999. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7), p.1397.

Jain, K.K. & Jain, K.K., 2010. *The handbook of biomarkers*, Springer.

Jani, B. & Rajkumar, C., 2006. Ageing and vascular ageing. *Postgraduate medical journal*, 82(968), pp.357–362.

Janssen, I., 2006. Influence of sarcopenia on the development of physical disability: the Cardiovascular Health Study. *Journal of the American Geriatrics Society*, 54(1), pp.56–62.

Janssen, I. et al., 2004. The healthcare costs of sarcopenia in the United States. *Journal of the American Geriatrics Society*, 52(1), pp.80–85.

Jessen, F. et al., 2011. Prediction of dementia in primary care patients. *PloS one*, 6(2), p.e16852.

Jin, G. et al., 2012. Validation of prostate cancer risk-related loci identified from genome-wide association studies using family-based association analysis: evidence from the International Consortium for Prostate Cancer Genetics (ICPCG). *Human genetics*, 131(7), pp.1095–103.

Johnson, W.E., Li, C. & Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), pp.118–127.

Judge, S. et al., 2005. Age-associated increases in oxidative stress and antioxidant enzyme activities in cardiac interfibrillar mitochondria: implications for the mitochondrial theory of aging. *The FASEB journal*, 19(3), pp.419–421.

Kaeberlein, M., Rabinovitch, P.S. & Martin, G.M., 2015. Healthy aging: the ultimate preventative medicine. *Science*, 350(6265), pp.1191–1193.

Kattan, M.W., 2004. Evaluating a New Marker's Predictive Contribution. *Clinical Cancer Research*, 10(3), pp.822–824.

Keller, P. et al., 2011. A transcriptional map of the impact of endurance exercise training on skeletal muscle phenotype. *J Appl Physiol*, 110(1), pp.46–59.

Kelly, R. et al., 1989. Noninvasive determination of age-related changes in the human arterial pulse. *Circulation*, 80(6), pp.1652–1659.

Kemper, T.L., 1994. Neuroanatomical and neuropathological changes during aging and dementia.

Kensinger, E.A. & Corkin, S., 2009. Cognition in aging and age related disease. *Handbook of the neuroscience of aging*, pp.249–256.

Kenyon, C.J., 2010. The genetics of ageing. *Nature*, 464(7288), pp.504–512.

Khaitovich, P. et al., 2004. Regional Patterns of Gene Expression in Human and Chimpanzee Brains. *Genome research*, 14(8), pp.1462–1473.

Kim, E.J. et al., 2007. Relationship between blood pressure parameters and pulse wave velocity in normotensive and hypertensive subjects: invasive study. *Journal of human hypertension*, 21(2), pp.141–148.

Kim, J. et al., 2002. Total-body skeletal muscle mass: estimation by a new dual-energy X-ray absorptiometry method. *The American journal of clinical nutrition*, 76(2), pp.378–383.

King, M.-C. & Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. *Essential Readings in Evolutionary Biology*, 188(4184), p.301.

Knudsen, S. et al., 2014. Development and validation of a gene expression score that predicts response to fulvestrant in breast cancer patients. *PloS one*, 9(2), p.e87415.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14(2), pp.1137–1145.

Kolarova, M. et al., 2012. Structure and pathology of tau protein in Alzheimer disease. *International journal of Alzheimer's disease*.

Kung, J.T.Y., Colognori, D. & Lee, J.T., 2013. Long Noncoding RNAs: Past, Present, and Future. *Genetics*, 193(3), pp.651–669.

De la Fuente, A., 2010. From 'differential expression' to 'differential networking'-identification of dysfunctional regulatory networks in diseases. *Trends in genetics*, 26(7), pp.326–333.

De La Grange, P. et al., 2010. Splicing factor and exon profiling across human tissues. *Nucleic acids research*, p.gkq008.

Laakso, M.P. et al., 2000. Hippocampus and entorhinal cortex in frontotemporal dementia and Alzheimer's disease: a morphometric MRI study. *Biological psychiatry*, 47(12), pp.1056–1063.

Van Der Laan, M.J. & Bryan, J., 2001. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2(4), pp.445–461.

Lacolley, P. et al., 2009. Genetics and pathophysiology of arterial stiffness. *Cardiovascular Research*, 81(4), pp.637–648.

Laird, P.W., 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3), pp.191–203.

Lajemi, M. et al., 2001. Angiotensin II type 1 receptor- 153A/G and 1166A/C gene polymorphisms and increase in aortic stiffness with age in hypertensive subjects. *Journal of hypertension*, 19(3), pp.407–413.

Lakatta, E.G., 2000. Cardiovascular aging in health. *Clinics in geriatric medicine*, 16(3), pp.419–443.

Lakatta, E.G. & Levy, D., 2003. Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises. *Circulation*, 107(1), pp.139–146.

Lange, E.M. et al., 2012. Early onset prostate cancer has a significant genetic component. *The Prostate*, 72(2), pp.147–56.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–359.

Larrick, J.W. & Mendelsohn, A., 2010. Applied Healthspan Engineering. *Rejuvenation research*, 13(2–3), pp.265–280.

Larrouy, D. et al., 2008. Gene expression profiling of human skeletal muscle in response to stabilized weight loss. *The American journal of clinical nutrition*, 88(1), pp.125–32.

Laske, C. et al., 2014. Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's & Dementia*, pp.1–18.

Laterza, O.F., Price, C.P. & Scott, M.G., 2002. Cystatin C: an improved estimator of glomerular filtration rate? *Clinical chemistry*, 48(5), pp.699–707.

Laurent, S., 2012. Defining vascular aging and cardiovascular risk. *Journal of hypertension*, 30, pp.S3--S8.

Lefebvre, R. & Goomar, A., 2005. Growing up: The Social and Economic Implications of an Aging Population.

Lei, R. et al., 2015. Diminishing returns in next-generation sequencing (NGS) transcriptome data. *Gene*, 557(1), pp.82–7.

De Lencastre, A. et al., 2010. MicroRNAs both promote and antagonize longevity in C. elegans. *Current Biology*, 20(24), pp.2159–2168.

Levine, M.E. & Crimmins, E.M., 2016. A genetic network associated with stress resistance, longevity, and cancer in humans. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 71(6), pp.703–712.

Lewohl, J.M. et al., 2000. Gene expression in human alcoholism: microarray analysis of frontal cortex. *Alcoholism: Clinical and Experimental Research*, 24(12), pp.1873–1882.

Libbrecht, M.W. & Noble, W.S., 2015. Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16(6), pp.321–332.

Liu, D. et al., 2013. Microarray Analysis Reveals Novel Features of the Muscle Aging Process in Men and Women. *The journals of gerontology. Series A, Biological sciences and medical sciences*, (14), pp.1–10.

López-Otín, C. et al., 2013. The hallmarks of aging. *Cell*, 153(6), pp.1194–217.

Lotz, M. et al., 2013. Value of biomarkers in osteoarthritis: current status and perspectives. *Annals of the rheumatic diseases*, 72(11), pp.1756–63.

Lovestone, S. et al., 2009. AddNeuroMed--the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences*, 1180, pp.36–46.

Lu, T. et al., 2004. Gene regulation and DNA damage in the ageing human brain. *Nature*,

429(6994), pp.883–891.

Lubbe, S.J. et al., 2012. Comprehensive evaluation of the impact of 14 genetic variants on colorectal cancer phenotype and risk. *American journal of epidemiology*, 175(1), pp.1–10.

Lunnon, K. et al., 2013. A blood gene expression marker of early Alzheimer's disease. *Journal Of Alzheimer's Disease*, 33(3), pp.737–753.

Lunnon, K. et al., 2012. Mitochondrial dysfunction and immune activation are detectable in early Alzheimer's disease blood. *Journal of Alzheimer's disease : JAD*, 30(3), pp.685–710.

Lutz, W., Sanderson, W. & Scherbov, S., 2008. The coming acceleration of global population ageing. *Nature*, 451(7179), pp.716–719.

De Magalhães, J.P., Curado, J. & Church, G.M., 2009. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7), pp.875–881.

Makinodan, T. et al., 1991. Cellular immunosenescence: an overview. *Experimental gerontology*, 26(2), pp.281–288.

Marioni, R.E. et al., 2015. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *International journal of epidemiology*, 44(4), pp.1388–1396.

McArdle, A., Vasilaki, A. & Jackson, M., 2002. Exercise and skeletal muscle ageing: cellular and molecular mechanisms. *Ageing research reviews*, 1(1), pp.79–93.

McCall, M.N., Bolstad, B.M. & Irizarry, R. a, 2010. Frozen robust multiarray analysis (fRMA). *Biostatistics (Oxford, England)*, 11(2), pp.242–53.

McCullagh, K.A. & Balian, G., 1975. Collagen characterisation and cell transformation in human atherosclerosis. *Nature*, 258(5530), pp.73–75.

Medley, T.L. et al., 2002. Fibrillin-1 genotype is associated with aortic stiffness and disease severity in patients with coronary artery disease. *Circulation*, 105(7), pp.810–815.

Meguro, K. et al., 2001. Cognitive function and frontal lobe atrophy in normal elderly adults: implications for dementia not as aging-related disorders and the reserve hypothesis. *Psychiatry and clinical neurosciences*, 55(6), pp.565–572.

Melov, S. et al., 2007. Resistance exercise reverses aging in human skeletal muscle. *PloS one*, 2(5), p.e465.

Menden, M.P. et al., 2013. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PloS one*, 8(4), p.e61318.

Mitschke, M.M. et al., 2013. Increased cGMP promotes healthy expansion and browning of white adipose tissue. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 27(4), pp.1621–30.

Montecino-Rodriguez, E., Berent-Maoz, B. & Dorshkind, K., 2013. Causes, consequences, and reversal of immune system aging. *The Journal of Clinical Investigation*, 123(3), pp.958–965.

Mormino, E.C. et al., 2014. Amyloid and APOE ε4 interact to influence short-term decline in preclinical Alzheimer disease. *Neurology*, 82(20), pp.1760–7.

Morrison, J.H. & Hof, P.R., 1997. Life and death of neurons in the aging brain. *Science*, 278(5337), pp.412–419.

Mutz, K.O. et al., 2013. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, 24(1), pp.22–30.

Myers, A.J. et al., 2007. A survey of genetic human cortical gene expression. *Nature genetics*, 39(12), pp.1494–1499.

Myers, J. et al., 2002. Exercise capacity and mortality among men referred for exercise testing. *N Engl J Med*, 346(11), pp.793–801.

Naj, A.C. et al., 2014. Effects of Multiple Genetic Loci on Age at Onset in Late-Onset Alzheimer Disease A Genome-Wide Association Study. *JAMA neurology*, 71(11), pp.1394–1404.

Najjar, S.S. et al., 2008. Pulse Wave Velocity Is an Independent Predictor of the Longitudinal Increase in Systolic Blood Pressure and of Incident Hypertension in the Baltimore Longitudinal Study of Aging. *Journal of the American College of Cardiology*, 51(14), pp.1377–1383.

Narita, M. et al., 2003. Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell*, 113(6), pp.703–716.

Nelson, M.R. et al., 2010. Noninvasive measurement of central vascular pressures with arterial tonometry: clinical revival of the pulse pressure waveform? In *Mayo Clinic Proceedings*. pp. 460–472.

Newberg, A.B. et al., 2005. Cerebral Blood Flow Effects of Pain and Acupuncture: A Preliminary Single-Photon Emission Computed Tomography Imaging Study. *Journal of Neuroimaging*, 15(1), pp.43–49.

Newgard, C.B., Sharpless, N.E. & others, 2013. Coming of age: molecular drivers of aging and therapeutic opportunities. *The Journal of clinical investigation*, 123(123 (3)), pp.946–950.

Niccoli, T. & Partridge, L., 2012. Ageing as a risk factor for disease. *Current biology*, 22(17), pp.R741–R752.

Nicholson, G. et al., 2011. Human metabolic profiles are stably controlled by genetic and environmental variation. *Molecular systems biology*, 7(525), p.525.

Nilsson, P.M., Boutouyrie, P. & Laurent, S., 2009. Vascular aging a tale of EVA and ADAM in cardiovascular risk assessment and prevention. *Hypertension*, 54(1), pp.3–10.

Nilsson, P.M., Olsen, M.H. & Laurent, S., 2015. *Early Vascular Aging (EVA): New Directions in Cardiovascular Protection*, Academic Press.

North, B.J. & Sinclair, D.A., 2012. The intersection between aging and cardiovascular disease. *Circulation Research*, 110(8), pp.1097–1108.

O'Bryant, S.E. et al., 2011. A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI. *PloS one*, 6(12), p.e28092.

Office for national statistics, 2012. Population Ageing in the United Kingdom, its Constituent Countries and the European Union.

Oldham, M.C. et al., 2008. Functional organization of the transcriptome in human brain. *Nature neuroscience*, 11(11), pp.1271–1282.

Pan, K.H., Lih, C.J. & Cohen, S.N., 2005. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National*

Academy of Sciences of the United States of America, 102(25), pp.8961–8965.

Park, C.B. & Larsson, N.-G., 2011. Mitochondrial DNA mutations in disease and aging. *The Journal of cell biology*, 193(5), pp.809–818.

Partridge, L., 2010. The new biology of ageing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1537), pp.147–154.

Passtoors, W.M. et al., 2012. Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PLoS One*, 7.

Passtoors, W.M. et al., 2012. Transcriptional profiling of human familial longevity indicates a role for ASF1A and IL7R. *PloS one*, 7(1), p.e27759.

Patnaik, S.K. et al., 2010. Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection. *Cancer Res*, 70(1), pp.36–45.

Pawelec, G., 2007. Immunosenescence comes of age. Symposium on Aging Research in Immunology: The Impact of Genomics. *EMBO Reports*, 8(3), pp.220–223.

Perez-Tur, J. et al., 1995. A mutation in Alzheimer's disease destroying a splice acceptor site in the presenilin-1 gene. *Neuroreport*, 7(1), pp.297–301.

Perry, E.K. et al., 1982. Neurochemical activities in human temporal lobe related to aging and Alzheimer-type changes. *Neurobiology of aging*, 2(4), pp.251–256.

Peters, M.J. et al., 2015. The transcriptional landscape of age in human peripheral blood. *Nat Commun*, 6.

Phillips, B.E. et al., 2013. Molecular Networks of Human Muscle Adaptation to Exercise and Age G. Gibson, ed. *PLoS Genetics*, 9(3), p.e1003389.

Piper, M.D.W. et al., 2008. Separating cause from effect: how does insulin/IGF signalling control lifespan in worms, flies and mice? *Journal of internal medicine*, 263(2), pp.179–191.

De Preter, K. et al., 2008. Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions. *Nucleic acids research*, 36(7), p.e43.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6), pp.841–2.

R Core Team, 2015. R: A Language and Environment for Statistical Computing.

Raichlen, D. a & Alexander, G.E., 2014. Exercise, APOE genotype, and the evolution of the human lifespan. *Trends in neurosciences*, 37(5), pp.247–55.

Rakyan, V.K. et al., 2010. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome research*, 20(4), pp.434–439.

Ramasamy, A. et al., 2014. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience*, 17(10), pp.1418–28.

Ramasamy, A. et al., 2013. Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies. *Nucleic acids research*, 41(7), p.e88.

Ramskold, D. et al., 2009. An abundance of ubiquitously expressed genes revealed by tissue

transcriptome sequence data. *PLoS Comput Biol*, 5(12), p.e1000598.

Rattan, S.I.S., 2006. Theories of biological aging: Genes, proteins, and free radicals. *Free Radical Research*, 40(12), pp.1230–1238.

Raue, U. et al., 2012. Transcriptome signature of resistance exercise adaptations: mixed muscle and fiber type specific profiles in young and old adults. *Journal of applied physiology (Bethesda, Md. : 1985)*, 112(10), pp.1625–36.

Ray, S. et al., 2007. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nature medicine*, 13(11), pp.1359–62.

Raz, N. et al., 2005. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex (New York, N.Y. : 1991)*, 15(11), pp.1676–89.

Rechel, B. et al., 2013. Ageing in the European Union. *The Lancet*, 381(9874), pp.1312–1322.

Reference Values for Arterial Stiffness' Collaboration, 2010. *Determinants of pulse wave velocity in healthy people and in the presence of cardiovascular risk factors:establishing normal and reference values*, Eur Soc Cardiology.

Reff, M.E. & Schneider, E.L., 1982. *Biological markers of aging*, Department of Health and Human Services.

Reuter-Lorenz, P.A., 2002. New visions of the aging mind and brain. *Trends in cognitive sciences*, 6(9), pp.394–400.

Ritchie, M.E. et al., 2011. BeadArray expression analysis using bioconductor. *PLoS Comput Biol*, 7(12), p.e1002276.

Rivas, D. a et al., 2014. Diminished skeletal muscle microRNA expression with aging is associated with attenuated muscle plasticity and inhibition of IGF-1 signaling. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, pp.1–15.

Rodier, F. & Campisi, J., 2011. Four faces of cellular senescence. *The Journal of cell biology*, 192(4), pp.547–556.

Rodwell, G.E.J. et al., 2004. A transcriptional profile of aging in the human kidney. *PLoS biology*, 2(12), p.e427.

Romijn, M.D.M. et al., 2014. Mild chronic kidney disease is associated with cognitive function in patients presenting at a memory clinic. *International journal of geriatric psychiatry*, 30(7), pp.758–765.

Roth, R.B. et al., 2006. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, 7(2), pp.67–80.

Roth, S.M. et al., 2002. Influence of age, sex, and strength training on human muscle gene expression determined by microarray. *Physiological genomics*, 10(3), pp.181–190.

Roubenoff, R. & Hughes, V.A., 2000. Sarcopenia current concepts. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 55(12), pp.M716--M724.

Rudman, D. et al., 1990. Effects of human growth hormone in men over 60 years old. *New England Journal of Medicine*, 323(1), pp.1–6.

Ruitenbeek, A.G. et al., 2008. Age and blood pressure levels modify the functional properties of

central but not peripheral arteries. *Angiology*.

Ruzankina, Y. & Brown, E.J., 2007. Relationships between stem cell exhaustion, tumour suppression and ageing. *British journal of cancer*, 97(9), pp.1189–1193.

Sabia, S. et al., 2012. Influence of individual and combined healthy behaviours on successful aging. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 184(18), pp.1985–92.

Saeys, Y., Inza, I. & Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), pp.2507–2517.

Saito-Hisaminato, A. et al., 2002. Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA research*, 9(2), pp.35–45.

Salloway, S. et al., 2014. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *The New England journal of medicine*, 370(4), pp.322–33.

Salvi, S.M., Akhtar, S. & Currie, Z., 2006. Ageing changes in the eye. *Postgraduate Medical Journal*, 82(971), pp.581–587.

Sandberg, R. & Larsson, O., 2007. Improved precision and accuracy for microarrays using updated probe set definitions. *BMC bioinformatics*, 8(1), p.1.

Sattlecker, M. et al., 2014. Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 10(6), pp.724–34.

Savage, S.A. et al., 2008. TINF2, a component of the shelterin telomere protection complex, is mutated in dyskeratosis congenita. *The American Journal of Human Genetics*, 82(2), pp.501–509.

Sawhney, V. et al., 2012. Current genomics in cardiovascular medicine. *Current genomics*, 13(6), pp.446–62.

Scahill, R.I. et al., 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Archives of neurology*, 60(7), pp.989–994.

Schnabel, R.B. et al., 2012. Next steps in cardiovascular disease genomic research-sequencing, epigenetics, and transcriptomics. *Clinical chemistry*, 58(1), pp.113–126.

Schunkert, H. et al., 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4), pp.333–338.

Schwanhäusser, B. et al., 2011. Global quantification of mammalian gene expression control. *Nature*, 473(7347), pp.337–342.

Sebastiani, P. et al., 2012. Genetic signatures of exceptional longevity in humans. *PloS one*, 7(1), p.e29848.

Segal, E. et al., 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34(2), pp.166–176.

Sekar, S. et al., 2014. Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiology of aging*, 36(2), pp.583–591.

Shao, L. et al., 2013. Determination of minimum training sample size for microarray-based cancer outcome prediction-an empirical assessment. *PloS one*, 8(7), p.e68579.

Shedden, K. et al., 2008. Gene expression – based survival prediction in lung adenocarcinoma : a multi-site , blinded validation study. *Nature medicine*, 14(8), pp.822–827.

Shehadeh, L.A. et al., 2010. SRRM2, a potential blood biomarker revealing high alternative splicing in Parkinson's disease. *PLoS One*, 5(2), p.e9104.

Shi, L. et al., 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature biotechnology*, 28(8), pp.827–38.

Simon, R., 2005. Development and Validation of Therapeutically Relevant Multi-Gene Biomarker Classifiers. *Journal of the National Cancer Institute*, 97(12), pp.866–867.

Simon, R. et al., 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1), pp.14–18.

Singh, K., 2004. Mitochondrial dysfunction is a common phenotype in aging and cancer. *Annals of the New York Academy of Sciences*, 1019(1), pp.260–264.

Sinnaeve, P.R. et al., 2009. Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS ONE*, 4(9).

Sinsheimer, J.S. et al., 2011. Epigenetic Predictor of Age. *PloS one*, 6(6), pp.1–6.

Slentz, C.A. et al., 2011. The Effects of Aerobic versus Resistance Training on Visceral and Liver Fat Stores, Liver Enzymes and HOMA from STRRIDE AT/RT: A Randomized Trial. *Am J Physiol Endocrinol Metab*, 301(5), pp.E1033-9.

Slonim, D.K., 2002. From patterns to pathways: gene expression data analysis comes of age. *Nature genetics*, 32, pp.502–508.

Smith, N.L. et al., 2010. Association of Genome-Wide Variation With the Risk of Incident Heart Failure in Adults of European and African Ancestry A Prospective Meta-Analysis From the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *Circulation: Cardiovascular Genetics*, 3(3), pp.256–266.

Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), p.3.

Snyder, H.M. et al., 2014. Developing novel blood-based biomarkers for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 10(1), pp.109–14.

Sohal, R.S. & Weindruch, R., 1996. Oxidative Stress, Caloric Restriction, and Aging. *Science (New York, N.Y.)*, 273(5271), pp.59–63.

Somel, M. et al., 2009. Transcriptional neoteny in the human brain. *Proceedings of the National Academy of Sciences*, 106(14), pp.5743–5748.

Sood, S. et al., 2015. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome biology*, 16, p.185.

Sowers, J.R., 2013. Diabetes mellitus and vascular disease. *Hypertension*, 61(5), pp.943–947.

Speed, T., 2003. *Statistical analysis of gene expression microarray data*, CRC Press.

Spreng, R.N. & Mar, R.A., 2012. I remember you: a role for memory in social cognition and the functional neuroanatomy of their interaction. *Brain research*, 1428, pp.43–50.

Statnikov, A. et al., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5), pp.631–643.

Steyerberg, E.W. et al., 2003. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of clinical epidemiology*, 56(5), pp.441–447.

Strand, A.D. et al., 2007. Conservation of regional gene expression in mouse and human brain. *PLoS Genet*, 3(4), p.e59.

Strong, R., 1998. Neurochemical changes in the aging human brain: implications for behavioral impairment and neurodegenerative disease. *Geriatrics*, 53, pp.S9--12.

Su, A.I. et al., 2002. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences*, 99(7), pp.4465–4470.

Tabassum, R. et al., 2014. A Longitudinal Study of Health Improvement in the Atlanta CHDWB Wellness Cohort. *Journal of personalized medicine*, 4(4), pp.489–507.

Tait, S.W.G. & Green, D.R., 2012. Mitochondria and cell signalling. *Journal of Cell Science*, 125(4), pp.807–815.

Talseth-Palmer, B.A. et al., 2013. Combined analysis of three Lynch syndrome cohorts confirms the modifying effects of 8q23.3 and 11q23.1 in MLH1 mutation carriers. *International journal of cancer. Journal international du cancer*, 132(7), pp.1556–64.

Tan, L.J. et al., 2012. Molecular genetic studies of gene identification for sarcopenia. *Human genetics*, 131(1), pp.1–31.

Taupin, P., 2006. Neurogenesis and Alzheimer ' s Disease. *Drug Target Insights*, 1(65), pp.1–4.

Taylor, R.C. & Dillin, A., 2011. Aging as an event of proteostasis collapse. *Cold Spring Harbor perspectives in biology*, 3(5), p.a004440.

Tchkonia, T. et al., 2013. Cellular senescence and the senescent secretory phenotype: therapeutic opportunities. *The Journal of clinical investigation*, 123(123 (3)), pp.966–972.

Thalacker-Mercer, A.E. et al., 2010. The skeletal muscle transcript profile reflects accommodative responses to inadequate protein intake in younger and older males. *The Journal of nutritional biochemistry*, 21(11), pp.1076–82.

The International HapMap Consortium, 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311), pp.52–58.

Thompson, H.J. & Voss, J.G., 2009. Health-And Disease-Related Biomarkers in Aging Research. *Research in Gerontological Nursing*, 2(2), pp.137–148.

Timmerman, L., 2013. GlaxoSmithKline Shuts Down Sirtris, Five Years After $720M Buyout | Xconomy. Available at: http://www.xconomy.com/boston/2013/03/12/glaxosmithkline-shuts-down-sirtris-five-years-after-720m-buyout/.

Timmons, J.A. et al., 2005. Human muscle gene expression responses to endurance training provide

a novel perspective on Duchenne muscular dystrophy. *Faseb J*, 19(7), pp.750–760.

Timmons, J.A. et al., 2010. Using molecular classification to predict gains in maximal aerobic capacity following endurance exercise training in humans. *Journal of applied physiology*, 108(6), pp.1487–96.

Tokuda, Y. et al., 2009. The role of trastuzumab in the management of HER2-positive metastatic breast cancer: an updated review. *Breast cancer (Tokyo, Japan)*, 16(4), pp.295–300.

DE Toledo-Morrell, L. et al., 2000. From healthy aging to early Alzheimer's disease: in vivo detection of entorhinal cortex atrophy. *Annals of the New York Academy of Sciences*, 911(1), pp.240–253.

Trevino, V., Falciani, F. & Barrera-Saldaña, H.A., 2006. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular medicine (Cambridge, Mass.)*, 13(9–10), pp.527–541.

Turner, D. et al., 1998. The Macroeconomic Implications of Ageing in a Global Context. *OECD Economics Department Working Papers*, 193(193).

Turner, S.T. et al., 2006. Genomic loci with pleiotropic effects on coronary artery calcification. *Atherosclerosis*, 185(2), pp.340–346.

UNFPA and HelpAge, 2012. Ageing in the Twenty-First Century:A Celebration and A Challenge. In *United Nations Population Fund*.

United Nations, Department of Economic and Social Affairs, P.D., 2013. World Population Ageing 2013. In *World Population Ageing 2013*. p. 114.

Van't Veer, L.J. et al., 2002. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871), pp.530–536.

Vas, V. et al., 2012. Aging of the microenvironment influences clonality in hematopoiesis. *PloS one*, 7(8), p.e42080.

Vermeersch, S.J. et al., 2008. Age and gender related patterns in carotid-femoral PWV and carotid and femoral stiffness in a large healthy , middle-aged population. *Journal of hypertension*, 26(7), pp.1411–1419.

Vijg, J. & Suh, Y., 2013. Genome instability and aging. *Annual review of physiology*, 75, pp.645–668.

Vogel, C. & Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics*, 13(4), pp.227–232.

Wagers, A.J. & Weissman, I.L., 2004. Plasticity of adult stem cells. *Cell*, 116(5), pp.639–648.

Walhovd, K.B., Fjell, A.M., Reinvang, I., Lundervold, A., et al., 2005. Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of aging*, 26(9), pp.1261–1270.

Walhovd, K.B., Fjell, A.M., Reinvang, I. & Lundervold, A., 2005. Neuroanatomical aging: Universal but not uniform. *Neurobiology of Aging*, 26(9), pp.1279–1282.

Wallace, D.C., 2005. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annual review of genetics*, 39, p.359.

Walne, A.J. et al., 2008. TINF2 mutations result in very short telomeres: analysis of a large cohort

of patients with dyskeratosis congenita and related bone marrow failure syndromes. *Blood*, 112(9), pp.3594–3600.

Wang, E.T. et al., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), pp.470–6.

Wei, M. et al., 1999. Relationship between low cardiorespiratory fitness and mortality in normal-weight, overweight, and obese men. *Jama*, 282(16), pp.1547–1553.

Weidner, C.I. et al., 2014. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome biology*, 15(2), p.R24.

Weindruch, R. et al., 2001. Microarray profiling of gene expression in aging and its alteration by caloric restriction in mice. *The Journal of nutrition*, 131(3), p.918S--923S.

Welle, S. et al., 2004. Skeletal muscle gene expression profiles in 20--29 year old and 65--71 year old women. *Experimental gerontology*, 39(3), pp.369–377.

Wennmalm, K., Wahlestedt, C. & Larsson, O., 2005. The expression signature of in vitro senescence resembles mouse but not human aging. *Genome Biology*, 6(13), p.R109.

West, R.L., 1996. An application of prefrontal cortex function theory to cognitive aging. *Psychological bulletin*, 120(2), p.272.

Westra, H.J. & Franke, L., 2014. From genome to function by studying eQTLs. *Biochimica et biophysica acta*, 1842(10), pp.1896–1902.

Wiesweg, M. et al., 2013. Feasibility of preemptive biomarker profiling for personalised early clinical drug development at a Comprehensive Cancer Center. *European journal of cancer (Oxford, England : 1990)*, 49(15), pp.3076–82.

Wind, A.W. et al., 1997. Limitations of the Mini-Mental State Examination in diagnosing dementia in general practice. *International journal of geriatric psychiatry*, 12(1), pp.101–108.

Winocur, G., 1998. Environmental influences on cognitive decline in aged rats. *Neurobiology of Aging*, 19(6), pp.589–597.

World Health Organization, 2011. World Report on Disability. In *World Health Organization*.

World of Work 67, 2009. Ageing societies: The benefits, and the costs, of living longer. , pp.1–5.

Wu, J. et al., 2014. The role of oxidative stress and inflammation in cardiovascular aging. *BioMed research international*, 2014.

Xiong, C., Yan, Y. & Gao, F., 2013. Diagnostic Utility of Gene Expression Profiles. *Journal of biometrics & biostatistics*, 4(1).

Yamashina, A. et al., 2002. Validity, reproducibility, and clinical significance of noninvasive brachial-ankle pulse wave velocity measurement. *Hypertension Research*, 25(3), pp.359–364.

Yan, J. & Gu, W., 2009. Gene Expression Microarrays in Cancer Research. In *Pharmaceutical Perspectives of Cancer Therapeutics*. Springer, pp. 645–672.

Yang, J. et al., 2015. Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific reports*, 5.

Yashin, A.I. et al., 2010. Joint influence of small-effect genetic variants on human longevity. *Aging*,

2(9), pp.612–620.

Yasuno, K. et al., 2010. Genome-wide association study of intracranial aneurysm identifies three new risk loci. *Nature genetics*, 42(5), pp.420–425.

Ye, S., 2006. Influence of matrix metalloproteinase genotype on cardiovascular disease susceptibility and outcome. *Cardiovascular research*, 69(3), pp.636–645.

Yoav Benjamini, 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multple Testing. *Journal of Royal Statistical Society*, 57(1), pp.289–300.

Yu, C.E. et al., 1996. Positional cloning of the Werner's syndrome gene. *Science*, 272(5259), pp.258–262.

Zahn, J.M. et al., 2007. AGEMAP: a gene expression database for aging in mice. *PLoS Genet*, 3(11), p.e201.

Zahn, J.M. et al., 2006. Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet*, 2.

Zaidi, A., 2008. Features and challenges of population ageing: The european perspective. *Policy Brief*, (1), pp.1–16.

Zethelius, B. et al., 2008. Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *The New England journal of medicine*, 358(20), pp.2107–16.

Zieman, S.J., Melenovsky, V. & Kass, D.A., 2005. Mechanisms, pathophysiology, and therapy of arterial stiffness. *Arteriosclerosis, thrombosis, and vascular biology*, 25(5), pp.932–943.

Zou, J. et al., 2014. Epigenome-wide association studies without the need for cell-type composition. *Nature methods*, 11(3), pp.309–311.

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
236278_at	HIST1H3E	down	Histone cluster 1, H3e	Replication-dependent histone; core component of nucleosome; reduced gene expression in aged mice in hippocampus
204974_at	RAB3A	down	RAB3A, member RAS oncogene family	GTPase/Ca+ signalling; age-related changes in human brain; Alzheimer's Disease link;
205050_s_at	MAPK8IP2	down	mitogen-activated protein kinase 8 interacting protein 2	AKA JIP2; scaffold protein that binds many JNK isoforms; regulates MAPK8; APP and Glucose - biochem of 'ageing diseases'
206416_at	ZNF205	down	zinc finger protein 205	DNA binding protein; regulates human M-LPH - potentially oxidative stress related
229730_at	SMTNL2	down	smoothelin-like 2	JNK substrate
226674_at	SHISA4	down	shisa homolog 4 (Xenopus laevis)	Secreted and transmembrane protein; in Xenopus Shisa proteins may inhibit Wnt and FGF signaling
234495_at	KLK15	down	kallikrein-related peptidase 15	Serine protease; upregulated in advanced tumours; snp associated with cancer risk; androgen regulated
240686_x_at	TFRC	down	transferrin receptor (p90, CD71)	Iron delivery to cells; previously identified as underexpressed with age (in meta-analysis)
234536_at	SARDH	down	sarcosine dehydrogenase	Mitochondrial matrix protein, catalyses oxidative demethylation of sarcosine
239446_x_at	DCBLD2	down	discoidin, CUB and LCCL domain containing 2	Cancer-linked
222197_s_at	LOC100128008	down	Similar to RIKEN 4933439F11	---
227738_s_at	ARMC5	down	armadillo repeat containing 5	Armadillo/beta-catenin-like repeats. A tandemly repeated sequence motif first identified in the Drosophila segment polarity gene armadillo; repeats also found in the mammalian armadillo homolog beta-catenin, the junctional plaque protein plakoglobin, the adenomatous polyposis coli (APC) tumor suppressor protein, and a number of other proteins
228876_at	BAIAP2L2	down	BAI1-associated protein 2-like 2	Binds phosphoinositides
234694_at	CNTROB	down	centrobin, centrosomal BRCA2 interacting protein	Cell division - centriole associated; cancer-related
203842_s_at	MAPRE3	down	microtubule-associated protein, RP/EB family, member 3	Microtubule-associated
217079_at	217079_at	down	q12.13 Homo sapiens unknown protein mRNA	---
217696_at	FUT7	down	fucosyltransferase 7 (alpha (1,3) fucosyltransferase)	Involved in creation of sialyl-Lewis X antigen
217700_at	CNPY4	down	Canopy 4 homolog (zebrafish)	AKA Prat4b; secreted; negative regulator of Toll-like receptor trafficking/cell surface expression
221309_at	RBM17	down	RNA binding motif protein 17	Part of spliceosome complex
230044_at	PCYT2	down	phosphate cytidylyltransferase 2, ethanolamine	Enzyme involved in phospholipid biosynthesis; mice with PCYT2 accumulate more DAG and TG with age
236091_at	HMGB2	down	high-mobility group box 2	Chromatin-associated; linked to osteoarthritis
212512_s_at	CARM1	down	coactivator-associated arginine methyltransferase 1	Transcription; methylates proteins including histones & chromatin-associated proteins; in skeletal muscle, linked to glycogen gene expression and differentiation
244707_at	HCN4	down	hyperpolarization activated cyclic nucleotide-gated potassium channel 4 (HCN4)	Potassium channel
215488_at	215488_at	down	Vitelliform macular dystrophy 2 (Best disease, bestrophin)	Bestrophins may form chloride channels or regulate voltage-gated L-type calcium channels; linked to Vitelliform macular dystrophy
202588_at	AK1	down	adenylate kinase 1	Cytosolic; energy metabolism
215844_at	TNPO2	down	transportin 2	RanGTP-binding nuclear transport receptors; HuR is TRN2 export substrate; RNA binding

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
228279_s_at	TNK2	down	Homo sapiens tyrosine kinase, non-receptor, 2	RAC related; tumour motility; cdc42hs?
238006_at	SIN3A	down	SIN3 transcription regulator homolog A (yeast)	Transcriptional corepressor activity; HDAC regulation/chromatin remodelling
240147_at	C7orf50	down	chromosome 7 open reading frame 50	Includes a pro-survival human ageing SNP
243906_at	243906_at	down	Organic solute transporter alpha	---
244504_x_at	ARF1	down	ADP-ribosylation factor 1 (microRNA 3620 within?)	Modulates cell surface Cdc42 dynamics; GTP-binding protein involved in protein trafficking; modulates vesicle budding/uncoating within Golgi complex
210483_at	TNFRSF10C	down	tumor necrosis factor receptor superfamily, member 10c, decoy without an intracellular domain	AKA DCR1/TRAILR3; cytokine related, tumour related; receptor has extracellular TRAIL-binding domain + TM domain but no cytoplasmic death domain - not able to induce apoptosis but thought to be antagonistic receptor that protects cells from TRAIL-induced apoptosis
216327_s_at	SIGLEC8	down	sialic acid binding Ig-like lectin 8	Adhesion molecule that mediates sialic-acid dependent binding to cells; mostly expressed in eosinophils and mast cells
217046_s_at	AGER	down	advanced glycosylation end product-specific receptor	AKA RAGE; transmembrane receptor of Ig superfamily; binds advanced glycation endproducts; linked to pro-inflammatory gene activation; alternatively spliced with 6 isoforms - some lack TM domain and thought to be secreted; linked to impaired skeletal muscle insulin action via AGEs; increased in Alzheimer's Disease; Diabetes linked
222080_s_at	SIRT5	down	Sirtuin 5	NAD-dependent protein acetylase associated with the mitochondria; involved in ammonia detoxification; deactivated by suramin; sirtuins linked to lifespan extension in rodents - resveratrol (possible SIRT activator) inhibits gene exp profile associated with muscle ageing; linked to brain ageing
227456_s_at	C6orf136	down	chromosome 6 open reading frame 136	---
227781_x_at	FAM57B	down	family with sequence similarity 57, member B	Transmembrane protein; PPARg responsive and linked to ceramides/adipogenesis regulation
229508_at	U2AF2	down	U2 (RNU2) small nuclear RNA auxiliary factor 2	Pre-mRNA splicing factor
211180_x_at	RUNX1	down	runt-related transcription factor 1	AKA AML1; transcription factor that binds to core elements of enhancers & promoters; regulates differentiation of hemopoietic stem cells into mature blood cells; leukemia link
213690_s_at	213690_s_at	down		---
213987_s_at	CDK13	down	cyclin-dependent kinase 13	Family members have roles as master switches in cell cycle control; impact on RNA processing/splicing
218063_s_at	CDC42EP4	down	CDC42 effector protein (Rho GTPase binding) 4	May be GTPase related; GTP Rho binding; organisation of actin cytoskeleton
219150_s_at	ADAP1	down	centaurin, alpha 1	Phospholipid binding protein; linked to Alzheimer's Disease
229607_at	LOC100652912	down	uncharacterized LOC100652912	---
236269_at	ZNF628	down	zinc finger protein 628	---
239125_at	SLC25A5-AS1	down	SLC25A5 antisense RNA 1 (non-protein coding)	Antisense RNA to mitochondrial ANT2; SLC25A5 is an inner mitochondrial membrane transport protein that translocates ADP

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
239422_at	GPC2	down	glypican 2 (cerebroglycan)	Cell surface proteoglycan; found in developing nervous system; role in cell adhesion
239837_at	ADAM11	down	ADAM metallopeptidase domain 11	Metalloprotease-like protein
240098_at	RIF1	down	RAP1 interacting factor homolog (yeast)	Maybe involved in DNA repair; telomere-associated (may regulate telomere length)
244182_at	244182_at	down	Homo sapiens, clone IMAGE:5756056, mRNA	---
89476_r_at	NPEPL1	down	aminopeptidase-like 1	May catalyze removal of unsusbstituted N-terminal AA from various peptides
202312_s_at	COL1A1	down	collagen, type I, alpha 1	Reduced in aging skin and bone; osteoporosis linked
208232_x_at	NRG1	down	neuregulin 1	Repairs nerve damage in the adult; high levels linked to longevity in rodents
209280_at	MRC2	down	mannose receptor, C type 2	Role in ECM remodelling; linked to tumourigenesis and metastasis
220482_s_at	SERGEF	down	secretion regulating guanine nucleotide exchange factor	Guanyl-nucleotide exchange factor activity - may be involved in secretion process; deafness-related
226871_s_at	ATG4D	down	autophagy related 4D, cysteine peptidase	Cysteine-type endopeptidase involved in autophagy
244164_at	FAM223B	down	Homo sapiens family with sequence similarity 223, member B (non-protein coding) (FAM223B), non-coding RNA	Non-protein coding
244591_x_at	RNF207	down	Ring finger protein 207	Variation in QT interval SNPS
211837_s_at	PTCRA	down	pre T-cell antigen receptor alpha	T cell development
214213_x_at	LMNA	down	lamin A/C	Linked to Hutchinson-Gilford Progeria Syndrome (HGPS), caused by a spontaneous mutation (truncated version), and characterized by premature aging. Nuclear membrane structural component- roles in cell cycle control, DNA Replication & chromatin organisation; cleaved during apoptosis; mice deficient have enhanced mTORC1 signaling linked to dystrophic pathology
214316_x_at	CALR	down	Calreticulin	Calcium binding/regulation; protein folding in ER; possible nuclear receptor modulation; reduced protein expression during aging in mouse skeletal muscle
223415_at	RPP25	down	ribonuclease P 25kDa subunit	Component of ribonuclease P, a protein complex that generates mature tRNA by cleaving their 5' ends; linked to developmental brain disorders
228677_s_at	RASAL3	down	RAS protein activator like 3	Ras GTPase activator activity
228684_at	ZNF503	down	zinc finger protein 503	AKA Nolz1; RAR signalling
236845_at	TRIM62	down	tripartite motif containing 62	E3 ubiquitin ligase; potential immune role
238046_x_at	PWWP2B	down	PWWP domain containing 2B	Histone modification biochemistry
207883_s_at	TFR2	down	transferrin receptor 2	iron homeostasis
209983_s_at	NRXN2	down	neurexin 2	Neurological role; loss of function disorders - neuronal cell surface protein with role in cell recognition and adhesion molecule binding
210364_at	SCN2B	down	sodium channel, voltage-gated, type II, beta subunit	Subunit of voltage-gated sodium channel; may be regulated by BACE1
219967_at	MRM1	down	mitochondrial rRNA methyltransferase 1 homolog (S. cerevisiae)	Mitochondrial ribosome complex; RNA methylation
220989_s_at	AMN	down	amnion associated transmembrane protein	Developmental gene; transmembrane protein thought to regulate bone morphogenetic protein (BMP) receptor function
231764_at	CHRAC1	down	chromatin accessibility complex 1	Histone-fold protein that binds DNA

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
233894_x_at	EMID2	down	EMI domain containing 2	AKA COL26A1
234003_at	ENOX2	down	ecto-NOX disulfide-thiol exchanger 2	Cell surface protein; two enzyme activities - catalysis of hydroquinone or NADH oxidation and protein disulfide interchange - may control physical membrane displacement for vesicle budding or cell enlargement; pro-growth in tumour cells
235671_at	235671_at	down	Homo sapiens BAC clone RP11-489G24	---
236746_at	GALNT1	down	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 1 (GalNAc-T1)	O-linked oligosaccharide biosynthesis; reduced with age of bone cell donor
206080_at	PLCH2	down	phospholipase C, eta 2	G-coupled protein receptor modulation; lipid catabolic processes
213433_at	ARL3	down	ADP-ribosylation factor-like 3	Member of ribosylation factor family of GTP-binding proteins; RP2 is a GTPase-activating protein (GAP) for ARL3
214209_s_at	ABCB9	down	ATP-binding cassette, sub-family B (MDR/TAP), member 9	Membrane-associated ATP-binding cassette transporter; may transport peptides from cytosol into lysosomal lumen; associated with antigen processing
215649_s_at	MVK	down	mevalonate kinase	Mevanolate kinase activity - key enzyme in isoprenoid and sterol biosynthesis; substrate for Geranylgeranylpyrophosphate leads to aberrant activation of the small GTPase Rac1
230693_at	ATP2A1	down	ATPase, Ca++ transporting, cardiac muscle, fast twitch 1	Catalyzes ATP hydrolysis coupled with Ca++ translocation, & is involved in muscle excitation and contraction
231520_at	SLC35F3	down	Solute carrier family 35, member F3	---
239523_at	TUSC5	down	tumor suppressor candidate 5	Expressed abundantly in WAT, BAT and peripheral afferent neurons; may be involved in fat metabolism - increased expression in response to PPARgamma agonist in 3T3-L1 cells
240116_at	240116_at	down	AT rich interactive domain 1B (SWI1-like)	---
241427_x_at	FBXW7	down	Homo sapiens F-box and WD repeat domain containing 7, E3 ubiquitin protein ligase (FBXW7)	Substrate recognition component of SCF (SKP1-CUL1-F-box protein) E3 ubiquitin-protein ligase complex - mediates ubiquitination of cyclin E; mutations linked to cancer
208129_x_at	RUNX1	down	runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)	AKA AML1; transcription factor that binds to core elements of enhancers & promoters; regulates differentiation of hemopoietic stem cells into mature blood cells; leukemia link
216980_s_at	SPN	down	sialophorin (gpL115, leukosialin, CD43)	Transmembrane protein found on surface of immune cells; Wiskott-Aldrich syndrome
218762_at	ZNF574	down	zinc finger protein 574	---
219756_s_at	POF1B	down	premature ovarian failure, 1B	Associated with premature ovarian failure
226141_at	CCDC149	down	coiled-coil domain containing 149	---
229047_at	PLEKHB1	down	Pleckstrin homology domain containing, family B (evectins) member 1	developmental; membrane-associated signal transduction activity
229343_at	GTSE1	down	G-2 and S-phase expressed 1	Expressed in S & G2 phase of cell cycle; microtubule-associated protein important in cell migration
237046_x_at	IL34	down	interleukin 34	Alternative ligand for Csf-1 receptor; vitD regulated in skeletal cells; increased in serum and synovial fluid from rheumatoid arthritis patients
239060_at	239060_at	down	Homo sapiens Chromosome 11q13 BAC Clone b79g17	---

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
240241_at	240241_at	down	RP11-269P11 on chromosome 9	---
207914_x_at	EVX1	down	even-skipped homeobox 1	Developmental; altered methylation in cancers
213052_at	PRKAR2A	down	protein kinase, cAMP-dependent, regulatory, type II, alpha	Subunit of cAMP-dependent protein kinase; age-dependent changes reported in rats
217410_at	AGRN	down	agrin	Large proteoglycan; Induces aggregation of signaling proteins in immune and nervous systems through a common lipid raft pathway; role in development of NMJ during embryogenesis; binds several proteins on skeletal muscle surface like MuSK receptor, laminin & dystroglycan - stabilise NMJ
225072_at	ZCCHC3	down	zinc finger, CCHC domain containing 3	---
234400_at	234400_at	down		---
203055_s_at	ARHGEF1	down	Rho guanine nucleotide exchange factor (GEF) 1	Rho guanine nucleotide exchange factor
206906_at	ICAM5	down	intercellular adhesion molecule 5, telencephalin	Transmembrane glycoprotein involved in adhesion
220529_at	FLJ11710	down	uncharacterized protein FLJ11710	---
231242_at	BHLHE41	down	basic helix-loop-helix family, member e41	AKA DEC2 or Sharp-1; regulator of aggressive breast cancer; antitumour promotes HIF1 deg.; negative regulator of transcription from RNA Pol II promoter; regulator of molecular clock
234342_at	FAM20C	down	family with sequence similarity 20, member C	Calcium-binding kinase that phosphorylates the caseins and several secreted proteins implicated in biomineralization; loss relates to bone disorders
234748_x_at	KIF20B	down	kinesin family member 20B	Plus-end-directed motor enzyme that is required for completion of cytokinesis
240325_x_at	SOX30P1	down	Homo sapiens SRY (sex determining region Y)-box 30 pseudogene 1 (SOX30P1)	Transcriptional activator; developmental
201592_at	EIF3H	down	eukaryotic translation initiation factor 3, subunit 3 gamma, 40kDa	Important in initiation of protein translation; cancer-linked
203876_s_at	MMP11	down	matrix metallopeptidase 11 (stromelysin 3)	AKA stromelysin 3; overexpressed in human tumours
223137_at	ZDHHC4	down	zinc finger, DHHC-type containing 4	---
223426_s_at	EPB41L4B	down	erythrocyte membrane protein band 4.1 like 4B	Cytoskeletal binding protein; highly expressed in melanoma cells; progression in breast cancer
227563_at	FAM27E3	down	family with sequence similarity 27, member E3	---
231402_at	231402_at	down	Homo sapiens BAC clone RP11-563K23 from 7	---
238125_at	ADAMTS16	down	ADAM metallopeptidase with thrombospondin type 1 motif, 16	Expressed in human cartilage and synovium; increased expression in tissues from osteoarthritis patients; increased by TGFbeta in chondrocytes; linked to hypertension
209097_s_at	JAG1	down	jagged 1	rs2273061 of JAG1 gene associated with high BMD; lower fracture risk; G allele rs2273061 higher JAG1 mRNA; ligand for Notch receptors; reduced expression in skeletal muscle from older men compared to young
214125_s_at	NENF	down	Neuron derived neurotrophic factor	Neuron differentiation and development
215026_x_at	SCNN1A	down	sodium channel, non-voltage-gated 1 alpha subunit	Subunit of non-voltage-gated, amiloride-sensitive, sodium channel; lung fluid homeostasis role

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
217074_at	SMOX	down	spermine oxidase	Polyamine oxidase; some link to SMAD signalling
220096_at	RNASET2	down	Homo sapiens ribonuclease T2 (RNASET2)	Ribonuclease; appears to suppress tumorigenicity
222323_at	CRYGEP	down	crystallin, gamma E, pseudogene	---
227211_at	PHF19	down	PHD finger protein 19	Polycomb repressive complex 2; Phf19 binds with H3K36me2 and H3K36me3; zinc bindng; transcriptional repressor; overexpressed in many types of cancer
227720_at	ANKRD13B	down	ankyrin repeat domain 13B	Related to EGF signalling
230345_at	SEMA7A	down	semaphorin 7A, GPI membrane anchor (John Milton Hagen blood group)	Axonal growth; t-cell function; binds to cell surfaces via GPI linkage; role in integrin-mediated signaling - promotes formation of focal adhesion complexes; promotes pro-inflammatory cytokine production by monocytes & macrophages
205224_at	SURF2	down	surfeit 2	---
212114_at	ATXN7L3B	down	hypothetical LOC552889	Linked to neurological disease
220849_at	LOC79999	down	uncharacterized LOC79999	---
223153_x_at	TMUB1	down	transmembrane and ubiquitin-like domain containing 1	Cell cycle progression, DNA repair, apoptosis; mouse wakefulness
230576_at	BLOC1S3	down	Biogenesis of lysosome-related organelles complex-1, subunit 3	Biogenesis of organelles of the endosomal-lysosomal system
238406_x_at	SEZ6L2	down	seizure related 6 homolog (mouse)-like 2	Cell surface protein; link to autism spectrum disorders & increased in lung cancers
224886_at	JMJD8	down	jumonji domain containing 8	Possibly epigenetic-related
225693_s_at	CAMTA1	down	calmodulin binding transcription activator 1	Tumour suppressor; transcriptional activator; cancer-related
226706_at	FAM20C	down		---
227287_at	CITED2	down	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	Regulates PPARg/PGC1a, HIF1; increased by fasting; downregulated in ageing rat tendon; transcriptional coactivator of p300/CBP-mediated transcriptional coactivator complex; positive regulator of TGFbeta signaling
239522_at	IL12RB1	down	interleukin 12 receptor, beta 1	Innate immunity; forms part of the IL-12R complex for high affinity binding of IL-12
244193_at	DNAJC22	down	DnaJ (Hsp40) homolog, subfamily C, member 22	Heat-shock family member; protein folding; wurst protein; endocytosis
220024_s_at	PRX	down	periaxin	Nerve development; interacts with dystroglycan complex; early onset Charcot-Marie-Tooth neuropathy
241563_at	241563_at	down	Homo sapiens chromosome 3 clone RP11-384L8	---
240550_at	OTUB2	down	OTU domain, ubiquitin aldehyde binding 2	De-ubiquitinating enzyme
235879_at	MBNL1	up	muscle blind like 1	RNA binding; regulates splicing; regulates insulin receptor splicing - affecting binding kinetics
239629_at	CFLAR	up	CASP8 and FADD-like apoptosis regulator	anti-apoptotic; inhibits TNFRSF6-mediated apoptosis; lacks caspase acivity; c-FLIP may be related to muscle ageing in mouse model - overexpression in TG mice affected satellite cell prolif & promoted SM ageing
241789_at	RBMS3	up	RNA binding motif, single stranded interacting protein 3	TGFbeta-related; links to bone mineral density and tumours
212649_at	DHX29	up	DEAH (Asp-Glu-Ala-His) box polypeptide 29	RNA helicase involved in translation initiation
219737_s_at	PCDH9	up	protocadherin 9	Calcium-dependent cell-cell adhesion and recognition protein; neural
230375_at	PNISR	up	PNN-interacting serine/arginine-rich protein	AKA splicing factor, arginine/serine-rich 18; regulated by ageing in mouse (PMID: 19968875)

Probeset_ID	Gene Symbol	Ratio of Y:0 muscle	Gene Title	Biology notes
204362_at	SKAP2	up	src kinase associated phosphoprotein 2	Adapter protein; linked to actin assembly & stress fibre formation - regulates HSF4b (linked to cataracts); increased with age in mice hearts - reversed by caloric restriction
231199_at	RP11-271C24.3	up	Mak3 homolog (S. cerevisiae)	---
221589_s_at	ALDH6A1	up	aldehyde dehydrogenase 6 family, member A1	AKA MMSDH; targeted in ageing rat heart; mitochondrial tetramer expressed at high levels in the liver, kidney and heart and at lower levels in muscle and brain; mitochondrial enzyme with role in valine and pyrimidine catabolic pathways
204731_at	TGFBR3	up	transforming growth factor, beta receptor III (betaglycan, 300kDa)	AKA betaglycan; cell surface proteoglycan that acts as co-receptor with other TGFB receptor superfamily members; reduced expression in various cancers
1556095_at	UNC13C	up	unc-13 homolog C (C. elegans)	Link to processing of phorbol esters processing: APP, receptor activation may reduce Beta-A: Learning?; unc13 genes in c.elegans linked to aging and longevity
242197_x_at	CD36	up	CD36 molecule (thrombospondin receptor)	Receptor for oxidised lipids; increases with age; contributes to obesity-related cardiac hypertrophy in mice

Dataset	Tissue	GEO Accession ID	Platform	Number of Samples	Gender	Description	Included in
Stockholm	Skeletal Muscle	GSE59880	HGU133Plus2	30	30M;0F	Training dataset for our healthy ageing prototype classifier. All the samples are disease free and matched for aerobic fitness.	Chapter-2
Campbell	Skeletal Muscle	GSE9419	HGU133Plus2	66	66M;0F	Dataset used as training for Independent Validation	Chapter-3 and Chapter-4
Trappe	Skeletal Muscle	GSE28422	HGU133Plus2	48	24M;24F	Test dataset used for Independent Validation	Chapter-3 and Chapter-4
Kraus	Skeletal Muscle	GSE47969	HGU133Plus2	33	16M;17F	Test dataset used for Independent Validation	Chapter-3 and Chapter-4
Hoffman	Skeletal Muscle	GSE38718	HGU133Plus2	22	11M;11F	Test dataset used for Independent Validation	Chapter-3 and Chapter-4
Derby	Skeletal Muscle	GSE47881	HGU133Plus2	26	15M;11F	Test dataset used for Independent Validation	Chapter-3 and Chapter-4
Berchtold	Brain	GSE11882	HGU133Plus2	120	97M;23F	Test dataset used for Independent Validation (tests robustness across tissue)	Chapter-3 and Chapter-4
Muther	Skin	E-TABM-1140*	Illumina HT-12 V3 Beadchip	279	0M;279F	Test dataset used for Independent Validation (tests robustness across tissue and platform)	Chapter-3 and Chapter-5
Ulsam	Skeletal Muscle	GSE48264	HuExonST	108	108M;0F	Longitudinal study of ~70y old swedish men with 20 year follow up period	Chapter-3
BrainEac	Brain	GSE60862	HuExonST	1231	905F;326F	Dataset used to study the 'healthy ageing gene score' in ten post-mortem brain regions from 134 subjects representing 1231 samples (free from neurological diseases)	Chapter-3 and Chapter-4
AddNeuromed Cohort 1 (AD vs CTL)	Blood	GSE63060	Illumina HT-12 V3 Beadchip	113	48M:75F	Dataset used to study the 'healthy ageing gene score' in AD patients vs controls.	Chapter-3 and Chapter-4
AddNeuromed Cohort 1 (MCI vs CTL)	Blood	GSE63060	Illumina HT-12 V3 Beadchip	106	41M;65F	Dataset used to study the 'healthy ageing gene score' in MCI patients vs controls.	Chapter-3 and Chapter-4
AddNeuromed Cohort 2 (AD vs CTL)	Blood	GSE63061	Illumina HT-12 V4 Beadchip	111	44M;67F	Dataset used to study the 'healthy ageing gene score' in AD patients vs controls.	Chapter-3 and Chapter-4
AddNeuromed Cohort 2 (MCI vs CTL)	Blood	GSE63061	Illumina HT-12 V4 Beadchip	102	35M;67F	Dataset used to study the 'healthy ageing gene score' in MCI patients vs controls.	Chapter-3 and Chapter-4
CAD study	Blood	GSE12288	HG-U133A	222	172M;50F	Gene-chip clinical study used for comparing blood RNA in people with and without coronary artery disease	Chapter-3 and Chapter-4
Diabetes Study	Blood	GSE49925	Illumina HT-12 V4 Beadchip	144	93M:51F	Gene-chip clinical study used to compare blood RNA in type II diabetes with control	Chapter-3 and Chapter-4

* Dataset available in arrayExpress

Variable	Number of obs.	Mean@70y	SD	R	R^2	P-value
Cystatin C calculated GFR (ml/min)	123	64	12	0.48	0.110	0.0006
BMI (kg/m2)	128	25.8	2.8	-1.43	0.052	0.0172
s-Albumin (g/l)	126	59.9	32.1	-0.12	0.045	0.0221
Weight (kg)	128	78.9	9.9	-0.37	0.042	0.0338
OGTT p-gluc 60 min (mmol/l)	128	9.6	2.6	-1.14	0.028	0.0834
s-Phosphate (mmol/l)	127	43.0	2.3	1.26	0.025	0.1036
OGTT p-insulin AUC	128	1.4	0.8	-3.38	0.023	0.1195
OGTT p-gluc 120 min (mmol/l)	128	7.2	2.7	-0.78	0.015	0.2164
Free fatty acids (mmol/l)	128	4,0	1,0	2.14	0.014	0.2270
OGTT p-gluc 30 min (mmol/l)	128	9.1	1.6	-1.26	0.013	0.2400
Interleukin-6 (ng/l)	122	3.9	4.9	0.40	0.014	0.2432
HDL cholesterol (mmol/l)	125	0.5	0.2	-8.25	0.015	0.2558
s-Cholesterol (mmol/l)	128	1.3	0.3	6.07	0.012	0.2577
Systolic blood pressure supine (mmHg)	128	145	19	-0.10	0.010	0.2969
Leisure time physical activity	125	3*		2.99	0.010	0.3221
u-Albumin excretion rate (µg/min)	122	11.8	37.1	-0.05	0.009	0.3393
s-Triglycerides (mmol/l)	128	6.0	1.1	1.43	0.008	0.3648
s-Insulin (pmol/l)	124	45.3	20.7	-0.08	0.008	0.3673
OGTT p-gluc 0 min (mmol/l)	128	5.5	1.0	1.20	0.004	0.5099
Diastolic blood pressure supine (mmHg)	128	84	9	-0.13	0.004	0.5143
Pulse rate (beats/min)	128	65	9	-0.13	0.004	0.5149
Mini Mental State examination	121	28*		0.07	0.002	0.6276
s-Creatinine (mol/l)	127	340	64	0.01	0.002	0.6474
s-Uric acid (mol/l)	125	1.0	0.3	2.04	0.001	0.7157
C-reactive protein (mg/l)	124	2.6	2.7	0.16	0.001	0.7972
LDL cholesterol (mmol/l)	126	80.2	30.8	0.01	0.0005	0.8272

Table A3.1: Univariate linear regression on baseline characteristics in ULSAM at 70 years of age versus healthy age gene score. *Number of obs. denotes the number of complete observations available for each variable. Mean and SD denote mean and standard deviation respectively, variables marked with * are categorical and hence reported using median. R denotes the regression-coefficient of the variable. R^2 and P-value denote r-squared and p-value of the univariate analysis.*

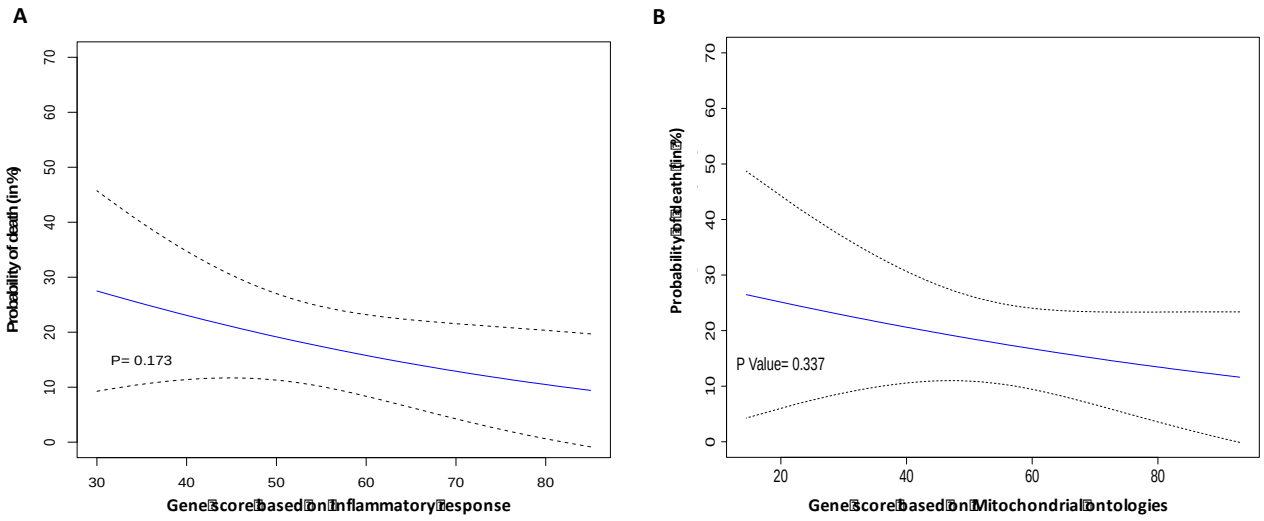


Figure A3.1 Logistic regression using genes involved in inflammatory response and mitochondrial ontologies respectively for ULSAM study with a 20 y follow-up period. *One-hundred and eight subjects provide a healthy tissue biopsy in 1992 that was suitable for RNA profiling and the fully annotated mortality-data, covering 2009-2011, was retrieved from the Swedish national health registry. Based on the literature premise that increased inflammation was bad for health and decreased mitochondrial gene expression was also bad for long term health, members of the Inflammatory response (GO:0006954) and Mitochondrion (GO:0005739) gene ontology families were selected from ENSEMBL (BioMart) and used to rank baseline samples by calculating gene expression score for these samples. A) A Logistic regression analysis performed using the genes involved in inflammation response showed no significant relationship between the median gene score and probability of death in the 20 y follow-up period (p=0.173). Here a high gene score implies higher value of expression for the members of inflammation response and vice-versa. B) Logistic regression analysis performed using the genes involved in mitochondrial biology showed no significant relationship between the median gene score and probability of death in the 20 y follow-up period (p=0.337). Here a high gene score implies higher value of expression for the members of mitochondrion ontology and vice-versa. However, using the cumulative ranking metric of top 150 genes from our original prototype was a good prognostic for mortality (Figure 3.3A).*

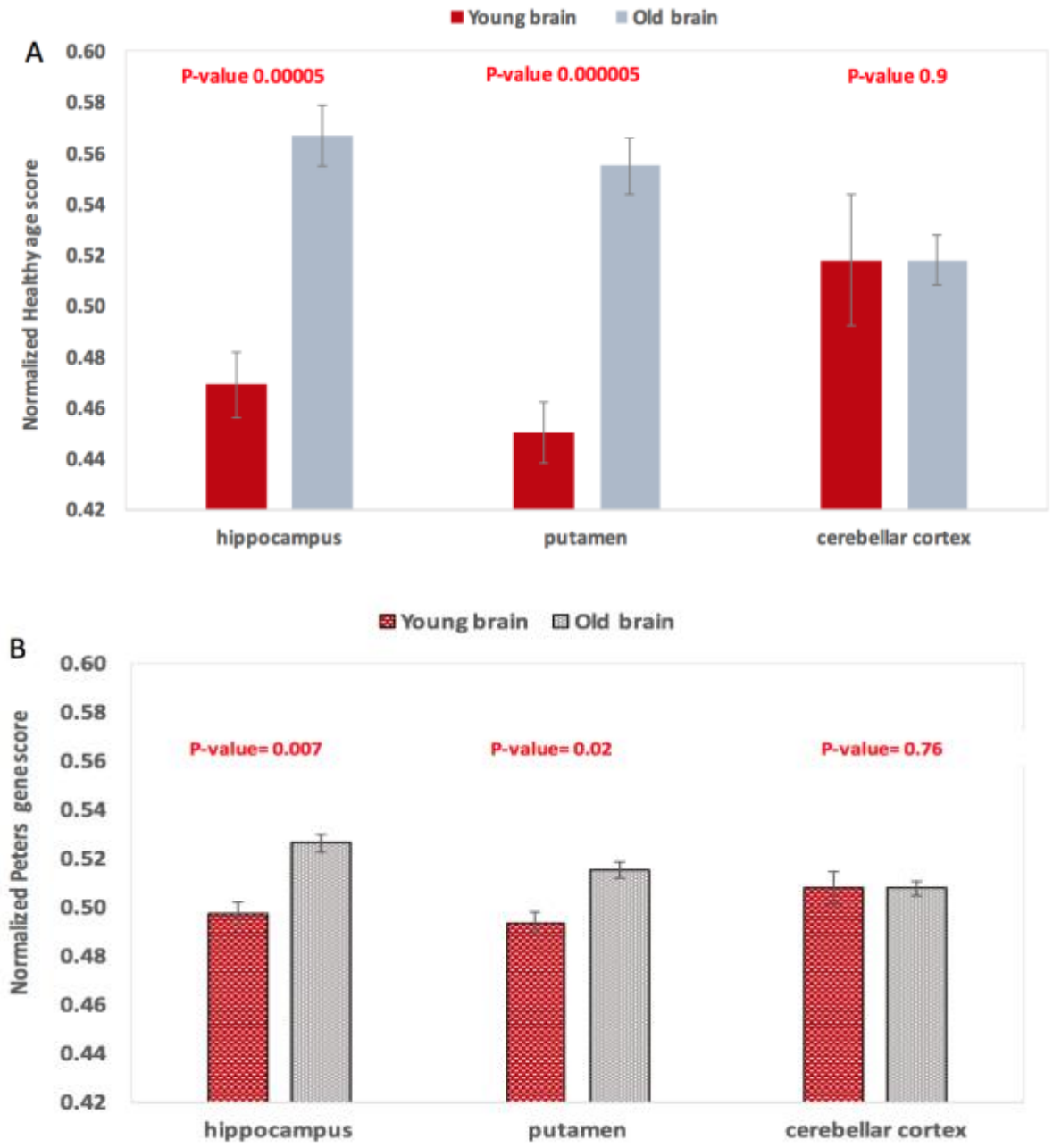


Figure A4.1 RNA signatures were studied across three anatomical human brain regions using BrainEac.org resource. *One hundred and thirty-four subjects were ranked for each brain region using the gene score method as discussed in section 4.5 and the median sum of the rank score was calculated for young and old brain regions. A) The RNA signature derived from healthy old muscle was highly regulated in regions associated with neurodegeneration. B) The Peters blood RNA signature also tracked human brain age, albeit to much lesser extent. Consistent with multiple published observations, human cerebellar cortex does not appear to be subject to substantial age-related changes.*

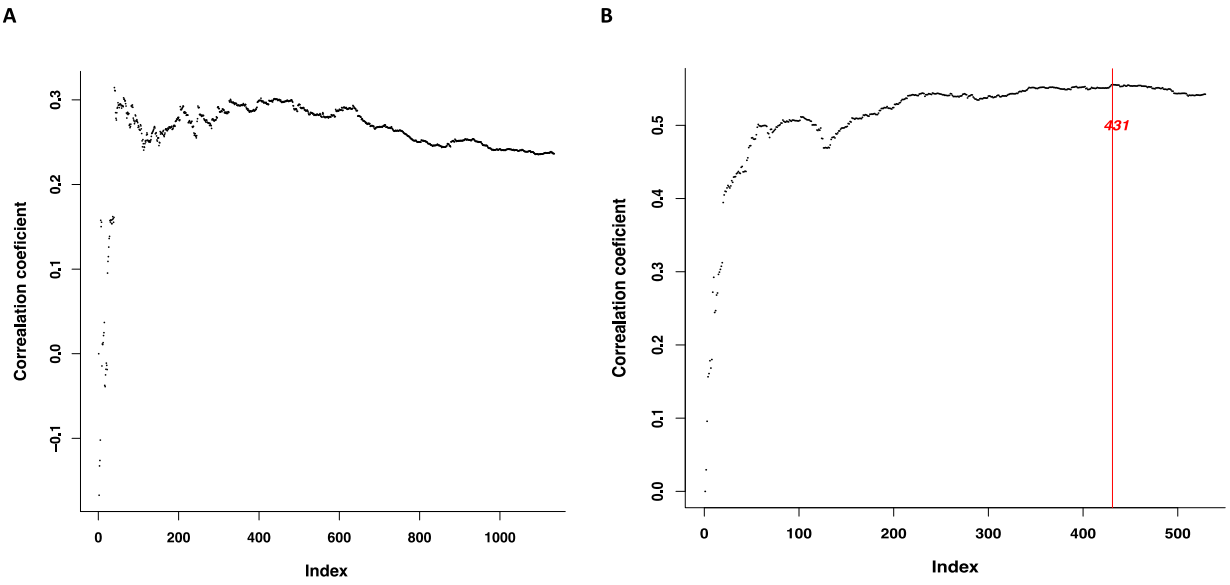


Figure A5.1 Selection criteria in 'model-selection' dataset that takes into account the effect each feature has on the model. *A) We iteratively add one feature at a time and compute the correlation coefficient of the gene set with PWV values. Then we record if adding the feature make the model better or worse and select one of the criteria from Table 5.2 B) Using the sub selection criteria we get the best model which in this case is a set of 431 features.*

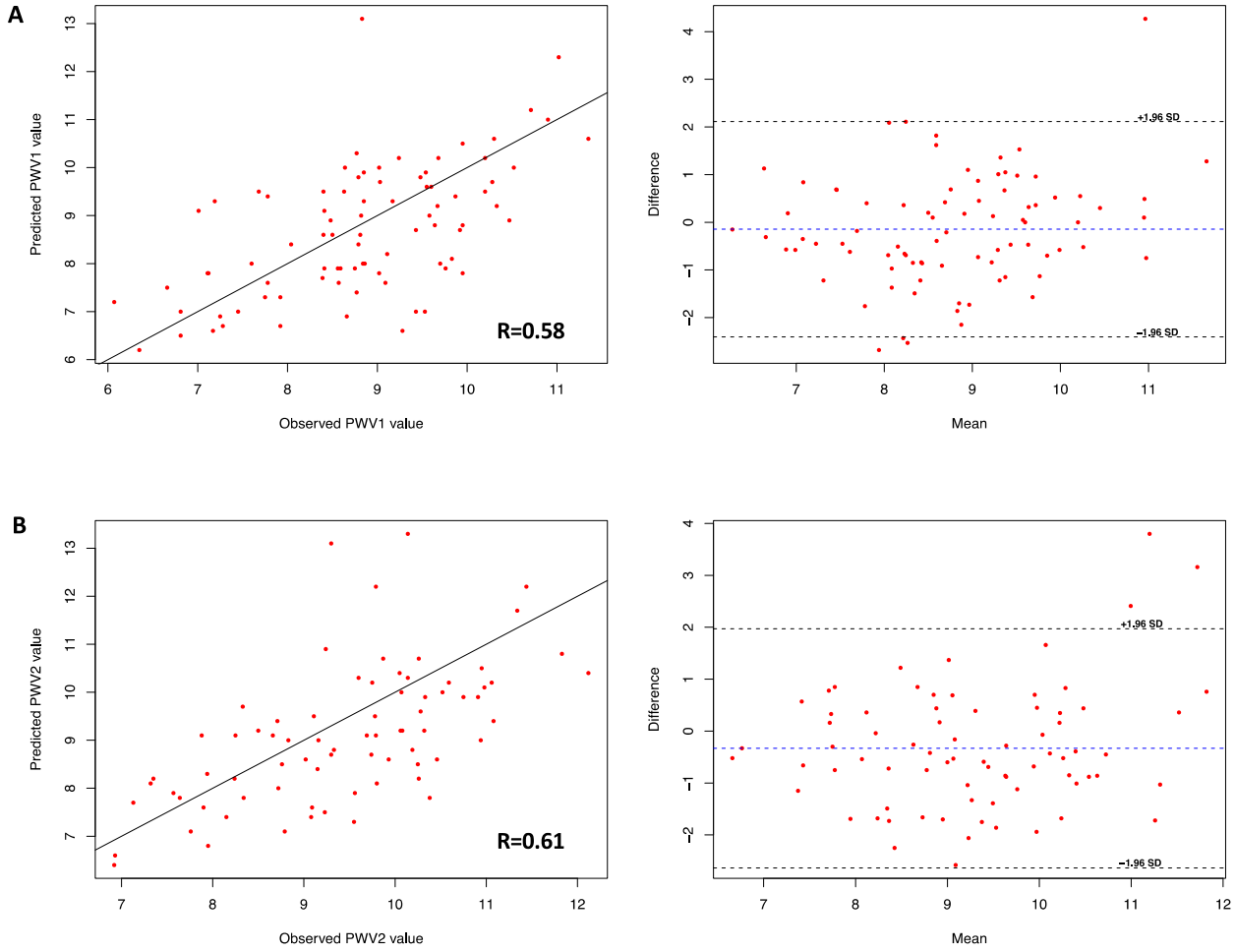


Figure A5.2 Validation of model 2 for vascular ageing. *Bland Altman plots (on the RHS) showing robustness of Model 2 in Validation datasets 1 (A) and 2 (B) respectively. Majority of observations for the model 2 were within the limit of agreement (dashed lines), thus implying that the difference between actual and predicted PWV values were marginal.*

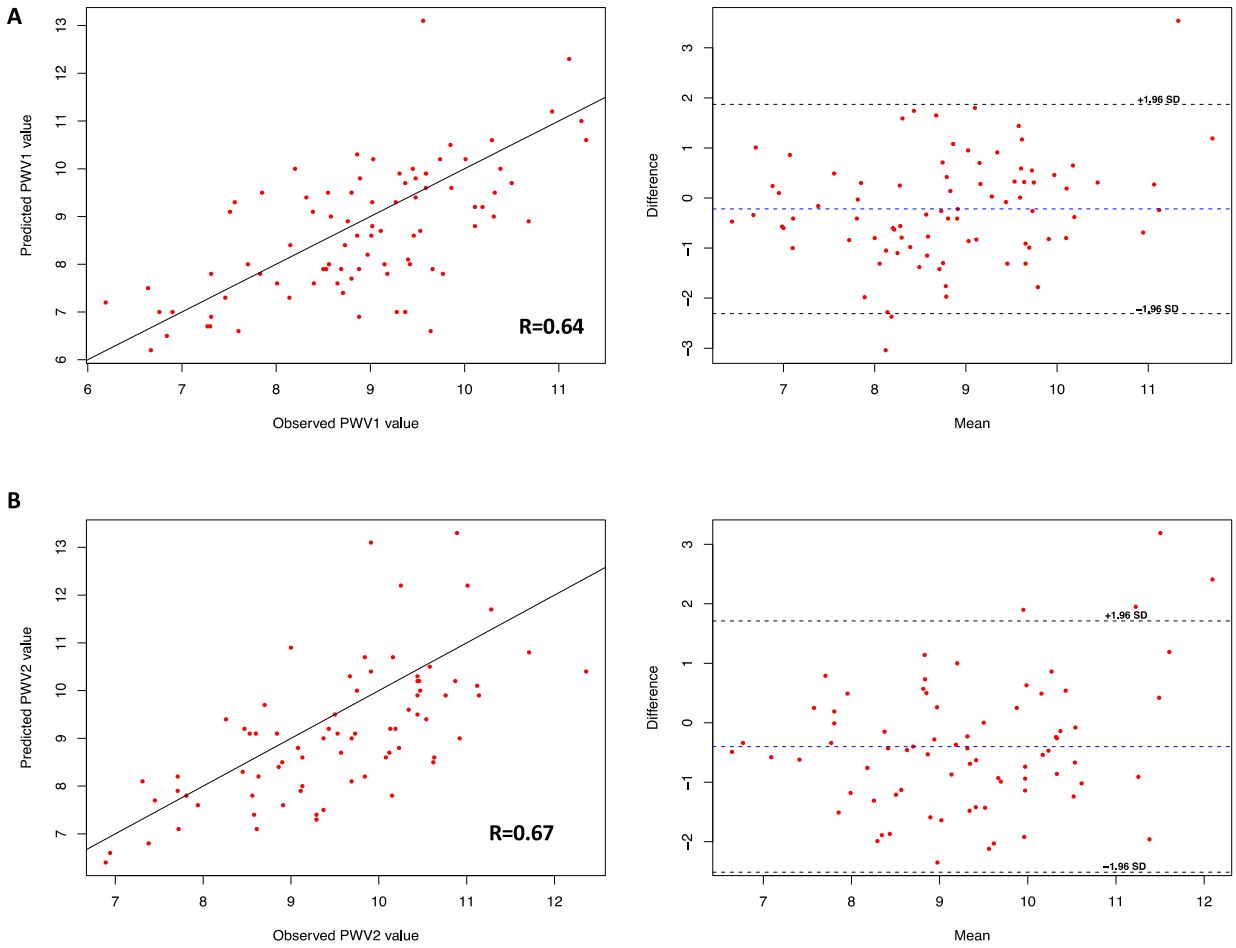


Figure A5.3 Validation of model 3 for vascular ageing. *Bland Altman plots (on the RHS) showing robustness of Model 3 in Validation datasets 1 (A) and 2 (B) respectively. Majority of observations for the model 3, like the previous two models, were within the limit of agreement (dashed lines), thus implying that the difference between actual and predicted PWV values were trivial. Model 3 was a regression model based solely on clinical data i.e. blood pressure and age and didn't take into consideration the gene expression data.*

3.1 Binary classifier reverse entry prototype

We first clear the workspace and load required libraries.

```
rm(list=ls())

library(affy)
library(class)
library(limma)
```

The data is loaded into R and the .CEL suffix stripped from each of the cel file names.

```
dataIn <- ReadAffy(celfile.path= CEL_files_path )
# remove .CEL from file names
sampleNames(dataIn) <- sub( \\.CEL$ , , sampleNames(dataIn))
```

The phenotype data is loaded from an external file. This file contains the array identifiers as well as phenotype information such as group membership. We use the array name to ensure that the rows of the phenotype data match the order of the arrays.

```
phenoTable <- read.table( Phenodata/dataset.csv , sep= , , header=T)
rownames(phenoTable) <- phenoTable$array

#make a vector that matches the cel names of the arrays to the row names of the pheno data
mt<-match(sampleNames(dataIn), rownames(phenoTable))

# attach the pheno data
phenoData(dataIn) = new( AnnotatedDataFrame , data = phenoTable[mt,])

# create the exp.group vector in the correct order.
exp.group <- factor(as.numeric(phenoData(dataIn)$group))
```

We next set up parameters for the Leave One Out Cross Validation (LOOCV) procedure. These are the smallest and greatest number of genes to use in the K-nearest neighbour (KNN) classifier as well as the number of neighbours used.

```
# largest number of genes used to classify
max.gene <- 200
# lower bound of genes used to classify
min.gene <- 2
# number of KNN neighbours to consider
knn.k <- 3
```

The gene expression is normalised using the RMA algorithm and then centered and scaled prior to KNN.

```
eset.data <- rma(dataIn)
expr.data <- scale(exprs(eset.data), center=TRUE)
```

We'll be collecting data as a result of the classification. In this code below we set up the data structures we will use to collect this data.

```
#Initialise results vector
summary.vector <- vector(length=max.gene-min.gene+1)
names(summary.vector) <- min.gene:max.gene

# Initialise the scorematrix
scorematrix <- matrix(rep(0,nrow(expr.data)*2), nrow=nrow(expr.data), ncol=2)
rownames(scorematrix) <- rownames(expr.data)

# Score for each gene list (of n=200-2+1=199)
list.score <- vector(length=max.gene-min.gene + 1)
list.score[] <- 0

# Score for each PS used in list.opt(initilaised in the loop) classification list.scorePS
<- matrix(rep(0,nrow(expr.data)*2), nrow=nrow(expr.data), ncol=3) colnames(list.scorePS)
<- c("Appearance Count","Successful Predictions", "Success Ratio" ) rownames(list.scorePS)
<- rownames(expr.data)
```

We now begin the selection of probesets for classification. This is a nested loop procedure. In the outer loop we hold out each array in turn. The middle loop uses KNN to examine the ability of selected probesets to classify the array. The innermost loop is used to select the probesets used in the middle loop.

The inner loop is used to select potentially useful probesets. Within the innermost loop an array is held out and limma is used to rank probesets in the remaining arrays by t-value. The top 200 probesets are taken forward in reverse order so that probesets with lower t-values are used for classification first. The ability of the selected probesets to classify the held out array is then tested using KNN. The results of the prediction and whether an individual probeset was used in that prediction are recorded. Each probeset is positively scored if it contributes to a correct classification.

This information is then used to select the top 150 performing probesets and these in turn go forward into a second KNN classification step on a second held out sample in the middle loop. Once again the classification performance of each probeset is recorded.

```
for (test.array in 1:ncol(expr.data)){
  #Loop 2
  for (list.opt in c(1:ncol(expr.data))[-test.array]){

    #Loop 3. Testing predictive power of 2-200 genes
   for (ngenes in min.gene:max.gene){

   #If first time through loop for any holdout array get sig genes first
     if (ngenes == min.gene){

       # set design matrix
       design <- cbind(1,exp.group[-c(test.array,list.opt)])

       # fit models
       fit <- lmFit(exprs(eset.data)[,-c(test.array, list.opt)], design = design)

       # do empirical Bayes
       fit2 <- eBayes(fit)

       # get t stats, abs value
       sig.genes <- abs(fit2$t[,2])
       # sorted in order of abs t-value, decreasing
```

```
  sig.genes <- sort(sig.genes, d=TRUE)
} # end eBayes here

# get top 200 sig probesets
top.200 <- sig.genes[c(1:max.gene)]

# REVERSE this list so lowest probesets go in first
top.200 <- rev(top.200)

# get the probesets to test this will vary with ngenes.
test.probes <- top.200[1:ngenes]

# get scaled expression data for ngenes (2 to 200), transpose for knn
candidate.genes <- expr.data[rownames(expr.data) %in% names(test.probes), ]

# transpose for knn (row-wise)
candidate.genes <- t(candidate.genes)

# take out the test.array and list.opt for training
training.data <- candidate.genes[-c(test.array, list.opt),]

# select test list.opt for testing
test.sample <- candidate.genes[list.opt, ]

# Prepare training, test and class data for knn
training.groups <- exp.group[-c(test.array, list.opt)]

# predict list.opt
predict <- knn(train=training.data, test=test.sample, cl=training.groups, k=knn.k)
# which probesets were used in this prediction
list.score.index <- which(rownames(list.scorePS) %in% names(test.probes))
# record the use (appearance) of that probeset
list.scorePS[list.score.index , 1] <- list.scorePS[list.score.index , 1] + 1

# print some details on classification
cat(paste("Predicting sample ", list.opt, "which is ", exp.group[list.opt],
          " and leaving out sample", test.array,"\tusing",
          ngenes, "genes \t"), sep="")

# so now we check result for list.opt for each ngenes
# if the prediction is right then add +1 to list.score position for ngenes
# this gives the list.size for a positive prediction
# if the prediction is right
# add +1 to the probe.performance position for included probes
if (predict==exp.group[list.opt]){
  cat ("CORRECT\n")
  list.score[ngenes-min.gene+1] <- list.score[ngenes-min.gene+1] + 1
  list.scorePS[list.score.index,2] <- list.scorePS[list.score.index,2]+1
}

else {
  cat ("Incorrect\n")
}
```

```
  }
  # END LOOP 3 - INNER LIST OPTIMISATION PREDICTIONS

  # Select the first 150 genes from sig.genes for kNN prediction
  test.length <- 150
  # get the scaled expression data for the genes that correctly predicted
  # test.array in loop 3
  candidate.genes.testing <- expr.data[rownames(expr.data) %in%
                                       names(sig.genes[1:test.length]),]
  # transpose for knn
  candidate.genes.testing <- t(candidate.genes.testing)
  # take out the test.array for training data
  training.data <- candidate.genes.testing[c(1:nrow(candidate.genes.testing))
                                           [-c(test.array)],]
  # get data for test.array only
  test.sample <- candidate.genes.testing[test.array,]
  # the true classes for training data only
  training.groups <- as.factor(exp.group[-c(test.array)])
  # do the prediction for the test.array
  predict.testing <- knn(train=training.data, test=test.sample,
                         cl=training.groups, k=knn.k)
  #print progress
  cat ("Predicting TEST ARRAY with ", test.length," genes\t")
  # if the prediction for test.array is right with ngenes
  if (predict.testing==exp.group[test.array]){
  cat ("CORRECT\n")
  # and add +1 to those contributing genes in 1st column of the scorematrix
  scorematrix[colnames(candidate.genes.testing),1] <-
                          scorematrix[colnames(candidate.genes.testing),1] + 1
  }
  else {
  cat ("Incorrect\n")
  }

  # add +1 to the second column of score matrix for each candidate gene tested
  # to count the number of times that gene is used in a prediction attempt
  scorematrix[colnames(candidate.genes.testing),2] <-
                          scorematrix[colnames(candidate.genes.testing), 2] + 1
  }
  # END LOOP 2 (predicting test.array 1 with all poss list.opt)
}
# END LOOP 1
```

We record the probesets used in classification and the number of appearances each probeset makes in the classification step above. For each probeset the ratio of correct predictions to total appearances is calculated and finally the data is written out.

```
# keep only probesets used in predictions
# i.e. column 2 on scorematrix does not equal 0
scorematrix <- scorematrix[which(scorematrix[,2]!=0),]
colnames(scorematrix) <- c("Correct Preds", "Appearances")

# Sort according to scoring first column (correct classifications); best predictors first
```

```
scorematrix_App_Sorted <- scorematrix[order(scorematrix[,2], decreasing=TRUE),]
scorematrix_Appearance_data <- scorematrix_App_Sorted[,2]

# keep only probesets which appeared in top.200 lists
# i.e. column 1 on list.scorePS does not equal 0
list.scorePS   <- list.scorePS[which(list.scorePS[,1]!=0),]

#calculates success ratio i.e correct predictions/ total appearance, for list.scorePS
list.scorePS[,3]  <- list.scorePS[,2]/list.scorePS[,1]

#Sort list.scorePS according to first column (appearance count);
list.scorePS_Sorted <- list.scorePS[order(list.scorePS[,1], decreasing=TRUE),]

# write out data
write.table(scorematrix, "scorematrix_appearances_reverse_entry_prototype.txt",
            sep="\t", quote=FALSE)

write.table(list.scorePS_Sorted,
            "FinallistofclassifiyingPS_individualPS_appearances_successrate_
            REVERSE_ENTRY_PROTOTYPE.txt",
            sep="\t", quote=FALSE
```

3.2 Independent Validation

Having previously identified a set of genes able to classify tissues as having a young or old profile we now examine the ability of this 'geneset' to classify independent datasets of young vs old tissue samples (samples not used in generating the classification model). The original data used to select the genes is not used at ANY stage of this subsequent process

We first clear the workspace, load required libraries and set some pointers to directories containing the data required.

```
rm(list=ls())

library(inSilicoDb)
library(inSilicoMerging)
library(affy)
library(class)
library(limma)
library(frma)
library(ROCR)
#Set pathway to CEL files
pathC = Path/training_set
pathM = Path/test_set
#Set names
Training_data<- training_set_name
Test_data<- test_set_name
```

The data to be classified is loaded. These are microarray cel files.

```
dataC <- ReadAffy(celfile.path=pathC)
dataM <- ReadAffy(celfile.path=pathM)
```

The phenotype data is loaded from an external file. This file contains the array identifiers as well as phenotype information such as group membership. We use the array name to ensure that the rows of the phenotype data match the order of the arrays.

```
sampleNames(dataM) <- sub( \\.CEL$ , , sampleNames(dataM))
phenoTableM <- read.table( Phenodata/training_set.csv , sep= , , header=T)
rownames(phenoTableM) <- phenoTableM$array
mtM <- match(sampleNames(dataM), rownames(phenoTableM))

sampleNames(dataC) <- sub( \\.CEL$ , , sampleNames(dataC))
phenoTableC <- read.table( Phenodata/test_set.csv , sep= , , header=T)
rownames(phenoTableC)<-phenoTableC$array
mtC<-match(sampleNames(dataC), rownames(phenoTableC))

# attach the pheno data
phenoData(dataM) = new( AnnotatedDataFrame , data = phenoTableM[mtM,])
phenoData(dataC) = new( AnnotatedDataFrame , data = phenoTableC[mtC,])
```

The microarray data we will use to validate our classifier were generated in di erent laboratories at di erent times and are independent biologically and from a technical perspective (including gene-chip format). The di erent sources of data can introduce technical variance that does not reflect the biological experiment. Below

we use the fRMA algorithm to limit the influence of technical variance e.g. different batches of microarrays. The technical manual for fRMA is here.

```
esetC <- frma(dataC)
esetM <- frma(dataM)
```

The fRMA datasets are adjusted using the COMBAT method which also corrects for batch effects across the the separate microarray datasets. After this treatment the adjusted datasets are prepared for the assessment of classification performance.

```
esets <- list(esetC, esetM)
esetMerge1 <- merge(esets, method = "COMBAT")
mtC <- which(esetMerge1$dataset=="training_set")
mtM <- which(esetMerge1$dataset=="test_set")
```

We use the knn classifier with a constant k=5 to examine the performance of our classifying geneset on independent microarray datasets. The strategy here is to use a NEW microarray dataset as the 'expression space' (named 'train' in the code) for predicting one sample at a time from the new 'test' batch of microarray data.

In this section we also create the data structures required for later receiver operator curve (ROC) analysis of the results. Specifically we set up a two column matrix with columns for the actual class (label) of each case and the prediction made by the knn classifier (predict). Label '1' is assigned to a case if it is 'young' and '-1' if it is 'old'.

Using the training and testing data we first extract only expression data for the previously selected age classifier geneset (which remains a fixed variable).

```
knn.k <- 5

#Load previously identified classification genes
scoreMatrix<-read.table("genes.txt",sep="\t", header=T)
rownames(scoreMatrix) <- scoreMatrix[,1]

#train data
expr.train <- exprs(esetMerge1[,mtC])
train.genes <- expr.train[rownames(expr.train) %in% rownames(scoreMatrix), ]
training.groups <- factor(as.numeric(pData(esetMerge1[,mtC])$group))

#test data
expr.test<- exprs(esetMerge1[,mtM])
test.genes <- expr.test[rownames(expr.test) %in% rownames(scoreMatrix), ]
testing.groups <-factor(as.numeric(pData(esetMerge1[,mtM])$group))

validation.score <- matrix(ncol=2,nrow=ncol(expr.test)) # create object to collect results

#Transpose for knn
training.data <- t(train.genes)

## set up matrix for ROC analysis
rocScore <- matrix(ncol=2,nrow=ncol(expr.test))
colnames(rocScore) <- c("label","predict")

labels <- testing.groups
```

```
for(i  in 1:ncol(expr.test)){
  if(labels[i]  == 1)
    rocScore[i,1]  <- -1
  else if(labels[i]  == 2)
     rocScore[i,1]  <- 1
}
```

Predictions rely on one 'test' sample at a time and in this scenario the 'training data' to examine which 5 members of the data are closest to a given member of the test data. The predictions made in this process are recorded in the `predict` column of the `rocScore` matrix we set up above. In addition we record the specific array tested and whether the prediction was correct or not for each that array (these are recorded in columns1 & 2 of the `validation.score` matrix respectively). If the prediction is correct we increment the `record` variable by 1 and use this to calculate the percentage of correct classifications.

```
record <- 0

for (validation.array  in 1:ncol(expr.test)){
  test.data  <-t(test.genes[,validation.array])
  predict  <- knn(train=training.data, test=test.data,cl=training.groups, k=knn.k)

  if(predict==1){
    rocScore[validation.array,2] <- -1
    }
  else if(predict==2){
      rocScore[validation.array,2] <- 1
      }

  if (predict==testing.groups[validation.array]){
    validation.score[validation.array,1] <- colnames(expr.test)[validation.array]
    validation.score[validation.array,2] <- "Correct Prediction"
    record <- record+1
  }
  else  {
    validation.score[validation.array,1] <- colnames(expr.test)[validation.array]
    validation.score[validation.array,2] <- "Incorrect Prediction"
  }
}

percentage <- 0
percentage <- ceiling((record/validation.array)*100)
```

We assess the results of the classification by ROC analysis in the code below. We classify the young as 'positive' and the old as 'negative' (these are arbitrary). To calculate the true positive rate (young classified as young) we first create a binary vector of true young. We then sum the `predict` values after filtering the `rocScore` dataframe by the binary vector.

Sensitivity is then calculated as the true positives divided by the actual number of young in the sample (i.e. the sum of the `binary.values` vector.

We calculate the false positives and false positive rate in a similar way to above, by inverting the binary labels. Finally we calculate the specificity.

```
threshold <- 0
# boolean vector of classes (young=TRUE; old=FALSE)
binary.labels <- rocScore[,1] == 1
# calculates total number of True Positives, i.e young classified as young
tp <- sum((rocScore[,2] > threshold) & binary.labels)
# calculates sensitivity i.e TP/total Positives from class labels
sensitivity <- tp/sum(binary.labels)

# calculates false positives for FPR/ 1-specificity
fp <- sum((rocScore[,2] > threshold ) & (!binary.labels))
# 1-specificity
fpRate <- fp/sum(!binary.labels)

sens <- round(sensitivity, digits = 3)
FPR <- round(fpRate, digits = 3)
spec <- 1-FPR

# Calculate area under the curve by using ROCR package
pred<-prediction(rocScore[,2], rocScore[,1])
auc<-attributes(performance(pred, auc ))$y.values[[1]]
```

Finally we write out a text file containing data on sensitivity and specificity.

```
write.table(validation.score,paste(Test_data,"_test_",Training_data,
                                   "_Train___BatchAdj_150PS_",percentage,"SR_Sens=",sens,
                                   "_Spec=",spec,".csv",sep=""),sep= , ,quote=
```

3.3 Gene ranking score calculation

This code calculates the tissue ageing Gene Score for each sample, as an median of all of the selected genes (the classification gene-set). It is applied in one of two scenarios. 1) to samples where all individuals have the same chronological age (birth year) or 2) to contrast cases versus controls where the chronological age and gender is equal in both groups. The ranking score can be standardised to the total number of samples being ranked to compare across studies.

Thus, if a gene was downregulated with age in the discovery data set, then the sample with the lowest expression in this new data set will be marked youngest and if upregulated the sample with highest score will be the youngest

Setting up preliminaries like clearing workspace and setting study names

```
rm(list=ls())
name_of_study<-"Study"
signature<-"study_signature"
```

Loading data - normalised intensities matrix

```
expr.data <- read.table("Expression_matrix.txt",sep="\t",header=TRUE,row.names=1)
```

Loading the list of genes with annoteted directionality calculated in training dataset: downregulated with 'down' sign and upregulated genes with 'up' values

```
genesUD <- read.delim("list_of_genes.txt",sep= \t , header=T)
```

Select only the geneID and directionality column from the file

```
genesUD<- subset(genesUD,select=c("geneID","Directionality"))
```

Select genes that are present on a paltform

```
genesUD<-genesUD[genesUD$geneID %in% rownames(expr.data),]
```

Seperate up and down regulated genes

```
down <- subset(genesUD,genesUD$Directionality=="down")
dReg<-down$geneID
```

```
up <- subset(genesUD,genesUD$Directionality=="up")
uReg<-up$geneID
```

Calculates score of each gene for all samples

```
geneRank <- matrix(nrow=ncol(expr.data), ncol=length(genesUD[,1]) )
rownames(geneRank) <- colnames(expr.data)
colnames(geneRank) <- c(as.character(dReg), as.character(uReg))
```

DownRegulated genes with aging: Scores the sample with highest expression value as youngest and values it 1, then the next sample maximum/higher value as 2 and so on

```
for(i in 1: length(dReg)){
  record <- 0
  PS <-  as.matrix(expr.data[which(rownames(expr.data)==dReg[i]),])
  PS <- t(PS)
  for(j in 1: ncol(expr.data))  {
    maxIn <-which.max(PS)
    maxIndex <- rownames(PS)[maxIn]
    PS   <- as.matrix(PS[-maxIn,])
    sample <- which(maxIndex==rownames(geneRank))
    geneRank[sample,which(colnames(geneRank)==dReg[i])] <- record
  }
}
```

Upregulated genes with aging: Scores the sample with lowest expression value as youngest and values it 1, then the next sample minimum/lower value as 2 and so on.

```
for(i in 1: length(uReg)){
  record <- 0
  PS <-  as.matrix(expr.data[which(rownames(expr.data)==uReg[i]),])
  PS <- t(PS)
  for(j in 1: ncol(expr.data))  {
    minIn <-which.min(PS)
    minIndex <- rownames(PS)[minIn]
    PS  <- as.matrix(PS[-minIn,])
    sample <- which(minIndex==rownames(geneRank)) record <-
    record + 1
    geneRank[sample,which(colnames(geneRank)==uReg[i])] <- record
  }
}
```

Calculate median value for each sample based on their individual genes score

```
cumulative_geneRank<- apply(geneRank,1,median)
geneRank <- cbind(geneRank,cumulative_geneRank)
```

Write out the ranking matrix

```
write.table(geneRank, paste(name_of_study,"_genescore_ranking_basedon_",signature,
                            ".csv",sep=""),sep= , ,quote=F)
```