Human Population Studies of Transcriptome-wide Expression in Age-related Traits

Mr Luke C. Pilling

Peninsula College of Medicine and Dentistry

PhD Thesis

2015

Submitted by Luke Christopher Pilling of the Peninsula College of Medicine and Dentistry Graduate School, to the Universities of Exeter and Plymouth, as a thesis for the degree of Doctor of Philosophy in Medical Studies, January 2015.

I certify that all material in this thesis which is not my own work has been identified and that no unchanged or acknowledged material has previously been submitted and approved for the award of a degree by this or any other University.

Signature:

Abstract

This thesis presents novel investigations of three common ageing phenotypes in human population studies, using microarray technology to assess 'transcriptome-wide' expression in whole blood to identify mechanisms and biomarkers.

Muscle strength is related to frailty and is predictive of disability in older persons. I assessed the association between transcript abundance in the InCHIANTI peripheral blood samples (N=695) and muscle strength. One gene (*CEBPB*) passed the multiple testing criteria, and is involved in macrophage-mediated repair of damaged muscle. I extended this work with a meta-analysis of over 7,781 individuals in four collaborating cohorts; expression of over 222 genes were significantly associated with strength, less than half of which have previously been linked to muscle in the literature. *CEBPB* did not replicate in these younger cohorts.

I then performed the first human analysis of gene expression and cognitive function (and separately with decline in cognitive ability over nine years) in the InCHIANTI cohort (N=681), and one gene was identified; *CCR2,* a chemokine receptor. Evidence in mice has implicated this gene in the accumulation of β-amyloid and cognitive impairment.

Finally, in a collaborative project with the Framingham Heart Study I studied age-related inflammation – another hallmark of ageing - using a novel approach to 'transcriptome-wide' analysis; each transcript was assessed for the proportion of the association between age and interleukin-6 (IL6) that it statistically mediated. Very few of the genes associated with IL6 alone also mediated the relationship with age. Findings include; *SLC4A10*, the strongest mediator, not previously linked to inflammation, and interleukin-1 beta and perforin, a cytokine and cytotoxic protein, respectively.

These novel analyses highlight key molecular pathways associated with age-related phenotypes in whole blood and provide links between mouse models and humans. They provide biological insight and directions for future research.

Table of Contents

Acknowledgements

I would firstly like to acknowledge the support and contributions of my supervisors Professor's David Melzer and Lorna Harries, and Dr David Studholme; your professionalism, expert advice, and encouragement have been motivating and inspiring throughout.

A number of people have made this journey possible, not least my parents Jon and Ann-Marie for supporting me throughout my life and academic journey, my fiancé Hazel for being so wonderful, and my brother Tom for, well, being my brother. Without the encouragement of my grandparents I would never have undertaken my master's degree and achieved this, so I am eternally grateful to them for their support.

At work there are many colleagues in the Epidemiology and Public Health group, the RNA group, and others, in Peninsula (now University of Exeter) Medical School who have helped me along the way, including Professor's William Henley and Tim Frayling, and their group members. Special mention should be made of Dr's Alice Holly and Ambarish Dutta for the many interesting discussions and mutual support in understanding the techniques and methods applied. I consider them both to be amongst my most excellent friends.

Much of the work of this thesis centres around epidemiological and bioinformatic analysis of the InCHIANTI cohort, so it would be remiss of me to not thank Dr's Luigi Ferrucci, Stefania Bandinelli, and Professor Andrew Singleton (amongst many others) who have made the data possible and available, and also the participants, without whom there would be no study.

I also owe many thanks to all the co-investigators on the projects I have been lucky enough to participate in, many of which are not comprised in this thesis. In particular all those involved in the CHARGE gene expression working group, especially Dr's Marjolein Peters and Andrew Johnson. The collaborative, friendly, and open nature of these projects has been very humbling; it is only through this that we can achieve the best science.

And last but not least, my friends, in Exeter and elsewhere. Who could have asked for a better bunch; Alfie Lethbridge, Brad "General Boat" Woodard, Patrick McGregor, Daniel Bartlett, Jacob "Charlie Murphy" Ringer, Patrick MacWatters, Lily Cooper, Jenny Watling, Andy "Stormbeard" Shepherd-Waring, Sam Hornett, Sam Pike, Ben Pears, Madeline Baker, Francesco Casanova, Pamela Gurney, Stuart Noble, Tom Palmer, Ben Cairns, Steph Brand, Billy McLennon, Hazel Beevers, Sophie Saint, Edgar Dawkins, Tim "Sparky" Mottram, David Hughes, Nicholas Harris, David Inness, Sam VN, Alex Busby, Katyrose Doherty, Ross Preston, Leigh Harris, and so many more over the years. Thanks for the great times.

Author's Declaration

I declare that this thesis and the work therein are my own original work and have been generated by me and the co-authors of the individual chapters. The nature of the work means many collaborators and co-authors are involved both in generating the data and analysing it, and I wish to thank once again those individuals, who are acknowledged by authorship in the chapters and published papers. I received invaluable assistance and guidance from my supervisors, colleagues and co-authors in every aspect of the PhD.

I confirm that; where work was done jointly or solely by others this is acknowledged at the start of each chapter; where work has already been published appropriate references are given; where I have referred to the published work of others appropriate references are given; this work is original and has not been copied from other students or academics.

List of Tables

List of Figures

List of Supplementary Tables

Some supplementary tables are truncated for print. The full versions can be found on the attached CD-ROM, or alternatively can be downloaded from the publishers' website.

List of Supplementary Figures

Abbreviations

5mC	5-methylcytosine
AD	Alzheimer's Disease
#bp	Number of Base-Pairs in a sequence of nucleotides, e.g. 50bp
BH	Benjamini-Hochberg
cDNA	Complementary DNA (usually from PCR, can be originally DNA or RNA)
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology
CNS	Central Nervous System
CpG	Cytosine-phosphate-Guanine
CRP	C-Reactive Protein
CVD	Cardiovascular Disease
DMR	Differentially Methylation Region
DNA	Deoxyribonucleic Acid
DNAm	DNA methylation
EST	Expressed Sequence Tag
FHS	Framingham Heart Study
FDR	False-Discovery Rate
GWAS	Genome-Wide Association Study
IL6	Interleukin 6
InCHIANTI	Invecchiare Nel Chianti – aging in the Chianti region of Italy
MMSE	Mini-Mental State Examination
NESDA	Netherlands Study of Depression and Anxiety
NIA	National Institute on Aging (part of the NIH)
NIH	National Institutes of Health
PCR	Polymerase Chain Reaction
QC	Quality Control
qRTPCR	Quantitative Real-Time PCR
RCT	Randomised Controlled Trial
RNA	Ribonucleic Acid
RS	Rotterdam Study
SASP	Senescence-Associated Secretory Phenotype
SHIP	Study of Health in Pomerania
SNP	Single Nucleotide Polymorphism
T2D	Type-2 Diabetes
TLDA	TaqMan Low Density Array

Chapter 1 – Introduction

Ageing is described as declining intrinsic physiological functioning over time, corresponding with increased mortality [1]. Tissues experience progressively decreased functioning over time, which manifest as age-related phenotypes and diseases including neurodegeneration, weakness and cardiovascular disease [2]. Experiments on rodents have shown that soluble factors found in the blood affect traits of ageing; rejuvenation of injured muscle, normally impaired in ageing, is improved by injecting mice with GDF11 (growth and differentiation factor 11) [3], and age-related cognitive decline can be ameliorated by countering the age-associated increase in systemic CCL11 (chemokine C-C motif ligand 11) [4], [5]. The implications for anti-ageing research are profound; circulating factors can promote or inhibit ageing phenotypes by influencing the resident tissues, which must therefore have the inherent potential to regenerate and function akin to youthful tissues. However, we lack understanding of comparable mechanisms in humans [6], [7].

In this thesis I will describe epidemiological projects that investigated gene expression data from human whole blood to identify molecular mechanisms and biomarkers related to three common and well-measured traits related to ageing; muscle strength, cognitive function and inflammation. To introduce the projects I have carried out and to place them in context I will give an overview of ageing and the known risk factors of age-related diseases, and move on to introduce three key age-related phenotypes prominent in this thesis. **Chapter 2** will give a detailed cohort description and methods that were used throughout the separate projects. **Chapters 3-6** are four studies presented as scientific papers investigating the three phenotypes. Finally **Chapter 7** summarises and discusses the findings of the studies and draws several conclusions.

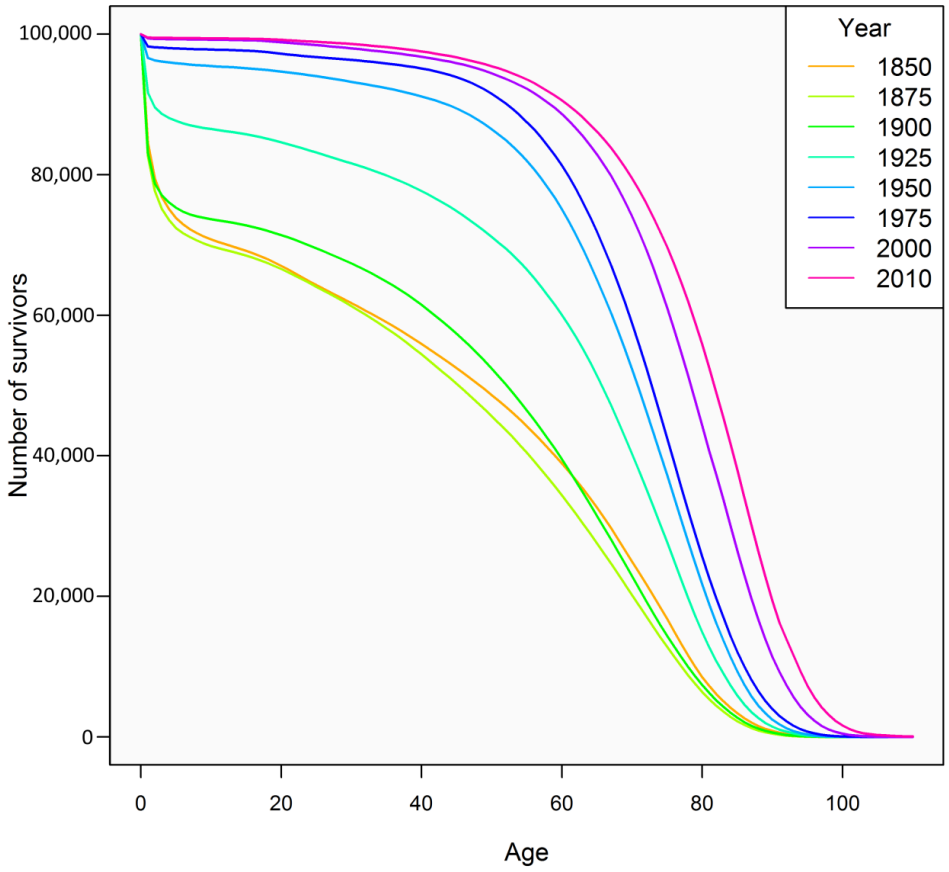
1.1 Ageing and age-related diseases

Substantial variation exists in the ageing process; although most people in the U.K. die between the ages of 60 and 80, individual members of the population die at strikingly different ages (**Figure 1.1**). It is common for older people to have one or more age-related disease, with corresponding increases in risk of mortality [8], yet there are those who "age well": the "successful agers" who live beyond middle age in good health and with freedom from impairment [9]. Therefore although ageing is associated with frailty and illness, the variability inherent in ageing provides hope of interventions [10].

The aim of most "anti-ageing" research is to decrease the proportion of one's life that is encumbered with age-related diseases; to increase peoples "healthspan" (as opposed to lifespan), the benefits of which are numerous and include, not least, improved quality of life for the individual and their family, increased productivity, and decreased burden on healthcare systems [11]. Considerable research effort is now focused on overcoming ageing entirely, with some proponents (notably Dr Aubrey de Grey) even claiming that the first 1000 year old human has already been born [12], although this view is not held by all.

In this thesis I will present original studies of age-related phenotypes and molecular mechanisms in an aged human population. This work is grounded in gerontology and epidemiology, and focuses on identifying biochemical factors (primarily expression of genes) that are related to ageing phenotypes, rather than "regenerative medicine" explicitly.

Figure 1.1 | Number of male survivors by age in England and Wales



Survival curves showing the number of male survivors by age for selected years between 1850 and 2010. Although average lifespan has increased, there is substantial individual variability in "age at death" as people experience ageing in different ways. Data from the Office of National Statistics [13].

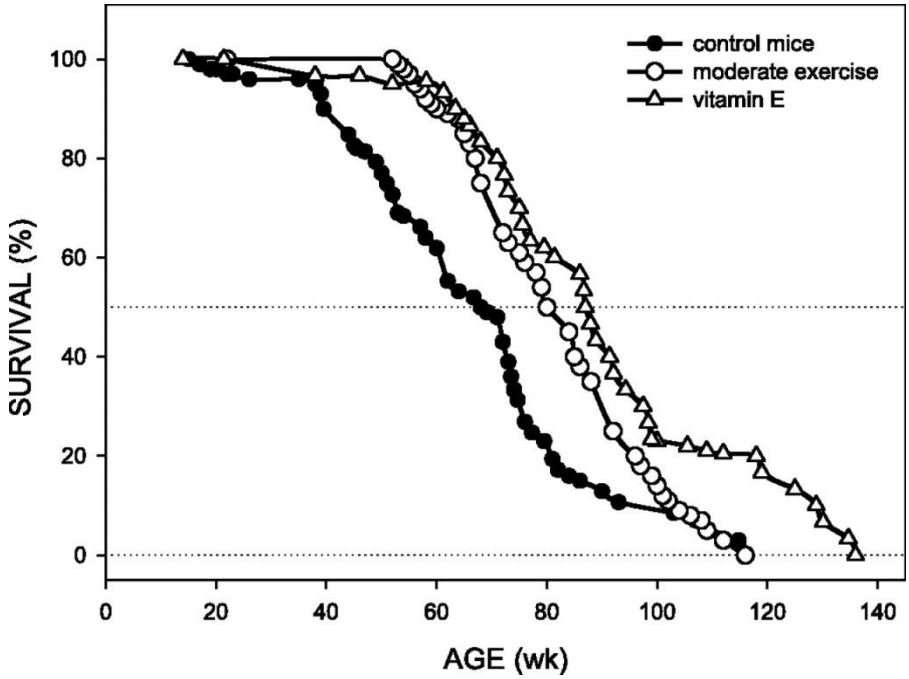
The "process" of ageing manifests in different ways; for instance two individuals born in the same hospital on the same day are unlikely to die on the same day from the same cause. Causes of variation in 'rate of ageing' can be broadly categorised into the intrinsic biology specific to each individual, and the environment in which they live [14]. Each individual has a unique genome (aside from twins; discussed later), and are exposed to different risk factors. Some "environmental" factors increase the rate of ageing (i.e. increase risk of death from an age-related disease), including exposure to pollutants [15] and smoking [16], whereas others

can decrease the rate of ageing, such as exercise [17] and diet [18]; the interaction between an individual's nature and their environment ultimately determines their rate of ageing (i.e. how quickly they are likely to succumb to age-related morbidity/mortality). Finally there is an element of randomness to the process that is most starkly presented in genetically identical model organisms, where a "normal" survival curve can be observed despite identical genetics and environment (illustrated in **Figure 1.2**). The curve is very similar to those in **Figure 1.1**, despite a controlled environment and identical genomes within the control mice. Random factors influence *individual* longevity, despite there being correlations between genetics, environment, and *population* average lifespan.

In humans, a recent meta-analysis found a relationship between age and DNA-damage [19] and thus, assuming that damage to DNA is related to reduced fitness, lifespan is inherently limited by the trade-off between the effectiveness of maintenance mechanisms and the accumulation of damage, which is highly dependent on environmental exposure and lifestyle. This is highlighted by the imperfect correlation in lifespan between twins: despite being genetically identical at conception their exposures and random damage differ – including divergence in genetics as mutations accumulate - leading to different disease risks [20]. This, at least in part, explains the existence of a normal survival curve in genetically identical mice under controlled conditions (**Figure 1.2**).

However, DNA damage is not the only cellular alteration associated with ageing; nine "hallmarks of ageing" were recently described in a review: genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication [2]. These are not mutually exclusive and are related to one another, but are characteristic cellular phenotypes of ageing. That ageing of peripheral tissues can to a degree be ameliorated in old mice by exposure to blood from young mice [6] is very intriguing, as it suggests that many of these hallmarks are potentially reversible.

Figure 1.2 | Survival curves of mice in laboratory conditions



The data shows three survival curves for genetically identical mice under 3 conditions. The purpose of this figure is to illustrate that even genetically identical organisms in a controlled environment do not have a fixed lifespan; in the control group the first mouse dies around 20 weeks of age, and the last is nearly 120 weeks. The *trend* in survival is altered by changing the environment (the exposures), yet the shape of the curve remains unchanged; there is still large variability (and unpredictability) in survival for individual mice. Figure from Navarro *et al* 2007 [21].

In nature there are evolutionarily conserved genetic pathways associated with longevity between related species [22]. Major differences between species are in the *use* of these pathways; for instance humans, chimpanzees and mice are closely related in the protein-coding regions of the genome, but have drastically different maximum lifespans (~90 years for humans, ~60 for chimps, ~3 for mice). Given the close genetic relatedness of the species

in the protein-coding regions the differences in rate of ageing could be in the *use* of the genes (i.e. when, where, and by how much the genes are expressed, rather than in their exact sequence per se) [23].

The research presented in this thesis examines gene expression data from whole blood samples drawn from human population study participants to examine the difference in use of genes between individuals with diverse phenotypes. These studies are exploratory and hypothesis generating, and sought to identify molecular biomarkers and elucidate mechanisms associated with ageing that may in the future help predict individuals at higher risk of age-related disease (who could be interpreted as “fast agers”) and may also highlight plausible mechanisms for future research. Before commencing an epidemiological study on age-related phenotypes it is important to understand the known factors that could influence or confound the relationship under examination; the following section describes some of the current known risk factors of ageing, health, and frailty.

1.2 Traditional risk factors and biomarkers

Determining predictive and modifiable risk factors of disease are central goals of epidemiological studies. Some are well known (tobacco exposure is the single largest modifiable risk factor for age-related diseases [24]), whereas others may have more subtle effects. Biochemical risk factors predictive of age-related disease include, but are not limited to; interleukin-6 (IL-6) and other immune markers [25]; uric acid [26]; and vitamin D [27]. However, this is not an exhaustive list (see the recent review by Sorrentino *et al* describing "gerontogens;" toxicants that accelerate ageing [28]), and there is much variation in longevity beyond the known biomarkers yet to be accounted for.

Social status is associated with health and can be predictive of disease [29]. It is typically measured as "income", "highest education level attained", or some similar definition. Social status encompasses many related exposures and behaviours, from "living conditions" to "availability to healthcare" to "physical activity/labour at work", which could all impact on exposure to risk factors and lifestyle.

Lifestyle factors often have large effects on health and longevity, many of which – including smoking as already mentioned – are easily modifiable, but require people to take a physically or mentally active role in 'administering' the treatment, be it giving up smoking, or being more physically active. The beneficial effects of exercise or physical activity are still debated in some areas; a recent systematic review and meta-analysis of randomised controlled trials (RCT's) found large inconsistencies in associations between resistance training, aerobic exercise, and Tai Chi, and indicators of cognitive performance/decline [30]. The inconsistencies could be due to differences in study design or the administration of exercise, as epidemiological studies have previously reported a protective effect of physical activity on cognitive function both in mid- and late-life [31], [32], as well as other age-related traits, such as cardiovascular function [33].

Many of the above quantifiable traits relate qualitatively to a one's *physiological reserve*, which affects the body's vulnerability to diseases and morbidities, and ultimately to mortality at older ages [34]. This relates quantitatively to 'frailty', recognised as more than simply the most severe disease an individual presents clinically, but instead is expressed as increased vulnerability over time and reduced ability to recover after a serious destabilising event [35]. Three traits in particular can be predictive (epidemiologically) of morbidity or mortality in later life by contributing to older peoples frailty, and will be the focus of this thesis; declining muscle strength [36], declining cognitive function [37], and increased (chronic) levels of peripheral inflammatory cytokines [38]. These traits are correlated cross-sectionally in the InCHIANTI cohort (**Table 1.1**).

Table 1.1 | Pearson correlations between age and three ageing traits

	Age	Strength	MMSE
Age	1		
Strength	-0.579	1	
MMSE	-0.462	0.447	1
IL-6	0.422	-0.307	-0.251

Strength: hand grip strength (kg),

MMSE: mini-mental state examination,

IL-6: interleukin-6 (pg/dL).

1.2.1 Muscle Strength

Low mid-life muscle strength (measured using grip strength) is strongly associated with reduced physical function and higher risk of mortality in later life [39], independent of adiposity (although other studies have found their relationship dependent in older persons [40]). Grip strength is a marker for overall strength and frailty [41], is commonly measured in population studies, correlates with overall body strength in older individuals [42], and is therefore utilised in the projects in this thesis. The definition of sarcopenia [43] also includes decreasing muscle mass with age, however changes to the composition of muscle tissue (for instance infiltration of intramuscular adipose and connective tissues) also contribute to

reduced strength [44], therefore strength can be lost without any apparent change in mass. The reasons for this are unclear, but impairment in the regeneration/homeostasis of muscle tissue could be due to reduced numbers of functional muscle stem cells available, or the aged muscle providing a non-favourable environment [45]. Evidence in mouse-models shows that muscle tissue from aged mice has the capacity to regenerate when transplanted into younger mice, and that muscle from young mice regenerates poorly in older mice [46]. This, coupled with the knowledge that the immune system plays a critical role in muscle maintenance [47] and is also impaired by ageing [48], suggests that one mechanism determining the success of muscle regeneration and maintenance is the circulating environment, which is known to change with age. Again, the implication is that aged tissues have the capacity to regenerate under normal (young) biological conditions, but the aged systemic milieu inhibits (or does not promote) this process. In **Chapter 3** a study of peripheral gene expression associations with muscle strength is reported, and in **Chapter 4** this is extended to a meta-analysis of multiple independent cohorts.

1.2.2 Cognitive Function

Declines in cognitive function are common in ageing; the prevalence of dementia – a syndrome which encompasses a range of symptoms from declines in memory to impaired reasoning and judgment [49] – increases dramatically with age (**Table 1.2**). Alzheimer's disease (AD) is the most common form and is pathologically characterised by neuronal loss, β-amyloid plaques, and neurofibrillary tangles [50]. Inflammation – both in the brain and outside the central nervous system (CNS) - is recognised as a key feature of cognitive dysfunction and dementia [51]. In particular there is evidence that with ageing there is a progressive decline in the effectiveness of the blood-brain-barrier (which separates the CNS from the peripheral blood) [52], although to what extent age-related chronic peripheral inflammation is involved is unclear. Mouse-model parabiosis experiments (where two organisms are attached allowing sharing of their blood) have shown that neurological and

cognitive declines with age could be due to changes in blood-borne factors [4]. Following publication of **Chapter 5** of this thesis (a study of peripheral gene expression associations with cognitive function) in 2012 evidence is increasing for an interplay between resident immune cells in the brain and the peripheral immune system [53], [54].

Table 1.2 | Prevalence rates of dementia in the UK

Age group, years	Male	Female
60-64	0.1%	0.2%
65-69	1.0%	1.5%
70-74	2.4%	3.1%
75-79	6.5%	5.1%
80-84	13.3%	10.2%
85-89	22.2%	16.7%
90-94	29.6%	27.5%
95+	34.4%	30.0%

In 2014 an estimated 850,000 people in the UK had dementia, with the prevalence increasing dramatically with age. Data from the Alzheimer's Society [55].

1.2.3 Chronic inflammation

Sustained low-level (chronic) systemic inflammation is common in ageing [56]. Inflammatory biomarkers IL-6 and C-reactive protein (CRP) are commonly used in such studies to indicate inflammation; both are associated with increased risk of numerous diseases including cognitive decline [57], and clinically relevant cut-points have been identified that are associated with poor physical performance and muscle strength in older people [58]. This age-related pro-inflammatory state is hypothesised to be detrimental by damageing tissues "directly" (with continued production of cytotoxic compounds by immune cells), "indirectly" (by altering the phenotype of peripheral tissues, for instance by promoting cellular senescence [59]), or by another means, such as disrupted insulin signalling [60]. In **Chapter 6** the results of the first study into the gene expression *mediators* of the association between age and inflammation in humans is reported; this is distinct to our study of gene expression associations with circulating IL-6 levels *independent* of age [61], published separately and does not form part of this thesis.

1.3 Molecular mechanisms

In this section I will review the genetic (and other molecular) risk factors of diseases, leading on to introduce the work included in this thesis.

1.3.1 Genetic risk factors

Many ageing traits are highly heritable, including muscle strength (30-85% heritability) [62], IL-6 levels (29-49% in an African population) [63], and cognitive function (35-74% depending on the specific domain of cognition studied) [64], suggesting that inherited genetic variation between people have a large impact on these traits. Studies of common genetic variation associated with disease predominantly focus on single-nucleotide polymorphisms (SNPs), which are single base-pair substitutions in the DNA (which account for 88% of the variants in dbSNP (build 135), the database of known genetic variation between individuals [65]). So far studies of associations with longevity have found very few replicable variants; the *APOE* locus is a notable exception [66], [67], although some promising candidates, including a variant previously associated with blood-pressure in region 5q33, have recently been identified [67]. There doesn't seem to be a single gene or pathway that confers longevity, with many plausible candidates identified in model organisms not translating to human populations [68]. I will here summarise the known genetic contributors to longevity in humans.

The United States National Human Genome Research Institute provides an up-to-date catalogue of published Genome-Wide Association Studies (GWAS), known as the "GWAS catalog" [69]. On 29th July 2014 this contained (after excluding associations with p-value's greater than $5x10^{-8}$, an often used cut-off for GWAS) 7609 records pertaining to 5005 unique SNPs associated with 684 diseases and/or traits. Just assessing age-related diseases/traits (81 at the time of writing, everything from Alzheimer's disease through cardiovascular

disease, diabetes and cancer) there are 789 unique SNPs. Although there are many variants that increase the risk of individual age-related diseases, the *APOE* locus is the only robust GWAS result associated with longevity specifically (though other candidates exist). DNA variants in *APOE* are implicated in a variety of age-related diseases including Alzheimer's disease (AD) and cardiovascular disease (CVD) [17] (this paper puts forward an interesting argument for the evolution of the e4 allele; that the negative effects of inheriting the risk allele may be ameliorated by exercise, and it is only recently that activity levels in the general population have dropped sufficiently for the negative effects of the e4 allele to manifest [17]). $p16^{INK4a}$ (9p21.3 locus) is also associated with a number of age-related diseases [70], but is not prominent in GWAS of longevity. No other genetic variant or gene in the catalogue has been robustly (i.e. with multiple replication studies) associated with >1 age-related disease or longevity itself (although there are some interesting candidates including *FOXO3a* and *LMNA*), so there is little evidence for common "ageing" (or longevity) variants in humans at the genetic level, albeit using current techniques which do not assess every nucleotide, and may rely on imputed variants that require further validation.

The above observations agree with a publication by the Leiden group (a study of long-lived individuals in the Netherlands) that long-lived individuals are not enriched with the protective alleles of GWAS-identified disease risk variants compared to controls [71]. Thus to achieve longevity you may need more than just not inheriting disease-associated genetic variants. Reports that offspring of long-lived parents have reduced risk of age-related diseases (including cancer and cardiovascular disease) compared to age-matched controls, independent of socio-economic status and a number of other factors [72], have a profound implication; that there may be "longevity alleles" yet to be discovered (*APOE* is the best candidate yet) that do more than just protect against disease, but actually extend lifespan. That these have not yet been identified may highlight the heterogeneity of the ageing process, lack of clear "phenotype" in studies of longevity, limited sample sizes available, inadequate analysis methods to take into account the additive effects of many small-effect or

rare variants, or that other factors such as epigenetics (introduced in **section 1.3.3**) play a more important role.

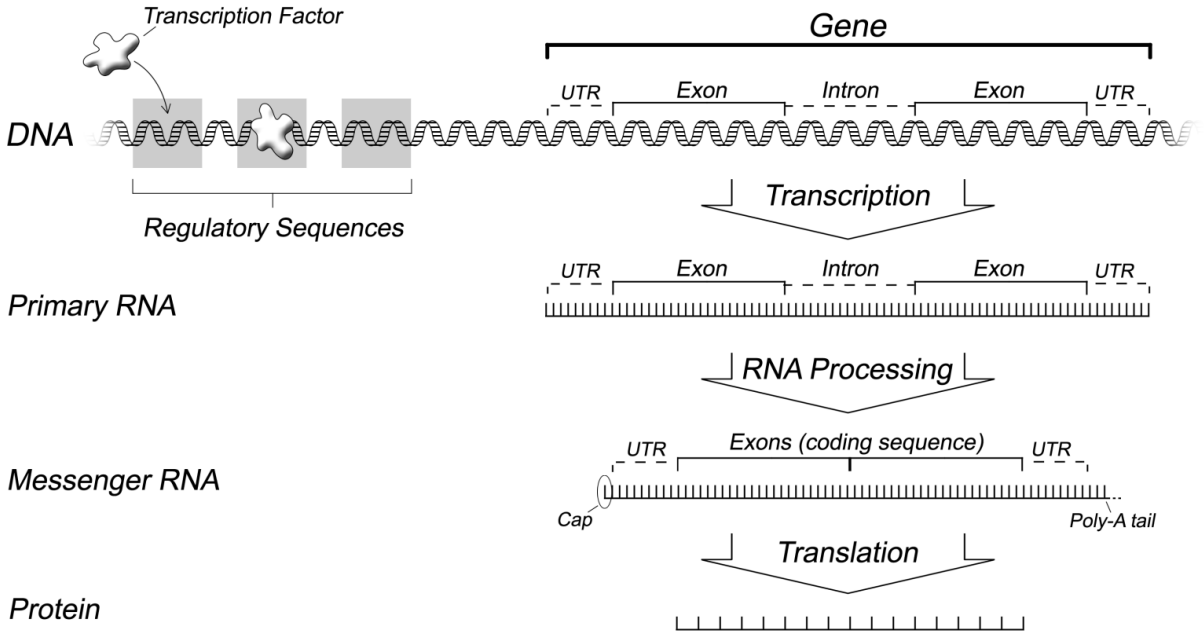
Of note are several rare monogenic diseases that present accelerated ageing (progeria) phenotypes, including Werner's syndrome and Hutchinson-Gilford Progeria Syndrome (HGPS) [73], the latter of which is caused by a mutation in the *LMNA* gene noted earlier as a candidate for longevity identified by GWAS. HGPS to some degree replicates phenotypes associated with normal ageing, in particular hair loss, skin problems and atherosclerosis, but also premature death due to stroke and coronary dysfunction [74] (mean age at death was 12.6 years in this study). Efforts have been made to study accelerated ageing phenotypes for clues to "normal" ageing mechanisms; lamin A (the protein product of *LMNA*, also called progerin) is expressed at high levels in individuals with HGPS-causing mutations and promotes atherosclerosis in vascular cells [75]; in individuals without a HGPS-causing mutation expression increases with age [76] suggesting that there may be a causal link.

1.3.2 Gene expression

The central dogma of molecular biology (**Figure 1.3**) is that information in the genome needs to be *transcribed* to RNA (a functional molecule, as well as an intermediate to proteins [77]) in order to be "used" by the cell – messenger RNAs are *translated* by ribosomes into proteins, and non-coding RNAs may function to regulate transcription of other genes. However, despite all cells in an organism containing the same genetic information (with the exception of a small few, e.g. B-lymphocytes have a hypervariable region to increase antibody diversity) the phenotypes can be drastically different. The temporal and spatial variation in *expression* (and processing) of gene transcripts – as the information in the DNA is transcribed to RNA in different cells and tissues at various time points – gives rise to the multitude of cellular phenotypes possible from the single genome available to organisms. By measuring levels of specific transcripts and asking which transcripts are related to diseases

or phenotypes we can gain insight into mechanism and pathways, as well as potentially identify biomarkers of disease [23].

Figure 1.3 | The central dogma of molecular biology



The central dogma of molecular biology is that “DNA makes RNA makes Protein”. This is over simplified but demonstrates that all functional molecules (RNA and proteins; RNA molecules can have many functions beyond simply “messenger” RNA) are derived from information in the DNA. Cellular and tissue specificity is dependent upon the exact product that is produced (when, where, how much, what form), which is reliant on not just the “coding” sequence that makes the final product, but the extensive regulatory elements that can be many kilobases away (only 64% of enhancers have at least one transcriptional-start-site within 500 kilobases [78]) and are regulated by epigenetic mechanisms. Figure from Pilling *et al* 2012 [23].

Large scale peripheral (blood-derived) gene expression studies in humans are now feasible due to the rapid expansion of methods such as microarray technology (described in the methods) in cohort studies, that assay the expression of many thousands of transcripts simultaneously. Studies by colleagues in our group which I have been involved in found that many gene expression profiles are associated with age and that the ratio of isoforms of some genes is age-related [79], and that RNA processing itself is related to ageing [80]. This suggests that regulation and "fine tuning" of gene expression transcripts and isoforms may be age-related, which is further supported by a study showing that common disease-associated variants are localised to regulatory DNA [81].

The whole blood transcriptome of an individual is stable over a time period of several months [82]. Although this study had a small sample size (n=22) the authors found that there was high within-subject correlation (>0.94 Spearman correlations) for all probes, while some differed seasonally. This re-enforces the suggestion that blood-based gene expression has the potential to become robust biomarkers, for example in the diagnosis of diseases such as Alzheimer's, where a panel of expression markers can identify cases and controls with reasonable (>75%) accuracy [83], [84]. Mechanistically, the blood-brain-barrier (that under normal conditions keeps the milieu in the brain separate from that in blood) degenerates in ageing and is correlated with cognitive decline [52]; therefore observing markers correlating with neurodegeneration in the blood is plausible. Gene expression has also been analysed in cancer cells (for instance in gastric cancer [85]) and can differentiate between different types and grades of cancer; this is typically done by measuring a panel (from tens to hundreds) of genes from the cancerous tissue. In breast cancer there are studies showing that using expression of just 30 ion channel genes has prognostic value [86].

The relevance of studying blood in relation to age-related diseases and phenotypes is clear; not only is chronic inflammation (immunosenescence) a widely accepted hallmark of ageing [87] but evidence from mouse-models of both age-related sarcopenia (loss of muscle function) and neurodegeneration (cognitive impairment) show that circulating factors drive

some of the changes, and that the tissues have the inherent capability to regenerate given the correct signal [4], [46]. Additionally, it is possible to estimate people's *biological age* (an individual's relative functioning to chronologically age-matched peers) using whole-blood gene expression only, and identify those who are biologically older than their age-matched peers [88], admittedly with limited precision. These individuals have lower muscle strength and higher levels of inflammation, consistent with the ageing phenotype. So the use of peripheral (whole blood) gene expression for observational analyses is logical and provides a novel avenue for research into biomarkers and molecular mechanisms of common diseases and traits in humans.

Two of the studies (**Chapter 3** and **Chapter 5**) described in this thesis were performed at a time when relatively few population cohorts had gene expression measures on many samples, so replication at the time was not possible. For **Chapter 4** a meta-analysis including over 7000 participants is presented, and in **Chapter 6** collaboration with the Framingham Heart Study was possible. At the time of writing in 2014 we are now in the position to participate in a number of these large meta-analyses, in particular in the CHARGE (Cohorts for Heart and Ageing Research in Genomic Epidemiology) consortium. The seminal CHARGE gene expression analysis of age included more than 15,000 participants in >12 studies [76], and many age-related phenotypes and diseases are currently under examination.

1.3.3 Epigenetics

In order to achieve the breadth of cellular phenotypes possible from a single genome, with huge amounts of adaptability and plasticity, there are numerous regulatory systems that tightly control when and by how much transcripts are expressed (substantially simplified in **Figure 1.3**), which ultimately regulate RNA polymerase II [89]. Epigenetics refers to the ability of various cells to maintain stable gene activity over time and through many replication

events without any change to the sequence of the DNA [90]. There are a variety of mechanisms used by cells to control whether genes are transcribed including, but not limited to; DNA methylation (DNAm), histone modifications, chromatin structure, and even non-coding RNA molecules that affect translation post-transcriptionally [91].

DNAm is the addition of a methyl group to the 5th carbon of a cytosine in DNA, making 5-methylcytosine (5mC). There are other modifications (such as 5-hydroxymethylcytosine, 5hmC) that could have distinct functions [92], but currently research is limited as to the importance of this modification. DNAm occurs at cytosine-guanine dinucleotides (abbreviated to CpG; cytosine-phosphate-guanine), and is part of the developmental process distinguishing lineages of cells. There are growing bodies of evidence linking methylation of specific CpG's with diseases and phenotypes, for instance in cancer (albeit with small sample sizes) [93], type-2 diabetes (T2D) [94] and age itself [95].

Chromatin (and histones) provide the three-dimensional scaffold around which DNA is wound, which not only allows compression of DNA but also regulates "accessibility" of DNA to translational machinery, which in turn affects expression [96]. Studies have shown that ageing is associated with reduced chromatin "organisation", possibly as a result of molecular damage [97], [98].

Although epigenetic data is not incorporated into these studies it remains entirely relevant to future work, as gene expression associations observed in the studies may be causally linked to specific epigenetic mechanisms.

1.4 Advancements in model organisms

"Model organism" is a term used to encompass any organism studied in a laboratory environment to dissect biological mechanisms [99]. There are several strengths to this approach, not least that model organisms live much shorter lives than humans (so experiments take less time), but also experiments not ethically viable in a human population can be performed on model organisms, and the experimental environment can be controlled and manipulated. Additionally, the genetics of the organisms can be controlled and manipulated, and access to samples from multiple tissues and time points are relatively easy to acquire and study. However, weaknesses include inconsistent translation to human data, at least partly due to the genetically homogeneous populations used as models that are not representative of the complex interactions present in wild populations, but also due to intrinsic inter-species biological differences [68].

1.4.1 Blood-borne factors affect ageing phenotypes

One of the most critical discoveries made possible by rodent models is that rejuvenation of tissues (particularly muscle, liver and brain) of aged animals can be achieved by exposure to blood from a younger animal, first discovered in 1972 [100]. This is possible using a technique known as parabiosis, where two organisms are surgically attached so they share a circulatory system, and are therefore exposed to the soluble factors therein [101]. By attaching an old mouse to a young mouse ("heterochronic parabiosis") it is possible to observe the physiological effects of exposing aged tissues to a young circulating environment [6]. There are a number of beneficial effects on the old mouse, including improved regenerative capacity of muscle stem cells [102], and negative effects on the young mouse, such as decreased synaptic plasticity and cognitive function [4]. These experiments in the genomic era now provide an unprecedented opportunity to discover the

exact factors in the circulatory system that rejuvenate old mice, and conversely which factors from old mice inhibit young tissues.

Reviewed in 2014 by Bitto and Kaeberlein [6], two factors in particular have been highlighted by a number of recent experiments; growth and differentiation factor 11 (GDF11) levels in the blood decrease with age, and exposing old mice to GDF11 improved muscle function, neurogenesis, and a number of other physiological measures [3]; conversely levels of C-C motif chemokine 11 (CCL11) in blood increase with age, and exposing young mice to CCL11 impairs neurogenesis, muscle function, and tissue regeneration [4]. Although human evidence is currently limited, studies such as those performed in this thesis hope to shed light on circulating mechanisms and biomarkers associated with ageing phenotypes.

1.4.2 Models of longevity

In ageing research analysis of many different species have highlighted conserved evolutionary pathways related to longevity between species [103], and not only this, that these mechanisms can be altered to influence the longevity of a species [104]. This suggests that these interventions are either altering the amount of “damage” done to the organism that causes ageing, or the “repair” of the damage is improved. However, the translation to human population genetics is very limited [68].

The most effective intervention that increases lifespan in model organisms (from fruit flies to mice) is “simply” restricting the caloric intake of the organisms [105]. This appears to affect a number of pathways, including mTOR and insulin signalling. Recently completed parallel experiments in Rhesus Macaques (a type of Old World monkey) at two institutes found conflicting results, possibly due to differences in control monkeys; the study that found a significant survival advantage in calorie-restricted monkeys fed the control animals ad libitum, whereas the study with no effect limited the food intake of the controls [106]–[108].

Human evidence is inconclusive, but there are suggestions of cardiovascular benefits in observational studies [109].

Senescence in cells - permanent cell-cycle arrest – is hypothesised to contribute to senescence at the organismal level (ageing) as cells begin to express a Senescence-Associated Secretory Phenotype (SASP) which is pro-inflammatory and potentially oncogenic [59], [110]. Senescence in cells can be caused by multiple factors including DNA damage, telomere shortening, and oxidative stress, and is hypothesised to have evolved as an anti-tumour-progression system that removes cells from the cell-cycle if sufficiently stressed [111]. In 2011 a seminal study was published in which the authors used a mouse-model of accelerated-ageing and a transgene which allowed inducible elimination of senescent cells (those which express high amounts of $p16^{INK4a}$) to show that removal of senescent cells from tissues *in vivo* delays ageing-associated phenotypes [112]. Whilst the study has limitations (the ageing-accelerated mice are not a natural population) and as yet no practical use (due to the nature of the genetically modified mice) the implication is quite clear; that removing senescent cells may present a viable treatment option for some age-related diseases or traits, after a lot more research. It is worth noting that although the mice were protected from some age-related pathologies (sarcopenia, cataracts, and loss of subcutaneous fat [112]) the mice did not actually live longer; this may be a manifestation of the age-accelerated nature of the mouse model, or it could be that reducing senescent cell build-up in peripheral tissues may increase healthspan (improved muscle function and eyesight, for instance) specifically, rather than lifespan [110].

Mouse models of particular phenotypes are used extensively in biomedical research to understand mechanisms; for instance in **Chapter 3** (examining muscle strength) the findings parallel a mouse model which has a particular pathway reduced in macrophages that causes a muscle-loss-like phenotype [113]. Similarly in **Chapter 5** (examining cognitive function) the literature contained information on mice with accelerated AD-like symptoms that were concordant with the results [114]. The utility of model organisms in research has clear value,

taken with the results of observational evidence in humans and an awareness of the limitations of both types of research.

1.5 Summary

The variability in the ageing process, affected by many factors including genetics, exposures, and lifestyle choices, make it inherently hard to predict; common indicators of ageing include loss of muscle strength, declines in cognitive ability, and sustained chronic inflammatory load. The specific molecular pathways and mechanisms remain to be elucidated in humans, although promising targets have been identified in mouse models, and epidemiological studies are necessary to understand what occurs in natural human populations during ageing.

Studies of gene expression in whole blood could identify biomarkers useful for determining "biologically old" individuals at highest risk of frailty, and highlight plausible biological mechanisms when accompanied by evidence from laboratory models and cell-culture experiments. Expression of genes is at the intersection between the physiological measures, genetic factors, and environmental exposures, and could prove crucial to many age-related traits and diseases. Of particular importance is the identification of biomarkers with prognostic value, and identifying mechanistic pathways driving or responding to age-related changes.

1.6 Aim and objectives

The aim of this thesis is to contribute to the existing knowledge base with exploratory analyses into whole blood transcriptome alterations in age-related traits, *in vivo* in human populations, aiming to elucidate mechanisms and discover biomarkers. I have analysed three common well-established ageing phenotypes that are easily measured, not disease-specific, and for which predictive markers would be of great value: muscle strength, cognitive function, and inflammation. The studies are hypothesis generating, therefore in each paper the majority of time is spent attempting to interpret the associations observed; hypothesising *why* the associations are observed. Many similar publications in the literature are sorely lacking in the interpretation of their results and simply report associations; this work aims to go further.

This was achieved by the completion of a number of epidemiological analyses in the InCHIANTI study, a human population study of ageing, and where possible with replication in other studies. This hypothesis-free approach accompanied by extensive reading of the literature has led to valuable contributions to the knowledge-base. The objectives of the individual studies were as follows, and correspond to the data chapters of this thesis;

3. To identify the transcripts in human whole blood samples most closely associated with muscle strength, a known marker of frailty,
4. To lead a meta-analysis within the CHARGE consortium investigating the transcriptome associations with muscle strength, and their relation to ageing,
5. To identify the transcripts most closely associated with cognitive ability, known to decline with age,
6. To identify the transcripts most closely associated with inflammation, specifically the age-related increase in chronic inflammatory load,

Chapter 2 – Methods

Each data chapter 3-6 is an individual body of work in the form of a published paper, and therefore have methodology sections relevant to that work. However, due to journal requirements the methods are described in brief; here I will discuss them in more detail, but may in part seem repetitive in order to be clear.

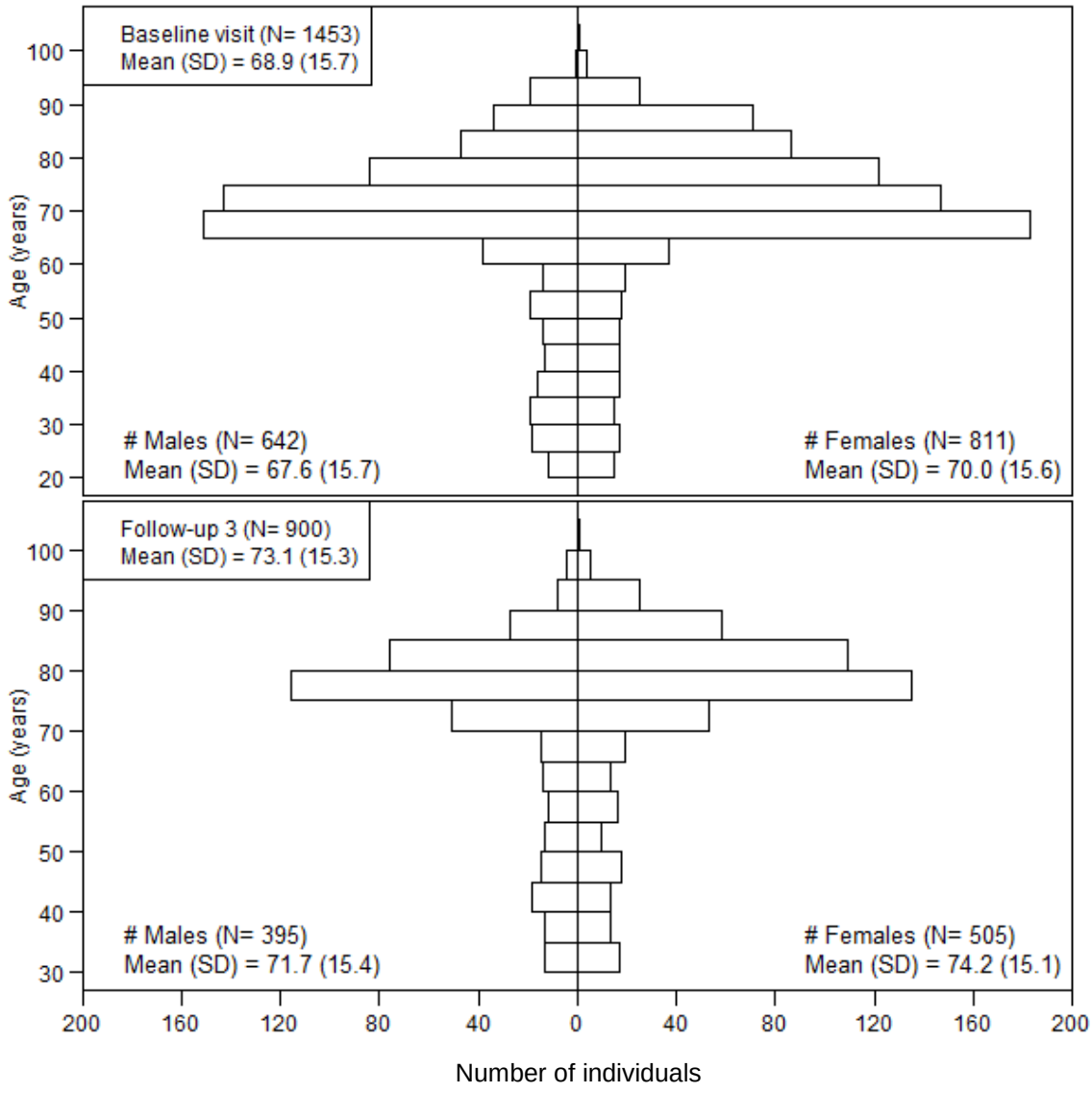
2.1 The InCHIANTI study (Invecchiare Nel Chianti)

The InCHIANTI study of ageing included 1453 members of the Italian population at two sites (Greve in Chianti, and Bagno a Ripoli) in the Tuscany region of Italy between 1998 and 1999 for their baseline visit [115]. The participants were invited back every three years, and in 2014 the participants returned for their fourth follow-up visit. From the onset this was a study designed to investigate the reduced mobility of elderly people, and as such at baseline "650 persons aged 65 years or more was drawn from the population registry […] then, random samples of 50 men and 50 women were drawn from the age strata 20-29, 30-39, 40-49, 50-59, 60-64 years" (**Figure 2.1a**) [116]. The analyses presented here utilise data from the third follow-up visit (2007/08), where the mean age of the participants is 73 years (range 30-104, **Figure 2.1b**). Once available, the 2014 data will used for prospective analyses.

A high response rate to surveys imply more accurate data, as using the results of analyses to generalize about the population requires a representative sample (that is, the responders do not differ from the non-responders). The response rate for the InCHIANTI study was very good, particularly in the elderly; of 1270 persons aged ≥65 who were chosen randomly from the population register, 1154 were eligible (i.e. had not died or moved away) and chose to participate (91.6%). The response rate was lower for people aged ≥85 (78%), possibly due to

mobility problems or illness, and was also lower for people aged <65 (69.4%), possibly due to other commitments (participants were required to spend several days with the investigators).

Figure 2.1 | The age distribution of males and females in the InCHIANTI study



The plot shows back-to-back histograms portraying the distribution of ages for males and females separately in the InCHIANTI study. The top panel (A) shows the baseline distributions (1998/1999), the bottom panel (B) shows the participants in 2007/2008 during their third follow-up visit after baseline. SD=standard deviation.

InCHIANTI is an ideal resource for the analyses presented in this thesis; the cohort covers a wide age range, and is rich in older individuals and phenotypes relevant to age-related research. These include cognitive assessments (to determine those with cognitive impairment/ dementia), "frailty"-related phenotypes (muscle strength, walk-speed, etc), and a battery of blood markers including interleukin-6 (IL-6) and broad cell counts (including neutrophils, lymphocytes and monocytes). The primary measures used in this thesis are presented with summary statistics in **table 2.1**, for the individuals who participated in the follow-up 3 examination, and have gene expression and cell-count data (n=695; see **section 2.2** for inclusion information).

Table 2.1 | InCHIANTI cohort summary statistics

		All participants	Aged <65	Aged 65:84	Aged ≥85	p-value
N	*n (%)*	695 (100)	161 (23.2)	429 (61.7)	105 (15.1)	
Sex (females)	*n (%)*	382 (55)	82 (51)	228 (53)	72 (69)	9.8E-03
Age	*min : max*	30 : 104	30 : 64	65 : 84	85 : 104	
	mean (SD)	72.2 (15.3)	47.6 (10)	77.5 (4.2)	88.4 (3.9)	
Smoking status	*Never, n (%)*	381 (54.8)	77 (47.8)	234 (54.5)	70 (66.7)	3.2E-03
	Former, n (%)	240 (34.5)	41 (25.5)	171 (39.9)	28 (26.7)	4.1E-01
	Current, n (%)	74 (10.7)	43 (26.7)	24 (5.6)	7 (6.7)	1.1E-09
Stature	*Height (cm)*	161 (9.8)	167 (8.9)	160 (9)	155 (9.3)	6.3E-24
mean (SD)	*Weight (Kg)*	70 (13.8)	74 (15.8)	71 (12.5)	62 (12.7)	1.3E-10
	Waist (cm)	95 (12.1)	91 (13.7)	97 (10.9)	93 (12.2)	2.1E-02
Cell counts	*Neutrophils (%)*	57.5 (9.1)	55.1 (8.4)	58.2 (9.1)	58.3 (9.6)	1.3E-03
mean (SD)	*Lymphocytes (%)*	30.9 (8.7)	34 (7.7)	30.1 (8.7)	28.9 (8.7)	2.9E-07
	Monocytes (%)	8 (2.1)	7.5 (1.8)	8.1 (2.2)	8.3 (2.1)	1.7E-03
	Eosinophils (%)	3.2 (2.1)	2.9 (1.8)	3.1 (1.9)	4 (3.1)	2.2E-04
Hand grip strength (Kg)	*Males: mean (SD)*	37.6 (11.6)	51 (9.5)	34.4 (7.8)	25.4 (7.4)	4.2E-46
	Females: mean (SD)	21.6 (6.9)	30.1 (5.3)	20.4 (4.9)	15.5 (5.1)	1.1E-49
MMSE *	*mean (SD)*	25.6 (5.6)	29.2 (1.6)	25.5 (4.6)	20.3 (8.6)	1.4E-39
	MMSE <24, n (%)	144 (20.9)	3 (1.9)	87 (20.4)	54 (54)	5.5E-23
	MMSE ≥24, n (%)	544 (79.1)	158 (98.1)	340 (79.6)	46 (46)	2.3E-26
Interleukin-6 (pg/mL) ^	*min : max*	0.38 : 13	0.38 : 13	0.58 : 13	1.2 : 13	
	mean (SD)	3.8 (3)	1.98 (1.6)	3.99 (2.9)	5.86 (3.6)	3.5E-28

* for MMSE 688 of 695 participants have this data; ^ for interleukin-6, the maximum value was capped at 13 pg/mL
p-value from ANOVA

2.2 Gene Expression Quantification

Upon collection of the whole blood samples from the participants, 2.5ml of whole blood was stored in PAXgene Blood RNA tubes to preserve the levels of mRNA transcripts, which would otherwise degrade [117]. PAXgene tubes contain reagents that protect RNA molecules from degradation by RNases and prevent further gene expression [118]. RNA was then extracted using the PAXgene Blood mRNA kit (Qiagen, Crawley, UK), which includes the following steps; centrifugation to remove cellular debris, addition of buffers and proteinase K to digest proteins and stabilise pH, addition of ethanol to improve binding conditions, and finally brief centrifugation in a RNA spin column which selectively binds the RNA, allowing contaminants to wash through. DNase enzymes digest any residual DNA to reduce contamination, and the resulting RNA was stored for later analysis [118]. All of the PAXgene processing was performing in Italy within days of collection (RNA is stable for at least 3 days after collection [117]). Although variation was minimised as much as possible, stabilisation of RNA has many possible sources of error; plastic and glassware was eliminated of any residual RNases, all work was performed in an RNase-free environment, and contamination by bacteria or fungi is possible so aseptic technique was followed [118].

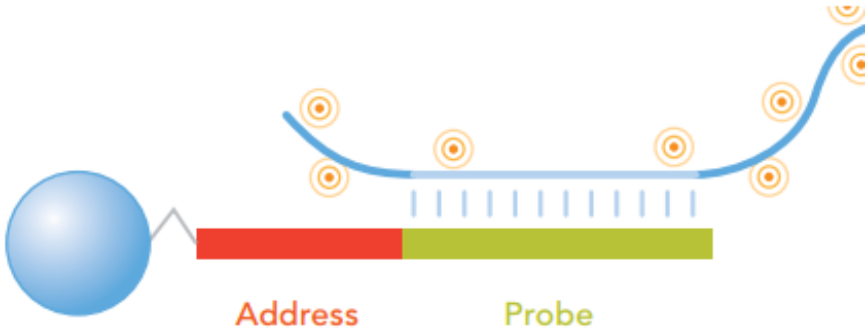
The development of array-based technologies has revolutionised the types of analysis that can be conducted using data from a single experiment. Although second-generation sequencing (or next-generation, as it was known) is now quicker, more accurate, and more affordable than ever before [119], the use of microarrays is still the prominent method for large human cohorts [69], due to their comparative cost and laboratory time. Gene expression arrays are used to determine the relative amounts of many different cDNA (complementary DNA, derived from RNA) molecules in a sample, and I have made extensive use of these during the studies described in this thesis.

2.2.1 Illumina HumanHT-12 v3 Expression BeadChip Kit

In 2008, during the third follow-up visit ("year 9") of the InCHIANTI study participants, the current Illumina gene expression array (the Illumina HumanHT-12 v3 Expression BeadChip) was used for relative abundance estimation of the RNA in the peripheral blood of the participants. Professor Andrew Singleton's group at the National Institute of Aging (NIA) in the USA carried out the wet lab work for the arrays. I will here describe the chip using information from the data sheets provided by Illumina [120].

Conceptually the arrays work as follows; there are a number of 'probes' bound to the array (each made of 50bp sequences of DNA) that hybridise (bind) to complementary strands of nucleotides, such as cDNA (PCR-amplified RNA) from a sample. Once hybridisation has completed the array is put onto a scanner that excites fluorescent markers only on those probes with bound cDNA; therefore the degree of fluorescence reflects the relative abundance of the cDNA in the sample (and therefore the relative amounts of RNA in the original blood sample). This is shown in **figure 2.2**.

Figure 2.2 | Showing a sequence of cDNA binding to a probe on an array.



The cDNA molecule (blue) matches the probe sequence, binds to it and fluoresces in the Illumina scanner. Additional "address" sequences are included for technical reasons (systematically identifying probes). "Beads", such as on the left, make up the "surface" of the array to which the probes are bound; many beads are used in a single array. Figure from Illumina Inc. [121]

Described next are the various processes and pipelines used to ensure data quality (compiled from [79], [122]), which were carried out by members of Professor Tim Frayling's group in Exeter; Illumina's BeadStudio software was used to 'normalise' the arrays to remove as much technical variability between the experiments (arrays) as possible. Each probe is given a corresponding 'detection p-value' which indicates whether the probe was determined to have fluoresced to a significant degree above the 'background' level (noise). P-values < 0.01 were taken as significant expression – values above this were not deemed to be significantly different from the background. We also excluded 'outlier' arrays (subjects) that appeared to have unusually high or low expression (fluorescence) intensities; arrays were excluded from analyses if the mean intensity for the whole array was greater or less than 3 standard deviations from the cohort median. Probes were excluded from further analysis if they were significantly expressed in <5% of the participants. Of the 710 subjects that had array profiling on their peripheral blood RNA, 698 passed this quality control procedure, with 16571 of the 48803 probes also going forward into future analyses. The array data is publicly available via the Gene Expression Omnibus (unique ID: GSE48152) [123].

The data that are then passed to analysts comes in matrix format, with 'Probe IDs' on one axis, and participant identifiers on the other. The probes are then used like any other variable in an epidemiological analysis.

2.2.2 Affymetrix microarray comparison

In chapters 4 and 6 of this thesis we collaborated with the Framingham Heart Study (FHS), which used a different microarray platform, the Affymetrix GeneChip Human Exon 1.0 ST. This differs from the Illumina HumanHT-12 array which has 48803 probe sequences; the Affymetrix Exon array has over 5 million, with multiple probes mapping to each exon in recognised genes that combine into probesets (www.affymetrix.com). This allows more robust estimates as the reliance on individual probe data is reduced, and allows gene-

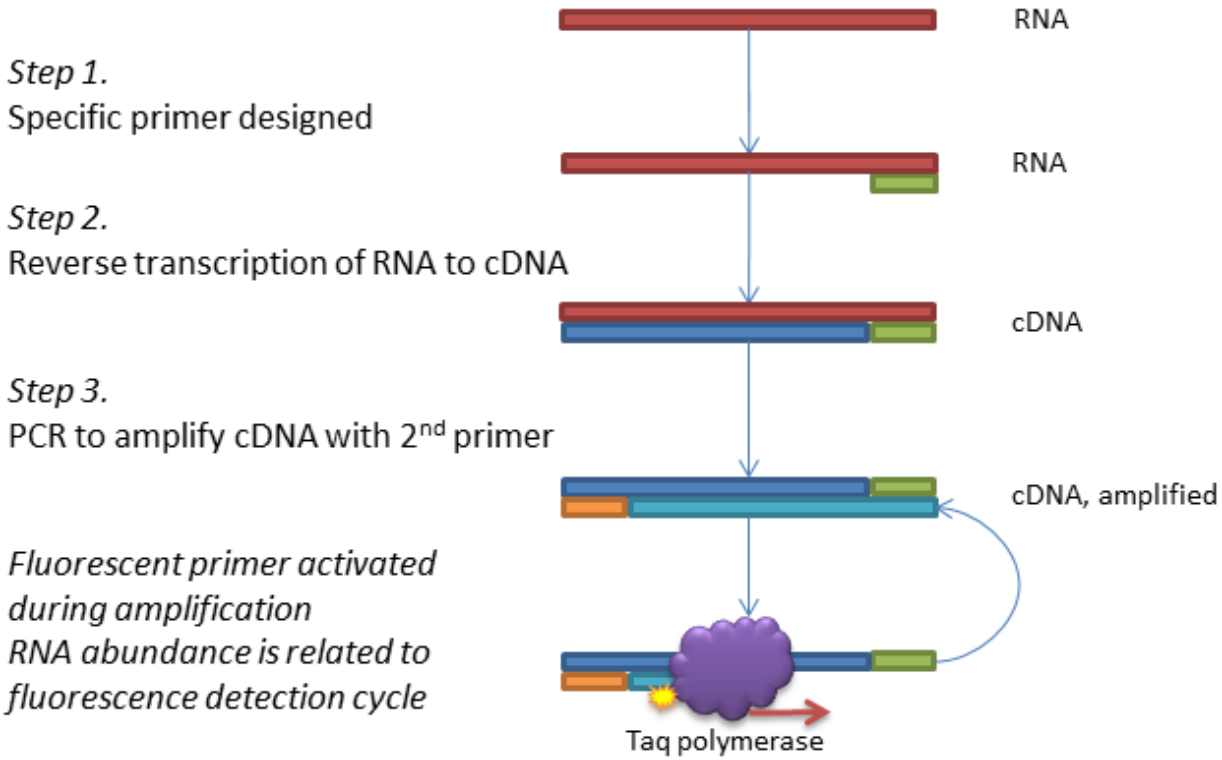
averages to be computed for comparison to the Illumina data. 17,876 gene-level probesets are available from this platform.

2.2.3 Quantitative Real-Time PCR

It is widely considered that technical replication of microarrays using quantitative real-time polymerase chain reaction (qRTPCR) techniques is essential, as the microarrays, although very useful, are inherently less accurate than qRTPCR (<80% agreement) [124]. qRTPCR concurrently amplifies (replicates) cDNA molecules in a sample and quantifies the abundance of specific transcripts using pre-designed sequences (probes). The Applied Biosystem's TaqMan Low Density Array (TLDA) qRTPCT system was utilised throughout the studies comprising this thesis, as it has up to 384 wells, so that many genes can be measured in multiple samples using qRTPCR, and it uses minimal resource (the isolated and archived RNA from previous participant visits) [125]. **Chapters 3**, **5** and **6** all used TLDA qRTPCR to validate results; **Chapter 4** was a meta-analysis with multiple cohorts and microarray platforms, so PCR validation was not sought.

TLDA quantifies RNA by first reverse-transcribing it into complimentary DNA (cDNA) then amplifying (replicating) the cDNA using polymerase enzymes and detecting the fluorescence of specifically designed probes that only fluoresce if a specific sequence is present and being processed by the polymerase (schematically represented in **figure 2.3**). This process is repeated (cycled) 40-50 times, and the stage at which fluorescence is detected (the “cross-over”, or Ct) is proportional to the amount of RNA in the sample. The comparative Ct approach was used to determine the relative expression of target genes compared to stable control genes, and finally normalisation allows multiple samples to be compared [126].

Figure 2.3 | Schematic representation of quantitative real-time polymerase chain reaction



This simplified schematic represents the process by which the relative abundance of a specific RNA transcript is estimated in a sample by TLDA qRTPCR. Fluorescence occurs in step 3 when the fluorescent probe is digested by the *Taq* polymerase during the elongation phase (when the polymerase is replicating the sequence). This PCR amplification step is repeated (cycled) through 40-50 times; RNA that is more abundant in the original sample will fluoresce at an earlier cycle than a less abundant RNA molecule. Information compiled from multiple sources [127], [128].

2.3 Statistical analysis

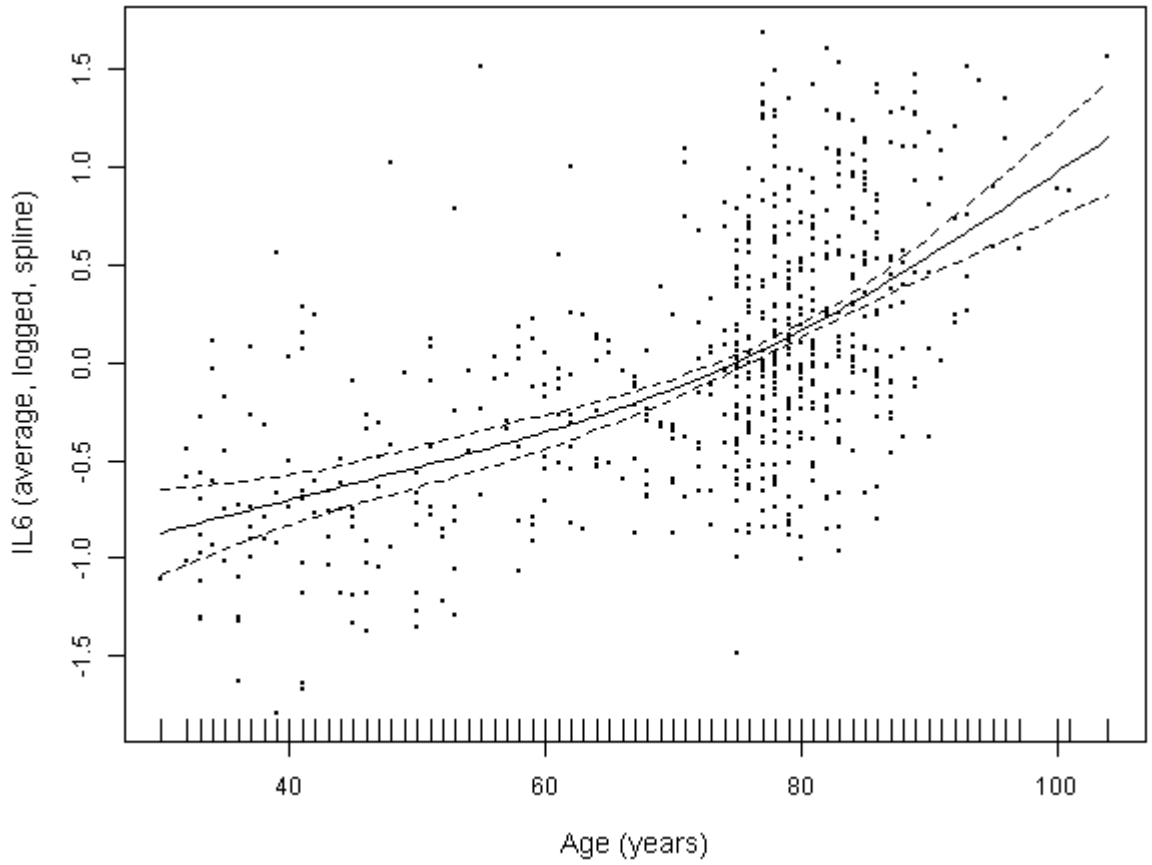
To gain insight into relationships in the large and complex data sets inherent to genomic epidemiology we need to apply statistical techniques. In this thesis I have focussed on cross-sectional (one time point), gene-expression array data (16,571 probes/variables) in a cohort of human study participants (n=698, age range 30-104).

2.3.1 Regression Analysis

For the primary analysis each gene (probe on the array) is assessed for its association with an outcome using linear or logistic regression, as appropriate (for continuous or binary outcome variables). It is often necessary to log-transform the expression probe and other biological traits (such as IL-6) to introduce a Gaussian distribution (many biochemical traits are right/positively skewed). Regression models then test for linear associations, which highlight correlations between specific gene expression probes and "gross" phenotypes of the individuals, such as muscle strength and circulating IL-6 levels. Findings highlighted by this approach are then re-analysed in more detail, such as returning to the raw data and performing various sensitivity analyses, for example stratification and checking for outlier individuals.

Regression analysis allows rigorous adjustment for potential confounding factors in the analysis, so it can be determined whether an observed association is simply due to correlation with another trait. For instance, in the analysis of gene expression and cognitive function we included age and sex as confounders (as well as technical covariates) so that the findings were independent of age and sex. In a number of analyses it has been necessary and interesting to investigate the linearity of the association; in a linear regression model it is assumed that the association between *x* and *y* is linear, but this may not truly be the case. A *spline* model can be fitted to the data which plots a series of linear lines (connected by points known as *knots*) through the values which is then "smoothed" to show a curve, if one exists. Confounding factors can also be taken into account (see **figures 2.4** and **5.3** for examples).

Figure 2.4 | A cublic spline plot showing the association between IL-6 and age in InCHIANTI.



The graph shows the association between age and IL-6 (logged) by cubic spline regression, adjusted for sex, waist circumference, smoking status (pack-years) and highest education level attained ($p=4.9x10^{-35}$). IL-6 is here expressed as "relative difference from the mean", as it has been adjusted for the confounders. N knots=4.

Choosing appropriate confounding factors to adjust for is important in an epidemiological analysis, which can be categorised into "biological" and "technical". Firstly there are biological differences between men and women with regards to ageing (e.g. muscle strength), so "sex" was taken into account. Due to the wide age-range included, "age" was also included as a continuous cofactor to determine the transcripts associated with the phenotype *independent* of age (except in models where mediation with age was tested).

Depending on the analysis, “stature” may also be taken into account (e.g. height and weight, or waist circumference). Smoking-status is known to significantly affect whole blood gene expression, so this is also taken into account [129].

Technical covariates include “study site” (the urban and rural sites for InCHIANTI), and batches; the blood RNA processing was performed in multiple batches, and there are significant differences in (for example) mean expression between some of the batches, so these are included as factor variables. Lastly I also adjust for “relative cell counts”; the rationale is that the RNA profile of a whole blood sample is dependent on both the expression in specific cell types and the relative abundance of the different cell types to one another. This is actually a strength of the study, as the expression changes we identify may be proxies for changing cell types with certain phenotypes; our future work aims to address this. There are also measures of absolute numbers of the cells, however because the gene expression values are calculated on the microarray as fluorescence relative to the baseline (control genes) expression on each specific array we reasoned that the proportions of the cell types to one another is more relevant than absolute values.

2.3.2 Multiple testing

When performing many parallel analyses in order to reduce the likelihood of false-positive associations we must adjust the threshold required for an association to be termed “significant”. Traditionally “statistical significance” is when the likelihood of observing a result as extreme or more so than the one observed is less than 5% [130], but this is only true for a single hypothesis. When testing 16571 associations if we used the 5% threshold we would expect 1 in 20 false-positive associations purely by chance (i.e. 829 of 16571 would be $p<0.05$).

Several methods for controlling the false-discovery rate (FDR) have been proposed, and two have been used in these analyses; firstly the Bonferroni correction provides a new adjusted threshold for significance (0.05 / N tests performed) which is highly conservative (that is, susceptible to false-negative findings; this is because it assumes all the tests are completely independent, when in fact expression profiles may be correlated [131]). Secondly the Benjamini-Hochberg (BH) method (or similar) is used in most of the analyses due to the improved trade-off between false-positive and –negative findings [132]; BH provides FDR-corrected p-values that are based on the p-value provided by the model, the *rank* of the p-value (when ordered by significance), and the number of tests overall. Therefore rather than just defining a new threshold (as per Bonferroni), the BH method is actually based on the p-values themselves, the distribution (ranking), *and* the number of tests.

2.3.3 Mediation Analysis

In analysis 4 (**Chapter 6**) the methods differ from the outline above; I determined the gene expression profiles associated with age-related inflammation by performing mediation analyses. There is a strong positive association between age and IL-6 which is associated with a number of age-related morbidities and mortality [38]; by assessing the proportion of the age~IL-6 association mediated by each gene transcript I aimed to determine the specific genes statistically related to this process. This is a novel analysis method in this context.

Methodologically it is straightforward to determine the proportion of an association that is mediated by another factor when using regression-based techniques. By performing two models and examining the change in the coefficient between the exposure and outcome after adjustment for the proposed mediator; if it is reduced, then a proportion of the association has been statistically accounted for by the additional independent variable. However this same procedure is also used to determine a significant confounder, so it is essential to have a biologically-informed conceptual model prior to the analysis.

For an individual proposed mediator this method is sufficient. However in our analysis we assessed the proportion of the age~IL-6 association mediator by each gene transcript individually, and thus desired a method that not only provided the "proportion mediated" but also an estimate of the statistical significance of the mediation, in order to adjust for multiple testing. Two methods were available for this; the Sobel test derives a p-value from the variance in coefficients and standard errors between the models [133]; the second available method uses re-sampling of the data (bootstrapping) to derive confidence intervals [134]. I used the bootstrapping method in **Chapter 6** of this thesis to determine confidence intervals because this empirically tests the robustness of the association in subsets of the whole sample, rather than making assumptions distributional assumptions, as in the Sobel test [135], [136]. In order to control for multiple testing we adjusted these confidence intervals according to the Bonferroni method, because it minimises the likelihood of a false-positive finding due to the highly conservative threshold, and because implementation of the BH method to confidence-limits is not, to our knowledge, possible. Bonferroni adjustment to the confidence intervals was applied using the following formula (after personal correspondence with the authors of the 'mediation' R package);

$$(1 - 0.05 / 16571) * 100 = 99.9997\%$$

Any gene expression mediator with these adjusted confidence intervals not crossing zero was considered significant and tested for replication.

2.4 Statistical packages: R and STATA

For statistical analysis of data I have used both R and STATA, as different situations have required. Both have their advantages and disadvantages; briefly, STATA is most useful when exploring cohort phenotype data given the easier command structure and graphical interface; R though has many advantages over STATA, not least its object-oriented nature that allows multiple datasets to be loaded, manipulated, and analysed in parallel. R therefore lends itself to use on servers (for “big data”, “-omics” scale analyses), and is highly flexible due to its open-source nature; many programs and functions have been written by others to perform complex analyses, each of which can be dissected by the user as desired. Additionally, its most natural environment is a Unix system, another highly flexible and practical working environment.

2.5 Linux and Perl

Unix is an operating system environment with commands and file structures which provide the user with a simple and efficient platform for processing files and running programs. Linux (one of the ‘flavours’ of Unix) is free, completely open source, and is widely used in academia. For instance the systems biology high-performance computer cluster on the University of Exeter Streatham campus (“Zeus”, which I was fortunate enough to have access to during the course of my research) runs a version of Linux known as “Scientific Linux v6.3” (maintained by research labs including CERN) [137]. Perl is a scripting language that is, in essence, a more intuitive and compact way of using Unix to parse data [138].

The combination of Linux running both Perl and R meant that all analyses, no matter the scale, were easily carried out using high-performance servers, and due to the script-based nature of the workflows these are 100% reproducible. Throughout my research career I have made particular effort to back up the scripts and data used for analyses, and to extensively

comment on my codes. A selection of scripts and workflows utilised and developed during these projects are available on the accompanying CD-ROM.

2.6 Text-mining the literature for genes

For the muscle strength meta-analysis (**Chapter 4**) I developed a program in Perl that systematically searches the literature for genes and specific search terms. This functions in two parts; firstly GeneCards (www.genecards.org [139]) hosts a list of all publications relating to every gene, including those with older or out-dated gene symbols (this can be very problematic as names are updated, so this resource is very valuable), then the script downloads this list of publications and uses NCBI PubMed (www.ncbi.nlm.nih.gov/pubmed) to save the title and abstract into a database. Once the database of all publications for each gene is created, the 2nd part of the pipeline can be used to mine this database for specific search terms using regular expressions (programmatically searching a piece of text for a specific string or pattern). For instance, in **Chapter 4** I created a database for each of the significant genes associated with muscle strength and searched for terms “muscle” and “inflam” to see, for each gene, how many publications also mention those terms in the title or abstract. The program can also search for pairs of terms, such as “muscle AND inflam”, to see how many publications for each gene contain both of those terms in the title or abstract. For each positive finding the abstract is extracted and a web-page created so that, for each gene, I can view the abstracts containing the search terms (e.g. “muscle”) and quickly build a summary of how they relate.

2.7 The CHARGE consortium

Large meta-analyses are now becoming possible as many population study cohorts take advantage of the relative ease and affordability offered by microarray technologies to measure many thousands of transcripts in hundreds or thousands of samples. The CHARGE consortium (Cohorts for Heart and Aging Research in Genomic Epidemiology [140]) was formally initiated in 2008 for GWAS meta-analyses with the approval of a set of principles and procedures to improve the reliability of GWAS findings. A number of working groups were initiated to study different outcomes, which now number over 40, with phenotypes including; adiposity, age of menopause, blood pressure, cardiovascular disease, depression, and many more. In 2011 a gene expression working group was formed with the aim of determining in detail the peripheral gene expression associations with age.

The analysis of gene expression and age involved over 14 cohorts (>15,000 samples, including 698 from InCHIANTI), across multiple countries and continents, and utilising different array technologies [76]. This should give the most robust results from the meta-analysis, as the "true" gene expression associations with age (irrespective of sample or technological bias) should be identified. Almost 1,500 genes were found to be associated with age, substantially more than the 295 found in the InCHIANTI analysis alone [79].

The CHARGE gene-expression and age analysis was seminal in the field and created guidelines for following analyses. A large number of phenotypes are now under investigation, including; C-reactive protein (CRP), tobacco exposure, blood pressure and lipid levels, not to mention the analysis of muscle strength included as part of this thesis (see **Chapter 4**). These cohorts are now extending their phenotyping to include DNA methylation arrays, which will lead on to the next generation of analyses and understanding of complex traits and diseases.

Chapter 3 – Analysis 1 – Muscle Strength

CCAAT-enhancer-binding protein-beta expression in vivo is associated with muscle strength

Aging Cell (2012); 11:2. 262-268

Harries, Lorna W

Pilling, Luke C

Hernandez, L Dena G

Bradley-Smith, Rachel

Henley, William

Singleton, Andrew B

Guralnik, Jack M

Bandinelli, Stefania

Ferrucci, Luigi

Melzer, David

3.1 Overview

In this analysis of peripheral gene expression associations with muscle strength we report a single positive association between the gene *CEBPB* and hand grip strength. Grip-strength was chosen as it is commonly used in the gerontological field as a broad measure of muscle strength – a trait strongly linked with frailty. The aim of the analysis was to discover any genes expressed in peripheral immune cells that were associated with “general” muscle strength.

For this analysis I undertook the bioinformatic analyses (transcriptome-wide screen, validation of phenotype, investigation of cell-type effects), performed the literature review, and worked with the co-authors to interpret the results and write the manuscript.

3.2 Summary

Introduction: Declining muscle strength is a core feature of aging. Several mechanisms have been postulated, including CCAAT/enhancer-binding protein-beta (C/EBP-β) triggered macrophage-mediated muscle fibre regeneration after micro-injury, evidenced in a mouse model. We aimed to identify *in-vivo* circulating leukocyte gene expression changes associated with muscle strength in the human adult population.

Methods: We undertook a genome wide expression microarray screen, using peripheral blood RNA samples from InCHIANTI study participants (ages 30-104 yrs). Logged expression intensities were regressed with muscle strength using models adjusted for multiple confounders. Key results were validated by real-time PCR. The Short Physical Performance Battery score (SPPB) tested walk speed, chair stand and balance performance.

Results: *CEBPB* expression levels were associated with muscle strength (beta coefficient = 0.20560, $p=1.03*10^{-6}$, false discovery rate q=0.014). The estimated handgrip strength in 70 year old men in the lowest *CEBPB* expression tertile was 35.2 kg compared to 41.2 in the top tertile. *CEBPB* expression was also associated with hip, knee, ankle and shoulder strength and the SPPB performance score (p=0.018). Near study-wide associations were also noted for *TGFB3* ($p=3.4*10^{-5}$, q=0.12) and *CEBPD* expression ($p=9.67E^{-5}$, q=0.18) but not for *CEBPA* expression.

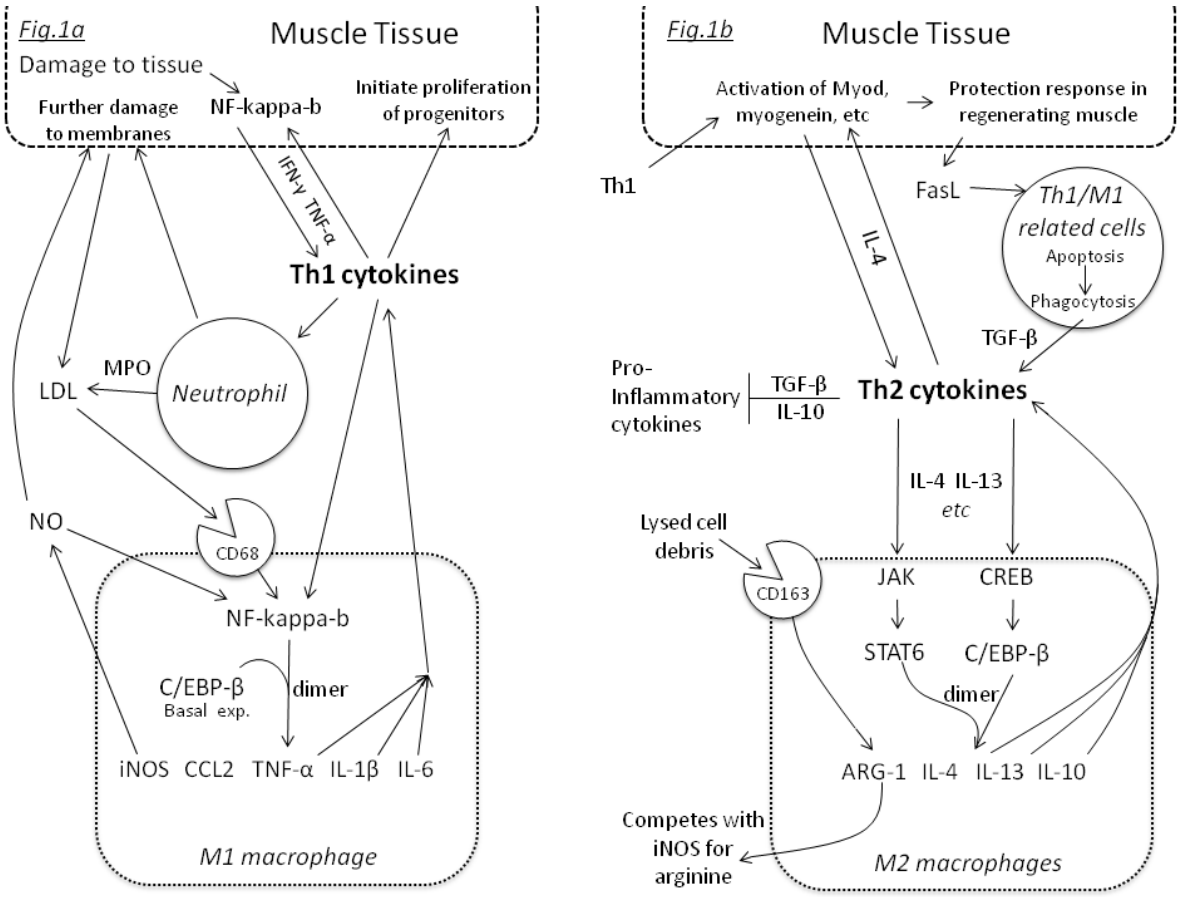
Conclusions: We report here a novel finding that raised *CEBPB* expression in circulating leukocyte derived RNA samples in-vivo is associated with greater muscle strength and better physical performance in humans. This association may be consistent with mouse model evidence of *CEBPB* triggered muscle repair: if this mechanism is confirmed it may provide a target for intervention to protect and enhance aging muscle.

3.3 Introduction

Aging is associated with a gradual loss of muscle mass and strength [141] accompanied by increasing susceptibility to injury and decreased capacity to repair and regenerate muscle [142]. Severe muscle loss (or 'sarcopenia') can eventually result in impairments and limitations to physical functioning. In 1995 Falkner et al [143] argued that 'everyday' contraction-induced injury contributed to muscle aging. Other suggested causes for sarcopenia include loss of motor unit and local motor neurone changes, but also pathways with systemic elements such as hormone / growth factor signalling, and inflammation [141].

Muscle repair after injury or change of use starts with an initial acute inflammatory response (typically from 0 to 48 hours) followed by a repair and regeneration phase, typically lasting from 48 hrs to 10 days [144]. Following injury, local muscle tissue secretes pro-inflammatory (Th1) cytokines (in particular TNF-α and IFN-γ, but also IL-1β, IL-6 and others) (see Fig. 3.1a for simplified representation of the pathways from the published literature) that instigates an innate immune response. Neutrophil invasion enhances inflammation at the injury site [144], [145], and is followed by local migration of macrophages, monocytes [146] and T-helper type 1 (Th1) lymphocytes. Monocytes are then available to differentiate into specific macrophage sub-types [147], which are essential for effective muscle regeneration [148].

Figure 3.1 | Macrophage polarization during muscle injury/repair



Macrophages have a repertoire of phenotypes ranging from classical activation, to several alternative activation states. These are generally termed pro-inflammatory / M1 (a) and anti-inflammatory / M2 (b) respectively. They are functionally distinct, and the correct differentiation at the right time is critical to an effective regeneration process. Drawn from the literature, this diagram represents that main factors involved in the classical (a) and alternative activation states (b) of macrophages as the time from wounding progresses. The review by Tidball et al, 2010, was particularly informative.

Key: LDL = low-density lipid, Th1 = T-helper type 1 / pro-inflammatory response, Th2 = T-helper type 2 / anti-inflammatory response.

At the late stage of the pro-inflammatory phase there is increased production of IL-4, IL-13 and Transforming Growth Factor-beta (TGF-β) from T-helper type 2 (Th2) lymphocytes; phagocytosis of necrotic tissue also influences this change in phenotype. These cytokines inhibit the production of Th1 cytokines and initiate the accumulation of alternatively activated (anti-inflammatory / M2) macrophages, which include wound-healing and regulatory phenotypes [149], thereby promoting tissue repair [150] (Fig. 3.1b). As inflammation is attenuated, myoblast differentiation is induced and new muscle fibers are formed.

CCAAT-enhancer-binding protein-beta (C/EBP-β) has a pivotal role in these processes. In 2009, Ruffell et al showed in a mouse model that deletion of two CREB binding sites from the *CEBPB* promoter prevented C/EBP-β induction and prevented the transition from M1 to M2 macrophage-specific gene expression during muscle repair [113]. In their model, the pro-inflammatory response was unaffected but muscle fiber regeneration was severely compromised, with the injured area showing a fibrotic appearance, with fewer, smaller and calcified myofibers. Although *CEBPB* is expressed at basal levels in M1 macrophages, significant (typically 10 fold) up-regulation of *CEBPB* above constitutive levels is required for M2 activation, and therefore muscle repair [151].

Another pathway suggested as pivotal in muscle regeneration is Wnt signalling; age-related loss of muscle strength due to increased fibrosis was associated with increased systemic Wnt signalling in aged mice [152]. Conversely, others have found that Wnt signalling induced the myogenic specification of resident myeloid stem cells during muscle regeneration [153], therefore it is likely that an optimal trade-off may be reached, but that it changes over the lifespan of an organism.

While muscle repair has been studied in depth ‘in-vitro’ and in animal models, “in-vivo” human studies are intrinsically more difficult. In theory, gene expression patterns could be studied in serial muscle biopsies but this would be too invasive and would interfere with repair processes. In this study, we instead examined expression in blood (leukocyte) samples collected with immediate stabilization of RNA, so that results reflect *in-vivo* mRNA

expression. Blood derived white cell transcriptome studies have already proved valuable in identifying signatures of major diseases and drug responses [154]. Studying systemic factors in muscle strength may be particularly informative, as experimental transplantation of old muscle to a young rat host results in regeneration as marked as that seen in young muscle, suggesting that circulating factors may be critical [46].

In this study, we aimed to test the relevance of *CEBPB* and other transcripts to muscle strength *in-vivo*. The InCHIANTI population study (www.inchiantistudy.net) is particularly useful for this as it includes a wide age-range and muscle strength measures at multiple sites.

3.4 Methods

3.4.1 Samples

InCHIANTI [115] is a population-based, prospective study in the Chianti area (Tuscany) of Italy. The participants were enrolled in 1998-2000, and were interviewed and examined every three years. Ethical approval was granted by the Instituto Nazionale Riposo e Cura Anziani institutional review board in Italy. Participants gave informed consent to participate.

3.4.2 RNA collection, extraction and the whole transcriptome scan

Peripheral blood specimens preserving *in-vivo* RNA expression were collected in 2008/9 from 733 participants. RNA was extracted from peripheral blood samples using the PAXgene Blood mRNA kit (Qiagen, Crawley, UK) according to the manufacturer's instructions. This ensures the transcript expression is as near to its *in-vivo* levels as possible. Whole genome expression profiling of the samples was conducted using the Illumina Human HT-12 microarray (Illumina, San Diego, USA) as previously described [155]. Data processing was done using the Illumina and Beadstudio software (Illumina, San Diego, USA) as previously described. All microarray experiments and analyses complied with MIAME guidelines. Data from 698 individuals and 16571 probes passed our quality control process; 12 subjects were excluded on the basis that their mean signal intensities across all probes with p = <0.01 were >3 standard deviations from the cohort mean; probes with <5% of subjects giving intensities with p >= 0.01 different from background were also excluded.

3.4.3 Hand-grip strength

Maximum hand-grip strength was measured using a grip dynamometer, recording the maximum force generated (kg) during two attempts with each hand. In our analysis we used the participant's highest measured grip strength for either hand.

3.4.4 Statistical analysis

Expression data were available for 707 individuals. Seven individuals were excluded following subject level QC steps, where all intensity data were missing or +/- 3 standard deviations from the mean (leaving 698 individuals). The relationship between gene expression and hand-grip and other muscle strength was tested mainly using linear regression models with natural log-transformed gene expression level as the dependent variable. Separate regression models were fitted for each of the 16,571 expressed probes. We used the false discovery rate (FDR) to account for multiple testing. R (statistical computing language) v2.8.1 was used for large-scale analyses and STATA v10.1 for confirmation and additional exploration.

Regression models were adjusted for potential confounding factors on gene expression: age; gender; lifetime pack-years smoked (in five categories: none, less than 20 years, 20 to 39 years, 40 plus years, and missing); waist circumference (as a continuous trait); highest level of education attained (in five categories: none, elementary, secondary, high school, and university / professional); study site (individuals were drawn from a rural village [Greve] and an urban population [Bagno a Ripoli]); and the proportion of leukocyte cell types (neutrophil %, lymphocyte %, monocyte %, eosinophil %). We also controlled for potential hybridization and/or amplification batch effects in all our analyses.

The Short Physical Performance Battery score is a summary of performance on walk speed, time to stand up from the sitting position several times (chair stands) and graded balance

tests [156], and SPPB scores are predictive of disability progression and mortality. Physical activity was classified based on the MET (Metabolic Equivalent) [157] approach into: Hardly any activity; Mostly sitting/some walking; Light exercise 2-4 hrs/week; Moderate 1-2 hrs or light >4 hrs/wk; Moderate exercise >3hrs/wk; Intense exercise many times/wk; Walks 5+ km/day, 5+days/wk.

3.4.5 TaqMan Low Density Array (TLDA) validation of microarray results

We selected a subsample of 100 male subjects with middle tertile lymphocyte percentages for validation of the *CEBPB* expression data. Total RNA (30 to 170 ng) was reversed transcribed in 20µl reactions using the Superscript III VILO kit (Invitrogen, Paisley, UK), according to the manufacturer's instructions. *CEBPB* expression levels for each target transcript were then measured using the Taqman Low Density Array (TLDA) platform (Applied Biosystems, Foster City, USA), using commercially available assay *CEBPB*-Hs00942496_s1. Reaction mixes contained 50µl 2x TaqMan universal master mix (no AMPerase) (Applied Biosystems, Foster City, USA), 30µl dH_2O and 20µl cDNA template. Cycling conditions were 50°C for 2 minutes, 94.5°C for 10 minutes followed by 50 cycles of 97°C for 30 seconds and 57.9°C for 1 minute. *CEBPB* expression was measured in triplicate for each sample. Relative gene expression levels were calculated using the comparative Ct technique using the StatMiner relative quantification software developed for analysis of TLDA plates (Integromics, UK). Gene expression levels were calculated relative to the geometric mean of the 3 genes – *COQ2*, *NFKB1* and *TNF* - whose transcript expression was most stable across all samples, as identified by the GeNorm function of the StatMiner software (Integromics, UK).

3.4.6 Gene Set Enrichment Analysis

GSEA is an alternative method of analysing RNA-expression data, in which gene-sets are specified by the user (can be custom, or downloaded from the GSEA site) and the association across the whole set of genes (usually related functionally) is tested against a phenotype. We downloaded the 16 Gene Ontology [158], [159] pathways in which CEBPB is involved, which included ‘inflammatory response’, ‘defense response’, ‘response to stress’ and ‘RNA metabolic process’. See [79] for detailed methods.

3.5 Results

Six hundred ninety eight people were included in the analysis, ranging in age from 30 to 104 yrs (**Table 3.1**). Nearly 45% of the sample was male, and 54% had been non-smokers. In linear models adjusted for multiple potential confounders (see **methods 3.4**), *CEBPB* transcript expression emerged as the most strongly associated with handgrip muscle strength: beta-coefficient = 0.20560, $p=1.03*10^{-6}$ (**table 3.2**) (for table of top 250 regression results see **Supplementary Table 1**; for QQ-plot see **Supplementary Figure 3.1**). *CEBPB* was also the only transcript to reach study wide significance accounting for multiple statistical testing of array data (false discovery rate FDR q-value=0.014). Associations with muscle strength for 6 other transcripts were likely to be robust (i.e. with a FDR q-value<0.2, indicating that >80% of these are valid associations). Notable in these was *TGFB3* (p=3.4*10-5, q=0.12) and *CEBPD* expression (p=9.67E-05, q=0.18). Relatively little is known about the four other transcripts *(TAAR8, EXOC1, CBWD5 and CA433556-CD40L).* There were no associations with other measured *CEBP*- transcripts, namely *CEBPA*, *CEBPZ*, *CEBPG* and *CEBPE* (for all p>0.18).

3.5.1 CEBPB expression and grip strength: effect sizes

We focussed on *CEBPB* expression, being the only gene to reach study wide significance with muscle strength. First, we explored the associations between *CEBPB* expression and white blood cell counts. In regression models adjusted for the same terms as previously but testing each white cell sub-type separately, we found a strong negative association between *CEBPB* expression and increasing lymphocyte percentage (standardised beta=-0.50, p<0.001) and a marginally less strong positive association with increasing neutrophil percentage (standardised beta=0.47, p<0.001). There was no association with circulating

monocyte percentage. We tested the model for interactions between *CEBPB* expression and lymphocyte percentage for handgrip muscle strength: the interaction term was not significant (p=0.250).

Table 3.1 | Characteristics of the study sample

Study sample characteristics		
	n	**%**
Age (years)		
30yrs – 65yrs	163	23.35
66yrs – 85yrs	444	63.61
86yrs - 104yrs	91	13.04
Gender		
Men	313	44.84
Women	385	55.16
Site		
Greve	342	49
Bagno a Ripoli	356	51
Education		
None / elementary	416	59.61
Secondary	93	13.32
High school	87	12.46
University / Professional	102	14.61
Pack years smoked (lifetime)		
None	379	54.30
<20	153	21.92
20-39	101	14.47
40+	55	7.88
missing	10	1.43
Waist Circumference (cm)		
Mean (sd)	698	95.3 (12.1)
Leukocyte fraction (%)	***Mean***	***SD***
Neutrophils	57.5	9.14
Lymphocytes	30.8	8.66
Monocytes	7.97	2.13
Eosinophils	3.17	2.11
Basophils	0.55	0.21
Leukocyte fraction (n K/µL)	***Mean***	***SD***
Neutrophils	3.68	1.30
Lymphocytes	1.94	0.76
Monocytes	0.50	0.16
Eosinophils	0.20	0.15
Basophils	0.01	0.03

Table 3.2 | Probes most closely associated with hand grip-strength (natural log kg), with a false discovery rate q value <20%

Rank	Probe	Gene	Strength Coefficient	95% CI	p-value	q-value
1	*ilmn_1693014*	*CEBPB*	0.20560	0.14 to 0.27	1.03E-06	0.014416
2	*ilmn_1772523*	*TAAR8*	1.06986	0.66 to 1.48	2.46E-05	0.123658
3	*ilmn_1687652*	*TGFB3*	1.25827	0.76 to 1.75	3.40E-05	0.13609
4	*ilmn_1745583*	*EXOC1*	1.30954	0.78 to 1.84	5.01E-05	0.148728
5	*ilmn_1652417*	*CBWD5*	-1.22926	-1.73 to -0.73	5.38E-05	0.150724
6	*ilmn_1890206*	*CD40L*	1.24223	0.73 to 1.76	7.74E-05	0.171274
7	*ilmn_1782050*	*CEBPD*	0.18314	0.11 to 0.26	9.67E-05	0.182561

Note: models adjusted for age, gender; lifetime pack-years smoked; waist circumference; education; study site; proportions of leukocyte cell types plus RNA hybridization and amplification batch.

To provide estimates of the unadjusted effect sizes, we show the distribution of handgrip muscle strength by tertile of *CEBPB* expression in men and women separately (**figure 3.2**) for those in the middle tertile of both neutrophil and lymphocyte percentages (n=160). Using our fully adjusted models, we estimated grip strength for 70 year old men or women (**table 3.3**). This showed similar results to the unadjusted data, with mean handgrip strength in men in the highest *CEBPB* expression tertile being 41.2 kg compared to 35.2 kg in the bottom tertile. In women, there was a similar spread of predicted mean grip strength (from 26.0 to 20.0 kg respectively).

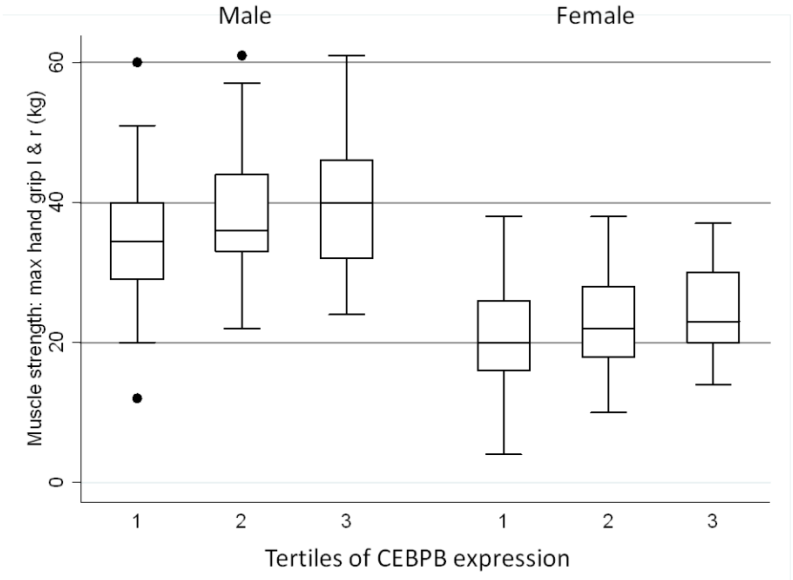
Advancing age is associated with lower hand-grip strength (beta=-0.5, p<1e-10) and separately with lower *CEBPB* expression (beta=-0.15, p<1e-4): however age (beta=-0.04, p>0.05) ceases to be significantly associated with grip-strength when CEBPB expression is included as an explanatory covariate.

Table 3.3 | Regression model based estimated hand grip strength for men and women aged 70 and with lymphocyte percentage at mean value (30.8%)

	Men		Women	
C\EBP-β expression tertile	Estimated mean strength (kg)	95% Prediction interval (kg)	Estimated mean strength (kg)	95% Prediction interval (kg)
1	35.2	24.1 to 46.3	20.0	9.1 to 31.0
2	38.2	37.1 to 39.4	23.1	22.1 to 24.0
3	41.2	29.4 to 53.0	26.0	14.0 to 37.9

Note: models adjusted for age, gender; lifetime pack-years smoked; waist circumference; education; study site; proportions of leukocyte cell types plus RNA hybridization and amplification batch.

Figure 3.2 | Demonstrating increased *CEBPB* expression with higher grip strength



Using a box-plot and tertiles of *CEBPB* expression, the association with grip strength is clearly visible. *CEBPB* is split by gender because of the natural disparity in muscle strength. Note: participants with 'normal' (mid-tertile) leukocyte percentages. Adjacency lines are 1.5*the interquartile range above/below the 75th and 25th percentiles. Key: CEBPB = CCAAT/enhancer-binding protein beta.

To confirm the microarray result we re-measured *CEBPB* expression in 100 male InCHIANTI subjects. Very similar results were found plotting tertile of *CEBPB* expression against grip strength (**Supplementary Figure 3.2**): those in the top expression tertile had higher mean strength compared to those in the bottom tertile (*CEBPB* expression correlation with handgrip strength p=0.048, excluding 5 outlier values).

3.5.2 Associations with strength in other muscle systems and physical performance

We next examined associations between *CEBPB* expression and muscle strength measured at other sites, in models adjusted as previously. There were significant positive associations with measures at the knee, hip, ankle and shoulder (**Supplementary table 3.2**).

A critical question is whether the changes in muscle strength seen are of functional significance. Those in the highest tertile of *CEBPB* expression were far less likely to have impaired performance on the Short Physical Performance Battery (SPPB) score (**Supplementary table 3.3**): the highest expression tertile group had an Odds Ratio=0.42 (95%CI 0.20 to 0.86, p=0.018) for having impaired performance (SPPB score<10, 22.35% of the sample). There were trends for the three subscales, with a significant association for walk speed impairment (OR=0.33, 95%CI 0.14 to 0.80, p=0.014).

3.5.3 Muscle strength and physical activity

In post-hoc sensitivity analyses, we explored whether the association of *CEBPB* expression with muscle strength was explained by higher levels of physical activity. Adjusting for seven levels of physical activity over the previous one or three years had limited impact on the association between *CEBPB* expression and hand grip, which remained strongly significant

(beta=0.13, 95%CI: 0.06 to 0.19, p=0.001 and beta=0.16, 95%CI: 0.09 to 0.23, p<0.001 respectively).

3.5.4 A potential mechanism underlying the CEBPB association with muscle strength

To identify, as far as possible, the likely general mechanisms mediating the *CEBPB* association with muscle strength, we searched the Gene Ontology database [160] for pathways in which *CEBPB* is involved: this yielded 16 results, most of which are inflammatory/immune related, including acute and chronic stages of inflammation, response to stress and wound-healing. Gene ontology does not include a specific pathway explicitly including both *CEBPB* and muscle, or macrophage activation states. A Gene-Set Enrichment Analysis (see methods), which determines whether sets of biologically related genes are more strongly associated with grip strength than might occur by chance, showed no significant associations (p-values>0.05). A leading alternative pathway involved in muscle repair is *WNT* signaling, but we found no evidence that β-catenin (downstream effector of Wnt) expression was associated with grip strength in our leukocyte samples (p>0.05).

3.6 Discussion

Loss of muscle strength is a core feature of aging, and a range of potential local and systemic mechanisms have been proposed to explain it [141]. Amongst the more promising is the theory that contraction-induced damage to muscle is more common and associated with impaired repair in old age [143]. Almost every activity of daily living involves lengthening contractions, the majority of which do not induce injury: with age, however, the increasing numbers of highly stretch-susceptible fibers place the muscles of older animals and humans at greater risk [161]. A recently published mouse model with reduced *CEBPB* expression after muscle injury prevented the transition from the pro-inflammatory M1 to the M2 macrophage mediated repair and regeneration phases, and resulted in sarcopenia like loss of muscle fibers and fibrosis of muscle tissue [113]. Until now, the relevance of this experimental demonstration to muscle strength in people in-vivo was unknown.

We set out to test blood leukocyte derived gene expression associations with muscle strength in an adult human population. We found, for the first time, that increased *CEBPB* expression was strongly associated with greater muscle strength in a human population. C/EBP-β is a transcription factor playing a role in 16 pathways classified within the Gene Ontology system [160]: however, we did not find evidence of deregulation of these general (mostly inflammatory) pathways. We also found no evidence that β-catenin (downstream effector of Wnt), an alternative potential explanatory pathway influencing proliferation and differentiation of muscle-residing monocytes [153], is differentially expressed. The most likely remaining explanation for our finding appears to be that higher *CEBPB* expression occurs with more effective M2 macrophage mediated regeneration in muscle tissue; C/EBP-β is involved in both M1 and M2 expression profiles. However additional activation above basal expression, via CREB, only occurs in M2; IL-10, produced by regulatory macrophages, is dependent upon C/EBP-β-mediated transcription [162], as is Arg-1 (produced by wound-healing macrophages) [163], and others.

Clearly, further work is needed to confirm that the M2 mediated regeneration process is indeed the source of the higher *CEBPB* expression in our blood derived RNA samples. The regeneration phase after injury is relatively lengthy (8 days typically) compared to the initial pro-inflammatory phase (2 days), and therefore regeneration related signaling can be expected to predominate in population level studies of participants unselected with respect to the timing of everyday muscle fiber injury.

C/EBP-β is a transcription factor and as such its specificity of action is tightly regulated by phosphorylation and competing cofactors. The precise nature of the promoters that C/EBP-β binds to and activates (i.e. M1 or M2) is dependent on upstream effectors of *CEBPB* expression, such as TGF-β, and on whether C/EBP-β is partnered with pro-inflammatory cofactors such as NF-κB or with anti-inflammatory cofactors such as STAT6 (fig. 3.1a and 3.1b). We found both TGFB3 and 1 of 3 isoforms of TGFBr2 were significantly associated with measures of strength ($p=3.4*10^{-5}$ and $p=0.037$ respectively). As TGF-B (produced by macrophages after phagocytosis of apoptotic cells (**fig. 3.1b**)) attenuates pro-inflammatory signals [164], it's up-regulation alongside *CEBPB* is a further indication that an enhanced anti-inflammatory (M2) phenotype is advantageous for maintaining muscle strength.

There is high homology between the C/EBP members; they can form heterodimers [165], and substitute for one-another to rescue functionality [166]. However, each is separated functionally by their response to stimuli, their location within the nucleus and their tissue specificity (i.e. macrophage or muscle cell). C/EBP-δ (which was also associated with muscle strength in our study) has previously been shown to reverse the inactivation of PPAR-γ [167]; this results in the stimulation of fatty acid oxidation, glycolysis, mTOR activation and anti-inflammatory actions in macrophages in order to increase hepatic lipogenesis in the liver and adipose tissue [168]. This is due to differing ratios of saturated to unsaturated fatty acids in the blood, which promote either M1 (pro-inflammatory) or M2 (anti-inflammatory) phenotypes respectively [169]–[171]. It is therefore interesting that we found a correlation between *CEBPD* expression and muscle strength in our cohort. This may possibly be due to an effect

of *CEBPD* on circulating saturated fatty acid levels in leaner people, producing a more M2-like response and better muscle repair. Again, this will require further work to elucidate the exact mechanism.

Large scale gene-expression studies are confounded by the high variability between healthy individuals. Results can be affected by a range of factors from ethnicity to employment, including details of blood-sample storage [154], and therefore we have used a collection procedure designed to stabilize RNA at the point of collection from the participants' veins. We found the proportion of different white cell sub-types to be a significant factor when analysing *CEBPB* expression. Due to the nature of the investigation (observing increased *CEBPB* expression hypothesized to be due to, or involved with the causation of, increased anti-inflammatory macrophages) it is not surprising that varying proportions of neutrophils (and lymphocytes) between people influences *CEBPB* expression: however, we found no evidence of interactions between *CEBPB* expression, leukocyte subtype proportions and muscle strength. In addition, the *CEBPB* expression association with strength was still present in those with 'normal range' mid-tertile lymphocyte and neutrophil counts. Neutrophils are upstream of macrophages; they are involved in the initial recruitment to sites of muscle injury, and amplify the chain of events leading to successful resolution of damage [47], [144].

Along with C/EBP-β, C/EBP-δ and TGF-β, 4 other probes were associated (genome-wide) with grip-strength. These were Trace Amine Associated Receptor 8 (*TAAR8*), EXOcyst Complex component 1 (*EXOC1*), COBW domain-containing protein 5 (*CBWD5*) and Cell-surface receptor 40 ligand (*CD40L*). CD40L is associated with neutrophil count and has previously been described to be involved in inflammation [172]. TAAR8 is a receptor for trace amines, which are important in neuronal function [173]. EXOC1 and CBWD5 are proteins necessary for exocytic vesicle targeting [174] and neurotransmission respectively. CD40L may therefore play a role in M1 macrophage activation, whilst the remaining genes may reflect improved motor neuron function in people with better muscle strength.

A simple association between *CEBPB* expression and muscle strength would be of limited interest without evidence that the effect size is big enough to influence everyday functioning, impairment and disability. Lower handgrip muscle strength in 45-68 year old men in the Honolulu Heart Program was predictive of disability 25 years later [39]. The short physical performance battery (SPPB) score is a well validated and widely used objective measure of functional impairment, and is predictive of subsequent disability and mortality [36]. Our finding of an association between higher *CEBPB* expression and a markedly lower risk of impairment on the SPPB suggests that our findings may be of clinical and practical significance.

Given the mouse model demonstrating that increased *CEBPB* expression is necessary for muscle regeneration, it is plausible that the *CEBPB* expression associations reported here with muscle strength reflect a causal relationship, although the statistical association may theoretically reflect an effect rather than a cause. As noted, future work is needed to confirm that higher *CEBPB* expression in peripheral blood derived samples does indeed reflect greater M2 macrophage mediated muscle repair. Studies of responses to experimentally induced muscle stresses could chart the temporal response from early reaction to injury through to the later repair phases in people of different ages and strengths. It may be possible to explore the effects of polymorphisms (SNPs) in the implicated genes on expression and strength. Eventually, experimental approaches to enhance *CEBPB* pathway expression and related repair may provide a means of protecting and enhancing aging muscle.

3.7 Conclusions

We have presented the first report of the association between raised *CEBPB* expression *in-vivo* and greater muscle strength in humans. This may be consistent with mouse model findings that C/EBP-β signalling is critical for regeneration of muscle fibres after injury. TGF-β, a key initiator and cofactor is also up-regulated. These changes are big enough to affect functioning. Independent replication and mechanistic studies linking *CEBPB* expression in leukocytes to muscle are needed.

3.8 Acknowledgements and Contributions

All contributions are acknowledged with authorship. The authors declare no conflict of interest. This analysis was supported by funding from the Peninsula College of Medicine and Dentistry (John Bull Building, Tamar Science Park, Research Way, Plymouth, UK, PL6 8BU).

3.9 Supplementary Information

Supplementary Table 3.1: Top 250 microarray probes (genes) associated with hand-grip strength (kg)

Probe-ID	P	B	Q	Gene	Description
ilmn_1693014	1.0E-06	0.206	0.0144	CEBPB	CCAAT/enhancer binding protein (C/EBP) /
ilmn_1772523	2.5E-05	1.070	0.1237	TAAR8	trace amine associated receptor 8
ilmn_1687652	3.4E-05	1.258	0.1361	TGFB3	transforming growth factor / beta 3
ilmn_1745583	5.0E-05	1.310	0.1487	EXOC1	exocyst complex component 1
ilmn_1652417	5.4E-05	-1.229	0.1507	CBWD5	COBW domain containing 5
ilmn_1890206	7.7E-05	1.242	0.1713	CA433556	
ilmn_1782050	9.7E-05	0.183	0.1826	CEBPD	CCAAT/enhancer binding protein (C/EBP) /
ilmn_1805973	1.3E-04	1.442	0.1949	GPR19	G protein-coupled receptor 19
ilmn_1699091	1.6E-04	1.365	0.2031	C14orf118	chromosome 14 open reading frame 118
ilmn_1717579	1.6E-04	1.114	0.2037	C17orf64	chromosome 17 open reading frame 64
ilmn_1811955	1.6E-04	1.108	0.2040	PRMT5	protein arginine methyltransferase 5
ilmn_1737847	2.1E-04	1.458	0.2274	RMND5B	required for meiotic nuclear divisi cerevisiae)
ilmn_1773228	2.4E-04	1.326	0.2390	DLST	dihydrolipoamide S-succinyltransferase
ilmn_2401987	2.8E-04	1.284	0.2547	DLEC1	deleted in lung and esophageal cancer 1
ilmn_1810805	2.9E-04	1.153	0.2560	HEATR5B	HEAT repeat containing 5B
ilmn_2179018	3.0E-04	-0.296	0.2599	NDUFAB1	NADH dehydrogenase (ubiquinone) 1 / alpha
ilmn_1724145	3.9E-04	0.819	0.2798	CBX4	chromobox homolog 4
ilmn_1718136	4.4E-04	-0.159	0.2889	hCG_25371	
ilmn_1750088	4.4E-04	-0.797	0.2893	VRK2	vaccinia related kinase 2
ilmn_1676895	4.6E-04	0.790	0.2911	BTNL8	butyrophilin-like 8
ilmn_2305407	4.9E-04	0.752	0.2962	ZBTB16	zinc finger and BTB domain containing 16
ilmn_1720604	5.0E-04	1.056	0.2971	ARSG	arylsulfatase G
ilmn_1787541	5.2E-04	0.716	0.2991	SPSB2	splA/ryanodine receptor domain and SOCS
ilmn_1734290	5.5E-04	1.235	0.3027	MAPRE3	microtubule-associated protein / RP/EB f
ilmn_1730054	5.8E-04	0.449	0.3063	GSTT1	glutathione S-transferase theta 1
ilmn_2396002	6.7E-04	-0.957	0.3141	MRPL10	mitochondrial ribosomal protein L10
ilmn_1663090	6.9E-04	1.134	0.3156	SON	SON DNA binding protein
ilmn_2081682	7.2E-04	0.152	0.3181	SMAP2	small ArfGAP2
ilmn_2376403	7.4E-04	0.106	0.3190	TSC22D3	TSC22 domain family / member 3

Note: Linear regression models adjusted for age, sex, waist circumference, smoking-status, highest education level, study site, hybridization-batch, amplification-batch and leukocyte proportions

Only top 30 shown in print. Full table available on the Aging Cell website: http://dx.doi.org/10.1111/j.1474-9726.2011.00782.x

Supplementary Table 3.2: Regression based estimated of CEBPB expression associations with measures of strength

Measure	N	Mean	SD	Regression Coefficient	Lower 95% CI	Upper 95% CI	p-value
Knee							
Flexion (kg)	626	17.77	5.62	1.533	0.404	2.662	0.008
Extension (kg)	626	18.99	5.83	1.713	0.527	2.899	0.005
Hip							
Flexion (kg)	629	21.17	6.82	2.522	1.111	3.933	0.000
Extension (kg)	616	19.90	5.46	0.889	-0.330	2.109	0.153
Abduction (kg)	655	21.02	6.26	1.856	0.688	3.024	0.002
Adduction (kg)	655	19.03	5.91	1.637	0.566	2.709	0.003
Ankle							
Flexion (kg)	626	17.99	5.43	1.254	0.005	2.504	0.049
Extension (kg)	626	28.06	7.76	2.663	1.016	4.310	0.002
Shoulder							
Abduction (kg)	658	18.78	5.87	1.587	0.511	2.663	0.004

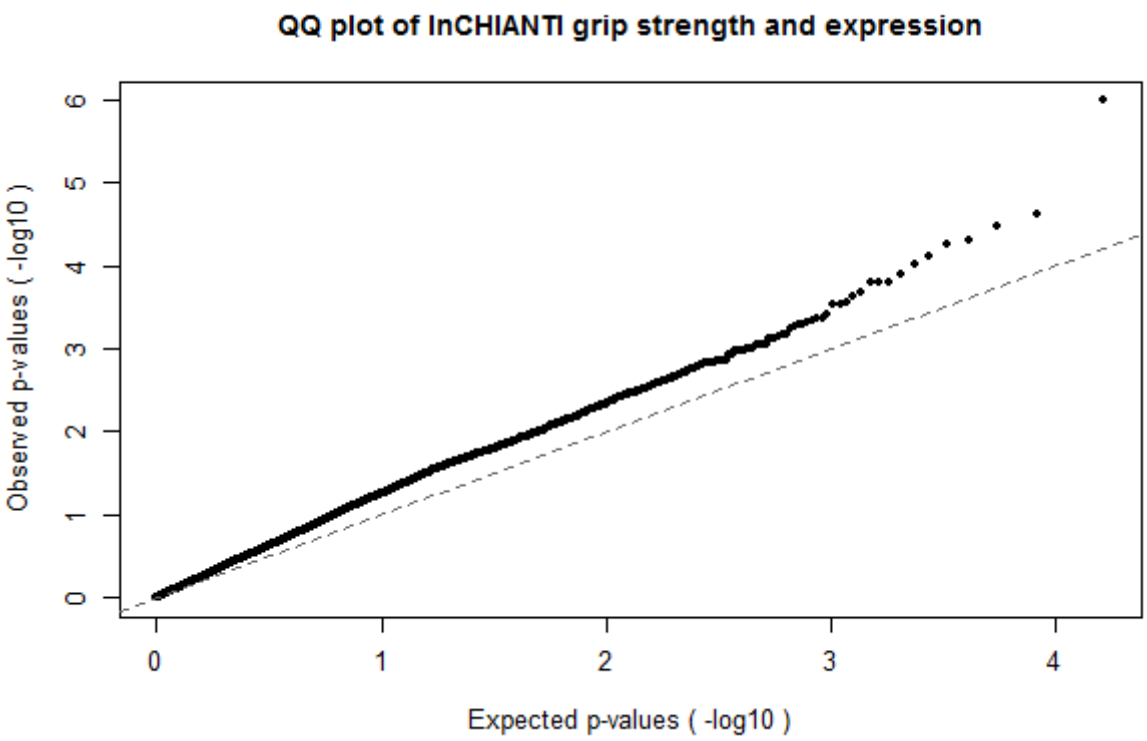
Note: models adjusted for age, gender; lifetime pack-years smoked; waist circumference; education; study site; proportions of leukocyte cell types plus RNA hybridization and amplification batch.

Supplementary Table 3.3: Odds ratio for having impaired performance on performance scores, comparing those with top tertile CEBPB expression to lowest tertile expression

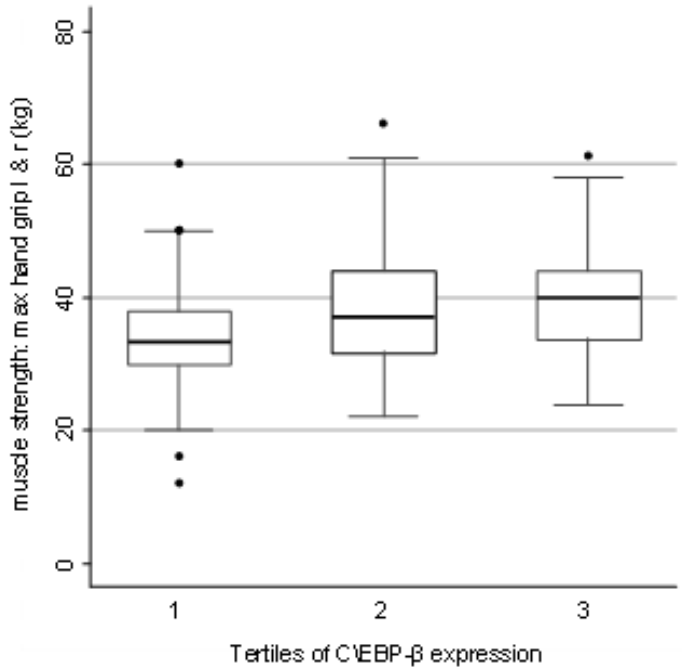
Measure	**N**	**% impaired performance**	**Odds Ratio** (top vs lowest tertile expression)	**95% Confidence interval**	**p-value**
Walking	680	15.04	0.33	0.14 to 0.80	0.014
Chair stands	690	18.19	0.63	0.29 to 1.40	0.253
Balance	690	16.62	0.46	0.19 to 1.11	0.084
Short Physical Performance Battery Score (SPPB)	690	22.35	0.42	0.20 to 0.86	0.018

Note: models adjusted for age, gender; lifetime pack-years smoked; waist circumference; education; study site; proportions of leukocyte cell types plus RNA hybridization and amplification batch.

Supplementary Figure 3.1: QQ plot for muscle strength and gene expression associations



Supplementary Figure 3.2: Microarray Validation



Hand Grip strength by tertiles of C\EBP-ß expression measured by TLDA, for 95 men (5 excluded as outliers) with middle tertile lymphocyte and neutrophil percentages (p=0.048).

Chapter 4 – Analysis 2 – Muscle Strength Meta-analysis

Gene transcripts associated with muscle strength: a CHARGE meta-analysis of 7,781 persons

Manuscript in preparation (2015)

Pilling, LC *

Joehanes, RJ *

Kacprowski, T *

Peters, MJ *

Jansen, R *

...

Pennix, BWJH #

van Meurs, JBJ #

Ferrucci, L #

Kocher, T #

Murabito, JM #

Melzer, D #

*,# *joint first and last authors, respectively*

Full author table included in ***supplementary table 4.0***

4.1 Overview

This meta-analysis of four independent cohorts assesses the association between whole-blood gene expression and muscle strength. The aim was to extend the original analysis in InCHIANTI (presented in **Chapter 3**) and identify novel associations that may provide insight into mechanism or biomarkers. 222 genes were associated with muscle strength, and more than 50% had not been linked to muscle in the literature previously. *CEBPB*, the gene identified in **Chapter 3**, did not replicate in the other cohorts.

In this analysis I performed the transcriptome-wide analysis in InCHIANTI, coordinated the collaborators to perform the analysis in their cohorts, performed the meta-analysis, developed novel methods to assess gene/phenotype associations in the literature, and worked with the co-authors to interpret the results and write the manuscript. All data acquisition was performed by the co-authors on this manuscript. Detailed author contributions are included in **Supplementary Table 6.0**.

4.2 Summary

Introduction: Muscle strength in midlife is predictive of disability and mortality in later life; loss of muscle strength is a key feature of aging. Blood factors, such as GDF11, influence muscle function and repair in mice, and inflammation is linked to muscle decline in humans. We identified circulating gene expression associations with muscle strength in adults.

Methods: Meta-analysis of whole blood gene expression and hand-grip strength in four independent cohorts (n=7,781, ages: 20-104 years, weighted mean 56 years), adjusted for confounders including age, sex, height, weight, and leukocyte subtypes. Three cohorts used Illumina HumanHT-12 microarrays, one used Affymetrix Exon-array data (gene level). Separate analyses were also performed in subsets (old/young, male/female).

Results: In the meta-analyses, expression of 222 unique genes were significantly associated with muscle strength (FDR<0.05). Significant genes include *ALAS2* (rate limiting enzyme in heme synthesis), *PRF1* (perforin, a cytotoxic protein previously associated with age-related chronic inflammation), and *IGF1R* and *IGF2BP2* (growth factor-related), but not *GDF11*. Gene Ontology pathway analysis found enrichment for “hemoglobin biosynthesis” and “innate immune activation”. 115 (of 222) genes have not previously been linked to muscle in the literature. Analyzing the Affymetrix and Illumina cohorts separately, 90 genes (40% of 222) showed independent replication (p<0.05).

Conclusions: Blood transcripts associated with muscle strength in adults include genes involved in innate immune pathways and hemoglobin synthesis, but also include novel markers needing characterization. The link between muscle strength and circulating growth factors may differ between humans and mice. Further work is needed on subgroup specific expression (especially in the frail elderly) and on long-term prediction of frailty and disease.

4.3 Introduction

Muscle strength correlates with health and physical function, and poor muscle strength in midlife is a strong, independent predictor of health status decline and mortality over 25 years [175]. Sufficient muscle strength in the hands, arms and legs is needed for everyday functioning; persons with poor strength are at high risk of disability, injury from falls, and other age-related morbidities [36], [141]. Mechanisms that lead to decline of muscle mass and strength with aging are unknown. Current theories emphasize the role of denervation not compensated by adequate re-innervation, mitochondrial dysfunction, cellular senescence, inflammation and changes in microenvironment [176], [177]. However, empirical evidence for these theories in humans is lacking. With the growing availability of genomic data it is now possible to determine if muscle strength is associated with the expression of certain genes, which might suggest pathways driving or secondary to the determinants of muscle strength in mid and later life.

Studies in mice using a technique known as heterochronic parabiosis – where an old and a young mouse are surgically attached so as to share a circulatory system – have shown that the lower muscle regenerative capacity in older compared to younger animals is due to circulating factors [46], [102], and that restoration of circulating GDF11 (growth and differentiation factor 11) improves muscle function in older mice [3]. Thus, insights into the relationship between circulating factors and muscle strength in adults could help identify biomarkers and develop interventions aimed at improving strength.

Previous studies of transcriptome associations with muscle strength in humans were conducted predominantly in muscle tissue, and are mostly based on a group of biologically plausible candidate genes and on relatively small sample sizes due to the challenges of obtaining human muscle tissue. For example one study investigated associations between circulating cytokines and expression of 44 candidate genes in muscle in 88 aged men [178], while another assessed >26,000 genes in only 51 men [179], raising severe challenges of

false positive statistical associations due to multiple testing. Pathways such as Wnt signaling [180], Kelch proteins [181], and insulin-like growth factors (IGF's) [182] are known to play crucial roles in muscle homeostasis in muscle tissue. The immune system also plays a critical role in muscle maintenance and repair after damage [47].

A transcriptome-wide study of whole blood transcript associations with grip strength conducted by the InCHIANTI Aging Study (n=679, mean age 72 years, 71% ≥72 years old) found only one gene, *CEBPB* (CCAAT/enhancer-binding protein beta, required for macrophage-mediated muscle repair in a murine model [113]), to be robustly associated with muscle strength in older humans after adjustment for confounders and multiple testing [183]; however this study had a limited sample size (n=679). A follow-up study in humans found that *CEBPB* expression increased following exercise-induced muscle damage [184].

In the present study we sought to test associations between transcripts expressed in whole blood and hand grip strength in multiple human cohorts. Our underlying aim was to identify the transcripts in blood driving or responding to muscle strength in adults, with subgroup analysis by age-group and gender. Using a robust meta-analysis framework within the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) consortium [140] we have analyzed data from four independent cohorts (n=7,781 participants) to identify the genes whose level of expression assessed by blood transcripts was associated with muscle strength.

4.4 Methods

4.4.1 Study Sample

Four cohorts were included in the discovery analysis; the Framingham Heart Study [185] (FHS, n=5,582, ages=24-90), the InCHIANTI study [115] (n=679, ages=30-104), the Rotterdam study [186] (RS, n=591, ages=46-89) and the Study of Health in Pomerania [187] (SHIP, n=986, ages=20-81) (total n=7781). The FHS study included two related generations of participants (accounted for in the statistical methodology); the FHS Generation 2 (n=2,427, ages=40-90) and Generation 3 (n=3,155, ages=24-78) cohorts were included. The Netherlands Study of Depression and Anxiety [188] (NESDA, n=1,989, ages=18-65) is also reported, but was not included in the meta-analysis due to unavailable data on white cell proportions necessary for the meta-analysis protocol. Detailed study design information can be found in the online methods and references.

4.4.2 Phenotype

The primary phenotype was hand-grip strength in kg (a normally distributed phenotype). In the FHS, hand grip strength was measured with a Jamar dynamometer with three trials performed in each hand, and the maximum of the six trials for each participant was used in the analysis. In InCHIANTI each participant recorded their maximum grip strength three times in each hand, and the maximum recorded value of the six trials was used. In the Rotterdam Study grip strength of the non-dominant hand was measured three times for each participant, and the maximum recorded value was used. In the SHIP cohort participants were asked to press the hand dynamometer firmly for several seconds, once per hand (left and right), and the maximum value was used.

Table 4.1 | Characteristics of the study cohorts

	Meta-analysis cohorts					
Variables	**FHS Gen2 ***	**FHS Gen3 ***	**InCHIANTI ~**	**RS3 ~**	**SHIP ~**	**NESDA**
N	2427	3155	679	591	982	1989
N >= 60 years (%)	1319 (54%)	48 (1%)	547 (81%)	264 (45%)	268 (27%)	158 (8%)
Sex (male), n (%)	1095 (45%)	1470 (47%)	311 (46%)	272 (46%)	435 (44%)	1328 (67%)
Age, mean±SD	66 ± 8.9	46 ± 8.8	72 ± 15	60 ± 7.9	50 ± 14	42 ± 13
Age, min:max	40 : 90	24 : 78	30 : 104	46 : 89	20 : 81	18: 65
WBC-counts	yes ł	yes	yes	yes	yes	no
Grip strength						
Mean ± SD (Kg)	31 ± 12	38 ± 12	29 ± 12	25 ± 9.4	38 ± 12	38 ± 12
Min : max (Kg)	1 : 76	5 : 84	3 : 76	2 : 55	11 : 73	10 : 90
Microarray platform	Affymetrix GeneChip Human Exon 1.0 ST	Affymetrix GeneChip Human Exon 1.0 ST	Illumina HumanHT-12 v3 BeadChip	Illumina HumanHT-12 v4 BeadChip	Illumina HumanHT-12 v3 BeadChip	Affymetrix Human Genome U219

FHS=Framingham Heart Study; Gen=generation; RS3=Rotterdam Study 3; SHIP=Study of Health in Pomerania; NESDA=Netherlands Study of Depression and Anxiety; SD=standard deviation; WBC=white blood cell

* FHS cohorts analysed together prior to overall meta-analysis

~ Illumina-based cohorts analysed together prior to overall meta-analysis

∞ cohort not included in meta-analysis due to data missing from analysis protocol

ł cell counts imputed in this dataset; see methods

4.4.3 Peripheral Gene Expression Data

Blood samples were drawn from participants and RNA was isolated, reverse-transcribed to cDNA, which was then amplified and hybridized to a microarray individually for each cohort; methods described in detail online. Briefly, the FHS used Affymetrix Human Exon 1.0 ST GeneChips, characterizing the expression of 16798 unique genes (after exclusion of probesets with relative log expression mean values <3).

The InCHIANTI and SHIP studies used the Illumina HumanHT-12 v3 Expression BeadChip Kit, and the RS used the Illumina HumanHT-12 v4 Expression BeadChip Kit, with 37,348

probes measured on both Illumina platforms (22,911 unique genes; after exclusions of probes expressed above background in <5% of participants this becomes 15,639 unique genes). Quantile normalization and log2 transformation was performed separately in each cohort according to published methods, and both probes and samples were z-transformed; details available in the online methods.

Systematic mapping of probe sequences to RefSeq transcripts found 26,746 pairs of probes (one Affymetrix Exon ST and one Illumina HumanHT-12 that mapped to the same RefSeq transcript) corresponding to 17,534 unique RefSeq gene symbols (see Online Methods for details of annotation sequence mapping and additional information on array normalization). Finally the NESDA cohort utilized the Affymetrix Human Genome U219 Array, with expression information available on 18,212 unique gene identifiers.

The Human Genome Nomenclature Committee [189] list 19,060 protein-coding genes (Sept 15th 2014), less than the total “unique identifiers” mapped by the two arrays used in the overall meta-analysis; this discrepancy is due to probes on the array mapping to non-protein-coding transcripts, which we have included under the term “unique genes” or “transcripts” in this manuscript.

4.4.4 Statistical Analysis

Using the R statistical software [190] and package “lme4” [191] each cohort performed a linear mixed effects model for each probe in their microarray data, using the probe as the outcome, muscle strength as an independent variable, and with the following confounders included as fixed effects; age, sex, height (cm), weight (kg), cell count estimates (neutrophils, monocytes, basophils and eosinophils), and fasting state (where applicable). The following covariates were included as random effects; batch (e.g. amplification and/or hybridization), study site (InCHIANTI), family structure (in FHS), and RNA integrity number (RIN), a

measure of RNA quality (where available). Empirical cell counts were only available in half of the FHS cohort; the rest of the cohort was imputed using partial least square regression methods (see online methods for more details).

4.4.5 Meta-analysis

A sample-size weighted meta-analysis method was used, where an overall p-value and Z-score for each probe are calculated which together describe the significance of the effect, and the direction and magnitude, respectively; this method was chosen over the effect size/standard error method because of the multiple array technologies and technical considerations that differed between the cohorts. The analysis was done using the Meta-Analysis Tool for Genome Wide-Association Scans (METAL) [192] which took the effect size, sample size, and p-values from the individual cohort results as input (we set the “minor allele”, “major allele”, “minor allele frequency”, and “strand” to the same for all cohorts and probes, as this method was developed for GWAS and these options are not relevant for gene expression data).

Five pre-specified analyses were performed; all study participants, those ≥60 years, <60 years, males only, and females only. For each analysis, the Illumina-based cohorts (InCHIANTI, RS, SHIP) were meta-analyzed together first, as these technologies are very similar, then a secondary meta-analysis was performed which used the FHS results and the Illumina results as the input; these are the final meta-analysis results reported. This reduced the heterogeneity in the meta-analysis due to array differences between the cohorts.

Before interpretation of the results, probes were excluded if they were expressed in <5% of the sample or if the heterogeneity p-value calculated by METAL was <0.05. The Benjamini-Hochberg (BH) [132] false-discovery rate (FDR) correction was applied to each analysis to

determine the statistically significant probes for each analysis. Validation was described as a gene with $p<0.05$ in the NESDA cohort.

4.4.6 Ontology enrichment analysis

The WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) online resource is a method for determining pathway enrichment [193]. We conducted a "Gene Ontology" analysis (database version: 11th Nov 2012) and a "Human Phenotype Ontology" analysis (database version: 20th May 2014), which uses a systematic approach to phenotype abnormalities to link them into ontologies [194]. Default analysis options were selected, including BH multiple testing adjustment, and the list of 17,534 genes included in the meta-analysis we used as the "background".

4.4.7 *A priori* genes associated with muscle function

We selected sets of genes known to influence muscle function for *a priori* analysis to highlight whether the pathways in muscle tissue are also associated with strength in whole blood. Kelch proteins, including KLHL19, KLHL31, KLHL39m, and KLHDC1, are involved in skeletal muscle function and development [181]; canonical and non-canonical Wnt signaling play crucial roles in maintenance and development of skeletal muscle [180]; insulin-like growth factors (IGF's) are also known to play roles in muscle growth and homeostasis [182], and finally supplementation of GDF11 in mice ameliorates sarcopenia [3]. In their 2007 study, Melov *et al* identified 586 unique genes expressed in muscle that were associated with endurance exercise training and differed between older and younger men [179], which we also checked for associations with muscle strength in this analysis.

4.4.8 Systematic literature search for genes

For each significant gene in the analysis a systematic search of literature was performing by accessing the “publications” list from GeneCards (www.genecards.org) [139], which has the advantage of including publications where the gene ID may have changed over time. From this list the title and abstract were downloaded from NCBI PubMed (www.ncbi.nlm.nih.gov/pubmed). Searches were then made within each publication for the text string “muscle”, and results counted.

4.5 Results

4.5.1 Cohort characteristics

Characteristics of the cohorts are presented in **Table 4.1**. Complete data for the planned meta-analysis were available for 7,781 participants from the four cohorts, namely the FHS, InCHIANTI, RS and SHIP. Sample sizes for the subgroup analyses were 1) older participants ≥60 years (n=2,402), 2) younger participants <60 years (n=5,379), 3) male participants (n=3,557), 4) female participants (n=4,224). Overall the cohorts were quite similar with respect to sex-distribution and sampling methods, differing only by age distribution and unexpectedly lower mean hand-grip strength in the RS. To cope with incomplete overlap between the two microarray platforms used (Affymetrix in the FHS cohorts, and Illumina in InCHIANTI, RS, and SHIP) we first analyzed the FHS generations 2 and 3 together, then the Illumina-based cohorts, and finally an overall meta-analysis combining the two platform-specific meta-analyses was performed on the genes that had coverage in both platforms.

4.5.2 Meta-analysis: genes associated with muscle strength

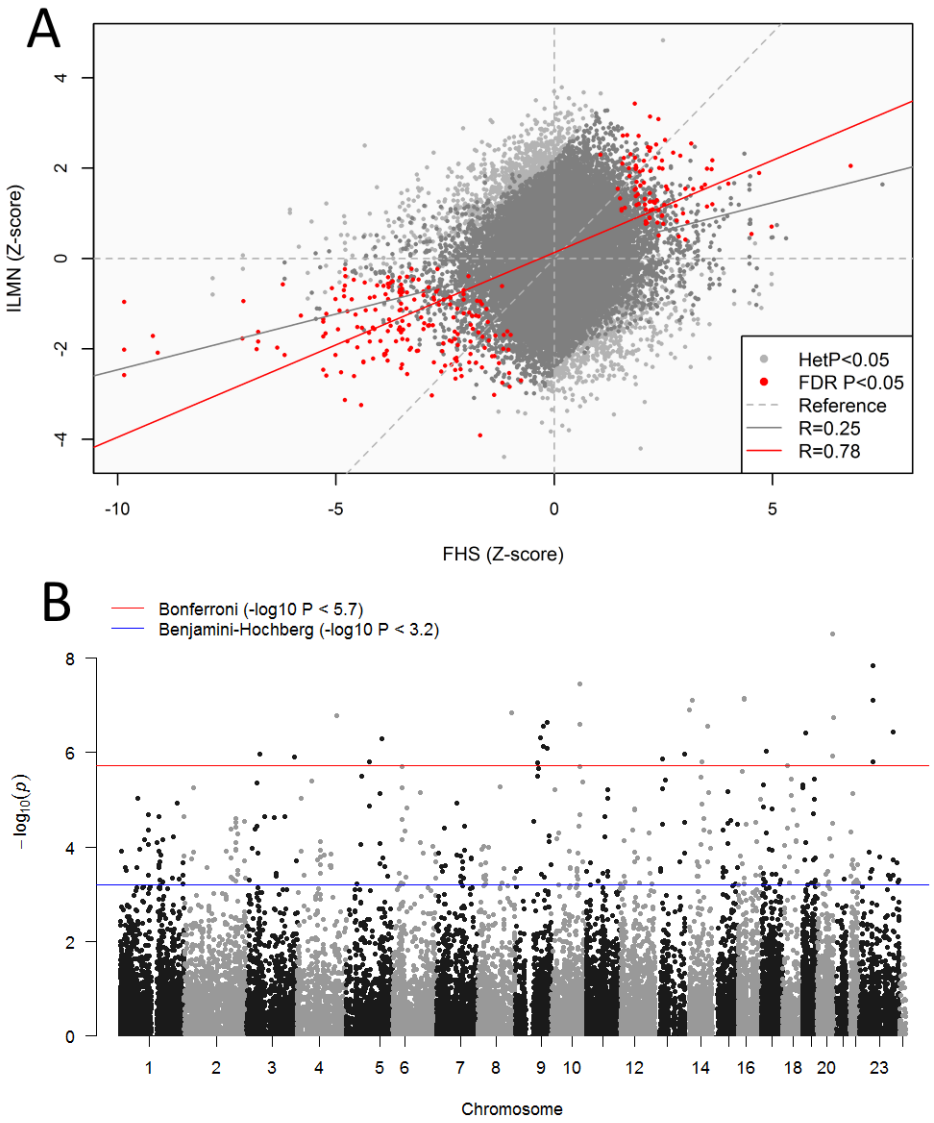
Overall, 26,746 probe-pairs (corresponding probes on the Affymetrix and Illumina platforms), mapping to 17,534 unique gene identifiers, were available for the meta-analysis. Across all five analyses (“all individuals”, “older”, “younger”, “males”, “females”) 264 probe-pairs (222 unique genes) were significantly associated with muscle strength (FDR<0.05; **Figure 4.1**; see **Supplementary Tables 4.1-4.5** for significant results in each analysis) after correction for multiple confounders, and excluding results with significant heterogeneity (het p<0.05). In the individual analysis including all participants (n=7,781) 246 probe-pairs (208 unique genes) were significant; after restricting to the older (≥60 years, n=2,402) participants no probe-pairs were significant; within the younger (<60 years, n=5,379) participants 39 probe-pairs (31 unique genes) were significant; within the male participants (n=3,557) 26 probe-

pairs (24 unique genes) were significant; and finally limiting to the female participants (n=4,224) 1 probe-pair (1 unique gene) was significantly associated with muscle strength.

Of the 222 unique genes associated with muscle strength in the meta-analysis all were significant (nominal $p<0.05$) in FHS alone, and 90 (40%) were also independently associated with muscle strength (nominally significant; $p<0.05$) in the Illumina meta-analysis. The proportion "independently replicated" was greater (21, of 30, 70%) in the top 30 most significant genes identified in the meta-analysis.

The overlap between the five analyses was moderate (**Figure 4.2**, **Supplementary Tables 4.6-4.7**). Of the 222 genes associated with muscle strength in any analysis, 176 (79%) were only associated in the "all individuals" meta-analysis at transcriptome-wide significance. In the "older" participants no genes were significantly associated (transcriptome-wide) with muscle strength, and in the "females" analysis only one gene was associated, *DEFA4*, which was only statistically significant in that subset. The other two analyses - "younger" and "males" - had 31 and 24 genes associated with strength respectively, most of which were also identified in the overall meta-analysis (see **Figure 4.2** for details). **Supplementary table 4.8** contains additional summarized information about the genes in **table 4.2** and those that overlap between subset analyses.

Figure 4.1 | Gene expression probe-pairs associated with muscle strength in all participants



A) compares the individual meta-analyses performed in the Illumina-cohorts and the FHS separately. The red points represent gene transcripts significantly associated with muscle strength (FDR<0.05). The light grey points were excluded due to significant heterogeneity (Cochran's Q-test p < 0.05 [192]). The darker grey points were not significant in this analysis. The grey line shows the trend across all the genes. The red line shows the trend across genes significantly associated with muscle strength (FDR<0.05).

B) Shows the meta-analysis results by Manhattan plot. The blue line indicates those probes significantly associated after Benjamini-Hochberg correction, the red line shows those significant after Bonferroni correction, for comparison.

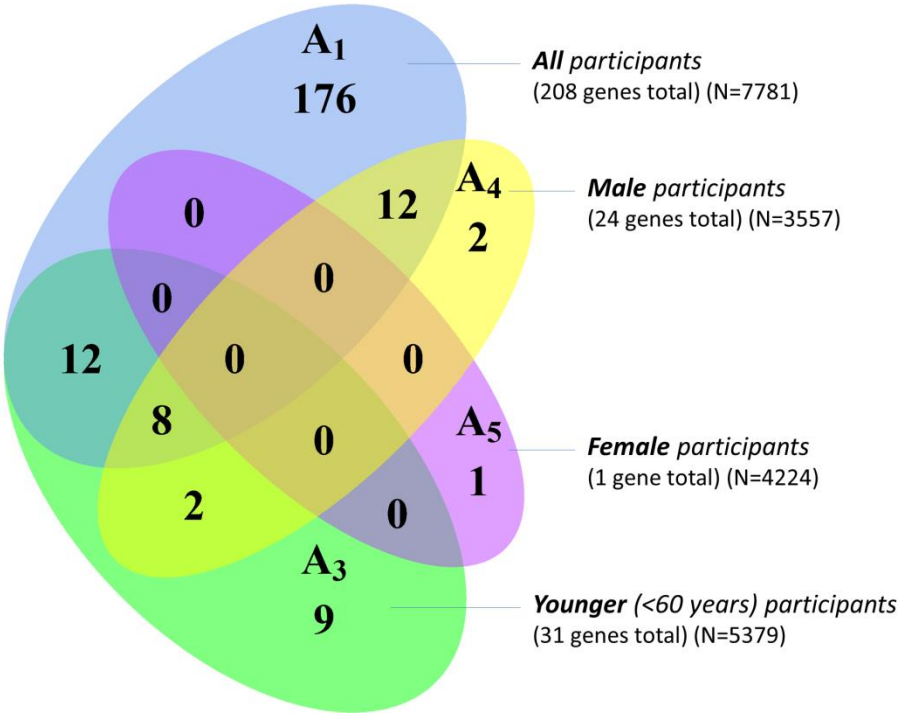
Table 4.2 | Top 20 unique genes associated with muscle strength in the meta-analysis of all participants in all four discovery cohorts

Meta-analysis			*P-values*				
Zscore	**P-value**	**BH P-value**	**FHS**	**Illumina**	**Gene**	**EntrezID**	**Name**
-5.67	$1.5x10^{-8}$	$2.0x10^{-4}$	$3.9x10^{-7}$	$9.9x10^{-3}$	*ALAS2*	212	aminolevulinate, delta-, synthase 2
5.37	$7.8x10^{-8}$	$3.1x10^{-4}$	$4.3x10^{-7}$	$4.0x10^{-2}$	*HEATR5A*	25938	HEAT repeat containing 5A
-5.28	$1.3x10^{-7}$	$3.9x10^{-4}$	$2.7x10^{-5}$	$1.2x10^{-3}$	*PNP*	4860	purine nucleoside phosphorylase
-5.17	$2.4x10^{-7}$	$4.7x10^{-4}$	$1.1x10^{-6}$	$5.0x10^{-2}$	*STOM*	2040	stomatin
5.14	$2.8x10^{-7}$	$4.7x10^{-4}$	$7.9x10^{-6}$	$1.1x10^{-2}$	*RPS6KA5*	9252	ribosomal protein S6 kinase, 90kDa, polypeptide 5
-5.09	$3.7x10^{-7}$	$5.8x10^{-4}$	$5.4x10^{-5}$	$1.8x10^{-3}$	*MBNL3*	55796	Muscleblind-Like Splicing Regulator 3
-5.07	$3.9x10^{-7}$	$5.8x10^{-4}$	$2.3x10^{-6}$	$4.4x10^{-2}$	*RAD23A*	5886	RAD23 homolog A (S. cerevisiae)
5.02	$5.1x10^{-7}$	$6.8x10^{-4}$	$7.7x10^{-5}$	$1.7x10^{-3}$	*HNRNPA0*	10949	heterogeneous nuclear ribonucleoprotein A0
-4.90	$9.5x10^{-7}$	$1.1x10^{-3}$	$2.0x10^{-5}$	$1.5x10^{-2}$	*GID4*	79018	GID Complex Subunit 4
-4.88	$1.1x10^{-6}$	$1.1x10^{-3}$	$9.3x10^{-6}$	$3.4x10^{-2}$	*TFDP1*	7027	transcription factor Dp-1
4.80	$1.6x10^{-6}$	$1.4x10^{-3}$	$9.9x10^{-6}$	$4.7x10^{-2}$	*ARRDC3*	57561	arrestin domain containing 3
-4.77	$1.7x10^{-6}$	$1.5x10^{-3}$	$1.8x10^{-5}$	$3.3x10^{-2}$	*RIOK3*	8780	RIO kinase 3
-4.71	$2.5x10^{-6}$	$1.8x10^{-3}$	$1.9x10^{-5}$	$4.1x10^{-2}$	*NTAN1*	123803	N-terminal asparagine amidase
4.66	$3.1x10^{-6}$	$2.2x10^{-3}$	$8.1x10^{-4}$	$6.0x10^{-4}$	*PDE4D*	5144	phosphodiesterase 4D, cAMP-specific
-4.62	$3.8x10^{-6}$	$2.4x10^{-3}$	$1.1x10^{-4}$	$1.2x10^{-2}$	*RGCC*	28984	Regulator Of Cell Cycle
4.61	$4.1x10^{-6}$	$2.5x10^{-3}$	$9.6x10^{-5}$	$1.4x10^{-2}$	*POLR2B*	5431	polymerase (RNA) II (DNA directed) polypeptide B
-4.59	$4.4x10^{-6}$	$2.5x10^{-3}$	$6.1x10^{-5}$	$2.4x10^{-2}$	*EIF1B*	10289	eukaryotic translation initiation factor 1B
4.57	$4.8x10^{-6}$	$2.7x10^{-3}$	$5.4x10^{-4}$	$2.0x10^{-3}$	*CIRBP*	1153	cold inducible RNA binding protein
-4.57	$4.8x10^{-6}$	$2.7x10^{-3}$	$6.2x10^{-5}$	$2.7x10^{-2}$	*ASGR2*	433	asialoglycoprotein receptor 2
-4.55	$5.3x10^{-6}$	$2.9x10^{-3}$	$3.8x10^{-5}$	$4.5x10^{-2}$	*CA1*	759	carbonic anhydrase I

BH=Benjamini-Hochberg. Ordered by meta-analysis P-value. Excluded: duplicate gene entries or genes with Illumina p>0.05.

Table continued in Supplementary Table 1.

Figure 4.2 | Venn diagram showing the overlap between the five meta-analyses



This plot shows a Venn diagram of genes with significant associations between whole-blood expression and muscle strength in the meta-analysis of 7,781 participants. Although five meta-analyses were performed, there were no significant associations in the "older" subset so it is excluded from this plot. 222 unique genes in total were associated with muscle strength in at least one analysis. Additional information is available in **Supplementary Tables 4.6/4.7**.

4.5.3 Significant genes only available on one array

Due to differences between the array technologies and relative abundance of transcripts not all the genes were eligible for the meta-analysis. In the analysis on all individuals, 1,123 probes (898 unique identifiers) were present on the Affymetrix Exon array that did not have a corresponding probe on the Illumina array; 21 of these probes (14 unique genes) were

significantly associated with hand-grip strength after BH adjustment for multiple testing (see **Table 4.3** for top 20 probes in the "all individuals" analysis and **Supplementary Tables 4.9-4.13** for list of significant probes in each of the five analyses in FHS only). In the Illumina array, 7,768 probes (6,119 unique gene identifiers) were available that did not map to a gene/transcript in the Affymetrix Exon array (after excluding lowly expressed probes). None of the probes were significantly associated with muscle strength after BH multiple testing correction.

4.5.4 Ontology enrichment of strength-associated genes

Using the 208 genes associated with muscle strength in the meta-analysis of all participants, two analyses were performed using the WebGestalt web resource;

1. *Gene Ontology* analysis found that 10 biological processes (**Table 4.4**; including "hemoglobin metabolic process" and "innate immune response") were significantly enriched (FDR<0.05), 10 molecular functions were enriched (including "protein binding genes"), and 10 cellular component pathways were enriched (including "intracellular membrane-bound organelle") (See **Supplementary Table 4.14** and **Supplementary Figure 4.1**).

2. *Human Phenotype Ontology* analysis found 10 phenotypes significantly enriched in the genes, including "Anemia due to reduced life span of red cells", "Hemolytic Anemia", and "Abnormality of erythrocytes" (See **Supplementary Table 4.15** and **Supplementary Figure 4.2**).

Table 4.3 | Top 20 probes in the FHS analysis that did not map to a corresponding Illumina probe, ordered by p-value

Estimate	P-value	BH P-value	Chr	Start	Gene	EntrezID	Name
-0.0037	$1.76x10^{-6}$	$1.49x10^{-3}$	1	144989319			
-0.0028	$4.89x10^{-6}$	$2.88x10^{-3}$	9	37800563	*DCAF10*	79269	DDB1 and CUL4 associated factor 10
0.0031	$8.57x10^{-6}$	$4.12x10^{-3}$	1	150522766	*ADAMTSL4*	54507	ADAMTS-like 4
0.0038	$9.29x10^{-6}$	$4.15x10^{-3}$	16	89980135			
0.0040	$1.95x10^{-5}$	$5.97x10^{-3}$	2	162412847	*SLC4A10*	57282	solute carrier family 4, ... member 10
0.0029	$3.24x10^{-5}$	$7.74x10^{-3}$	16	9186734			
0.0031	$3.93x10^{-5}$	$8.51x10^{-3}$	1	44440179	*ATP6V0B*	533	ATPase, H+ transporting, lysosomal 21kDa, V0 subunit b
-0.0027	$4.10x10^{-5}$	$8.64x10^{-3}$	3	63819562	*THOC7*	80145	THO complex 7 homolog (Drosophila)
0.0032	$5.38x10^{-5}$	$9.81x10^{-3}$	20	3898284			
0.0019	$1.49x10^{-4}$	$1.70x10^{-2}$	20	1316212			
0.0018	$1.89x10^{-4}$	$1.91x10^{-2}$	3	128628719	*ACAD9*	28976	acyl-CoA dehydrogenase family, member 9
-0.0017	$2.21x10^{-4}$	$2.08x10^{-2}$	1	8021733	*PARK7*	11315	parkinson protein 7
0.0036	$3.18x10^{-4}$	$2.69x10^{-2}$	12	57809458			
0.0010	$3.20x10^{-4}$	$2.70x10^{-2}$	1	153901987	*DENND4B*	9909	DENN/MADD domain containing 4B
-0.0019	$3.56x10^{-4}$	$2.87x10^{-2}$	3	10157370	*BRK1*	55845	BRICK1, SCAR/WAVE actin-nucleating complex subunit
-0.0028	$3.84x10^{-4}$	$2.97x10^{-2}$	2	95517671	*TEKT4*	150483	tektin 4
-0.0014	$3.95x10^{-4}$	$3.03x10^{-2}$	1	247937996	*OR9H1P*	81439	olfactory receptor, family 9, subfamily H, member 1 pseudogene
0.0018	$6.01x10^{-4}$	$3.89x10^{-2}$	9	124042152	*GSN-AS1*	57000	GSN(gelsolin) antisense RNA 1
0.0028	$6.41x10^{-4}$	$4.03x10^{-2}$	6	20451305			
-0.0021	$6.90x10^{-4}$	$4.19x10^{-2}$	12	8024178	*NANOG*	79923	Nanog homeobox

BH=Benjamini-Hochberg. Blank gene symbols were not annotated to a specific gene. Table continued in Supplementary Table 9. Ordered by P-value. Not all probes map to gene ID's.

4.5.5 *A priori* genes associated with muscle function

Of 20 *IGF*-genes tested in this meta-analysis (from *IGF1* to *IGFL4*) two were significantly associated with muscle strength: *IGF1R* (positively associated) and *IGF2BP2* (negatively associated) (meta-analysis FDR=$3.2x10^{-2}$ and FDR=$1.2x10^{-3}$, respectively). *GDF11* expression was not associated with muscle strength (FDR>0.05). 40 unique Kelch genes were tested in the meta-analysis; none were associated with muscle strength in whole blood (FDR>0.05). 18 unique Wnt genes (from *WNT1* to *WNT9B*) were tested in the meta-analysis; none were associated with muscle strength in whole blood (FDR>0.05). All 10 Frizzled genes (*FZD1-10*, receptors for the Wnt pathway) were also available to test; none were associated with muscle strength. Similarly all three Dishevelled genes were available to test (*DVL1-3,* acts directly downstream of the Frizzled receptors) and none were associated with muscle strength. Of 586 genes identified by Melov *et al* that were differentially expressed in muscle tissue between old and young men following endurance training [179] four were associated with muscle strength in this analysis: *ANP32B*, *CIRBP*, *MCM7* and *MGST1*.

4.5.6 Most associated genes are not previously linked to muscle in the literature

By systematically searching the literature for each of the 222 genes significantly associated with muscle strength we determined that 115 have not been mentioned in the same title or abstract as the search term “muscle” (as of 12th Nov 2014; **Supplementary Table 4.16**).

Table 4.4 | Ten biological processes were enriched in the genes associated with muscle strength in the overall analysis

Biological Process	C	O	rawP	adjP	Top Genes
hemoglobin metabolic process (GO:0020027)	14	5	$3.2x10^{-7}$	0.0004	*ALAS2, AHSP, FECH, EIF2AK2, EPB42*
hemoglobin biosynthetic process (GO:0042541)	9	3	$1.0x10^{-4}$	0.0139	*ALAS2, FECH, EIF2AK2*
innate immune response (GO:0045087)	539	18	$3.8x10^{-5}$	0.0139	*RPS6KA5, IFI27, TLR5, DUSP3, FCGR1B*
negative regulation of protein metabolic process (GO:0051248)	460	16	$6.5x10^{-5}$	0.0139	*IGF2BP2, CIRBP, BANP, GCLC, DUSP3*
posttranscriptional regulation of gene expression (GO:0010608)	371	14	$8.0x10^{-5}$	0.0139	*HNRNPA0, IGF2BP2, EIF1B, CIRBP, ASGR2*
response to arsenic-containing substance (GO:0046685)	20	4	$6.5x10^{-5}$	0.0139	*GCLC, GSTO1, FECH, DDX3X*
regulation of translational initiation by eIF2a ... (GO:0010998)	2	2	$1.0x10^{-4}$	0.0139	*FECH, EIF2AK2*
regulation of eIF2 alpha phosphorylation by heme (GO:0010999)	2	2	$1.0x10^{-4}$	0.0139	*FECH, EIF2AK2*
response to stress (GO:0006950)	2952	55	$4.7x10^{-5}$	0.0139	*ALAS2, RPS6KA5, RAD23A, HNRNPA0, UBQLN1*
protoporphyrinogen IX metabolic process (GO:0046501)	11	3	$2.0x10^{-4}$	0.0167	*FECH, ALAS2, EIF2AK2*

C=number of reference genes in the category; O=number of genes in the gene set and also in the category; rawP=p value from hypergeometric test; adjP=p value adjusted by the multiple test adjustment (Benjamini-Hochberg); Top Genes=top five genes from pathway based on meta-analysis P-value

4.5.7 Few genes replicate in the NESDA cohort

NESDA was not included in the meta-analysis due to data limitations, notably the lack of empirically determined or reliably imputed white cell count data, the use of a different microarray technology (a predecessor to the Exon array used by FHS, much more dissimilar than the v3/v4 Illumina arrays are to one another), and a younger population than the other cohorts included in the meta-analysis (max age=65, see **Table 4.1**).

As noted above, in total 222 unique genes were associated with muscle strength across all the meta-analyses performed. Of 208 genes significantly associated with muscle strength in analysis 1 (all participants) it was possible to test 144 in the NESDA cohort; 7 genes were also associated with muscle strength ($p<0.05$) in the NESDA cohort with the same direction of effect (*ACSL6, ALDH5A1, CARHSP1, FGL1, NRG1, PIGB, SIGLEC7*). No genes were significantly associated with muscle strength in the older subset. Of 31 unique genes associated in the younger subset, 23 were available to test in NESDA; none replicated. Of 24 genes significant in the males only analysis 12 were available to test in NESDA, one of which was significant (*ALAS2*, B=-0.009, p=0.01). Finally in the analysis on females only 1 gene was significant in the meta-analysis but was not associated with strength in the NESDA cohort.

4.6 Discussion

In this discovery study we set out to determine whether specific transcript levels in blood are associated with muscle strength in multiple human cohorts including mostly middle-aged volunteers. Previous cross-sectional (and longitudinal) studies have shown that the degree and *rate* of loss of strength (and muscle mass) is greater in older (male) participants [195]. We therefore performed stratified analyses by gender and age to determine whether transcripts or pathways associated with muscle strength in whole blood differ between these groups. 222 unique genes were associated with muscle strength in at least one analysis, 115 of which were not previously linked to the term “muscle” in the published literature cataloged on GeneCards and PubMed (as of 12th Nov 2014).

Expression of *GDF11*, a protein that can reverse age-related muscle dysfunction in mice [3], was not associated with muscle strength in this analysis. Instead, we observe significant associations between muscle strength and expression of *IGF1R* and *IGF2BP2* (positive and negative directions of association with muscle, respectively), growth factors involved in skeletal muscle growth [182]; it may be that the circulating growth factor pathways required in humans are different to those in mice. No genes from the Wnt or Kelch pathways (both known to be important for muscle function [180], [181]) were associated with muscle strength in whole blood in this analysis, potentially due to the specificity of these particular pathways to muscle tissue. We also examined 586 genes identified by Melov *et al* that differ in expression in muscle between old and young men after endurance training [179], four of which were associated with grip strength in this study: *MGST1* (negative direction), an immune mediator [139] which may also protect against oxidative stress [196]; *MCM7* (positive direction), regulates DNA replication during proliferation [197] and is up-regulated in peripheral leukocytes following IGF-1 administration [198]; *CIRBP* (positive direction), with promotes inflammation in response to shock and sepsis [199]; and *ANP32B* (negative direction), a cell-cycle progression and anti-apoptosis factor [139]: however lack of strength measures in the original study make direct comparison challenging.

The gene *ALAS2,* which is known to be a rate-limiting step in heme biosynthesis [200] and is found in erythrocytes in blood [201], was most strongly associated with muscle strength in the meta-analysis (negative direction) and was also associated in the NESDA cohort. Other genes of note include *CCR6* and *PRF1* (both positively associated with muscle strength and age [79]); *CCR6* is implicated in B-cell maturation and recruitment of other immune cells during immune responses [202]; perforin (*PRF1*), a protein secreted by cytotoxic T-cells creates pores in membranes to permit apoptosis-inducing granzyme into the target cell [203], was also significantly associated with strength. In a separate study *PRF1* expression mediated age-related inflammation [204]. *NANOG* expression is positively associated with strength in this analysis (**Table 4.2**), can reverse ageing of some stem cells [205] and, in combination with three other genes (OCT4, SOX2, and LIN28, not significant in this analysis), can induce pluripotency of somatic cells [206]. This may suggest that differentiation (of whole blood cells) is inversely correlated with muscle strength, but the mechanisms are unclear.

4.6.1 Enrichment analysis

In ontology enrichment analyses using the online tool WebGestalt we identified statistically significant enrichment for genes in the biological process “Hemoglobin Metabolic Process” (5 of 14 genes in the biological process were associated with strength) and the phenotypic abnormality “Hemolytic Anemia” (8 of 54 genes), amongst others. Hemoglobin levels are positively associated with muscle strength and density [207]. Circulating reticulocytes (erythrocyte precursors with some residual RNA present) were not adjusted for in this analysis and are likely the source of the associations with genes such as *ALAS2*, which appears to be exclusively expressed by erythrocytes in blood [201].

“Innate Immune Response” (18 of 539 genes) is also enriched in the Gene Ontology annotation results from WebGestalt; this includes cells such as macrophages which are

essential for maintenance of muscle strength due to their anti-inflammatory, wound-healing response [47]. *CEBPB*, the gene implicated in the macrophage wound-healing response [113] and significantly associated with muscle strength in the 2012 study by InCHIANTI [183], did not replicate in the other cohorts. This could be due to methodological differences between the previous study and this meta-analysis, as well as differences in age distribution (81% of the InCHIANTI cohort is aged >=60 years, compared to only 31% of the meta-analysis participants, which includes the InCHIANTI cohort; **Table 4.1**). Whether the original finding was a false-positive is unclear given the mouse model evidence of plausible biological mechanism [113] and evidence from a study in humans showing that exercise-induced muscle damage is associated with change in *CEBPB* expression in peripheral blood cells [184]. Further work on this marker will be required in older and frail groups. It is important to note that although the haematology pathways were more statistically significant, the immune pathway actually contained more genes associated with muscle because the significance is related to the proportion of the total genes in the pathway that are present. Examination of **Supplementary Table 4.14** is necessary for a full appreciation of this and the biologic implications.

4.6.2 Heterogeneity in subset analyses

Five analyses were performed; one in all participants, and four in subsets. The majority of genes (79%) were only associated in the analysis including all study participants. Overlap between the subsets of analyses was variable (see **Figure 4.2** and **Supplementary Tables 4.6/4.7** for details) and no genes were identified at transcriptome-wide significance when just analyzing individuals over 60 years of age. Although the sample size was still reasonably high (2,402 participants) variability in the strength phenotype as individuals age, and development of different co-morbidities and chronic inflammation (for instance multiple co-morbidities reduce the association between anemia and disability [208]) may reduce the power to detect associations. However in the younger group (n=5,379) we find 31 unique

genes associated in the meta-analysis; this analysis could have more power firstly due to the sample size, but also the participants have a greater range of strength values, with lower levels of chronic inflammation and disease. Defensin, Alpha 4, Corticostatin (*DEFA4*, negative strength association in the analysis of females only) is a cytotoxic peptide that has antimicrobial activity against Gram-negative bacteria (predominantly) [202], and there appears no obvious explanation for this finding in females. These results confirm previous work indicating that the immune response is strongly associated with muscle strength [176], [177].

4.6.3 Limitations

There are several potential limitations of this study including its cross-sectional design. As such it is not possible to determine if changes in muscle strength lead to differential gene expression or whether gene expression affects muscle strength. Future longitudinal studies should address this by testing gene expression associations with subsequent changes in grip strength. A second potential limitation is the “whole blood” nature of the tissue studied, which makes it impossible to determine the specific cellular origins of gene expression associations. Third, the microarray technology used across the participating cohorts was not the same; although 21 (70%) of the top 30 meta-analysis results were independently replicated between the platforms, replication of the less strong results was limited. However, this does suggest that our top results are very robust to cohort and array differences.

4.6.4 Conclusions

We have identified robust associations between the expression of 222 genes muscle strength in human whole blood. For 115 genes this is the first published evidence in the literature linking them to muscle. Genes identified include key growth factor-related genes, although not *GDF11*, the innate immune response, and hemoglobin metabolism. Further work is needed to characterize expression patterns in the older groups and to investigate the cellular origins of gene expression signals. Future research should also include independent replication and longitudinal data to assess long-term prediction of frailty and disease. Perhaps most importantly, this analysis suggests that there are heterogeneous processes involved in muscle strength in mid- versus later-life, and separate analyses are needed of muscle strength at advanced ages.

4.7 Funding and Acknowledgements

All authors declare no competing financial interests.

FHS gene expression profiling was funded through the Division of Intramural Research (Principal Investigator, Daniel Levy), National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD. Dr. Murabito is supported by NIH grant R01AG029451. Dr. Kiel is supported by NIH R01 AR41398. The Framingham Heart Study is supported by National Heart, Lung, and Blood Institute contract N01-HC-25195.

The InCHIANTI study was supported in part by the Intramural Research Program, National Institute on Aging. D.M. and L.W.H. were generously supported by a Wellcome Trust Institutional Strategic Support Award (WT097835MF). W.E.H. was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for the South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health in England.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. The generation and management of RNA-expression array data for the Rotterdam Study was executed and funded by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. We thank Marjolein Peters, MSc, Ms. Mila Jhamai, Ms.

Jeannette M. Vergeer-Drop, Ms. Bernadette van Ast-Copier, Mr. Marijn Verkerk and Jeroen van Rooij, BSc for their help in creating the RNA array expression database.

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network ‘Greifswald Approach to Individualized Medicine (GANI_MED)’ funded by the Federal Ministry of Education and Research (grant 03IS2061A). The University of Greifswald is a member of the 'Center of Knowledge Interchange' program of the Siemens AG and the Caché Campus program of the InterSystems GmbH.

4.8 Supplementary Information

Due to their number and size, supplementary figures and tables are not included in the printed version of this thesis. Please see the accompanying CD-ROM.

Supplementary Table 4.0 | Authors list

Order	*Position*	*Initials*	*Name*	*Titles*	*Cohort*
Starred first authors / analysts (ordered by inclusion in meta-analysis, then sample size)					
1	First-first	L.C.P.	Luke C. Pilling	*MSc*	InCHIANTI
2	First	R.J.	Roby Joehanes	*PhD*	FHS
3	First	T.K.	Tim Kacprowski	*MSc*	SHIP
4	First	M.J.P.	Marjolein J. Peters	*MSc*	RS
5	First	R.J.2	Rick Jansen	*PhD*	NESDA
Middle, active co-authors ranked alphabetically by cohort, then surname					
6	Middle	D.K.	David Karasik	*PhD*	FHS
7	Middle	D.P.K.	Douglas P. Kiel	*PhD*	FHS
8	Middle	L.W.H.	Lorna W. Harries	*PhD*	InCHIANTI
9	Middle	A.T.	Alexander Teumer	*PhD*	SHIP
Middle co-authors are ranked alphabetically by cohort, then by surname					
10	Middle	D.L.	Daniel Levy	*PhD*	FHS
11	Middle	H.L.	Honghuang Lin	*PhD*	FHS
12	Middle	K.L.	Kathryn Lunetta	*PhD*	FHS
13	Middle	P.J.M.	Peter J. Munson	*PhD*	FHS
14	Middle	S.B.	Stefania Bandinelli	*MD*	InCHIANTI
15	Middle	W.E.H	William Henley	*PhD*	InCHIANTI
16	Middle	D.G.H	Dena G. Hernandez	*MSc*	InCHIANTI
17	Middle	A.B.S.	Andrew B. Singleton	*PhD*	InCHIANTI
18	Middle	T.T.	Toshiko Tanaka	*PhD*	InCHIANTI
19	Middle	G.G.	Gerard van Grootheest	*Msc*	NESDA
20	Middle	A.H.	Albert Hofman	*PhD*	RS
21	Middle	A.G.U.	André G. Uitterlinden	*PhD*	RS
22	Middle	R.B.	Reiner Biffar	*MD*	SHIP
23	Middle	S.G.	Sven Gläser	*MD*	SHIP
24	Middle	G.H.	Georg Homuth	*PhD*	SHIP
25	Middle	C.M.	Carolin Malsch	*MSc*	SHIP
26	Middle	U.V.	Uwe Völker	*PhD*	SHIP
Starred last authors / group leaders					
27	Last	B.W.J.H.P.	Brenda W.J.H. Penninx	*PhD*	NESDA
28	Last	J.B.J.M.	Joyce B.J. van Meurs	*PhD*	RS
29	Last	L.F.	Luigi Ferrucci	*MD, PhD*	InCHIANTI
30	Last	T.K.2	Thomas Kocher	*MD*	SHIP
31	Last	J.M.M.	Joanne M. Murabito	*MD, ScM*	FHS
32	Last-last	D.M.	David Melzer	*MBBCh, PhD, FFPH*	InCHIANTI

Chapter 5 – Analysis 3 – Cognitive Function

Leukocyte CCR2 Expression Is Associated with Mini-Mental State Examination Score in Older Adults

Rejuvenation Research (2012); 15:4. 395-404.

Harries, Lorna W *

Bradley-Smith, Rachel M *

Llewellyn, David J

Pilling, Luke C

Fellows, Alexander

Henley, William

Hernandez, Dena

Guralnik, Jack M

Bandinelli, Stefania

Singleton, Andrew

Ferrucci, Luigi

Melzer, David

* These authors contributed equally

5.1 Overview

In this study we analysed the expression of genes in InCHIANTI for their association with cognitive function – the Mini-Mental State Examination (MMSE) score. It is a general feature of ageing that general cognitive ability decreases, so we chose to undertake this analysis and found a single marker (gene *CCR2*) to be negatively associated with MMSE score (therefore to have higher expression in those individuals with some form of cognitive impairment). We also assessed gene expression associations with change in MMSE score over 9 years, hoping to discover a marker for cognitive decline, but found only the *CCR2* gene again.

I undertook the bioinformatic analysis (transcriptome-wide screen, validation of phenotype, gene-set enrichment analysis, investigation of confounders, interrogation of online databases), and worked with the other authors in reviewing the literature, discussing and interpreting the results, and writing the manuscript. This project was initiated as a lab-based investigation led by Rachel, however the emphasis of the study changed dramatically after the bioinformatic analysis revealed the *CCR2* finding.

5.2 Summary

Introduction: Circulating inflammatory markers may play an important role in cognitive impairment at older ages. Mice deficient for the chemokine (C-C motif) receptor 2 (*CCR2*) develop an accelerated Alzheimer's-like pathology, and *CCR2* is also important in neurogenesis. To identify human gene transcripts most closely associated with Mini Mental State Examination (MMSE) scores we undertook a genome-wide and inflammation specific transcriptome screen in circulating leukocytes from a population-based sample.

Methods: We measured *in-vivo* transcript levels by microarray analysis in 691 subjects (mean age 72.6 yrs) in the InCHIANTI study. We assessed expression associations with MMSE performance at RNA collection and prior 9 year change in MMSE score in linear regression models.

Results: In genome-wide analysis, raised *CCR2* expression was cross-sectionally the most strongly associated transcript with lower MMSE score (beta=-0.16, $p=5.1x10^{-6}$, False discovery rate 'FDR' q=0.077). Amongst inflammatory transcripts, only *CCR2* expression was associated with both MMSE score and accelerated decline in score over the preceding 9 years (beta=-0.16, $p=5.1x10^{-6}$, q=0.003; and beta=-0.13, $p=5.5x10^{-5}$, q=0.03; respectively). *CCR2* expression was also positively associated with ApoE e4 Alzheimer's disease risk haplotype.

Conclusions: We show for the first time that *CCR2* expression is associated with lower MMSE scores in an older human population. Laboratory models of *Ccr2*-mediated □-amyloid removal and regulation of neurogenesis affecting cognitive function may be applicable in humans. *CCR2*-mediated pathways may provide a possible focus for intervention to potentiate protective reactions to Alzheimer's pathology in older people, including for people with adverse ApoE haplotype.

5.3 Introduction

Cognitive impairment and dementia at older ages commonly results from accumulating vascular and neurodegenerative pathology in the brain [209] and is experienced by over half of adults aged 85 or over [50]. Major risk factors for cognitive impairments in later life include *APOE* e4 haplotype [210], but the underlying biological mechanisms are still unclear. Gene expression arrays offer a new approach to identifying the most important molecular mechanisms causing or responding to the pathologies underlying cognitive decline. The most accessible tissue for gene expression array analyses in large numbers of older people is circulating blood leucocytes, which are likely to be most sensitive to inflammatory and related mechanisms.

Chronic low-level inflammation has been proposed as a key mechanism underlying cognitive decline and dementia, and has been implicated in the neuropathological cascade leading to late onset Alzheimer's disease (LOAD) [211]. Inflammatory factors have been further implicated in cognitive impairment and dementia using mouse models [212]. These include transgenic animals in which chemo-attractant proteins or receptors such as *Ccl2* or *Ccr2* have been abolished [212]. These genes, associated with the migration of phagocytic and inflammatory macrophages, have been associated with Alzheimer's pathology and peripheral atherosclerosis [212]–[214]. *Ccr2*-deficient mice show early accumulation of β-amyloid, with premature mortality [214]. Blood-borne systemic factors including CCR2 and CCL11 were also recently implicated in the negative regulation of neurogenesis and cognitive function in rodent studies [4].

The importance of CCR2 and related signaling in human age-related cognitive impairment is unclear. Maes et al (2007) identified leukocyte expression differences between 14 Alzheimer's Disease and control individuals in a microarray experiment [215], while Grunblatt et al identified five peripheral blood leucocyte genes whose mRNA correlated significantly with MMSE (Mini Mental State Examination) score in a smaller series of Alzheimer's and

control individuals [216]. These studies present initial evidence that brain changes in pathologically-specific dementia patients may be detectable using peripheral leukocyte gene expression, but it is currently not known if this also applies to age-associated cognitive decline.

In the current study, we used a genome-wide and inflammation-focused approach to identify the most strongly associated *in-vivo* transcript levels in circulating leukocytes associated with MMSE score or rate of change in MMSE score in a general population sample of predominantly older people. The MMSE is a widely used measure of cognitive function in prospective epidemiological studies of elderly populations, and is sensitive to moderate or severe cognitive declines, often due to dementia [217], [218]. To ensure population relevance, no exclusions for co-morbidity were made in the main analysis.

5.4 Methods

5.4.1 Study cohort

We used InCHIANTI, a population-based study of aging [115], which has followed older persons over a 9 year period to assess 'normal' aging using both interviews (conducted at the participants home by experienced interviewers) and blood samples (in the study clinic, all patients fasted for 8 hours prior to collection). Peripheral blood samples for RNA extraction were collected from participants at the 9 year follow-up (2008/9). Cohort demographics are given in table 1. Ethical approval was granted by the Instituto Nazionale Riposo e Cura Anziani institutional review board in Italy. Participants gave written informed consent to participate and for sample collection after having received an extensive description of the procedures, purposes and potential risks of the study. RNA was extracted from each sample using the PAXgene Blood mRNA kit (Qiagen, Crawley, UK) according to the manufacturer's instructions. To ensure population relevance, no exclusions for co-morbidity were made in the main analysis.

5.4.2 Whole transcriptome scan

Whole genome expression profiling of the samples was conducted using the Illumina Human HT-12 microarray (Illumina, San Diego, USA), data processing was carried out using the Illumina and Beadstudio software (Illumina, San Diego, USA) as previously described [155]. Baseline intensities were calculated as mean and SD computed over all beads for a particular probe. QC steps included correction for local background effects, removal of outlier beads, computation of average bead signal and SD for each probe and gene, calculation of detection p-values using negative controls present on the array, quantile normalization across arrays, check of outlier samples using a clustering algorithm, and checks of positive controls. Subject level QC steps included removing individuals where the expression

intensity was +/- 3 standard deviations from the mean. All microarray experiments and analyses complied with MIAME guidelines [219]. Following microarray data QC steps, 16,571 transcripts gave reliable signals above background ($p<0.01$) in >5% of the sample population and were therefore eligible for analysis.

5.4.3 Assessment of cognitive function

Participant interviews and clinical assessment, including the Mini-Mental State Examination of global cognitive function (MMSE), were undertaken at baseline, and at years 3, 6 and 9 of follow-up. The MMSE is an assessment of global cognitive function widely used in both hospital and community settings as a dementia screening tool [220], [221].

5.4.4 Included sample and statistical analysis

From 733 blood samples collected, RNA quality and microarray QC steps resulted in loss of 35 participants from the analysis. From the remaining 698 participants, 7 were excluded due to absence of MMSE data, yielding 691 individuals. Three further individuals were dropped from the regression models due to incomplete leukocyte data. (See **table 5.1** for summary statistics of the cohort included in the analyses). All the remaining 688 individuals were included in the analyses, irrespective of other pathologies.

Table 5.1 | This table summarises the population statistics for the 691 participants eligible for our study.

Summary statistics	
Age (years)	**n**
30-49	86 (12.5%)
50-69	98 (14.2%)
70-89	479 (69.3%)
90-104	28 (4%)
Mean age at RNA-collection	72.6 (SD: 15.3)
Gender	**%**
Men	44.8
Women	55.2
MMSE score	
Mean score at RNA-collection	25.58 (SD: 5.66)
Mean change in MMSE (9 years)	-1.31 (SD: 4.71)
Frequency of scores at RNA collection	**%**
0-18	8.14
19-23	12.79
24-27	30.52
28-30	48.55
Education	**%**
None	13.31
Elementary	46.02
Secondary	13.46
High school	12.59
University / Professional	14.62
Pack years smoked (lifetime)	**%**
None	55.72
0.1-20	22.87
20-39	14.33
40+	7.09

Associations between gene expression and MMSE score at RNA collection were analysed using multiple regression models adjusted for the following potential confounding factors: age in years (as a continuous variable), gender, highest level of education received (in five categories: none, elementary, secondary, high school, and university/professional), lifetime pack-years smoked (in four categories: none, 0.1 to 20 years, 20 to 39 years, and 40 plus years), blood leukocyte type (neutrophil, lymphocyte, monocyte, eosinophil percentages) each as continuous variables, hybridisation and amplification batch (in 10 and 14 categories respectively), and study site; participants lived in either a rural village (Greve) or an urban site (Bagno a Ripoli).

Separate linear regression models were fitted for each of the full set of 16,571 probes which passed QC in the discovery dataset. We have expressed the effect sizes of the associations in standardized betas to aid in the interpretation (expression intensities vary from probe to probe, and the normalization procedures further affect the interpretability of coefficients). We controlled for the effect of multiple testing by measuring the statistical significance of each association using both the p-value and the q-value. The q-value quantifies significance in terms of the false discovery rate (FDR) rather than the false positive rate [222], and forms a measure of how likely a particular p-value is to represent a genuine association. Expression levels were taken to have a significant association with MMSE score at RNA collection, or cognitive decline from baseline 9 years previously, if the association achieved a nominal p-value ≤0.05 and a q-value <0.1. Our study was powered to detect expression differences of <0.25 SD for the 16,571 transcripts studied, allowing us to detect moderate expression differences between groups.

5.4.5 Inflammatory gene scan

The large number of transcripts tested in the whole genome scan yields a very stringent requirement for statistical significance, resulting in a substantial risk of type II (false negative) error. Since leukocytes are an inflammation-related tissue, we carried out a sub-analysis focusing on a smaller subset of genes involved in inflammatory response only. The genes of interest were identified using the Molecular Signatures Database (MSigDb) [159], with search terms for inflammatory/immune-related gene pathways as defined by the Gene Ontology (GO) project (www.geneontology.org). 44 inflammation-related gene-sets were returned, comprising 635 unique probe IDs, which equate to 425 unique genes with available data in our cohort (Supplementary table 4.1). Linear regression models were carried out as described above. This study is powered to detect expression differences of <0.22 SD for the inflammation transcript set.

5.4.6 Real-time PCR validation

The expression of the *CCR2* gene in a subcohort of 100 individuals selected at the extremes of the MMSE spectrum was validated by the use of a custom real-time PCR assay (probe and primer sequences available on request). Reaction mixes included 5μl 2x TaqMan universal master mix (no AMPerase) (Applied Biosystems, Foster City, USA), 30μl dH_2O and 2μl cDNA template. PCR amplifications were performed on the ABI 7900HT platform (Applied Biosystems, Foster City, USA). Cycling conditions were 50°C for 2 minutes, 94.5°C for 10 minutes followed by 40 cycles of 97°C for 30 seconds and 57.9°C for 1 minute. The expression of each gene was measured in triplicate for each sample. Gene expression relative changes were quantified using the 2^{-DDCt} method [223] relative to the geometric mean of the *GUSB*, *B2M* and *PPIA* endogenous controls. The correlation between quantifications achieved using microarray and by real-time PCR was then assessed using linear regression.

5.4.7 Gene Set Enrichment Analysis (GSEA)

We also performed GSEA to identify gene sets/pathways associated with cognitive function, according to the method of Subramanian *et al* [159]. Molecular or biological function pathways were identified using Gene Ontology gene sets from the molecular signature database (MSigDB) (http://www.broadinstitute.org/gsea/msigdb/index.jsp). Gene set size filtering excluded gene sets containing less than 15 or more than 500 genes. Genes were ranked according to the magnitude of their association with cognitive function, and representation of each gene set within this ranked list was analysed. Gene sets significantly overrepresented at the top or bottom of the ranked list were taken to be significantly associated with cognitive function. A signal-to-noise metric was used to rank genes, and gene set enrichment scores were calculated using a weighted enrichment statistic. One thousand random permutations of the phenotype label were used to calculate the empirical p-values of each pathway compared to the p-values that would be ascertained by chance. Gene sets with a nominal p-value<0.01 and q-value<0.1 were considered associated with cognitive impairment or cognitive decline.

5.4.8 Penalized cubic regression spline

Using R package 'mgcv' [224] we fitted an adjusted generalized additive model (GAM) using a smoothed penalized cubic regression spline; the smoothing parameter was chosen automatically by cross-validation. The same data and covariates that made up the original generalized linear regression (GLM) screen were used. GAM is applied here because it allows non-parametric fits, and thus if an aberrant relationship between expression and age had existed it could be highlighted.

5.5 Results

5.5.1 Cohort Details

Participant characteristics are given in **table 5.1**; the study sample had a mean age of 72.6 years (SD: 15.3, range 30 to 104) and 55.2% were female. At RNA collection the mean MMSE score was 25.58 (SD 5.66), 21% of the sample had MMSE scores of 23 or less (8.1% with MMSE of 18 or less), and the mean change in MMSE scores over 9 years was -1.31 (SD = 4.71). The correlation between MMSE change and MMSE at baseline is 0.14, and the correlation between MMSE change and MMSE at year 9 is 0.89.

5.5.2 Genome wide analysis

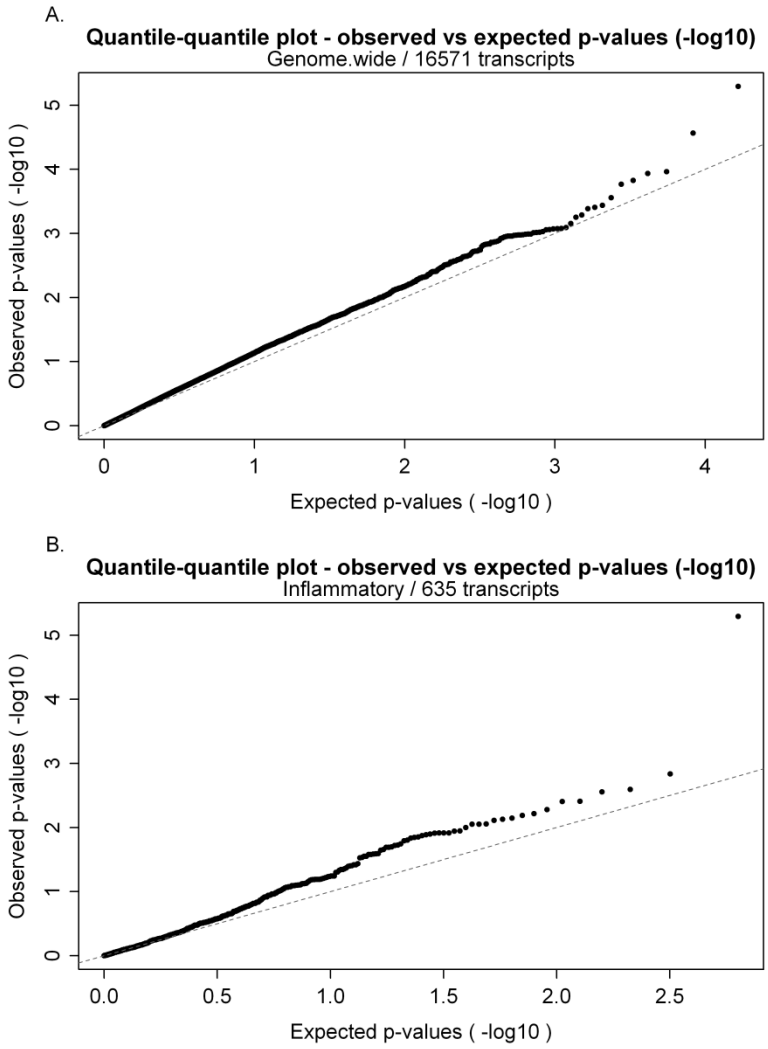
In genome wide analysis, only one transcript (*CCR2*, measured via probe ilmn_1774761) showed a very near significant genome wide association with MMSE score at RNA collection (beta=-0.16, $p=5.1x10^{-6}$; q=0.076; **table 5.2**). No probes were significantly associated ($q < 0.1$) with change in MMSE score over the preceding 9 years (**table 5.2**), although CCR2 was the most strongly associated transcript (beta=-0.13, $p=5.5x10^{-5}$, q=0.70). Quantile-Quantile (QQ) plots comparing the observed p-values ($-\log_{10}$) with those that would be expected by chance alone (**figure 5.1a**) confirmed that large scale disruption to gene expression levels in blood was not a feature of MMSE score at RNA collection or change in score over the previous nine years, although some small deviations from the expected pattern were noted.

Table 5.2 | The ten probes most closely associated with MMSE score at wave 9 and change in MMSE score (baseline to year 9), ordered by false discovery rate q-value (n observations = 688).

Probe ID	p-value	Coefficient	Beta	95% CIs		q-value	Gene
MMSE score at year 9							
ilmn_1774761	5.1E-06	-0.0042	-0.1616	-0.0056	-0.0027	0.076	CCR2
ilmn_2374362	2.7E-05	-0.0015	-0.1709	-0.0021	-0.0009	0.204	FAM108B1
ilmn_1791912	1.1E-04	0.0052	0.1475	0.0030	0.0074	0.356	SIDT2
ilmn_1758457	1.2E-04	0.0008	0.1473	0.0005	0.0011	0.361	TBC1D16
ilmn_1796094	1.5E-04	-0.0044	-0.1486	-0.0063	-0.0025	0.381	CD36
ilmn_1693949	1.7E-04	0.0011	0.1153	0.0006	0.0015	0.390	UNQ1944
ilmn_1724266	2.8E-04	0.0017	0.1382	0.0009	0.0025	0.418	LYPD2
ilmn_1772821	3.7E-04	-0.0015	-0.1561	-0.0022	-0.0008	0.430	KIAA1671
ilmn_2313434	3.9E-04	0.0013	0.1476	0.0007	0.0019	0.433	TCP1
ilmn_2393573	4.1E-04	-0.0019	-0.1073	-0.0027	-0.0010	0.434	RASSF1
Change in MMSE score (9 years)							
ilmn_1774761	5.5E-05	-0.0040	-0.1305	-0.0056	-0.0024	0.700	CCR2
ilmn_2374362	1.5E-04	-0.0015	-0.1345	-0.0022	-0.0009	0.700	FAM108B1
ilmn_1796094	2.2E-04	-0.0047	-0.1146	-0.0067	-0.0026	0.700	CD36
ilmn_1761941	2.6E-04	-0.0017	-0.1135	-0.0024	-0.0009	0.700	C4orf18
ilmn_2313434	2.7E-04	0.0015	0.1273	0.0008	0.0022	0.700	TCP1
ilmn_1712684	3.1E-04	-0.0011	-0.1245	-0.0015	-0.0006	0.700	FAM20C
ilmn_1866887	3.4E-04	-0.0011	-0.1290	-0.0016	-0.0006	0.700	BX537605
ilmn_1772821	5.6E-04	-0.0016	-0.1277	-0.0023	-0.0008	0.700	KIAA1671
ilmn_1703314	5.7E-04	0.0013	0.1167	0.0007	0.0020	0.700	KLHL36
ilmn_1789751	6.1E-04	-0.0039	-0.0985	-0.0058	-0.0020	0.700	MFSD1

Models adjusted for: age, gender, highest education level, smoking-status, site, hybridization-batch, amplification-batch, leukocyte proportions (and MMSE-baseline score, in 'change' models).

Figure 5.1 | Quantile-Quantile plots for gene expression analysis of MMSE score at RNA collection



A. The Quantile-Quantile plot for the genome wide analysis of MMSE at RNA collection is shown. The actual P-values ($-\log_{10}$) obtained are given on the y-axis, plotted against expected P-values ($-\log_{10}$) given on the X-axis. This graph shows potential deviations to the p-value distribution that might be expected by chance. **B.** The Quantile-Quantile plot for the focused analysis of inflammatory genes is shown. The actual P-values ($-\log_{10}$) obtained are given on the y-axis, plotted against expected P-values ($-\log_{10}$) given on the X-axis. This graph shows potential deviations to the p-value distribution that might be expected by chance. The positive association with *CCR2* transcript is the outlier.

5.5.3 Specific analysis of inflammation related transcripts

In a targeted analysis on inflammation-related transcripts only, the *CCR2* transcript was associated with MMSE score at RNA collection (beta=-0.16, $p=5.1x10^{-6}$, q=0.003; **table 5.3**) and also with cognitive decline score over the previous nine years, using a false-discovery rate (FDR) of q<0.1 (beta=-0.13, $p=5.5x10^{-5}$, q=0.03; **table 5.3**; **figure 5.2**). As in the genome wide analysis, other large-scale alterations in the expression of inflammatory genes were not present (after accounting for multiple testing) for either cognitive function or decline, although one large effect (*CCR2*) was evident on the QQ plot (**figure 5.1b**).

5.5.4 Modeling of relationship between CCR2 transcript expression and MMSE/delta MMSE using spline analysis

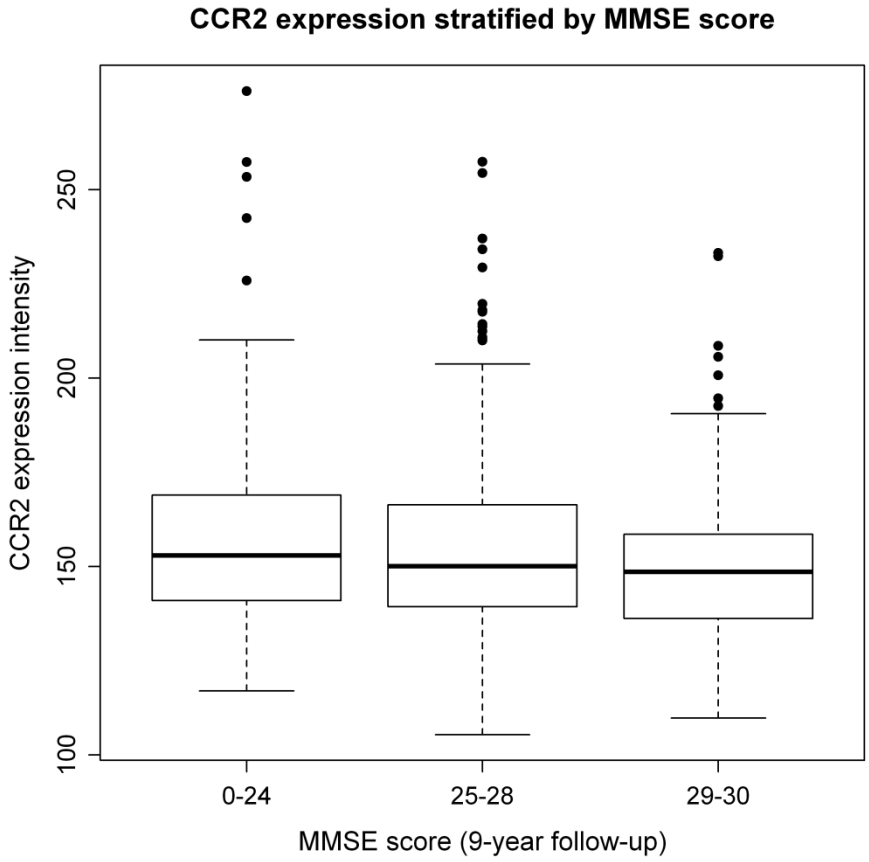
To understand any non-linearity in the relationship between *CCR2* expression and MMSE (and change in MMSE over 9 years, delta-MMSE) we fitted a penalized cubic spline regression model for; A. *CCR2* expression and MMSE at RNA collection, and B. *CCR2* expression and delta-MMSE over the preceding nine years (see **figure 5.3** for spline plots). We found the relationships to be approximately linear throughout the MMSE or delta-MMSE ranges.

Table 5.3 | The ten inflammation-related probes most closely associated with MMSE score at wave 9 and change in MMSE score (baseline to year 9), ordered by false discovery rate q-value (n observations = 688).

Probe ID	p-value	Coefficient	Beta	95% CIs		q-value	Gene
MMSE score at year 9							
ilmn_1774761	5.1E-06	-0.0042	-0.1616	-0.0056	-0.0027	0.003	CCR2
ilmn_2366212	1.5E-03	0.0083	0.1267	0.0040	0.0125	0.232	CD79B
ilmn_1677440	2.6E-03	-0.0055	-0.1194	-0.0084	-0.0025	0.262	ATP6AP2
ilmn_2276996	2.8E-03	-0.0009	-0.1071	-0.0014	-0.0004	0.266	CCR2
ilmn_1785439	3.9E-03	0.0042	0.1239	0.0018	0.0066	0.279	CD79B
ilmn_1710017	3.9E-03	0.0045	0.1302	0.0019	0.0070	0.279	CD79B
ilmn_1764396	5.3E-03	-0.0011	-0.0725	-0.0017	-0.0004	0.288	HDAC4
ilmn_1763875	6.1E-03	0.0022	0.0842	0.0009	0.0035	0.291	ABCF1
ilmn_1738767	6.5E-03	-0.0012	-0.0780	-0.0019	-0.0005	0.293	PLP2
ilmn_1747227	7.1E-03	0.0008	0.1412	0.0003	0.0013	0.295	ADORA1
Change in MMSE score (9 years)							
ilmn_1774761	5.5E-05	-0.0040	-0.1305	-0.0056	-0.0024	0.033	CCR2
ilmn_2366212	2.1E-03	0.0087	0.1047	0.0041	0.0134	0.305	CD79B
ilmn_1710017	2.2E-03	0.0052	0.1013	0.0024	0.0079	0.310	CD79B
ilmn_1677440	2.9E-03	-0.0059	-0.0946	-0.0091	-0.0027	0.325	ATP6AP2
ilmn_1747227	3.2E-03	0.0009	0.1169	0.0004	0.0015	0.332	ADORA1
ilmn_1737398	3.3E-03	-0.0012	-0.0938	-0.0018	-0.0005	0.332	PTPLAD1
ilmn_1785439	4.1E-03	0.0046	0.1024	0.0020	0.0072	0.349	CD79B
ilmn_1682312	6.5E-03	-0.0025	-0.0726	-0.0040	-0.0010	0.377	CYBB
ilmn_1811049	6.8E-03	0.0012	0.0999	0.0005	0.0019	0.379	POU2AF1
ilmn_1771333	6.9E-03	-0.0035	-0.0877	-0.0057	-0.0014	0.380	CD47

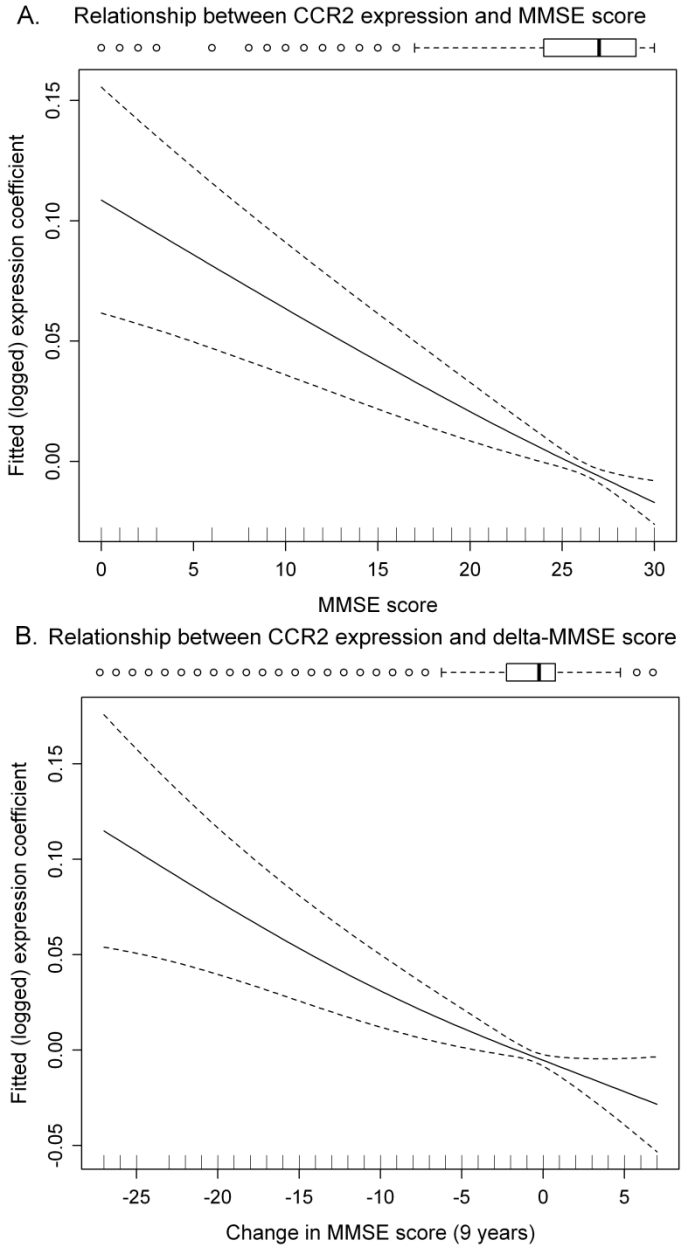
Models adjusted for: age, gender, highest education level, smoking-status, site, hybridization-batch, amplification-batch, leukocyte proportions (and MMSE-baseline score, in 'change' models).

Figure 5.2 | Boxplot of CCR2 expression by MMSE score at RNA collection



The boxplot shows CCR2 transcript expression levels (relative units) as plotted on the Y-axis by MMSE score RNA collection on the X-axis.

Figure 5.3 | Cubic Spline plots for linearity of association between CCR2 expression and MMSE



A. CCR2 expression and MMSE at RNA collection, and **B**. CCR2 expression and delta-MMSE over the preceding nine years. These penalized cubic regression splines visualize the fitted (adjusted) relationship between CCR2 gene expression and MMSE score, and separately with change in MMSE over 9 years. The relationships are approximately linear after normalization (log-transformed expression data) and adjustment for multiple confounders.

5.5.5 Quantitative real-time PCR validation of CCR2 levels.

To validate our microarray results, we also quantified *CCR2* expression in a subset of our cohort using quantitative real-time PCR (QRT-PCR). We found that the expression level of *CCR2* transcripts as measured by QRT-PCR correlated well with the expression levels measured by microarray analysis (r^2 = 0.5, p=2 X 10^{-4}). PCR-measured logged *CCR2* expression is significantly associated with MMSE is a one-sided test (B = -0.018, p = 0.096), however in this subset of participants the Illumina microarray probe is only modestly significant (B = -0.005, p = 0.013), suggesting power in 100 participants is a limiting factor.

5.5.6 Gene set enrichment analysis

No gene sets showed evidence of deregulation in association with MMSE score at RNA collection or change in score using GSEA (Gene Set Enrichment Analysis), at a FDR q-value of <0.1 (**table 5.4**).

Table 5.4 | Biological and molecular functional pathway enrichments for negative correlation with MMSE are presented

Pathway	Size	ES	p-value	q-value
Aromatic compound metabolic process	17	-1.82	0	0.85
Regulation of DNA metabolic process	29	-1.74	0.008	0.95
T cell activation	27	-1.56	0.031	1
Transferase activity transferring glycosyl groups	68	-1.42	0.034	1
Cellular defense response	38	-1.52	0.035	1
Anatomical structure formation	27	-1.55	0.036	1
Positive regulation of lymphocyte activation	15	-1.58	0.039	1
Coenzyme metabolic process	30	-1.44	0.045	1
RNA export from nucleus	17	-1.52	0.045	1
DNA recombination	32	-1.5	0.047	1

ES refers to the non-specific enrichment score.

5.5.7 Post-hoc Sensitivity analyses

To determine if the observed association is dependent on the inclusion of some younger people in the predominantly older sample, or dependent on the oldest old, we carried out a sensitivity analysis in subjects aged ≥70 and <90 years old at RNA-extraction; MMSE score and (logged) *CCR2* expression were still strongly associated (beta=-0.15, p = $2e^{-4}$) within this age-range.

ApoE genotype was available on n=480 (excluding e2e4 n=4) in our sample: *CCR2* expression was positively associated with ApoE risk haplotype (trend test across e2e2, e2e3, e3e3, e3e4, e4e4 groups: coef=0.02, CIs: 0.007 to 0.03, p=0.001). The association between *CCR2* expression and MMSE at RNA collection was attenuated but remained significant after additional adjustment for ApoE status (beta = -0.1, p = 0.018). There was no statistical interaction between ApoE status and *CCR2* expression, and ApoE status accounted for 1.6% of the variance in *CCR2* expression. ApoE status is not significantly associated with MMSE (cross sectionally) or with change over 9 years in these 480 participants (p>0.05).

Co-morbidity is very common with cognitive impairment in later life in the general older population, and therefore our approach has been to avoid disease specific exclusions: however, we did examine the effect of removing those diagnosed as having had a stroke (n=76 removed): *CCR2* remained associated with MMSE score (beta = -0.14, p=0.001) in the stroke free group.

5.6 Discussion and Conclusions

In this study, we have investigated associations between blood leukocyte-derived mRNA expression and MMSE scores in an ageing population. *CCR2* expression reached inflammation-specific significance (and only narrowly missed genome wide significance after accounting for multiple testing), providing the first human population evidence of likely consistency with the *Ccr2* mouse models. Our finding that ApoE haplotype status, the major inherited genetic risk factor associated with Alzheimer's pathology, is associated with CCR2 expression is also consistent with CCR2 signaling having a direct role in cognitive decline in later life in human populations.

The MMSE score is a widely used measure of cognitive impairment in later life, although it has a ceiling effect that limits its ability to detect early subtle cognitive changes, particularly in younger highly educated participants [217], [225] and is not considered an appropriate measure to diagnose mild cognitive impairment or related prodromal diagnoses without supplementing it with other cognitive measures, although the population relevance of these clinically-derived constructs remains disputed [226]. Over several years follow-up the MMSE does however provide a useful measure of cognitive impairment and cognitive decline in participants with and without dementia [218], and is the most widely used measure to monitor cognitive function and decline in elderly adults in clinical settings. We therefore consider that it is an appropriate tool to assess age-related differences in cognitive function in our cross sectional population.

Our functional, rather than pathological, classification of participants was important for several reasons. Cognitive decline leads to impairments in activities of daily living, and is a major public health threat. It is also rarely determined purely by one pathophysiological mechanism - mixed dementia is now considered to be the most common form of dementia [227] and functional decline has only a moderate association with the pathological changes of Alzheimer disease [228]. Finally, our use of a large, population-based cohort, rather than

from specialist clinics, gave us power to detect moderate biomarker effects that could be of clinical application in population testing and risk reduction: biomarkers of functional cognitive change may hold more clinical application in the population than markers of single pathological processes.

We investigated blood leukocytes in order to identify clinically useful, minimally invasive markers of cognitive function since cortical tissue is inappropriate for the identification of cognitive decline risk or progression in the population. Furthermore, recent studies suggest that circulating systemic factors such as chemokines and their receptors have an important role in neurogenesis and cognitive function in animal models [4]. Peripheral inflammation has also been shown to impact negatively on human cognitive function [229] with chronic immune activation in the brain having an association with neurodegenerative disease and cognitive decline [230], [231].

There are both similarities and differences between our data and previous studies [215], [216], [232], [233]. It has been suggested that genes involved in cytoskeletal maintenance, cellular trafficking, cellular stress response, redox homeostasis, transcription and DNA repair may be associated with Alzheimer's Disease [226] and proteomics studies have suggested that circulating lymphocytes may be promising biomarkers for this disorder [234]. The differences could arise from different patient selection criteria; we used population data rather than data from clinical settings. Differences with prior proteomic findings in peripheral blood in cognitive decline [235], [236] may also be due to small study sizes or the possibility that circulating cytokines previously found associated with cognitive decline may be expressed by activated CNS inflammatory cells and released into the circulation, rather than being expressed by circulating leukocytes [237].

The *CCR2* receptor binds CCL2, a protein that is abundantly expressed in macrophage-rich areas of atherosclerotic plaques and in brain microglia [214] resulting in migration of macrophages and brain microglial cells to their site of action. In mouse models, vascular disruption of *Ccr2* or *Ccl2* leads to a reduction in atherosclerotic plaque formation in ApoE-

or LDL-R null mice, even when fed a high fat diet [213], [234]. This may be mediated by *Ccr2* depletion reducing macrophage infiltration. In mouse brain, *Ccr2* knockout has been shown to result in accelerated disease progression, increased mortality and an increase in soluble antibody assemblies [114], [214]. Work on *CCR2* gene knockout mice bred on a background of chimeric mouse/humanβ-amyloid precursor/presenilin over-expression (APP^{Swe}/PS1/$CCR2^{-/-}$) suggested that AD might be associated with a decreased expression of *CCR2* [114]. Although our results at first appear to conflict with this, we suggest that the increased *CCR2* levels we note in cognitively impaired individuals probably reflects a reactive increase in the need for chemoattractants in subjects with increased β-amyloid deposition. This is supported by the observation that the ligand for CCR2, CCL2, has previously been reported to be up-regulated in the brain of patients suffering from AD [238]. Recent studies have also suggested that CCR2 is with a key modulator of negative regulation of neurogenesis and cognitive function in mice [4]. Our data supports a role for CCR2 in the aetiology of age-related cognitive decline, and indicates that the CCR2 mouse models [114], [214] may have particular relevance to the human population. Although much more work is needed to confirm causality for this mechanism in human populations the mouse models alongside the statistical evidence provide good biological plausibility.

Key limitations of our study include the possibility that that there are gene expression correlates of cognitive function in specific white cell subtypes which we have not measured separately. These are not however likely to be very marked or common, as we have found only limited overall expression changes. Similarly, as noted above the MMSE score, a widely used clinical measure of cognitive function in older people, suffers from a relative insensitivity to frontal-executive dysfunction and visuo-spatial deficits [239], and a ceiling effect inhibiting sensitive differentiation between medium and high cognitive performers [240]. Our cohort may also be subject to informative loss to follow up, and reflect a higher functioning group, since people with very impaired cognitive function may not have reported for blood sampling in year 9.

We note also that several probes to *CCR2* transcripts were present on the chip, but only one showed associations with MMSE score. Differences in the relationship between phenotype and alternative specific probes for the same gene are not atypical in microarray studies. These discrepancies can arise from several factors including differences in binding dynamic of the individual probes, differences in the inter-individual intensities between probe signal which can add noise to the data and reduce power and the presence of alternatively-processed isoforms which may bind probes differentially. In the case of *CCR2*, three alternatively-expressed isoforms exist, with only a single probe, ilmn_1774761, capable of binding to all isoforms. If the effect we note is driven by a specific isoform, only those probes that identify the transcript in question will show a significant association. Since up to 90% of all genes are alternatively spliced, and the full transcriptomic output from any gene is not fully known at present, this can lead to apparent differences in the association of particular probes with disease phenotype in association studies.

Future work should first seek to replicate our *CCR2* finding, and explore associations with the full range of isoform specific probes for the main target genes. Work should also seek to gather more evidence on which of the near significant probes may in fact be associated with cognitive function. Future prospective studies will also provide valuable information as to the role of chemo-attractant proteins in age-related cognitive decline. Functional work in humans is also needed to clarify the mechanisms involved in the raised expression of *CCR2* we observed in those with lower or declining MMSE scores. Studies of the role of CCR2 in specific forms of dementia (Alzheimer's, vascular, Lewy Body etc, or mixed) would also help characterise the role of CCR2. If raised expression of *CCR2* is a helpful but insufficient response to the accumulation of β-amyloid, then attempts to increase this response further may potentially be effective. Studies of the role of *CCR2* in vascular disease and pro-inflammatory response would also be informative. Our findings raise the possibility that *CCR2* expression levels may also be associated with early, pre-symptomatic cognitive changes, although a more sensitive approach than change in MMSE score would be required to detect this.

To conclude, we have carried out the largest assessment of in-vivo leukocyte human gene expression alterations in conjunction with MMSE measured cognitive function in a predominantly older population, to date. We identified associations between MMSE score and *CCR2* transcript levels, which may reflect the key role *CCR2* plays in the removal of β-amyloid and in regulation of neurogenesis. Work is now needed to confirm our findings, establish whether these reflect the proposed mechanisms and relate the circulating transcriptome changes to specific forms of brain pathology.

5.7 Acknowledgements

This study was supported in part by the Intramural Research Program, National Institute on Aging, and the U.S. National Institutes of Health. We thank the many people who contributed to the InCHIANTI study, including all of the anonymous participants. The US National Institute on Aging intramural research program contributed the array data and supported this analysis. William Henley completed this work while seconded to PenCLARHC which is funded by the National Institute of Health Research (NIHR). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

5.8 Supplementary Information

Supplementary table 5.1 | This table shows the inflammatory gene set as defined by the Gene Ontology (GO) project (www.geneontology.org). N refers to the number of genes in the pathway

GO - Pathway Name	N[genes]	N[genes in data]	%
Activation of immune response	17	12	70.6
Acute inflammatory response	11	5	45.5
Adaptive immune response	25	14	56.0
Adaptive immune response go 0002460	24	13	54.2
Antigen binding	28	15	53.6
B cell activation	19	13	68.4
Chemokine activity	42	15	35.7
Cytokine activity	111	38	34.2
Cytokine and chemokine mediated signaling pathway	23	13	56.5
Cytokine binding	46	33	71.7
Cytokine biosynthetic process	40	24	60.0
Cytokine metabolic process	41	25	61.0
Cytokine production	72	46	63.9
Cytokine secretion	18	10	55.6
Cytokinesis	19	8	42.1
Humoral immune response	31	17	54.8
Immune effector process	38	22	57.9
Immune response	232	149	64.2
Immune system development	80	52	65.0
Immune system process	326	218	66.9
Immunological synapse	11	9	81.8
Inflammatory response	124	81	65.3
Innate immune response	23	8	34.8
Interleukin 1 secretion	10	8	80.0
Jak stat cascade	31	23	74.2
Jnk cascade	47	32	68.1
Leukocyte activation	67	42	62.7
Lymphocyte activation	60	38	63.3
Negative regulation of cytokine biosynthetic process	11	5	45.5
Negative regulation of immune system process	14	7	50.0
Nuclear speck	11	7	63.6
Positive regulation of cytokine biosynthetic process	24	16	66.7
Positive regulation of cytokine production	15	11	73.3
Positive regulation of cytokine secretion	10	6	60.0
Positive regulation of immune response	29	16	55.2
Positive regulation of immune system process	51	31	60.8
Production of molecular mediator of immune response	13	7	53.8
Regulation of cytokine biosynthetic process	37	22	59.5
Regulation of cytokine production	25	16	64.0
Regulation of cytokine secretion	16	8	50.0
Regulation of immune effector process	15	9	60.0
Regulation of immune response	33	16	48.5
Regulation of immune system process	67	41	61.2
T cell activation	44	27	61.4
	2031	1228	Mean = 59%

Chapter 6 – Analysis 4 – Age-related Inflammation

Gene expression markers of age-related inflammation: Transcriptome-Wide Analysis in Two Human Cohorts

(In submission) *Experimental Gerontology* (December, 2014)

Pilling, Luke C *

Joehanes, Roby *

Melzer, David *

Harries, Lorna W

Henley, William

Dupuis, Josée

Lin, Honghuang

Mitchell, Marcus

Hernandez, Dena

Ying, Sai-Xia

Lunetta, Kathryn L

Benjamin, Emelia J

Singleton, Andrew

Levy, Daniel

Munson, Peter

Murabito, Joanne M #

Ferrucci, Luigi #

*,# = contributed equally (first/last respectively)

6.1 Overview

Interleukin-6 (IL-6) is a circulating cytokine and is a marker of chronic inflammation, the levels of which increase with age. In this analysis of peripheral (leukocyte) gene expression the aim was to determine which genes statistically mediated (cross-sectionally) a proportion of the association between age and IL-6. The rationale for a "mediation" analysis is that genes that are statistically related to this association are likely to be markers of age-related inflammation, rather than inflammation *independent* of age per se.

For this analysis I determined the methods to use, undertook the bioinformatics analysis (transcriptome-wide screen) in InCHIANTI (and worked with Roby Joehanes who undertook the analysis of the FHS data). I also organised and chaired the calls with the FHS, helped perform the literature review, interpreted the results, and co-wrote wrote the manuscript (with Professor David Melzer). The other co-authors generated the population-level and lab-based data for analysis, and contributed to discussions during the conference calls.

6.2 Summary

Introduction: Chronically elevated circulating inflammatory markers are common in older persons but mechanisms are unclear. Many blood transcripts (>800 genes) are associated with interleukin-6 protein levels (IL-6) independent of age. We aimed to identify gene transcripts statistically mediating the age~IL-6 association, to define a signature specific to age-related inflammation.

Methods: Blood derived in-vivo RNA from the Framingham Heart Study (FHS, n=2422, ages 40-92 yrs) and InCHIANTI study (n=694, ages 30-104 yrs), with Affymetrix and Illumina expression arrays respectively (>17,000 genes tested), were tested for statistical mediation of the age~IL-6 association using resampling techniques, adjusted for confounders and multiple testing.

Results: In FHS, *IL6* expression was not associated with IL-6 protein levels in blood. 102 genes (0.6% of 17,324 expressed) statistically mediated the age~IL-6 association of which 25 replicated in InCHIANTI (including 5 of the 10 largest effect genes). The largest effect gene (*SLC4A10*, coding for NCBE, a sodium bicarbonate transporter) mediated 19% (adjusted CI 8.9 to 34.1%) and replicated by PCR in InCHIANTI (n=194, 35.6% mediated, p=0.01). Other replicated mediators included *PRF1* (perforin, a cytolytic protein in cytotoxic T lymphocytes and NK cells) and *IL1B* (Interleukin 1 beta): few other cytokines were significant mediators.

Conclusions: This transcriptome-wide study on human blood identified a small distinct set of genes that statistically mediate the age~IL-6 association. Findings are robust across two cohorts and different expression technologies. Raised IL-6 levels may not derive from circulating white cells in age related inflammation.

6.3 Introduction

Chronically elevated levels of pro-inflammatory biomarkers are a core feature of aging, and a risk factor for many diseases and adverse phenotypes that are frequent in older persons [38]. Increased levels of pro-inflammatory markers in blood and other tissues with aging are paralleled by a progressive decline in overall immune responsiveness [60].

Elevated Interleukin-6 (IL-6) and C-reactive protein (CRP) have emerged as robust age-related risk factors for multiple adverse outcomes including several diseases, disability, cognitive impairment and death [241]. IL-6 is a pleiotropic cytokine produced by several cell types, including immune cells, hepatocytes, vascular endothelial cells, adipocytes and skeletal muscle [241]. IL-6 levels are often undetectable in young individuals but increase with advancing age, even in the absence of detectable causes, including cardiovascular risk factors and disease [242]. High levels of IL-6 predict all major age-associated diseases, physical and cognitive disability and mortality [243]–[246]. The mechanisms that cause and sustain high levels of inflammatory markers in aging are largely unknown.

The overproduction of pro-inflammatory markers may occur in many sites beyond circulating leukocytes; candidates include tissue resident macrophages, adipocytes, endothelial cells, cells within atherosclerotic plaques and muscle cells [247]. The accumulation of cells expressing a senescence-associated secretory phenotype [110], [248] in different tissues, possibly induced by age-related NF-kB dysregulation [249], may also be a possible cause. Data collected in animal models have provided inconsistent results and there is a scarcity of data in humans.

In a recent study, the authors determined the whole blood gene expression transcripts associated with IL-6 levels in 2 human populations, independent of age [61]. That analysis identified 4139 genes that were significantly associated with interleukin-6 levels (FDR<0.05), independent of age, sex and blood cell components, of which 807 genes replicated in the smaller InCHIANTI cohort. Many of the top genes generally associated with blood IL-6

protein levels (independent of age) are in inflammation-related pathways or erythrocyte function, including the JAK/Stat signalling pathway and interleukin-10 signalling pathway.

In the study presented here, we aimed to identify the subset of IL-6 associated gene transcripts that might be specific to the chronic inflammation of aging. To do this we utilized data from 2 independent community-based cohorts in a discovery (Framingham Heart Study, FHS) and replication (InCHIANTI) analysis. The cohorts utilize two different microarray platforms, allowing replication of findings robust to cohort and microarray differences. For each gene transcript measured on the microarray we tested whether the expression levels statistically mediate “age-related inflammation” (i.e. account for a significant portion of the statistical association between age and IL-6), as these may reflect the most relevant molecular pathways in aging, and additionally may help to identify the specific cell subtypes most closely involved.

6.4 Methods

6.4.1 The participants

Our study was performed in two independent and well-characterized human cohorts. The discovery cohort was the Framingham Heart Study (FHS) Offspring cohort, USA [185], with replication of the significant mediators in the InCHIANTI (Invecchiare in Chianti, aging in the Chianti area) study, a community-based study cohort study of aging in Florence, Italy [115].

FHS participants targeted in this analysis were from the Offspring Cohort enrolled in 1971 as the offspring (and offspring spouses) of the Original FHS cohort. Offspring participants who attended examination 8 (2005-2008) and had blood-derived RNA prepared were included in this analysis. This study was approved by the Institutional Review Boards at Boston University Medical Center, and all participants gave written informed consent [250].

InCHIANTI participants were originally enrolled in 1998-2000, and were interviewed and examined every 3 years. Ethical approval was granted by the Instituto Nazionale Riposo e Cura Anziani Institutional Review Board in Italy. Participants gave informed consent to participate. RNA was available at wave 4 (year 9) of the study, with IL-6 also measured at year 9. All the data required for the full analyses were available for 694 individuals (see **Table 6.1** for further cohort details).

6.4.2 RNA collection and extraction

FHS: The methods for gene expression profiling were previously published [251]. Briefly, peripheral blood samples were extracted using the PAXgene Blood mRNA kit (PreAnalytiX, Hombrechtikon, Switzerland), and amplified by the WT-Ovation Pico RNA Amplification System (NuGEN, San Carlos, CA), according to manufacturer's instructions. cDNA was then hybridized to the Human Exon 1.0 ST Array (Affymetrix, Inc., Santa Clara, CA) for

quantification. The raw data were quantile-normalized and natural-log transformed, followed by summarization using Robust Multi-array Average [252]. The gene annotations were obtained from Affymetrix NetAffx Analysis Center (version 31). We excluded transcript clusters that were not mapped to RefSeq transcripts, resulting in 17,873 distinct transcripts (17,324 unique gene identifiers) for downstream analysis.

Table 6.1 | Summary statistics of the Framingham and InCHIANTI cohort characteristics

Framingham Heart Study Offspring				**InCHIANTI**		
Sex	***N***	***%***		**Sex**	***N***	***%***
Men	1,093	45.1		Men	313	45.1
Women	1,329	54.9		Women	381	54.9
Age, years				**Age, years**		
40-54	199	8.2		30-49	88	12.7
55-69	1,359	56.1		50-69	100	14.4
70-84	817	33.7		70-89	477	68.7
85-92	47	1.9		90-104	29	4.2
Mean (SD)	2422	66.4 (9.0)		Mean (SD)	694	72.2 (15.3)
Tobacco Exposure				**Tobacco Exposure**		
None	775	32.0		None	380	54.8
Former Smoker	1444	59.6		Former Smoker	240	34.6
Current Smoker	203	8.4		Current Smoker	74	10.7
	N	***Mean (SD)***			***N***	***Mean (SD)***
BMI (kg/m2)	2422	28.5 (5.4)		**BMI (kg/m2)**	694	27.1 (4.3)
Interleukin 6 (pg/mL)	2422	2.7 (3.0)		**Interleukin 6 (pg/mL)**	694	3.8 (2.9)
Leukocyte Composition				***Leukocyte Composition***		
Neutrophils	2422	59.8 (7.9)		Neutrophils	694	57.5 (9.1)
Lymphocytes	2422	27.0 (7.5)		Lymphocytes	694	30.8 (8.7)
Monocytes	2422	9.2 (1.9)		Monocytes	694	8.0 (2.1)
Eosinophils	2422	3.3 (1.6)		Eosinophils	694	3.2 (2.1)
Basophils	2422	0.8 (0.2)		Basophils	694	0.6 (0.2)

InCHIANTI: Peripheral blood samples were also extracted using the PAXgene Blood mRNA kit according to the manufacturer's instructions [117], which preserves transcript expression levels as at the time of blood sampling. Samples were collected in 2008/9 (wave 4) from 733 participants. Whole genome expression profiling of the samples was conducted using the Illumina Human HT-12 microarray (Illumina, San Diego, USA) as previously described [155]. Data processing was done using the Illumina and BeadStudio software (Illumina, San Diego, USA) as previously described. All microarray experiments and analyses complied with MIAME guidelines [219]. Participants were excluded if mean signal intensities across all probes with $p \leq 0.01$ were >3 standard deviations from the cohort mean; probes with <5% of participants giving intensities with $p \leq 0.01$ different from background also were excluded, 3 further exclusions were made due to missing leukocyte data. More detailed methods have been previously published [79]. Data from 695 individuals and 16,571 probes (11,393 unique gene identifiers) passed our quality control process and went forward into our analyses. The expression data were first normalized using natural log transformation and standardized using z-scores.

6.4.3 Serum Protein Measures

FHS: The interleukin-6 concentration was assayed by the quantitative enzyme-linked immunosorbent assay according to the manufacturers' protocols (R&D Systems, Minneapolis, MN, USA). The minimum detectable concentration was 0.039 pg/ml. Ten percent of measures were run in duplicate. Standard quality control was performed, and the mean intra-assay coefficient of variation was 4.0% and the inter-assay coefficient of variation was 3.7% [253] (http://www.framinghamheartstudy.org/researchers/description-data/vascular-manuals/offspring_exam8_omni1_exam3_marker_manual.pdf).

InCHIANTI: Serum IL-6 was quantified using a quantitative sandwich enzyme-linked assay (BioSource Cytoscreen UltraSensitive kits, BioSource International Inc., Camarillo, CA,

USA). The lowest detectable concentration was 0.10 pg/ml and the interassay coefficient of variation (CV) was 7%.

6.4.4 Blood Cell count Measurement

FHS: We imputed the percentage of the immune cells comprising different cell types: neutrophil, lymphocyte, monocyte, eosinophil and basophil. The percentages of each cell type were then normalized, where the negative predicted values were set to 0 and the sum of the percentages for all cell types was set at 100% [personal communication, RJ].

InCHIANTI: Assessments of the number of red blood cells, white blood cells, platelets, the hemoglobin concentration were performed through an automated system at the Laboratory of Clinical Chemistry and Microbiological Assays, SS. Annunziata Hospital, Azienda Sanitaria 10, Florence, Italy, using a Coulter LH 750 Hematology Autoanalyzer (Beckman Coulter Inc, Brea, CA, USA).

6.4.5 Statistical analyses

In all models using IL-6 as the outcome we used natural log of the serum levels to reduce the skewness inherent to the measure.

Mediation analysis: Using the R statistical software application (v2.14.1) [190] and package 'mediation' [254] (v4.2) we individually assessed whether each transcript statistically mediated the association between logged interleukin 6 serum levels (the 'treatment' as it is referred to in the package) and age (the 'exposure' that affects the treatment, and may be statistically mediated by the expression of genes in white cells – although no causality is inferred here). In each model (1 per gene) the mediating variable was the expression level of

that particular gene. The common approach to testing the effect of a potential mediator is to adjust for the proposed mediator in a linear/logistic model, and determine by how much the effect size is changed. Whilst this tells the analyst whether an association is affected by a mediator, it does not provide an estimate of statistical significance or confidence intervals for the mediation effect; hence we are using the package 'mediation' as this re-samples the data to determine confidence intervals for the mediation estimate.

Using parametric linear regression models throughout, we modeled the mediating effects for each transcript on the association between IL-6 and age. Each model was adjusted for the following independent variables; sex, technical covariates (including microarray batches and study site – these are specific to each cohort) and, leukocyte sub-type proportions. The cell type proportions included lymphocytes, monocytes, eosinophils and basophils, but neutrophil percentage was not included (as it was closely negatively associated with lymphocyte proportion), to avoid these measures totaling 100% and leaving no degrees of freedom in the model. The package uses quasi-Bayesian resampling simulations to estimate the uncertainty of the effects derived from our parametric models. After exploration of the package it was decided that 10,000 simulations per model was sufficient to give relatively stable estimates whilst keeping computation time realistic.

Although the package provides a p-value estimate, this was not of high enough resolution to control for multiple testing and was susceptible to sample distribution biases (following correspondence with package authors – who agreed with the following solution). We have therefore used adjusted confidence-intervals (CIs) to reflect the number of tests performed, akin to the Bonferroni adjustment [255]. In FHS the number of array probes (and thus models) available was 17,873, so the CI were set to 99.99972% (1-(0.05/17873)). Significant mediators are those with CIs that do not cross the null (0) value. When we refer to "genome-wide significant mediators", or "significant mediators in FHS", these were determined by this CI-adjusted (Bonferroni) method.

6.4.6 Concordance with InCHIANTI

Genes that mediate a proportion (≥5%) of the association between age and IL-6 after adjustment for multiple-testing in FHS were then assessed in the InCHIANTI study. Each test was conducted as described in the previous section for the FHS, except 95% confidence intervals that did not cross the null (0) were considered significant replication (rather than the multiple-testing adjusted confidence intervals applied to the discovery analysis in the FHS cohort). We are therefore using a discovery/replication procedure, where any findings significant in the discovery analysis after adjustment for multiple testing are considered independent tests for replication, and thus we accepted $p<0.05$ (i.e. 95% confidence intervals) as significant.

6.4.7 InCHIANTI: PCR analysis of gene expression

A number of genes including *IL6* and *SLC4A10* did not have gene expression information available in the InCHIANTI microarray data. We selected 2 random subsamples of 200 InCHIANTI participants for additional analysis and validation using polymerase-chain-reaction (PCR) techniques. Subsets of the participants were chosen because these are historic samples and the RNA is a finite resource.

Firstly, PCR 383-well plates were used to assess the expression levels of *SLC4A10, DAAM2* and *FLT3*; total RNA (45–85 ng) was reverse transcribed in 13-uL reactions using the Superscript III VILO kit (Invitrogen, Paisley, UK), according to the manufacturer's instructions. Expression levels for each target transcript were then measured in triplicate 5ul reactions using 384-well plates on the ABI Prism 7900HT platform (Life Technologies, Foster City, CA, USA), using commercially available assays (*DAAM2* assay Id Hs00322497_m1, *FLT3* assay Id Hs00174690_m1, *SLC4A10* assay Id Hs00222849_m1). Gene expression levels were calculated relative to the mean crossing point of the

endogenous control genes *IDH3B*, *GUSB* and *18s*, and normalized to the median *DAAM2*, *FLT3* or *SLC4A10* expression level across the sample.

Secondly, TaqMan Low Density Arrays were used to measure *IL6* and *CDKN2A* expression data. Total RNA (30–170 ng) was reverse transcribed in 20-lL reactions using the Superscript III VILO kit (Invitrogen, Paisley, UK), according to the manufacturer's instructions. Expression levels for each target transcript were then measured using the TaqMan Low Density Array approach on the ABI Prism 7900HT platform (Life Technologies, Foster City, CA, USA), using commercially available assays (*IL-6* assay Id Hs00985639_m1, *CDKN2A* assay Id Hs00923894_m1). Gene expression levels were calculated relative to the mean crossing point of the endogenous control genes *GUSB* and *18s*, and normalized to the median *IL6* or *CDKN2A* expression level across the 200 samples.

6.4.8 Mediation Analysis in InCHIANTI PCR Subset

To assess the statistical significance of the genes unavailable in the InCHIANTI microarray data but were selected for PCR analysis, we used the Sobel test, as it provides higher power in small samples sizes when compared to bootstrapping methods [135].

6.4.9 Analysis of significant mediators

Clustering: Principal components analysis (R package: psych) was performed to determine the correlation structure of the significant mediators in FHS. Default options were used (including "Varimax" rotation). The number of components to extract was determined by Scree analysis.

Pathways: DAVID functional annotation tools [256] were used to test for enrichment of pathways or processes in the significant mediators identified by FHS.

6.4.10 Sensitivity Analyses

We ran the analysis using a binary phenotype of IL-6 using an empirically determined clinically relevant cut-point (>3.5pg/mL) [58]. To assess the effect of adjusting for whole white blood cell (WBC) count (number of cells in sample) in addition to the proportions of major white cell subtypes, we ran the analysis with and without WBC count.

6.5 Results

6.5.1 Characteristics of the samples

The cohorts differ in size and age-distributions, but are otherwise broadly similar (**Table 6.1**). Overall, 2422 participants from FHS (age-range: 40-92 years) were eligible for the discovery analysis. 694 participants (age-range: 30-104 years) from the InCHIANTI study had complete data for the replication analysis.

6.5.2 Serum IL-6 protein levels associated with age but not with leukocyte IL-6 transcript abundance

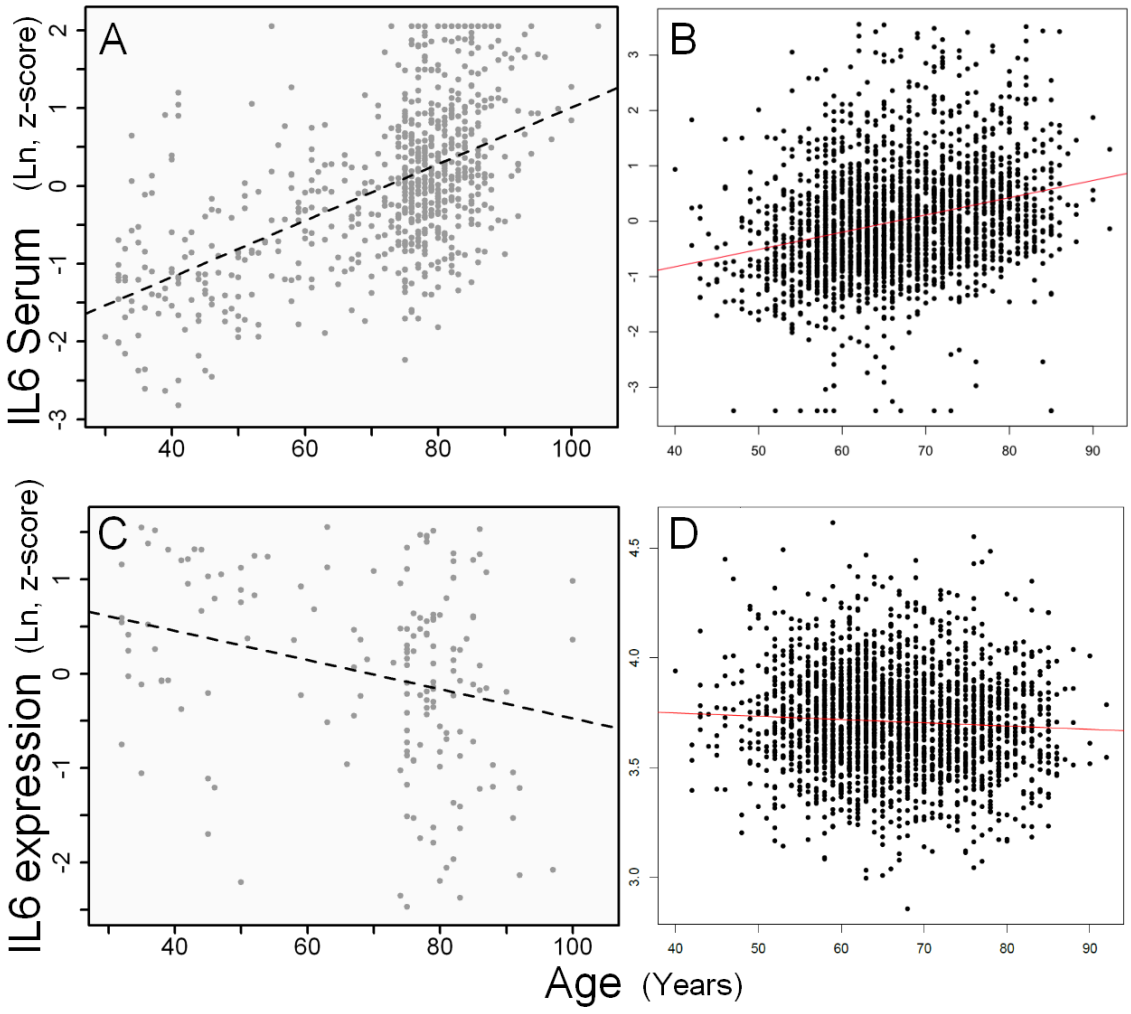
Interleukin-6 serum (protein) levels (IL-6) were strongly positively associated with advancing age both in FHS (N=2422; unadjusted: beta=0.031, $p=1x10^{-44}$) and InCHIANTI (N=694; unadjusted: beta=0.038, $p=1\ x10^{-61}$, adjusted: beta=0.036, $p=1.1\ x10^{-36}$) (**Figure 6.1 A** and **B**). Circulating IL-6 receptor (a pro-inflammatory factor that enhances IL-6 activity) is associated with age independent of sex (coef = 0.0029, p = 2.2x10-5) but not when adjusted for IL-6 levels (coef = 0.0016, p = 0.054). Expression of IL-6 transcripts (from circulating blood leukocytes, see methods) was not associated with age in FHS and mildly negatively associated with age in InCHIANTI (Figure 1C & D) (FHS: N=2422, coefficient=0.013, p=0.12; InCHIANTI: N=186, coefficient=-4.7, $p=5.6\ x10^{-4}$). Surprisingly, *IL6* transcript levels also were not significantly correlated with IL-6 protein levels in serum in either study (FHS: coefficient=0.0064, p= 0.26; InCHIANTI: coefficient=0.021, p=0.76).

6.5.3 Immune cell sub-type associations

In InCHIANTI, directly measured lymphocyte, neutrophil, monocyte and eosinophil percentages were each associated with higher mean age (Beta= -0.1, 0.09, 0.02, 0.02; P=

$9x10^{-9}$, $1x10^{-4}$, $2x10^{-4}$, $3x10^{-3}$, respectively); lymphocyte, neutrophil and basophil percentages were associated with serum IL-6 adjusting for age (Beta= -2.9, 2.8, -0.03; P= $4x10^{-14}$, $5x10^{-12}$, $1x10^{-3}$, respectively); lymphocyte and neutrophil percentages mediated the association between age and IL-6 (proportion of effect mediated= 0.09, 0.06 respectively); the ratio between lymphocyte and neutrophil percentages also significantly mediated the age~IL-6 association (proportion of effect mediated= 0.06).

Figure 6.1 | The associations between interleukin-6 and age



A) IL-6 protein serum levels and age in 694 InCHIANTI samples.
B) IL-6 protein serum levels and age in 2422 FHS samples.
C) *IL-6* gene expression and age in leukocytes in 186 InCHIANTI samples.
D) *IL-6* gene expression and age in leukocytes in 2422 FHS samples.

6.5.4 A small set of transcripts partially mediate the serum IL-6 – age association

Seven gene transcripts statistically mediate more than 5% of the association between age and IL-6 in both studies, although the approach is quite conservative and there may be other mediators we did not have the power or coverage to detect:

Transcripts for 17,324 unique genes could be tested in FHS, of which 4693 (27.1%) were associated with age (FDR<0.05) and 4140 (23.9%) were associated with IL-6 [61]. However, expression of only 102 (0.6% of the 17,324) genes significantly mediated a proportion of the association between age and IL-6 (after adjustment for confounders – see methods - and multiple testing using 99.99972% confidence intervals, and after excluding 3 invalidly annotated transcripts, **Supplementary Table 6.1**). Twenty-nine genes mediated ≥5% of the association, with *SLC4A10* (solute carrier family 4, sodium bicarbonate transporter) mediating 19% of the association (CI 8.9 to 34.1%) in FHS (**Table 6.2**).

The InCHIANTI Illumina array provided data on n=11,393 genes, substantially fewer than the 17,324 measured in FHS. Data on 9,971 unique genes were available in both cohorts with an additional 7,345 measured in FHS only and 1,422 in InCHIANTI only. Of the 102 genes partial mediators of the age~IL-6 association identified in FHS, 88 were available (**Supplementary Table 6.1**) in InCHIANTI and 17 of them significantly mediated the association of age with IL-6 after adjustment for confounders. Of the 29 genes mediating ≥5% in FHS (**Table 6.2**), 22 were also in the InCHIANTI microarray data. Significant associations in the same direction were found for *IL1B*, *VCAN*, *PRF1* and *LTB*. No genes mediating ≥5% had significant associations in both cohorts but different directions of effect. Four of those mediating <5% were significant and with opposite directions of effect (**Supplementary Table 6.1**).

Table 6.2 | Transcripts mediating ≥5% of the age~IL-6 association in FHS, with array based replication estimates from the InCHIANTI study

				FHS			InCHIANTI			
N	**Entrez**	**Gene**	**Description**	**Prop ł**	**99.9% CIs ˜**		**Prop ł**	**95% CIs**		*****
1	57282	*SLC4A10*	solute carrier family 4, sodium bicarbonate transporter	0.191	0.089	0.341				
2	2322	*FLT3*	fms-related tyrosine kinase	-0.099	-0.211	-0.04				
3	9839	*ZEB2*	Zinc Finger E-Box Binding Homeobox 2	0.098	0.032	0.192	0	-0.008	0.008	
4	8460	*TPST1*	tyrosylprotein sulfotransferase 1	-0.09	-0.182	-0.036	-0.025	-0.07	0.012	
5	1432	*MAPK14*	mitogen-activated protein kinase 1	0.086	0.03	0.172	0.001	-0.01	0.015	
6	3553	*IL1B*	interleukin 1, beta	-0.081	-0.171	-0.032	-0.052	-0.093	-0.023	*
7	23500	*DAAM2*	dishevelled associated activator of morphogenesis 2	-0.079	-0.176	-0.027				
8	2153	*F5*	coagulation factor V (proaccelerin, labile factor)	0.075	0.026	0.148	0.014	-0.007	0.039	
9	1462	*VCAN*	versican	0.073	0.024	0.15	0.043	0.018	0.076	*
10	3820	*KLRB1*	killer cell lectin-like receptor subfamily B, member 1	0.068	0.019	0.135	0.016	-0.007	0.043	
11	5551	*PRF1*	perforin 1 (pore forming protein)	-0.064	-0.14	-0.001	-0.057	-0.104	-0.021	*
12	3267	*AGFG1*	ArfGAP with FG repeats 1	0.064	0.02	0.147	-0.003	-0.022	0.013	
13	51099	*ABHD5*	abhydrolase domain containing 5	-0.062	-0.145	-0.017	0.002	-0.004	0.015	
14	400360	*C15orf54*	chromosome 15 open reading frame 54	0.062	0.017	0.123				
15	116369	*SLC26A8*	solute carrier family 26, member 8	0.062	0.005	0.128	0.012	-0.008	0.037	
16	768211	*RELL1*	RELT-like 1	-0.06	-0.173	-0.011				
17	26253	*CLEC4E*	C-type lectin domain family 4, member E	-0.059	-0.135	-0.016	-0.012	-0.037	0.008	
18	2204	*FCAR*	Fc fragment of IgA, receptor for	0.059	0.019	0.12	0.005	-0.004	0.021	
19	55350	*VNN3*	vanin 3	-0.056	-0.133	-0.01	-0.003	-0.019	0.008	
20	147947	*ZNF542*	zinc finger protein 542	0.054	0.016	0.115				
21	285533	*RNF175*	ring finger protein 175	0.053	0.012	0.125	0	-0.006	0.005	
22	330	*BIRC3*	baculoviral IAP repeat-containing 3	-0.053	-0.119	-0.017	-0.01	-0.033	0.007	
23	4609	*MYC*	v-myc myelocytomatosis viral oncogene homolog (avian)	-0.052	-0.128	0	-0.048	-0.103	0	
24	1374	*CPT1A*	carnitine palmitoyltransferase 1A (liver)	0.052	0.004	0.13	-0.001	-0.011	0.005	
25	22901	*ARSG*	arylsulfatase G	-0.052	-0.124	-0.012	0.001	-0.007	0.012	
26	317	*APAF1*	apoptotic peptidase activating factor 1	-0.051	-0.133	-0.013	0.013	-0	0.034	
27	84255	*SLC37A3*	solute carrier family 37 (glycerol-3-phosphate transporter), 3	0.051	0.014	0.109	-0.004	-0.019	0.004	
28	55356	*SLC22A15*	solute carrier family 22, member 15	-0.05	-0.153	-0.001	0.001	-0.004	0.013	
29	4050	*LTB*	lymphotoxin beta (TNF superfamily, member 3)	-0.05	-0.118	-0.006	-0.079	-0.138	-0.031	*

ł Proportion of the IL-6~age association mediated; * significant at p<0.05 in InCHIANTI

˜ Bonferroni-adjusted confidence intervals (ie. False-discovery-rate adjusted)

Three genes (*SLC4A10, FLT3, DAAM2*) with larger mediating effects (>7%) in the FHS analysis did not pass the background inclusion criteria in InCHIANTI microarray data. These were selected for confirmation by PCR in a random subset of the InCHIANTI sample. In 194 samples (after exclusions / data cleaning) there was statistically significant partial mediation of the association between age and IL-6 for all 3 genes (*SLC4A10* % mediated=35.6%, p=0.01; *DAAM2* %=-7.92%, p=0.0001; *FLT3* %=-1.03%, p=0.0002), although this small sample size of the replication makes comparisons of effect sizes difficult.

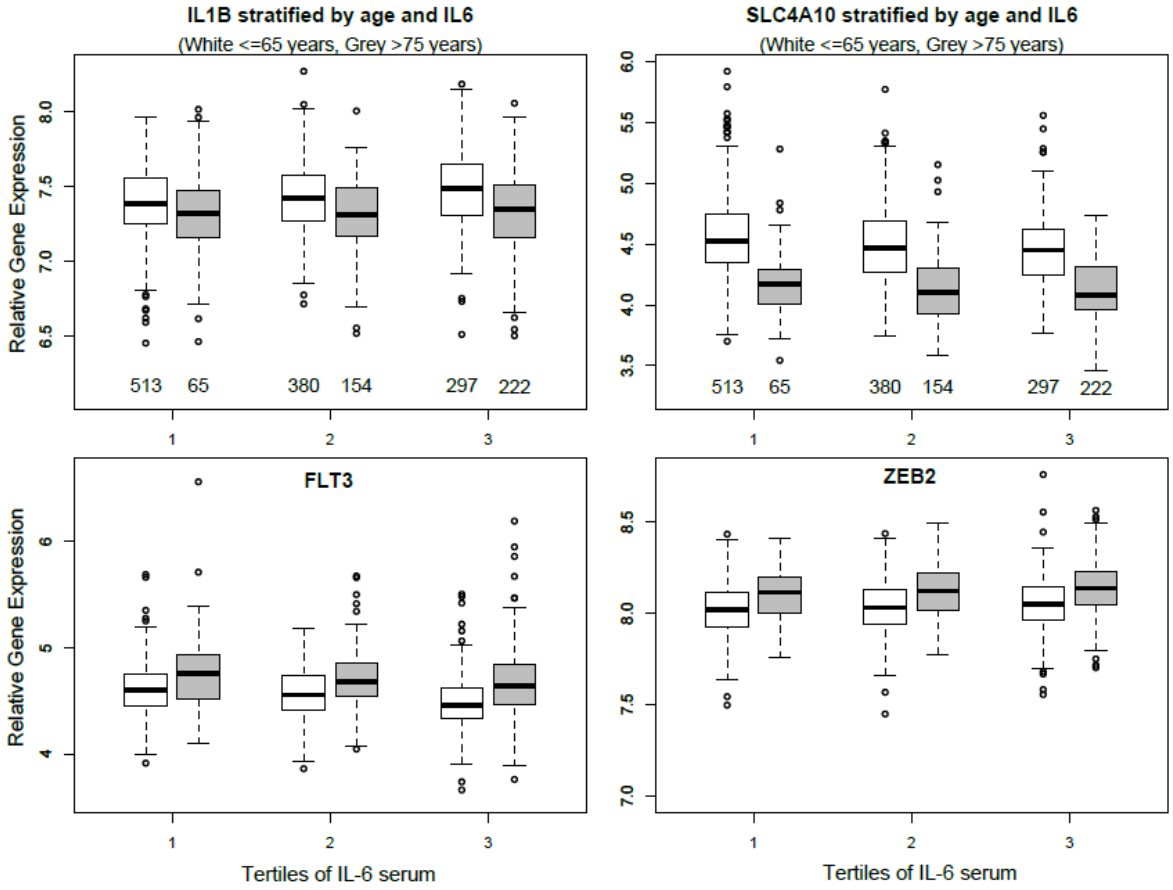
As noted previously, both positive and negative mediators were identified. **Figure 6.2** shows boxplots of expression (in FHS) for 4 genes that have been stratified by age (young ≤65years, old >75) and IL-6 expression (tertiles); *IL1B* (negative mediator), *SLC4A10* (positive mediator), *FLT3* (negative) and *ZEB2* (positive). The plot illustrates that different mediators were positively or negatively associated with age and / or IL-6. **Figure 6.2** also shows that *SLC4A10* expression reduces with increasing IL-6 concentration, and also has markedly lower expression in the older group in general.

6.5.5 Consistent negative findings

Overall, very few interleukins or cytokines emerged as significant mediators (**Figure 6.3**) with all mediation effect sizes for these genes in FHS being <5%, with the sole exception of *IL1B*. Similarly there were few larger effect CD marker mediators (Cluster of Differentiation, used to identify and characterize leukocyte subtypes) other than notable exceptions with updated gene symbols, such as *KLRB1* (CD161).

CDKN2A (*$p16^{INK4a}$*) expression is increased in-vivo with circulating lymphocyte cell senescence [257]. In InCHIANTI, *CDKN2A* array expression was not above background, so TaqMan Low Density Array real-time PCR measured expression in a subsample. Expression of *CDKN2A* was not a significant mediator of the IL-6~age association in either cohort.

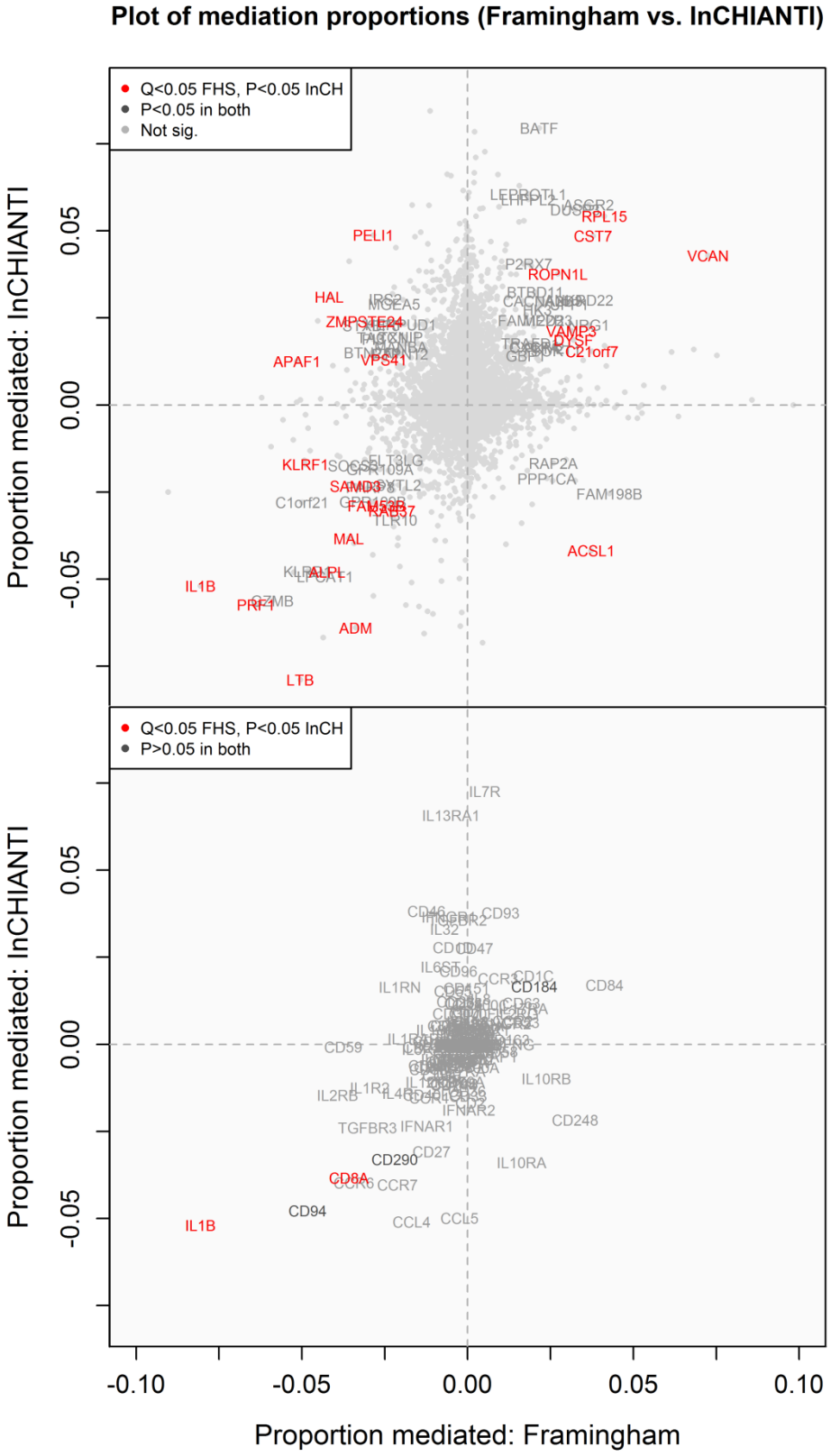
Figure 6.2 | The relationships between expression of four genes, IL-6 protein, and age



Boxplots of gene expression stratified by tertile of IL-6 serum protein concentrations, split by age group (white boxes are individuals <=65 years, grey boxes are individuals >75) in the FHS cohort. To be a statistically significant mediator of the age~IL-6 association, the genes are independently associated with both traits.

Figure 6.3 | Comparing the mediation effects of genes between FHS and InCHIANTI

Scatter plots of the proportion of the age/serum-IL-6 association mediated by each gene expression transcript in FHS and InCHIANTI for (A) all genes, (B) cytokines, interleukin-related and CD (cluster of differentiation) molecules.



6.5.6 Clustering and gene ontology enrichment of mediators

To determine which genes "cluster" together (have similar expression profiles) we used principal components analysis. Eighteen clusters were identified by the Scree analysis of the 102 significant mediators in FHS (**Supplementary Figure 6.1**). Of particular note, *SLC4A10*, the largest single mediator in FHS, only clusters strongly with *KLRB1* (highlighted in yellow - also called CD161, which inhibits natural killer cell (NK) toxicity and is also expressed in some T cells). *IL1B* and *ADM* were also closely related.

Gene ontology enrichment analyses of the 102 FHS expression probes were explored in the DAVID bioinformatics tool. Only 1 KEGG pathway was significant in the functional annotation analysis; the NOD-like receptor signalling pathway ($p=4.9 \times 10^{-5}$, Benjamin-Hochberg multiple testing adjusted $p=3.2 \times 10^{-3}$). In an analysis restricted to the 7 replicated genes only, no output was returned by DAVID.

6.5.7 Sensitivity Analyses

To test the consistency of the results of the associations of age and genes with linear IL-6 concentrations, we also performed an analysis using a binary IL-6 trait as the outcome. The correlation between the proportions mediated by the genes was highly consistent, the R^2 = 0.81 (**Supplementary Figure 6.2**).

Adjusting for WBC count slightly altered the ordering (by proportion mediated) in the list of significant mediators in FHS, however the final list of genes mediating >5% found to replicate in InCHIANTI did not change from the original list of 7.

There were no significant interactions ($p>0.05$) between age, IL-6 and depression (measured by the CESD - Center for Epidemiological Studies Depression – scale).

6.6 Discussion

Our study presents the first larger-scale transcriptome-wide analysis of age-related inflammation in human blood. Very strong associations were present in both study cohorts between age and serum IL-6 protein concentrations [61]. Examining statistical mediators of the age~IL-6 association, we identified and independently replicated 7 partial mediators with effect sizes ≥5%, including one larger effect transcript, namely *SLC4A10*. Our results include known immune markers and several novel genes not previously linked to aging. These results are very different from the many associations with general IL-6 levels independent of age (807 replicated genes) [61] and represent a mediator signature specific to age-related inflammation.

Increased Interleukin 1 beta protein concentrations (IL1B), a cytokine involved in initiation of most inflammatory responses and a product of inflammasome activation, is a core feature of the pro-inflammatory state of aging [258]. *IL1B* expression emerged as the only replicated cytokine mediator in our analyses. The negative mediation value reflects *IL1B* expression increasing with IL-6, but decreasing with age (as shown in **Figure 6.2**). Perforin (*PRF1*) is one of the main proteins of cytolytic granules and is a key effector molecule for T-cell and natural killer-cell-mediated cytolysis. Perforin expression per cell falls with advanced age [259]. A high proportion of CB8+ T cells express Granzyme B but not PRF1 in response to an influenza challenge, and the absence of PRF1 results in Granzyme B degradation of the extracellular matrix and inflammation [260]. Alterations to T-cell populations (and other immune cell types) with aging leads to reduced antibody response and increased incidence of infectious disease, contributing to age-related inflammation [261]. Our finding of *IL1B* and perforin amongst the replicated mediators shows that our study was able to identify primary mechanisms involved in age-related inflammation.

Our largest effect mediator *SLC4A10* (19.1% in FHS, 35.6% in InCHIANTI) codes for NCBE (also called NBCn2) and is a sodium bicarbonate transporter, part of a gene family involved

in intracellular acid – base homeostasis. *SLC4A10* expression is highest in the brain, and *SLC4A10* has been linked to autism, epilepsy and mental retardation [262]. There are little published data on *SLC4A10* in the immune system, but the BioGPS database [201] shows similar levels of expression in all studied leukocyte cell sub-types in humans, suggesting that *SLC4A10* mechanisms may be widespread. Inflammation is associated with a local drop in pH resulting from infiltration of inflammatory cells, and extracellular acidosis in inflammation may be a 'danger signal' activating immune responses [263]. Farwell and Taylor [264] showed that a higher anion gap and lower bicarbonate level were associated with a higher leukocyte count and higher C-reactive protein level, in the NHANES study. Additionally, other members of the SLC superfamily are known to influence mTOR activity, which in turn affects inflammation [265]. Clearly, much more work is needed to clarify the mechanisms of involvement of *SLC4A10* in age-related inflammation.

FTL3 (also called *CD135*) emerged as the second largest effect gene in FHS (replicated in InCHIANTI), and plays a key role in lymphocyte (B and T cell) development but not for the development of other blood cells. The third larger effect gene, *ZEB2,* is mainly expressed in monocytes and is involved in epithelial-mesenchymal transition (ETM), which is characterized by the down-regulation of cellular adhesion complexes and is a key step in cancer development [266]. The *ZEB2* mediation effect was not significant in the InCHIANTI array data, but further work using gold standard lab methods is needed to be sure that the InCHIANTI estimate is not a 'false negative' due to differences in array technology or sample size.

An important finding of our study is that although we confirmed a strong positive relationship between serum interleukin-6 (IL-6) levels and age, we found strikingly few interleukins, cytokines or interferon gene transcripts mediating the age~IL-6 association in the studied blood white cells (**Figure 6.3**), with the exception of *IL1B*, based on an analysis checking synonyms for these genes. These findings suggest that the primary origin of age-related inflammation may not be circulating immune cells. This is further supported by the lack of

association between *IL6* mRNA and IL-6 protein, although this could also be due to post-transcriptional regulation or enhancement of expression at the level of translation, or regulation by moderation of the half-life of protein or RNAs.

6.6.1 Limitations and Future Work

Although moderate (50%) replication of the most significant mediators was observed, there were many small effect genes that were not significant in the InCHIANTI cohort. Our study includes two separate expression arrays (plus three PCR validations), with the arrays having different gene coverage and sensitivities, imposing some limitations. However the alternate technologies do provide evidence for the robustness of our positive findings across our two study cohorts. The Affymetrix Human Exon 1.0 ST array (used by FHS) uses 25-mer probes, typically grouped in sets of four per exon, to estimate exon-specific levels of expression, while the (InCHIANTI) Illumina Human HT12v3 array uses 50-mer probes biased towards the 3′ end of mRNA transcripts to estimate whole-gene levels of expression [267]. The latter array appears less sensitive in our blood derived RNA samples as far fewer gene expression measures passed quality control in InCHIANTI than FHS, and thus some of the genes that appeared not to replicate may be found to be mediators in further studies. The analysis is cross-sectional, so more work will be needed to establish the direction of causation. The study subjects were of European origin, so further work will be needed to establish whether the results are applicable to other ethnic groups. The RNA samples were from whole blood containing a mixture of white cells with results reflecting overall expression patterns. Our results should be seen as starting points for establishing underlying mechanisms, including changes in cell subtype proportions not already adjusted for in our analyses.

Future work should include studies to establish the cellular subtype origins of the novel mediators reported, and their mechanistic roles. This should account for associations with age-related diseases; a recent study found that white-blood cell counts, but not IL-6, is

associated with cardiovascular diseases in >70 year-olds [268]. As longitudinal follow-up becomes available, dynamic changes in expression and the predictive value of expression changes for age related health outcomes can be examined. Follow-up analyses, including analysis of C-reactive protein, are being initiated with collaborating cohorts.

6.6.2 Conclusions

To conclude, we have performed the first larger-scale transcriptome-wide analysis of age-related inflammation in two human population cohorts. A small set of genes appear to have partially mediated the age~IL-6 association, several of which have not previously been linked to the pro-inflammatory state of aging. We have identified a novel larger effect partial mediator (*SLC4A10*, coding for NCBE), thought to be involved in pH homeostasis. We did not find increased expression of *IL6* and cytokines with age in our blood derived data, suggesting that the raised IL-6 protein concentrations in blood may be predominantly generated from other body compartments. Further work is needed to distinguish mediators of age-related inflammation from the much larger numbers of age correlated expression changes in blood.

6.7 Acknowledgements

FHS gene expression profiling was funded through the Division of Intramural Research (Principal Investigator, Daniel Levy), National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD. Measurement of interleukin-6 was funded through RO1 HL 064753, RO1 HL076784, and 1 R01 AG028321. The Framingham Heart Study is supported by contract N01-HC-25195. JMM was supported by RO1 AG29451.

This study was supported in part by the Intramural Research Program, National Institute on Aging. **The analysis was generously supported by a Wellcome Trust Institutional Strategic Support Award (WT097835MF).** WH was funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) for the South West Peninsula. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health in England. This work has made use of the resources provided by the University of Exeter Science Strategy and resulting Systems Biology initiative. Primarily these include high-performance computing facilities managed by Konrad Paszkiewicz and Robin Batten of the College of Environmental and Life Sciences and the University of Exeter IT service.

6.8 Supplementary Information

Supplementary Table 6.1: 102 gene transcripts mediate the age IL-6 association in FHS

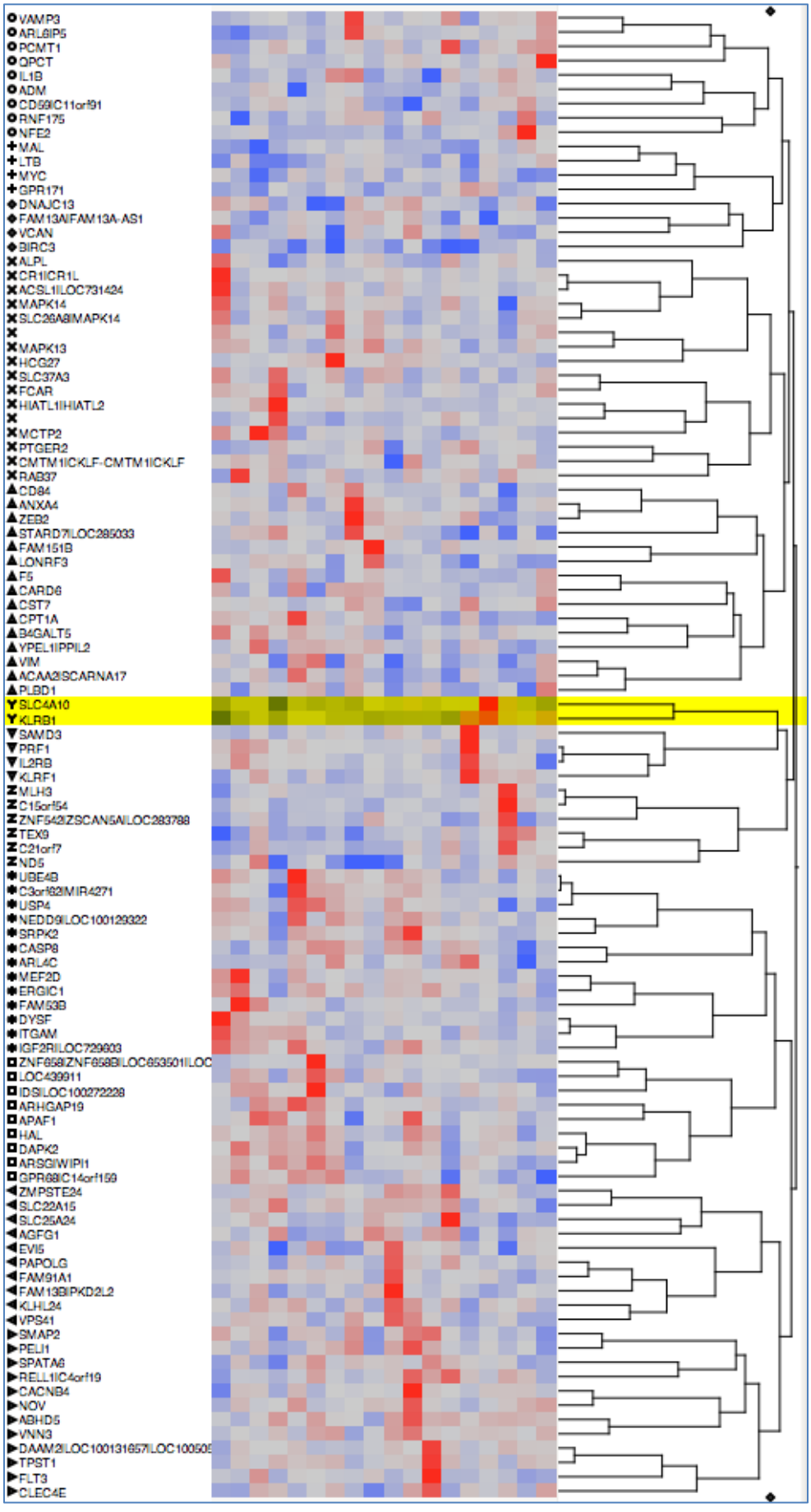
			FHS			InCHIANTI			
N	**Entrez**	**Gene**	**Prop ǂ**	**99.9% CIs ~**		**Prop ǂ**	**95% CIs**		**Significant?**
1	57282	SLC4A10	0.191	0.089	0.341				
2	2322	FLT3	-0.099	-0.211	-0.040				
3	9839	ZEB2	0.098	0.032	0.192	0.000	-0.008	0.008	
4	8460	TPST1	-0.090	-0.182	-0.036	-0.025	-0.070	0.012	
5	1432	MAPK14	0.086	0.030	0.172	0.001	-0.010	0.015	
6	3553	IL1B	-0.081	-0.171	-0.032	-0.052	-0.093	-0.023	*
7	23500	DAAM2	-0.079	-0.176	-0.027				
8	2153	F5	0.075	0.026	0.148	0.014	-0.007	0.039	
9	1462	VCAN	0.073	0.024	0.150	0.043	0.018	0.076	*
10	3820	KLRB1	0.068	0.019	0.135	0.016	-0.007	0.043	
11	5551	PRF1	-0.064	-0.140	-0.001	-0.057	-0.104	-0.021	*
12	3267	AGFG1	0.064	0.020	0.147	-0.003	-0.022	0.013	
13	51099	ABHD5	-0.062	-0.145	-0.017	0.002	-0.004	0.015	
14	400360	C15orf54	0.062	0.017	0.123				
15	116369	SLC26A8	0.062	0.005	0.128	0.012	-0.008	0.037	
16	768211	RELL1	-0.060	-0.173	-0.011				
17	26253	CLEC4E	-0.059	-0.135	-0.016	-0.012	-0.037	0.008	
18	2204	FCAR	0.059	0.019	0.120	0.005	-0.004	0.021	
19	55350	VNN3	-0.056	-0.133	-0.010	-0.003	-0.019	0.008	
20	147947	ZNF542	0.054	0.016	0.115				
21	285533	RNF175	0.053	0.012	0.125	0.000	-0.006	0.005	
22	330	BIRC3	-0.053	-0.119	-0.017	-0.010	-0.033	0.007	
23	4609	MYC	-0.052	-0.128	0.000	-0.048	-0.103	0.000	
24	1374	CPT1A	0.052	0.004	0.130	-0.001	-0.011	0.005	
25	22901	ARSG	-0.052	-0.124	-0.012	0.001	-0.007	0.012	
26	317	APAF1	-0.051	-0.133	-0.013	0.013	0.000	0.034	
27	84255	SLC37A3	0.051	0.014	0.109	-0.004	-0.019	0.004	
28	55356	SLC22A15	-0.050	-0.153	-0.001	0.001	-0.004	0.013	
29	4050	LTB	-0.050	-0.118	-0.006	-0.079	-0.138	-0.031	*
30	23604	DAPK2	-0.049	-0.118	-0.003	-0.003	-0.017	0.005	
…	*Continued on digital version only*								

*** Abridged gene name from Entrez**

ǂ Proportion of the IL-6~age association mediated by this gene/probe

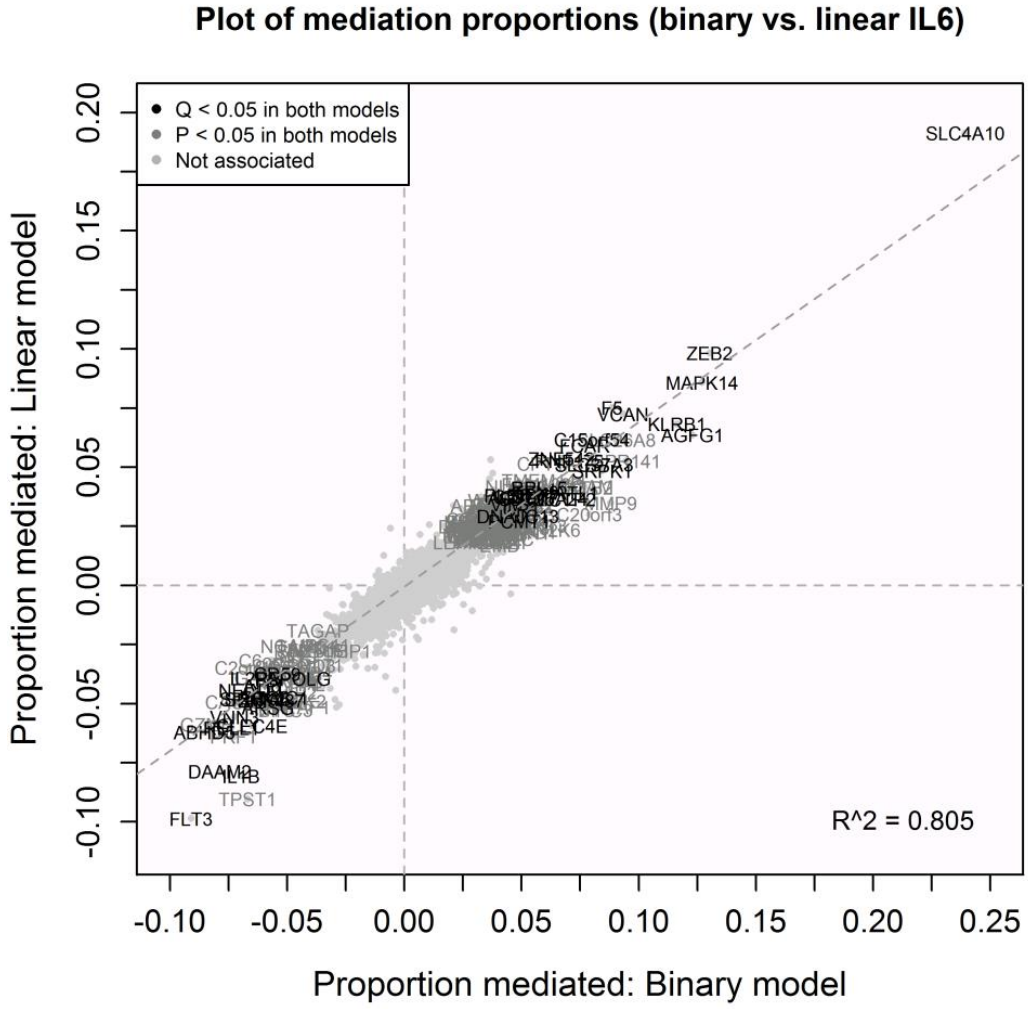
~ 99.9997% CIs were used in the discovery analysis in the FHS, to account for multiple testing

Supplementary Figure 6.1: Principal Components Analysis of 102 significant mediator gene transcripts in FHS.



18 components of variance were identified for the 102 significant mediators in FHS. In the heatplot the components are the x-axis, and each gene is on the y-axis. The colour represents the correlation between the gene and the components (red = positive, blue = negative). The genes have been hierarchically clustered with respect to correlation with the component. The highlighted genes load to the same component, and in fact are the only two genes to load to it, indicating they have a different expression profile to the other genes.

Supplementary Figure 6.2: Sensitivity analysis of transcripts mediating the age binary IL-6 ('high' defined as >3.5pg/mL of IL-6) association



This figure shows the proportion of the age~IL-6 association mediated by each gene when using first continuous IL-6 (y-axis) and using a binary IL-6 ('high' defined as >3.5pg/mL of IL-6) variable (x-axis). Although there is slight variation in the absolute proportion mediated, the trend is very strong, particularly for the stronger effects.

Chapter 7 – Discussion and Conclusion

7.1 Summary of Thesis

Human ageing exhibits significant inter-individual variation; some individuals succumb to age-related disease much later than others. Efforts to understand the processes involved and determine predictive factors and mechanisms to "ageing well" are progressing, but much remains to be elucidated; few robust biomarkers exist in humans – IL-6 being one of the few – yet ageing traits can be reversed in experiments in mice, giving hope that interventions could be identified in humans. Biomarkers in blood could not only be used to identify individuals at the greatest risk, but may also have mechanistic implications leading to treatments to improve healthspan.

I have performed analyses on whole blood gene expression profiles to gain insight to mechanisms of ageing and age-related disease. During the course of my PhD I have performed the first population-based studies of whole-blood gene expression and ageing phenotypes in the InCHIANTI ageing study, including;

- In depth analysis of muscle strength, including identifying transcripts associated with muscle strength, investigating the robustness with other strength phenotypes, confounders and cell-type adjustment, and presenting plausible biological insight utilising published experimental models.

- A follow-up analysis of gene expression and muscle strength involving meta-analysis of four independent cohorts, where I co-ordinated and performed a meta-analysis of gene expression and muscle strength, involving not only the transcriptome-wide scan, but also sensitivity and subset analyses, pathway analysis, investigation of the phenotype and confounders, and systematically screening the literature.

- An analysis of cognitive function involved not only the cross-sectional transcriptome-wide screen, but also an analysis of cognitive decline over 9 years, subset and pathways analysis, in depth investigation of the robustness of the association with respect to linearity and confounders.

- Finally I performed an analysis of age-related inflammation in collaboration with the Framingham Heart Study (FHS), in which we applied mediation analysis to genomic epidemiology in a novel way; I determined the genes whose expression statistically mediated a proportion of the association between age and IL-6, providing novel insight into the specific inflammatory pathways that are age-related. The analysis included investigation and validation of the phenotypes, extensive literature follow-up, clustering analysis in the FHS, and analysis of PCR data.

In the following sections I will summarize the key findings and conclusions of each individual analysis, and draw the ideas together in a final discussion.

7.2 Summary of Data Chapters

7.2.1 Chapter 3 – Muscle Strength

To identify the gene transcripts most closely associated with muscle strength in the InCHIANTI ageing study I utilised microarray data available on 698 of the participants who also had hand-grip strength measured using a dynamometer. After adjustment for selected confounding factors and multiple testing a single gene – CCAAT/enhancer-binding-protein beta (*CEBPB*) - was statistically significant. This association was confirmed using qRTPCR, a gold-standard lab method for quantifying nucleic acid sequences. *CEBPB* was already known to be a transcription factor with multiple functions, with various binding partners providing specificity.

Evidence of a conditional knock-out mouse model provided a biological link; by specifically inhibiting the activation of *CEBPB* by *CREB* the up-regulation of *CEBPB* required to initiate the 'alternative' macrophage function (which is characterised by anti-inflammatory secretions and promotion of wound-healing) could not occur, whilst allowing basal expression of *CEBPB* to be maintained. These mice had impaired repair of muscle following injury, as occurs during the normal use of muscle; repair is a normal part of muscle maintenance and development. It is very thought provoking that *CEBPB* was the only gene found to be robustly associated with muscle strength from peripheral blood samples taken from the InCHIANTI population. Of note is that the damage to muscle in the mice was artificially induced by cardiotoxin, which caused local necrosis and infiltration of pro-inflammatory macrophages; in the mice with altered *CREB*-domain in the *CEBPB* promoter there was no subsequent anti-inflammatory response and the muscle tissue appeared highly fibrotic [113]. This is in contrast to the "everyday" contraction-induced injury that causes damage in humans [269], however the response should be comparable (pro-inflammatory M1 phase followed by anti-inflammatory M2 phase).

The importance of this finding is the link between plausible biological mechanism and observational data in humans using a highly relevant phenotype, maximum hand grip strength, which is a well-established marker of poor prognosis and disability [270]. Although using additional strength measures (e.g. quadriceps) does not provide additional predictive value of ageing phenotypes beyond just using measures of "maximum grip strength" [271], other measures of strength (such as knee strength) were measured in the InCHIANTI participants and replicated the finding with *CEBPB*. Therefore the use of hand grip in this study is applicable to the wider ageing and sarcopenia research field.

Reflecting on the limitations of this study; firstly, muscle strength (an age-related phenotype) was assessed for associations with gene expression using linear regression models adjusted for a number of potential confounding factors such as sex and adiposity, and also for age. Therefore the conclusion was that *CEBPB* is associated with muscle strength in an elderly population (InCHIANTI is an ageing study, and as such the majority of the participants (>70%) are over 72 years of age), not that *CEBPB* was associated with age-related muscle loss; this is a hypothesis drawn from the results and from the literature. In **Chapter 6** I assessed which transcripts *mediate* the association between age and inflammation. Performing a similar analysis to assess which transcripts mediate a proportion of the association between age and muscle strength I found that only 18 genes were significant mediators. *CEBPB* mediated the largest proportion of the age~muscle strength association of any measured gene (5%). Secondly, we were not in a position to get independent replication in the same paper. Thirdly, the data is from whole blood, a heterogeneous mixture of cells that could affect the gene-expression associations; we could only investigate and control for the effects of "broad" cell counts (e.g. "lymphocytes", but not the proportion of T- or B-cells), which was a crucial adjustment to identify the association between *CEBPB* expression and muscle strength, and future studies will investigate this further.

Future work: the replicability of this study is discussed in **Chapter 4**, however more studies of older people are needed to dissect the interaction with age. Data on specific cell subtypes is

essential to determine the origins of the associations observed. Developments in microarray and sequencing technology will allow a much more detailed analysis of isoforms, genes, and non-coding transcripts not available to test in the data.

7.2.2 Chapter 4 – Muscle Strength – CHARGE Meta-analysis

In chapter four I coordinated and performed a CHARGE meta-analysis on four independent cohorts (including InCHIANTI), and found that expression of 222 unique genes were associated with muscle strength in 7781 participants, after adjustment for confounding factors and multiple testing. Subsequent bioinformatic analyses included gene ontology enrichment analysis, text mining the literature, and utilisation of databases such as BioGPS [201] which displays the relative expression of genes across human tissues.

The genes found to be significantly associated with muscle strength covered a variety of biological functions, and although “haemoglobin metabolic process” and “innate immune response” pathways were significantly enriched, only 33% of genes associated with muscle were in an enriched biological process. Genes hypothesised to be associated with muscle strength *a priori* included *GDF11* (mouse model showed that supplementation of GDF11 protein restored muscle function [3]), which was not associated with muscle strength, possibly highlighting different growth factor pathways for muscle in the circulation between mice and humans. *IGF1R* and *IGF2BP2* (insulin-like growth factor-related genes) were significantly associated with strength in opposite directions (positive and negative, respectively). Both are implicated in glucose homeostasis and diabetes [69], [272]. I found that of the 222 genes, 115 had not been associated with the term “muscle” before in publications; this analysis therefore identified many novel genes to follow-up.

CEBPB expression was not significantly associated with muscle strength in this meta-analysis. Using a power calculate (R package "pwr") and following parameters (r=0.04 – this is the correlation coefficient between *CEBPB* and strength in InCHIANTI – alpha=0.05 and desired power of 0.8) suggests that a sample size of 4,903 should be sufficient to replicate the association. The lack of replication could be due to the lower average age in the meta-analysis (weighted mean = 56 years, compared to 72 years in InCHIANTI, with >70% of the sample aged over 72). However, given the robust associations found in the primary analysis (**Chapter 3**; [183]), subsequent pilot study [184], and compelling biological plausibility from the mouse model [113], *CEBPB* deserves further investigation in ageing (in **section 7.2.1** I note that *CEBPB* mediates the largest proportion of the association between age and muscle strength of any gene in the InCHIANTI study).

The meta-analysis was performed in two parts; firstly the two FHS generations (which used Affymetrix exon arrays) were analysed together, secondly the three cohorts using Illumina-based arrays (InCHIANTI, RS and SHIP) were meta-analysed together, and finally a meta-analysis was performed on the summary statistics of these two models; the aim was to reduce the heterogeneity in the final meta-analysis by combining cohorts using the same array technology first. It is both a strength and a limitation of these studies that cohorts have used different array technologies; significantly replicating results are robust to array differences, giving confidence to the reported results, but heterogeneity in microarray design mean the measures may not be directly comparable, and in some cases genes/transcripts are only measured on one array, meaning replication is not possible. Future studies using new technologies can address this and discover more novel associations.

Further investigation is warranted not only into the difference between mice and men regarding *GDF11* and muscle strength, but also the 115 genes not previously linked to muscle strength in the literature. The role of *IGF1* in skeletal muscle mitogenesis and glucose homeostasis is known [272], but the relation to *IGF1R* expression in whole blood is unclear. More detailed expression data and knowledge of cellular composition and origin

would help to shed light on the genes found. The results of complimentary CHARGE analyses of genetic and whole blood DNA methylation associations with muscle strength (the latter of which I will be coordinating) may provide further mechanistic insight.

7.2.3 Chapter 5 – Cognitive Impairment

In this analysis of whole blood gene expression and cognitive function in the InCHIANTI study I used gene ontologies and gene-set enrichment analysis to investigate inflammatory pathways, as well as utilising regression spline models to investigate the associations more thoroughly. This transcriptome-wide analysis of gene expression and cognitive function (measured using the MMSE criteria for identifying cognitively impaired individuals) found one near-significant association between *CCR2* and MMSE in a linear regression model.

There was an already-published lab model of *CCR2*-deficient mice displaying an accelerated Alzheimer's-like pathology [214]. It was on the strength of the already published evidence and the stringent multiple-testing criteria that this study was published. Since publication there has been further evidence published – reviewed by ourselves and others [53] – that quite convincingly shows (albeit in mice) that bone marrow–derived microglia are CCR2+ and can restrict cerebral amyloidosis [273]. We note in **Chapter 5** that the association we see in blood (*CCR2* expression is negatively correlated with MMSE; cognitively impaired individuals express more) appears in contrast to the reports in mice. In collaboration with Serge Rivest – the principal investigator behind much of the *CCR2* work in mice - we published an editorial in 2013 addressing this point and suggested that "Increased *CCR2* expression in the cognitively impaired may be due to an increased demand for macrophage-mediated clearance of β-amyloid, caused by impaired phagocytosis by microglia and also the self-perpetuating nature of the pro-inflammatory state of aging" [53].

The importance of the finding that *CCR2* expression in peripheral blood was associated with cognitive impairment is clear, given the biological literature available; changes in expression of *CCR2* may reflect differences in underlying cell-type composition, or instead reflect differential activation of monocytes, either of which may be interacting with peripheral and neuro-inflammation as the blood-brain-barrier becomes less distinct and permeable. Future studies need to identify whether *CCR2* expression in peripheral blood has predictive value for those at highest risk, and whether steps can be taken to ameliorate the changes; here evidence from mouse models and human observational studies will be essential once again.

This study is limited by the lack of follow-up data to help address causality, and by the lack of detailed cell-count information to determine the origin of the signal observed; as more data becomes available this will be investigated in a future research programme. The MMSE is designed to detect cognitive impairment, rather than cognitive *ability* per se (as it is the loss of function that is age-related, the MMSE is still useful), yet it is quite crude; 80% accuracy in identifying participants with no impairment vs. any impairment [274]. Future studies should use multiple measures of cognitive function and impairment. We plan to extend these findings to a longitudinal setting in the InCHIANTI study as more follow-up data of these elderly participants becomes available, and seek replication (and/or meta-analysis) with other cohorts.

7.2.4 Chapter 6 – Age-related Inflammation

Here I present the results of an analysis into gene expression transcripts from whole blood in the FHS study, with replication in the InCHIANTI study, and whether they statistically mediate the association between age and inflammation (interleukin-6, IL-6). This is distinct from the analysis of gene-expression transcripts associated with IL-6 that we participated in, which found 807 genes significantly associated with IL-6 *independent of age* [61]. This is a novel

analysis method in genomic research into age-related inflammation, and demonstrated that only a small number of specific genes mediated the age~IL-6 association.

Twenty-five of the ninety significant genes in FHS replicated in InCHIANTI, with increasing replication as the "effect size" increased (5 of the top 10 genes replicated). I therefore focussed on genes mediating >5% of the age~IL-6 association. This analysis provided the first evidence that expression of *SLC4A10* (a sodium bicarbonate transporter involved in pH homeostasis) in whole blood is associated with age-related inflammation, although the biology is not clear; mice with disrupted *SLC4A10* showed reduced neuronal network activity and increased intracellular acidification [275], with other neuronal implications from a human patient with cognitive impairment and epilepsy [276]. However, other significant genes including *IL1B* (interleukin-1 beta) and *PRF1* (perforin) are known to be involved in the pro-inflammatory immune response, showing that mediation analysis is a valid analysis method.

By constructing a mediation model rather than a simple linear regression adjusted for age I have discovered a small, distinct subset of IL-6-associated genes that statistically mediate the IL-6 association with age. Given the importance of age-related inflammation to ageing phenotypes and diseases [60] methods such as these that investigate the interaction between inflammation, age, and genes need further use and development to overcome the limitations encountered during this project.

Regarding the mediation analysis methods, the package chosen to perform the resampling to determine confidence intervals for the estimates was not designed for multiple-testing adjustment; in order to adjust for multiple-testing the confidence intervals were modified from the traditional 95% to Bonferroni-adjusted confidence limits. As mentioned in the methods, this is a conservative multiple-testing adjustment and may have led to false-negative findings. In future analyses we will determine if the method can be modified to compute p-values so that a Benjamini-Hochberg (or similar) adjustment could be made. Additionally the term "mediation" may be misleading in this context as it could imply causality; I stress that no causal inferences are drawn, as this is cross-sectional research and the mediation is purely

statistical. The only thing distinguishing a mediator and a confounder in this analysis method is the conceptual model developed at the start of the project.

There are also significant limitations when comparing data from two different microarray platforms; it is impossible with this data to determine whether the low replication rate (25 of 90 genes) is due to biological or technical differences between the cohorts (no doubt both contribute, but the degree of each is unknown). Future work should focus on dissecting the novel genes determined by this analysis (in particular *SLC4A10*) and determining the reason for the low replication rate; whether this difference is biological (cohort differences) or technical (differences in RNA handling, or sensitivity between the array platforms).

7.3 Limitations

Studying RNA (and other biochemical markers) in whole blood has clear advantages, in particular to provide insights into mechanism (both up- and down-stream effects) and the possibility of easy-to-access biomarkers of disease. Yet the limitations must be considered when drawing any conclusions from this research. Firstly, a criticism of this research is the "whole blood" nature; that the tissue is heterogeneous with regard to its cell composition, and that the different subtypes of immune cell will differ in relative proportions between individuals, and could be related to the phenotype studied. This is an important consideration, and the results are only valid when bearing this in mind; whilst the methods adjust for the relative proportion of lymphocytes, neutrophils, monocytes and eosinophils, the association observed could be due to the ratio of B- to T-lymphocytes (for instance) also being associated with the outcome. This is still an interesting and relevant finding; that transcripts and cell-types are associated with ageing phenotypes. Future studies (including our own – see the next "Future Research" section) will aim to investigate the mechanisms behind gene expression associations, starting with whether the association is independent of or related to cell-type, and if the latter from which cell-type the signal originates.

It is worth noting that as the concentration of cell-types in the sample decreases (e.g. B-cells comprise 9% of total lymphocytes in the InCHIANTI baseline data, and lymphocytes make up only 30% of total white blood cells) the contribution of that cell-type to the total RNA decreases, and a larger effect would therefore be required to have sufficient power to detect it in whole blood; therefore the signals we detect will mostly derive from expression in the larger fractions of the data, such as neutrophils (~60% of leukocytes in the InCHIANTI data). However, if a gene/transcript signal is *specific* to a particular cellular subtype in the whole blood sample, changes in expression (or relative proportion of the cells that are this subtype) could still be identified in smaller subsets; this is one reason why investigating genes in the BioGPS database [201] can be useful to suggest cellular origins.

Another important consideration often overlooked is the assumption that RNA levels correspond to protein levels. In an unpublished manuscript in 2012 Dr Nancy Kendrick [277] reports that in 4 separate studies examining the relationship between mRNA and protein levels the R^2 values are below 0.4. The reasons suggested are many, and include post-transcriptional regulation (e.g. degradation of transcript prior to translation), variations in half-life of mRNA/protein, and the variety of functions of the proteins themselves, for example some proteins are highly stable (such as collagen, a structural protein) whereas others will have a much higher and more variable turnover, such as antibodies. Yet the study report by Kendrick only studied the amount of protein, let alone that which is *active*. Protein function can be greatly affected by the presence/absence of co-factors, or modifications such as acetylation; a recent study by Lozada *et al* demonstrated that de-acetylation of Werner syndrome protein affects it's binding affinity to DNA structures, directly impacting its DNA damage response ability [278]. Therefore the research published in this thesis (RNA concentrations statistically associated with phenotypes) must be considered in this context; although production of RNA may be associated with a phenotype, for protein-coding transcripts this may not correspond directly to differential protein production or activation. However, given the complex relationship between RNA and protein (and taking into account that many transcripts do not code for proteins but are functional molecules in their own right), including not least the effects of alternative splicing and isoform usage, the variability in protein activation (be it through phosphorylation or co-factor) and half-life, it still seems a valid assumption that increased production of RNA is related to *demand* for that RNA at that time-point.

Population studies themselves are not without limitations and systematic errors, some of which are overcome when seeking independent replication in other cohorts. These include, but are not limited to, residual confounders (unaccounted factors that confound relationships [279]), learning effects (as participants make repeat visits they may perform better in some tests as they get better at them, for example MMSE [280]), selection bias (not everyone in the population will participate in a study, for instance particularly unwell or disabled

individuals [281]), and, similarly, loss to follow-up (when participants are followed for significant periods of time, those who become unwell or disabled may no longer participate) [282]. The ideal design for a study in molecular epidemiology would be a large population with a good response rate, with DNA, RNA and a host of other biochemical and cell-type (as well as general) phenotypes available at baseline, with a number of follow-up visits, including links to medical records. In this way cause and effect could be suggested without a number of bias' discussed in this section.

There are also technical and analytic considerations, in addition to the biological ones. Linear regression models make several key assumptions; that the association between the dependent and independent variables are linear; that the residuals (error in prediction) are normally distributed; and that the variance of the prediction errors are consistent [283]. In practical terms when performing large-scale analysis it is optimal to perform the analysis on normalised and quality controlled data, and then explore in further depth the output of such analyses to assess the validity of the model. If multiple different tests (e.g. non-linear regression variants) were performed on each gene expression probe this would increase the multiple testing problem dramatically (a challenge inherent when using frequentist-based statistics reliant on p-values, where performing a large numbers of parallel tests – such as on many gene expression probes - could produce many false-positive results without adjusting the p-value threshold [132]).

Gene ontology enrichment (pathways) analysis also has important considerations. The methods for testing enrichment tend not to take into account *direction* or *magnitude* of association, nor do they weight genes relative to one another or assign particular functions to gene interactions (e.g. gene A inhibiting gene B is not taken into account) [284]. Additionally, strength of association (one of the key tenets of causality in epidemiology [285]) is not always apparent, for instance in **Chapter 4** the haematology pathways were the most statistically significant but did not contain the largest number of genes associated with

muscle strength. However, these methods are still useful for providing additional and simplified biological insight to a list of genes, albeit with caveats.

Finally, microarray data is inherently limited to the specific probes present on the array; not only does this mean *de novo* transcripts will not be measured, but the probes have differing binding qualities (affinities) to transcripts, potentially leading to bias. In two chapters there are results from two array platforms, which may differ in their expression profile capture. Unfortunately it was not possible with this data to compare the two platforms on samples from the same individual. To have a truly transcriptome-wide, hypothesis free approach, we need a different kind of data, possibly utilising second-generation sequencing approaches such as the Illumina RNA-Seq platforms, however this was not within the scope of this project and will be addressed by future studies, and of course come with their own bias' and considerations. Many RNAs transcribed by cells are not protein-coding, and the ENCODE project is revealing that regulation of transcription of mRNAs and non-coding RNAs is complex [286]; multiple levels of data are required to get even a basic understanding of the factors involved in regulating transcription and translation.

Microarrays correlate only moderately to "gold-standard" PCR approaches, both in the literature (R=0.5 for Illumina-based microarray and PCR, albeit in mice [287]) and in the InCHIANTI study, for example the Illumina probe for *CEBPB* was significantly correlated (pairwise comparison R=0.3, $p<0.05$) with the PCR data in **Chapter 3**, but only modestly. This could be due to cross-hybridization of probes, low-sensitivity, or measurement of different isoforms/transcripts [287]. Additionally there are other potential sources of technical variation, including the collection of the RNA (PAXgene tubes used in InCHIANTI to stabilize RNA prevent degradation [288]), the choice of normalisation procedures to account for systematic technical variability [289] (although much of this is handled by Illumina's "GenomeStudio" there are other choices, including how to account for batch effects, that are less clear), the unavailability of a measure of RNA integrity in the InCHIANTI participants

(although the PAXgene extraction kits leave very pure samples, with A_{260}/A_{280} values between 1.8 and 2.2, from www.qiagen.com), and the lack of globin reduction in InCHIANTI.

The alternative splicing of gene transcripts may also be obscured by the use of microarrays; work in our group (which I have contributed to) found that expression of specific splicing-factors associated with age and senescence [80], and that the ratio of isoforms of genes can also be age-associated [79]. This is relevant to the microarray probes if, for instance, a particular probe only captures one isoform, or if the probe captures all indiscriminately; the former would result is a negative/positive association which may not be the whole story if an undetected isoform is expressed instead, and the latter would identify a gene where the ratio of different isoforms was associated with the phenotype, but not "total expression."

Despite the limitations these kinds of studies have value; the opportunity to discover biomarkers and gain insight into mechanisms related to ageing in humans is unprecedented, and over the last three years the field has been transformed and now many studies can work together in large and detailed meta-analyses. However, caution must be used to not "over interpret" the data; these studies may be confounded, and make no claims as to the final protein abundance. Nor are any causal inferences made; links between mechanism and association are hypotheses based on existing knowledge and observations.

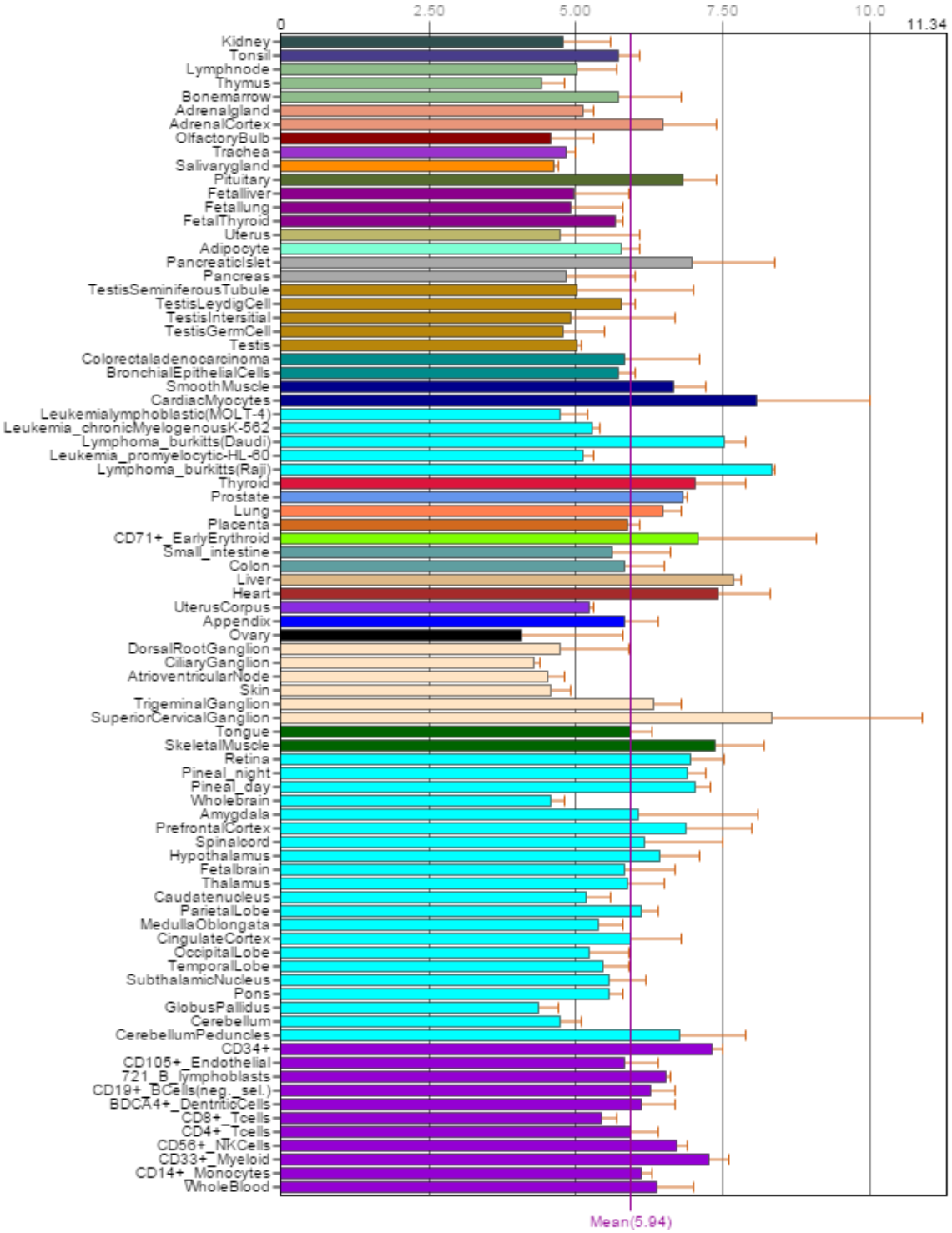
7.4 Discussion of Thesis

This thesis reports four data chapters that have explored the relationships between ageing phenotypes and molecular markers, specifically gene expression. This includes not only transcriptome-wide scans of gene/phenotype associations, but meta-analysis, mediation models, gene-set enrichment, pathways, clustering, text-mining, and other bioinformatic techniques. The *de novo* discovery of genes associated with ageing phenotypes in human whole blood samples has made an important contribution to the literature, and foundations are in place to continue this work on a much larger, collaborative scale, as more diverse cohorts are able to replicate and extend these analyses prospectively (future research discussed in **section 7.5**).

Mouse model experiments have demonstrated that ageing phenotypes such as neurodegeneration can be reversed by exposing aged animals to young blood [5] (and vice versa). An experiment comparable to the heterochronic parabiosis mice (surgically attached so that they share a circulatory environment) in humans is not ethically feasible, so we cannot test whether the older tissues can be rejuvenated by exposure to young blood (or specific factors therein) and must rely on observational data until suitable model systems are generated. In **Chapter 4** I examined the relationship between expression of *GDF11* – supplementation in mice rejuvenates aged muscle [3] – and muscle strength and found that there was no association (nominal $p>0.05$). Although as suggested in that chapter it may be that the growth factor pathways are different in humans to mice, equally it could be that circulating GDF11 protein is not derived from blood cells; it is currently not known what tissue produces circulating GDF11 [3], but examining the BioGPS database of expression in many tissues types shows that *GDF11* expression is ubiquitous across all assayed cell types, including many blood cell types (**Figure 7.1**). This suggests that if expression of *GDF11* in whole blood is associated with muscle strength in humans we should have detected it in this analysis. Therefore GDF11 is either not expressed by blood cells in humans in relation to strength, or alternative circulating factors to GDF11 may drive the ageing of human tissues,

and could include the genes identified in **Chapter 4** such as insulin-like growth factors. Follow-up studies are needed.

Figure 7.1 | Relative *GDF11* expression in different human cell types, data from BioGPS



GDF11 expression data from BioGPS [201] dataset "GeneAtlas U133A gcrma" [290] in which 79 human tissues were profiled for the relative expression of 15,491 human RefSeq genes.

It is clear that extensive reviewing of existing literature should be a priority for all such studies; for both muscle strength and cognitive function the hypothesis-free approach of assessing all available genes found evidence in animal models for mechanistic associations with the phenotypes. Novel links between genes and muscle strength and age-related inflammation have been discovered and, where possible, suggestions made as to the causal pathways. Many other published epidemiological studies of genetic and transcriptomic associations do little more than report associations, yet utilising the literature on experiments in model organisms and using services such as BioGPS [201] to investigate the relative expression of genes in many cell types is of great value when seeking insight into mechanism. In particular, it provides evidence for biological plausibility of the statistical associations observed in humans, an important component required to draw conclusions regarding causality.

Although subject to limitations (discussed in **section 7.3**), the conclusions drawn in each chapter are based on empirical human observational data combined with experimental data from the literature. These studies have taken us further along the road to understanding, and have provided direction for future researchers (including myself and others in the group). I have presented associations between genes, muscle strength, cognitive function and age-related inflammation in an elderly human population, and discussed plausible biological mechanisms based on existing literature. These studies show that observational studies of whole blood in humans add value to the literature; it is becoming increasingly apparent that circulating factors can have a huge impact on systems as seemingly disparate as strength and cognition [3], [4], and human observational evidence can highlight which pathways identified in model organisms are of the highest priority.

In 2005 it was proposed by John Ioannidis that the majority of publications in the medical literature are false [291] - that many positive findings do not translate to the “real world” (the paper was in particular focussing on trials and drug discovery, and has subsequently been replicated and extended by Bayer Healthcare [292]). Replication and generalizability of

findings is therefore a priority, emphasised both by the differences between cohorts in the meta-analysis of muscle strength (**Chapter 4**), and the disparity of the results between the microarray platforms in the analysis on inflammation (**Chapter 6**).

Follow-on studies not included in this thesis, but which relate to these studies and enhance the interpretation, are important, not only for validating observational findings but understanding more of the biology. We collaborated with colleagues in the Sport and Health Sciences department and found that, in a pilot study of 16 participants, *CEBPB* expression in whole blood increases in response to muscle damage following exercise [184]. This pilot shows that it is possible to observe the relationship between *CEBPB* expression and muscle repair over time, which coincided with the time course proposed by Tidball *et al* [47]. Next, studies should investigate whether this *CEBPB* expression response to exercise-induced muscle damage changes with age. Additionally, we collaborated with Rivest *et al* in an editorial discussing the implications of the *CCR2* expression association with cognitive function in InCHIANTI, and the possible mechanisms [53], as noted in **Section 7.2.3**.

I have assessed three ageing phenotypes (muscle strength, cognitive impairments, and increased inflammation) in separate manuscripts, yet viewing ageing holistically these are interconnected processes; maintenance of muscle strength requires an orchestrated immune response, and inflammation is associated with cognitive decline. Throughout the course of these projects it has become increasingly apparent that studying whole blood in the context of ageing and humans is valuable; not only is the tissue easily accessible and already routinely collected, but the immune system and circulating factors are central to many ageing processes. Luigi Ferrucci (central to our involvement with the InCHIANTI study) and Claudio Franceschi, among others, have been studying this for some time [48], [242], and have provided great insight into the associations with age-related diseases and the possible evolutionary causes (respectively). In summary, the immune system is optimized to respond efficiently to *acute* infections but does not respond appropriately to long-term *chronic* conditions, such as those in ageing; the latter results in sustained innate immune activation

which suppresses the anti-inflammatory response and leads to progressive tissue damage [48]. Ferrucci *et al* found that a number of pro-inflammatory cytokines increase with age and are related to cardiovascular risk and morbidity, although the causal direction is unclear [242]. This is a central point; although age-related inflammation is damaging to multiple tissues and is associated with risk of disease, I show in **Chapter 6** that expression of *IL6* in blood is not positively associated with IL-6 (protein) in elderly individuals. This supports other studies suggesting that circulating IL-6 is produced by senescing peripheral tissues [110], and further suggests that the innate immune activation may be due to, rather than causing, the initial increase in pro-inflammatory cytokines; the inability of the immune system to respond with an anti-inflammatory resolution then creates a positive feedback loop as it causes even more damage via the pro-inflammatory processes.

Given the relationship between the immune system, systemic inflammation, ageing, and declining function including loss of strength and cognitive ability, determining the mechanisms of these interactions and developing therapeutics to postpone age-related phenotypes and morbidities are of high importance. Currently it is not known whether reducing inflammation could improve health and functioning in elderly people [247]. Studies such as those presented in this thesis are of great importance to the wider research field by identifying novel genes in age-related phenotypes, and highlighting experimental models of particular relevance.

7.5 Future Research

There are clear advantages and limitations to studying RNA levels in the blood, as noted in the previous sections. In future research the relationship of specific proteins must also be sought. For instance, secreted GDF11 - a cytokine (member of the TGF-β superfamily) - is reported as being age-associated (negatively), and supplementation in mice ameliorates the age-related dysfunction of skeletal muscle [3]. Yet in the InCHIANTI study, probes for *GDF11* mRNA are not associated with age or muscle strength ($p>0.05$ in both cases). It could be that we are looking in the wrong tissue/cell type, that the age-/muscle-related regulation occurs post-transcriptionally, or that the effect was specific to mice and does not directly translate to humans. Although mRNA levels are a proxy for demand for proteins, this was not empirically tested in these studies and it would be of great benefit in future studies to determine the accuracy of this assumption (that mRNA is a proxy for protein); it would also be of great benefit to compliment this with sequencing-based approaches to RNA quantification, to capture novel and non-coding RNAs too.

Meta-analysis of gene-expression and muscle-strength associations in the CHARGE consortium highlights the variability of gene-expression associations, and the necessity to replicate. We are contributing to a wide-range of phenotypes within this consortium, the results of which are yet to be published and are not included in this thesis, although they broadly fall into this area of research. Within the gene-expression working group analyses include inflammation (C-reactive protein, CRP), blood pressure, lipid levels, fasting glucose, tobacco exposure, and others. The strength of these studies are increased power and confidence in the result; if a gene is found to be associated with a phenotype across diverse populations, using different technologies, this is much less likely to be a false-positive; the field is now benefitting from many populations having this data and who are willing to collaborate on meta-analyses – this is a recent phenomenon, not possible even 3 years ago.

Similar approaches are being taken investigating the associations of epigenetic – particularly DNA methylation – associations with common diseases of ageing. Within CHARGE the epigenetics working groups are investigating multiple ageing phenotypes, including many of those listed previously. The InCHIANTI working group, including myself and others, are contributing to and will lead some of these analyses.

Naturally, the next questions once we have established expression and/or epigenetic profiles associated with common diseases of ageing in human peripheral blood include (but are certainly not limited to);

1. Do they have any prognostic value?

2. Are the associations driven by specific immune subtypes?

3. What are the causal mechanism(s) behind the associations?

4. Can interventions be developed for the processes indicated by these associations to prevent or delay adverse outcomes?

These questions provide an order to which the answers may arise; as more follow-up data is available in cohorts with gene-expression/epigenetic data question 1 becomes more feasible; questions 2 and 3 are similar, and ask why the association is observed (if the association is due to a specific immune subtype we would then go on to ask what the causal mechanism is); and finally if a causal mechanism can be determined, could this be an option for treatment/prevention/postponement of disease?

Many of these questions are the subject of current studies, whereas others will depend on the results of these. For instance in InCHIANTI, the participants visited the clinic again in 2014, 5-years after the gene-expression measurement, so we will soon be able to test the predictive nature of whole blood gene expression in an aged population. In the Baltimore Longitudinal Study of Aging (BLSA – a project with even more extensive phenotyping than

InCHIANTI, to which we will have access courtesy of Professor Melzer's long-standing collaboration with Dr Luigi Ferrucci at the United States National Institute on Aging (NIA)) a subset of participants with RNA measurements are also having their immune cell-types quantified for ~60 different cell-types, so we can hopefully address (to some degree, at least) which cell types are statistically interacting with observed associations with ageing phenotypes.

Follow-on or parallel experiments in model organisms are also a possibility; the Epidemiology and RNA groups in Exeter are forging collaborations with the United States NIA mouse centre. This will give unique opportunities to not only retrospectively look at what has already been done based on our results, but to design longitudinal experiments to test (in multiple tissues) the nature of the associations we observe. The translation and synergy between mice and human experiments will provide the greatest insight and efficiency to take this research forward.

7.6 Conclusion

Ageing is regarded as declining physiological function with age, accompanied by increased susceptibility to age-related diseases. Ageing research aims to improve healthspan in older humans. Mouse-model experiments suggest that the ageing process is plastic – that tissues retain the capability to rejuvenate – and can be modified to improve healthspan, in particular via specific circulating factors; however translation to humans remains unclear.

In these epidemiological studies I have designed analyses that investigate transcriptome-wide data derived from whole blood samples from human population study participants, aiming to determine biomarkers and gain mechanistic insight. This includes whole-transcriptome scans to determine gene transcripts associated with age-related phenotypes, and includes a meta-analysis of multiple human cohorts, mediation analysis to dissect age-related inflammation, and pathway-based and clustering methods to gain insight into relatedness between genes in terms of function and expression profile.

I have provided evidence supporting the link between age-related inflammation, loss of muscle strength, and loss of cognitive function. I have highlighted links between mouse-models and human phenotypes, with the potential for these genes to be used as predictive markers once follow-up data becomes available. In particular; *CCR2*-monocytes may be key for cognitive ageing and dementia; innate-immune activation is associated with muscle strength across a wide age-range; and age-related chronic inflammation is mediated by a small number of genes in whole blood, some of which are also related to ageing phenotypes.

The findings presented in this thesis provide a platform to continue research into molecular mechanisms and biomarkers of ageing phenotypes, including longitudinal and epigenetic data. I have demonstrated that overlap exists between model organisms and human ageing phenotypes and that observational molecular epidemiology is essential to highlight those of the greatest relevance.

Chapter 8 – Appendix

8.1 Contributions to other projects

During the course of the PhD I have contributed to a number of related projects within the group and in collaboration with others;

8.1.1 mRNA splicing and ageing

- Holly *et al* (2013) Changes in splicing factor expression are associated with advancing age in man. [80]
- Holly *et al* (2014) Splicing Factor 3B1 hypomethylation is associated with altered *SF3B1* transcript expression in older humans. [293]
- Holly *et al* (2015) Comparison of senescence-associated miRNAs in primary skin and lung fibroblasts [294]

8.1.2 mTOR pathway analysis

- Harries *et al* (2012) Advancing age is associated with gene expression changes resembling mTOR inhibition: evidence from two human populations. [295]
- Fellows *et al* (2012) Age related changes in mTOR-related gene expression in two primary human cell lines. [296]

8.1.3 CHARGE (and similar) meta-analyses

- Hek *et al* (2013) A genome-wide association study of depressive symptoms. [297]
- Lin *et al* (2014) Whole Blood Gene Expression and Interleukin-6 Levels. [61]
- Peters *et al* (2015) The transcriptional landscape of human age. [76]
- Huan *et al* (2015) A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension [298]
- Ibrahim-verbaas *et al* (2015) GWAS for executive function and processing speed suggests involvement of the CADM2 gene [299]

8.1.4 Ecotoxicology

- Fletcher *et al* (2013) Associations between PFOA, PFOS and changes in the expression of genes involved in cholesterol metabolism in humans. [300]
- Galloway *et al* (2015) PFOA and PFOS are associated with reduced expression of the parathyroid hormone 2 receptor (*PTH2R*) gene in women. [301]

8.1.5 Reviews and editorials

- Pilling *et al* (2012) Genomics and Successful Aging: Grounds for Renewed Optimism? [23]
- Harries *et al* (2013) Alzheimer's pathology: should peripheral monocytes and CCR2 take center stage? [53]
- Melzer *et al* (2013) Gene Expression Biomarkers and Longevity. [302]

8.1.6 Other studies

- Dutta *et al* (2013) Uric Acid Measurement Improves Prediction of Cardiovascular Mortality in Later Life. [26]
- Holly *et al* (2013) Towards a gene expression biomarker set for human biological age. [88]
- Winiarski *et al* (2014) Clinical Relevance of Increased Endothelial and Mesothelial Expression of Proangiogenic Proteases and VEGFA in the Omentum of Patients with Metastatic Ovarian High-Grade Serous Carcinoma. [303]
- Blackwell *et al* (2014) Changes in *CEBPB* Expression in Circulating Leukocytes following Eccentric Elbow-Flexion Exercise. [184]

8.2 References

[1] T. Flatt, “A new definition of aging?,” *Front. Genet.*, vol. 3, p. 148, Jan. 2012.

[2] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The hallmarks of aging.,” *Cell*, vol. 153, no. 6, pp. 1194–217, Jun. 2013.

[3] M. Sinha, Y. C. Jang, J. Oh, D. Khong, E. Y. Wu, R. Manohar, C. Miller, S. G. Regalado, F. S. Loffredo, J. R. Pancoast, M. F. Hirshman, J. Lebowitz, J. L. Shadrach, M. Cerletti, M.-J. Kim, T. Serwold, L. J. Goodyear, B. Rosner, R. T. Lee, and A. J. Wagers, “Restoring systemic GDF11 levels reverses age-related dysfunction in mouse skeletal muscle.,” *Science*, vol. 344, no. 6184, pp. 649–52, May 2014.

[4] S. a. Villeda, J. Luo, K. I. Mosher, B. Zou, M. Britschgi, G. Bieri, T. M. Stan, N. Fainberg, Z. Ding, A. Eggel, K. M. Lucin, E. Czirr, J.-S. Park, S. Couillard-Després, L. Aigner, G. Li, E. R. Peskind, J. a. Kaye, J. F. Quinn, D. R. Galasko, X. S. Xie, T. a. Rando, and T. Wyss-Coray, “The ageing systemic milieu negatively regulates neurogenesis and cognitive function.,” *Nature*, vol. 477, no. 7362, pp. 90–4, Sep. 2011.

[5] S. A. Villeda, K. E. Plambeck, J. Middeldorp, J. M. Castellano, K. I. Mosher, J. Luo, L. K. Smith, G. Bieri, K. Lin, D. Berdnik, R. Wabl, J. Udeochu, E. G. Wheatley, B. Zou, D. A. Simmons, X. S. Xie, F. M. Longo, and T. Wyss-Coray, “Young blood reverses age-related impairments in cognitive function and synaptic plasticity in mice.,” *Nat. Med.*, vol. 20, no. 6, pp. 659–63, Jun. 2014.

[6] A. Bitto and M. Kaeberlein, “Rejuvenation: it’s in our blood.,” *Cell Metab.*, vol. 20, no. 1, pp. 2–4, Jul. 2014.

[7] T. Murphy and S. Thuret, “The systemic milieu as a mediator of dietary influence on stem cell function during ageing.,” *Ageing Res. Rev.*, Dec. 2014.

[8] G. G. Fillenbaum, C. F. Pieper, H. J. Cohen, J. C. Cornoni-Huntley, and J. M. Guralnik, “Comorbidity of five chronic health conditions in elderly community residents: determinants and impact on mortality.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 55, no. 2, pp. M84–9, Feb. 2000.

[9] J. P. Butler, M. Fujii, and H. Sasaki, “Balanced aging, or successful aging?,” *Geriatr. Gerontol. Int.*, vol. 11, no. 1, pp. 1–2, Jan. 2011.

[10] T. Fulop, a Larbi, J. M. Witkowski, J. McElhaney, M. Loeb, a Mitnitski, and G. Pawelec, “Aging, frailty and age-related diseases.,” *Biogerontology*, vol. 11, no. 5, pp. 547–63, Oct. 2010.

[11] M. Bagley, T. Davis, J. Latimer, and D. Kipling, “The contribution of biogerontology to quality ageing,” *Qual. Ageing Older Adults*, vol. 12, no. 1, pp. 26–32, Mar. 2011.

[12] A. D. N. J. de Grey, “‘We will be able to live to 1,000,’” *BBC News Channel*, 03-Dec-2004. [Online]. Available: http://news.bbc.co.uk/1/hi/uk/4003063.stm.

[13] Office of National Statistics, “Mortality in England and Wales: Average Life Span, 2010,” 2012. [Online]. Available: http://www.ons.gov.uk/ons/dcp171776_292196.pdf.

[14] Y. Shi and S. Zhong, “From genomes to societies: a holistic view of determinants of human health.,” *Curr. Opin. Biotechnol.*, vol. 28C, pp. 134–142, Mar. 2014.

[15] A. Koulova and W. H. Frishman, “Air pollution exposure as a risk factor for cardiovascular disease morbidity and mortality.,” *Cardiol. Rev.*, vol. 22, no. 1, pp. 30–6, 2014.

[16] M. L. Levin, H. Goldstein, and P. R. Gerhardt, “Cancer and tobacco smoking; a preliminary report.,” *J. Am. Med. Assoc.*, vol. 143, no. 4, pp. 336–8, May 1950.

[17] D. a Raichlen and G. E. Alexander, “Exercise, APOE genotype, and the evolution of the human lifespan.,” *Trends Neurosci.*, pp. 1–9, Mar. 2014.

[18] S. Davinelli, D. C. Willcox, and G. Scapagnini, “Extending healthy ageing: nutrient sensitive pathway and centenarian population.,” *Immun. Ageing*, vol. 9, no. 1, p. 9, Jan. 2012.

[19] J. Soares, A. Cortinhas, T. Bento, and J. Leitão, “Aging and DNA damage in humans: a meta-analysis study,” *Aging (Albany. NY).*, vol. 6, no. 6, pp. 1–8, 2014.

[20] T. Leskinen, S. Sipilä, J. Kaprio, H. Kainulainen, M. Alen, and U. M. Kujala, “Physically active vs. inactive lifestyle, muscle properties, and glucose homeostasis in middle-aged and older twins.,” *Age (Dordr).*, vol. 35, no. 5, pp. 1917–26, Oct. 2013.

[21] A. Navarro and A. Boveris, “The mitochondrial energy transduction system and the aging process.,” *Am. J. Physiol. Cell Physiol.*, vol. 292, no. 2, pp. C670–86, Feb. 2007.

[22] J. P. de Magalhães, J. Curado, and G. M. Church, “Meta-analysis of age-related gene expression profiles identifies common signatures of aging.,” *Bioinformatics*, vol. 25, no. 7, pp. 875–81, Apr. 2009.

[23] L. C. Pilling, L. W. Harries, J. Powell, D. J. Llewellyn, L. Ferrucci, and D. Melzer, “Genomics and Successful Aging: Grounds for Renewed Optimism?,” *journals Gerontol. Ser. A, Biol. Med. Sci.*, Mar. 2012.

[24] WHO, “Tobacco: totally avoidable risk factor of CVD.” [Online]. Available: http://www.world-heart-federation.org/press/fact-sheets/tobacco-totally-avoidable-risk-factor-of-cvd/.

[25] A. Wikby, F. Ferguson, J. Strindhall, R. J. Forsey, T. Fulop, S. R. Hadrup, P. Straten, and G. Pawelec, “Immune Risk Phenotypes and Associated Parameters in Very Old Humans: A Review of Findings in the Swedish NONA Immune Longitudinal Study,” in *Immunosenescence*, G. Pawelec, Ed. Landes Bioscience and Springer Science, 2007.

[26] A. Dutta, W. Henley, L. C. Pilling, R. B. Wallace, and D. Melzer, “Uric Acid Measurement Improves Prediction of Cardiovascular Mortality in Later Life.,” *J. Am. Geriatr. Soc.*, pp. 319–326, Mar. 2013.

[27] M. Shardell, C. D’Adamo, D. E. Alley, R. R. Miller, G. E. Hicks, Y. Milaneschi, R. D. Semba, A. Cherubini, S. Bandinelli, and L. Ferrucci, “Serum 25-Hydroxyvitamin D, Transitions Between Frailty States, and Mortality in Older Adults: The Invecchiare in Chianti Study,” *J. Am. Geriatr. Soc.*, vol. 25, p. n/a–n/a, Jan. 2012.

[28] J. a Sorrentino, H. K. Sanoff, and N. E. Sharpless, “Defining the toxicology of aging.,” *Trends Mol. Med.*, vol. 20, no. 7, pp. 375–384, Jul. 2014.

[29] J. M. Adams and M. White, “Biological ageing: a fundamental, biological link between socio-economic status and health?,” *Eur. J. Public Health*, vol. 14, no. 3, pp. 331–334, 2004.

[30] M. E. Kelly, D. Loughrey, B. A. Lawlor, I. H. Robertson, C. Walsh, and S. Brennan, “The impact of cognitive training and mental stimulation on cognitive and everyday functioning of healthy older adults: A systematic review and meta-analysis.,” *Ageing Res. Rev.*, vol. 15C, pp. 28–43, May 2014.

[31] N. Zhu, D. R. Jacobs, P. J. Schreiner, K. Yaffe, N. Bryan, L. J. Launer, R. A. Whitmer, S. Sidney, E. Demerath, W. Thomas, C. Bouchard, K. He, J. Reis, and B. Sternfeld, “Cardiorespiratory fitness and cognitive function in middle age: the CARDIA study.,” *Neurology*, vol. 82, no. 15, pp. 1339–46, Apr. 2014.

[32] A.-M. Tolppanen, A. Solomon, J. Kulmala, I. Kåreholt, T. Ngandu, M. Rusanen, T. Laatikainen, H. Soininen, and M. Kivipelto, “Leisure-time physical activity from mid- to late life, body mass index, and risk of dementia.,” *Alzheimers. Dement.*, Apr. 2014.

[33] J. Myers, “Cardiology patient pages. Exercise and cardiovascular health.,” *Circulation*, vol. 107, no. 1, pp. e2–5, Jan. 2003.

[34] T. Rantanen, K. Masaki, Q. He, G. W. Ross, B. J. Willcox, and L. White, “Midlife muscle strength and human longevity up to age 100 years: a 44-year prospective study among a decedent cohort.,” *Age (Dordr).*, vol. 34, no. 3, pp. 563–70, Jun. 2012.

[35] J. Walston, E. C. Hadley, L. Ferrucci, J. M. Guralnik, A. B. Newman, S. A. Studenski, W. B. Ershler, T. Harris, and L. P. Fried, “Research agenda for frailty in older adults: toward a better understanding of physiology and etiology: summary from the American Geriatrics Society/National Institute on Aging Research Conference on Frailty in Older Adults.,” *J. Am. Geriatr. Soc.*, vol. 54, no. 6, pp. 991–1001, Jun. 2006.

[36] J. M. Guralnik, L. Ferrucci, E. M. Simonsick, M. E. Salive, and R. B. Wallace, “Lower-extremity function in persons over the age of 70 years as a predictor of subsequent disability.,” *N. Engl. J. Med.*, vol. 332, no. 9, pp. 556–61, Mar. 1995.

[37] K. B. Rajan, N. T. Aggarwal, R. S. Wilson, S. A. Everson-Rose, and D. A. Evans, “Association of cognitive functioning, incident stroke, and mortality in older adults.,” *Stroke.*, vol. 45, no. 9, pp. 2563–7, Sep. 2014.

[38] M. De Martinis, C. Franceschi, D. Monti, and L. Ginaldi, “Inflammation markers predicting frailty and mortality in the elderly.,” *Exp. Mol. Pathol.*, vol. 80, no. 3, pp. 219–27, Jun. 2006.

[39] T. Rantanen, J. M. Guralnik, D. Foley, K. Masaki, S. Leveille, J. D. Curb, and L. White, “Midlife hand grip strength as a predictor of old age disability.,” *JAMA*, vol. 281, no. 6, pp. 558–60, Feb. 1999.

[40] L. A. Schaap, A. Koster, and M. Visser, “Adiposity, Muscle Mass, and Muscle Strength in Relation to Functional Decline in Older Persons.,” *Epidemiol. Rev.*, Dec. 2012.

[41] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke, and M. A. McBurnie, “Frailty in older adults: evidence for a phenotype.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 56, no. 3, pp. M146–M156, 2001.

[42] S. W. and A. B. Tietjen-Smith, Steve W. Smith, Malissa Martin, Ruth Henry, “Grip Strength in Relation to Overall Strength and Functional Capacity in Very Old and Oldest Old Females,” *Phys. Occup. Ther. Geriatr.*, vol. 24, no. 4, pp. 63–78, 2006.

[43] D. Legrand, B. Vaes, C. Matheï, C. Swine, and J. M. Degryse, “The prevalence of sarcopenia in very old individuals according to the european consensus definition: Insights from the belfrail study,” *Age Ageing*, vol. 42, no. 6, pp. 727–734, 2013.

[44] R. Csapo, V. Malis, U. Sinha, J. Du, and S. Sinha, “Age-associated differences in triceps surae muscle composition and strength - an MRI-based cross-sectional comparison of contractile, adipose and connective tissue.,” *BMC Musculoskelet. Disord.*, vol. 15, no. 1, p. 209, Jan. 2014.

[45] S. Carosio, M. G. Berardinelli, M. Aucello, and A. Musarò, “Impact of ageing on muscle cell regeneration.,” *Ageing Res. Rev.*, vol. 10, no. 1, pp. 35–42, Jan. 2011.

[46] B. Carlson and J. Faulkner, “Muscle transplantation between young and old rats: age of host determines recovery,” *Am. J. Physiol. Physiol.*, vol. 256, no. 6, p. C1262, 1989.

[47] J. G. Tidball and S. A. Villalta, “Regulatory interactions between muscle and the immune system during muscle regeneration.,” *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 298, no. 5, pp. R1173–87, May 2010.

[48] C. Franceschi, M. Capri, D. Monti, S. Giunta, F. Olivieri, F. Sevini, M. P. Panourgia, L. Invidia, L. Celani, M. Scurti, E. Cevenini, G. C. Castellani, and S. Salvioli, “Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans.,” *Mech. Ageing Dev.*, vol. 128, no. 1, pp. 92–105, Jan. 2007.

[49] Alzheimer’s Association, “What Is Dementia?” [Online]. Available: http://www.alz.org/.

[50] N. A. Bishop, T. Lu, and B. A. Yankner, “Neural mechanisms of ageing and cognitive decline.,” *Nature*, vol. 464, no. 7288, pp. 529–35, Mar. 2010.

[51] U. Träger and S. J. Tabrizi, “Peripheral inflammation in neurodegeneration.,” *J. Mol. Med. (Berl).*, Apr. 2013.

[52] F. Marques, J. C. Sousa, N. Sousa, and J. A. Palha, “Blood-brain-barriers in aging and in Alzheimer’s disease.,” *Mol. Neurodegener.*, vol. 8, no. 1, p. 38, Jan. 2013.

[53] L. W. Harries, L. C. Pilling, A. Lampron, S. Rivest, and D. Melzer, “Alzheimer’s pathology: should peripheral monocytes and CCR2 take center stage?,” *Neurodegener. Dis. Manag.*, vol. 3, no. 1, pp. 9–12, Feb. 2013.

[54] I. Trapero and O. Cauli, “Interleukin 6 and cognitive dysfunction.,” *Metab. Brain Dis.*, Apr. 2014.

[55] “Alzheimer’s Society report 2014.” [Online]. Available: http://www.alzheimers.org.uk/dementiauk. [Accessed: 21-Apr-2015].

[56] C. Franceschi, M. Bonafè, S. Valensin, F. Olivieri, M. De Luca, E. Ottaviani, and G. De Benedictis, “Inflamm-aging. An evolutionary perspective on immunosenescence.,” *Ann. N. Y. Acad. Sci.*, vol. 908, pp. 244–54, Jun. 2000.

[57] P. Palta, Q.-L. Xue, J. a Deal, L. P. Fried, J. D. Walston, and M. C. Carlson, “Interleukin-6 and C-Reactive Protein Levels and 9-Year Cognitive Decline in Community-Dwelling Older Women: The Women’s Health and Aging Study II.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, no. 4, pp. 1–6, Aug. 2014.

[58] M. Cesari, B. W. J. H. Penninx, M. Pahor, F. Lauretani, A. M. Corsi, G. Rhys Williams, J. M. Guralnik, and L. Ferrucci, “Inflammatory markers and physical performance in older persons: the InCHIANTI study.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 59, no. 3, pp. 242–8, Mar. 2004.

[59] F. Rodier and J. Campisi, “Four faces of cellular senescence.,” *J. Cell Biol.*, vol. 192, no. 4, pp. 547–56, Feb. 2011.

[60] C. Franceschi and J. Campisi, “Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 69 Suppl 1, pp. S4–9, Jun. 2014.

[61] H. Lin, R. Joehanes, L. C. Pilling, J. Dupuis, K. L. Lunetta, S.-X. Ying, E. J. Benjamin, D. Hernandez, A. Singleton, D. Melzer, P. J. Munson, D. Levy, L. Ferrucci, and J. M. Murabito, “Whole Blood Gene Expression and Interleukin-6 Levels.,” *Genomics*, vol. In press, Oct. 2014.

[62] L.-J. Tan, S.-L. Liu, S.-F. Lei, C. J. Papasian, and H.-W. Deng, “Molecular genetic studies of gene identification for sarcopenia.,” *Hum. Genet.*, vol. 131, no. 1, pp. 1–31, Jan. 2012.

[63] I. Miljkovic, A. L. Kuipers, C. M. Kammerer, X. Wang, C. H. Bunker, A. L. Patrick, V. W. Wheeler, L. H. Kuller, R. W. Evans, and J. M. Zmuda, “Markers of inflammation are heritable and associated with subcutaneous and ectopic skeletal muscle adiposity in African ancestry families.,” *Metab. Syndr. Relat. Disord.*, vol. 9, no. 4, pp. 319–326, 2011.

[64] T. A. Greenwood, M. S. Beeri, J. Schmeidler, D. Valerio, H. Raventós, L. Mora-Villalobos, K. Camacho, J. R. Carrión-Baralt, G. Angelo, L. Almasy, M. Sano, and J. M. Silverman, “Heritability of cognitive functions in families of successful cognitive aging probands from the central valley of Costa Rica,” *J. Alzheimer’s Dis.*, vol. 27, no. 4, pp. 897–907, 2011.

[65] D. C. Koboldt, “The Current State of dbSNP,” 2012. [Online]. Available: http://www.massgenomics.org/2012/01/the-current-state-of-dbsnp.html.

[66] L. Broer, A. S. Buchman, J. Deelen, D. S. Evans, J. D. Faul, K. L. Lunetta, P. Sebastiani, J. a Smith, A. V Smith, T. Tanaka, L. Yu, A. M. Arnold, T. Aspelund, E. J. Benjamin, P. L. De Jager, G. Eirkisdottir, D. a Evans, M. E. Garcia, A. Hofman, R. C. Kaplan, S. L. R. Kardia, D. P. Kiel, B. a Oostra, E. S. Orwoll, N. Parimi, B. M. Psaty, F. Rivadeneira, J. I. Rotter, S. Seshadri, A. Singleton, H. Tiemeier, A. G. Uitterlinden, W. Zhao, S. Bandinelli, D. a Bennett, L. Ferrucci, V. Gudnason, T. B. Harris, D. Karasik, L. J. Launer, T. T. Perls, P. E. Slagboom, G. J. Tranah, D. R. Weir, A. B. Newman, C. M. van Duijn, and J. M. Murabito, “GWAS of Longevity in CHARGE Consortium Confirms APOE and FOXO3 Candidacy.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, pp. 1–9, Sep. 2014.

[67] J. Deelen, M. Beekman, H.-W. Uh, L. Broer, K. L. Ayers, Q. Tan, Y. Kamatani, A. M. Bennet, R. Tamm, S. Trompet, D. F. Guðbjartsson, F. Flachsbart, G. Rose, A. Viktorin, K. Fischer, M. Nygaard, H. J. Cordell, P. Crocco, E. B. van den Akker, S. Böhringer, Q. Helmer, C. P. Nelson, G. I. Saunders, M. Alver, K. Andersen-Ranberg, M. E. Breen,

R. van der Breggen, A. Caliebe, M. Capri, E. Cevenini, J. C. Collerton, S. Dato, K. Davies, I. Ford, J. Gampe, P. Garagnani, E. J. C. de Geus, J. Harrow, D. van Heemst, B. T. Heijmans, F.-A. Heinsen, J.-J. Hottenga, A. Hofman, B. Jeune, P. V Jonsson, M. Lathrop, D. Lechner, C. Martin-Ruiz, S. E. Mcnerlan, E. Mihailov, A. Montesanto, S. P. Mooijaart, A. Murphy, E. A. Nohr, L. Paternoster, I. Postmus, F. Rivadeneira, O. A. Ross, S. Salvioli, N. Sattar, S. Schreiber, H. Stefánsson, D. J. Stott, H. Tiemeier, A. G. Uitterlinden, R. G. J. Westendorp, G. Willemsen, N. J. Samani, P. Galan, T. I. A. Sørensen, D. I. Boomsma, J. W. Jukema, I. M. Rea, G. Passarino, A. J. M. de Craen, K. Christensen, A. Nebel, K. Stefánsson, A. Metspalu, P. Magnusson, H. Blanché, L. Christiansen, T. B. L. Kirkwood, C. M. van Duijn, C. Franceschi, J. J. Houwing-Duistermaat, and P. E. Slagboom, “Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age.,” *Hum. Mol. Genet.*, vol. 23, no. 16, pp. 4420–32, Aug. 2014.

[68] J. P. de Magalhães, “Why genes extending lifespan in model organisms have not been consistently associated with human longevity and what it means to translation research,” *Cell Cycle*, vol. 13, no. 17, pp. 2671–2673, Oct. 2014.

[69] L. Hindorff, J. MacArthur, A. Wise, H. Junkins, P. Hall, A. Klemm, and T. Manolio, “A Catalog of Published Genome-Wide Association Studies.” [Online]. Available: www.genome.gov/gwastudies.

[70] W. R. Jeck, A. P. Siebold, and N. E. Sharpless, “Review: a meta-analysis of GWAS and age-associated diseases.,” *Aging Cell*, vol. 11, no. 5, pp. 727–31, Oct. 2012.

[71] M. Beekman, C. Nederstigt, H. E. D. Suchiman, D. Kremer, R. van der Breggen, N. Lakenberg, W. G. Alemayehu, A. J. M. de Craen, R. G. J. Westendorp, D. I. Boomsma, E. J. C. de Geus, J. J. Houwing-Duistermaat, B. T. Heijmans, and P. E. Slagboom, “Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 42, pp. 18046–9, Oct. 2010.

[72] A. Dutta, W. Henley, J.-M. Robine, K. M. Langa, R. B. Wallace, and D. Melzer, “Longer Lived Parents: Protective Associations With Cancer Incidence and Overall Mortality.,” *journals Gerontol. Ser. A Biol. Sci. Med. Sci.*, 2013.

[73] F. Coppedè, “The epidemiology of premature aging and associated comorbidities.,” *Clin. Interv. Aging*, vol. 8, pp. 1023–1032, Jan. 2013.

[74] R. C. M. Hennekam, “Hutchinson-Gilford progeria syndrome: review of the phenotype.,” *Am. J. Med. Genet. A*, vol. 140, no. 23, pp. 2603–24, Dec. 2006.

[75] M. I. Miyamoto, K. Djabali, and L. B. Gordon, “Atherosclerosis in Ancient Humans, Accelerated Aging Syndromes and Normal Aging,” *Glob. Heart*, vol. 9, no. 2, pp. 211–218, Jun. 2014.

[76] M. J. Peters, R. Joehanes, L. C. Pilling, C. Schurmann, K. N. Conneely, J. Powell, E. Reinmaa, G. L. Sutphin, A. Zhernakova, K. Schramm, Y. A. Wilson, S. Kobes, J. Yang, T. Tukiainen, M. A. Nalls, Y. F. Ramos, H. H. H. Göring, M. Fornage, Y. Liu, S. A. Gharib, B. E. Stranger, P. L. De Jager, A. Aviv, D. Levy, J. M. Murabito, P. J. Munson, A. Hofman, A. G. Uitterlinden, F. Rivadeneira, J. van Rooij, L. Stolk, L. Broer, M. M. P. J. Verbiest, M. Jhamai, P. Arp, A. Metspalu, L. Tserel, L. Milani, N. J. Samani, P. Peterson, S. Kasela, V. Codd, A. Peters, C. Herder, M. Waldenberger, M. Roden, P. Singmann, S. Zeilinger, T. Illig, G. Homuth, H.-J. Grabe, H. Völzke, L. Steil, T. Kocher, D. Melzer, D. G. Hernandez, H. Yaghootkar, S. Bandinelli, E. K. Moses, J. E. Curran, M. P. Johnson, S. Williams-Blangero, H.-J. Westra, A. F. McRae, J. A.

Smith, S. L. R. Kardia, I. Hovatta, M. Perola, S. Ripatti, V. Salomaa, A. K. Smith, D. Mehta, E. B. Binder, K. M. Nylocks, T. Klengel, J. Ding, A. M. Suchy-Dicey, D. A. Enquobahrie, J. Brody, J. I. Rotter, Y.-D. I. Chen, J. Houwing-Duistermaat, M. Kloppenburg, P. E. Slagboom, Q. Helmer, W. den Hollander, S. Bean, J. Hardy, M. R. Cookson, M. Ryten, R. J. Gibbs, T. Raj, N. Bahkshi, Q. P. Wang, L. J. Oyston, B. M. Psaty, R. P. Tracy, G. W. Montgomery, S. T. Turner, J. Blangero, I. Meulenbelt, K. J. Ressler, A. B. Singleton, L. Franke, J. Kettunen, P. M. Visscher, G. G. Neely, R. Korstanje, R. L. Hanson, H. Prokisch, L. Ferrucci, T. Esko, A. Teumer, J. B. J. van Meurs, and A. D. Johnson, “The transcriptional landscape of human age,” *Nat. Commun.*, p. In submission, 2015.

[77] A. E. Kornienko, P. M. Guenzl, D. P. Barlow, and F. M. Pauler, “Gene regulation by the act of long non-coding RNA transcription.,” *BMC Biol.*, vol. 11, no. 1, p. 59, Jan. 2013.

[78] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, a M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. a Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin, “An atlas of active enhancers across human cell types and tissues.,” *Nature*, vol. 507, no. 7493, pp. 455–61, Mar. 2014.

[79] L. W. Harries, D. Hernandez, W. Henley, A. R. Wood, A. C. Holly, R. M. Bradley-Smith, H. Yaghootkar, A. Dutta, A. Murray, T. M. Frayling, J. M. Guralnik, S. Bandinelli, A. Singleton, L. Ferrucci, and D. Melzer, “Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing.,” *Aging Cell*, vol. 10, no. 5, pp. 868–78, Oct. 2011.

[80] A. C. Holly, D. Melzer, L. C. Pilling, A. C. Fellows, T. Tanaka, L. Ferrucci, and L. W. Harries, “Changes in splicing factor expression are associated with advancing age in man.,” *Mech. Ageing Dev.*, vol. 134, no. 9, pp. 356–66, Sep. 2013.

[81] M. Maurano, R. Humbert, and E. Rynes, “Systematic localization of common disease-associated variation in regulatory DNA,” *Science (80-.).*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[82] P. De Boever, B. Wens, A. C. Forcheh, H. Reynders, V. Nelen, J. Kleinjans, N. Van Larebeke, G. Verbeke, D. Valkenborg, and G. Schoeters, “Characterization of the peripheral blood transcriptome in a repeated measures design using a panel of healthy individuals.,” *Genomics*, vol. 103, no. 1, pp. 31–9, Jan. 2014.

[83] P. Fehlbaum-Beurdeley, A. C. Jarrige-Le Prado, D. Pallares, J. Carrière, C. Guihal, C. Soucaille, F. Rouet, D. Drouin, O. Sol, H. Jordan, D. Wu, L. Lei, R. Einstein, F. Schweighoffer, and L. Bracco, “Toward an Alzheimer’s disease diagnosis via high-resolution blood gene expression.,” *Alzheimer’s Dement. J. Alzheimer's Assoc.*, vol. 6, no. 1, pp. 25–38, Jan. 2010.

[84] K. Lunnon, M. Sattlecker, S. J. Furney, G. Coppola, A. Simmons, P. Proitsi, M. K. Lupton, A. Lourdusamy, C. Johnston, H. Soininen, I. Kłoszewska, P. Mecocci, M. Tsolaki, B. Vellas, D. Geschwind, S. Lovestone, R. Dobson, and A. Hodges, “A blood gene expression marker of early Alzheimer’s disease.,” *J. Alzheimers. Dis.*, vol. 33, no. 3, pp. 737–53, Jan. 2013.

[85] J. Cui, F. Li, G. Wang, X. Fang, J. D. Puett, and Y. Xu, “Gene-expression signatures can distinguish gastric cancer grades and stages.,” *PLoS One*, vol. 6, no. 3, p. e17819, Jan. 2011.

[86] J.-H. Ko, E. A. Ko, W. Gu, I. Lim, H. Bang, and T. Zhou, “Expression profiling of ion channel genes predicts clinical outcome in breast cancer.,” *Mol. Cancer*, vol. 12, no. 1, p. 106, Jan. 2013.

[87] C. Franceschi, “Inflammaging as a major characteristic of old people: can it be prevented or cured?,” *Nutr Rev*, vol. 65, no. 12, pp. 173–6, 2007.

[88] A. C. Holly, D. Melzer, L. C. Pilling, W. Henley, D. G. Hernandez, A. B. Singleton, S. Bandinelli, J. M. Guralnik, L. Ferrucci, and L. W. Harries, “Towards a gene expression biomarker set for human biological age.,” *Aging Cell*, Jan. 2013.

[89] V. M. Weake and J. L. Workman, “Inducible gene expression: diverse regulatory mechanisms.,” *Nat. Rev. Genet.*, vol. 11, no. 6, pp. 426–37, Jun. 2010.

[90] NIH, “NCBI Epigenomics resource,” 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/epigenomics.

[91] A. Brunet and S. L. Berger, “Epigenetics of Aging and Aging-related Disease.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 69 Suppl 1, pp. S17–20, Jun. 2014.

[92] G. P. Pfeifer, S. Kadam, and S.-G. Jin, “5-Hydroxymethylcytosine and Its Potential Roles in Development and Cancer.,” *Epigenetics Chromatin*, vol. 6, no. 1, p. 10, Jan. 2013.

[93] V. Kirn, R. Shi, S. Heublein, J. Knabl, M. Guenthner-Biller, U. Andergassen, C. Fridrich, W. Malter, J. Harder, K. Friese, D. Mayr, and U. Jeschke, “Estrogen receptor promoter methylation predicts survival in low-grade ovarian carcinoma patients.,” *J. Cancer Res. Clin. Oncol.*, Jun. 2014.

[94] V. D. F. de Mello, L. Pulkkinen, M. Lalli, M. Kolehmainen, J. Pihlajamäki, and M. Uusitupa, “DNA methylation in obesity and type 2 diabetes.,” *Ann. Med.*, vol. 46, no. 3, pp. 103–13, May 2014.

[95] C. Weidner, Q. Lin, C. Koch, L. Eisele, F. Beier, P. Ziegler, D. Bauerschlag, K.-H. Jöckel, R. Erbel, T. Mühleisen, M. Zenke, T. Brümmendorf, and W. Wagner, “Aging of blood can be tracked by DNA methylation changes at just three CpG sites,” *Genome Biol.*, vol. 15, no. 2, p. R24, 2014.

[96] N. Sen, “Epigenetic Regulation of Memory by Acetylation and Methylation of Chromatin: Implications in Neurological Disorders, Aging, and Addiction.,” *Neuromolecular Med.*, Apr. 2014.

[97] R. J. O’Sullivan and J. Karlseder, “The great unravelling: chromatin as a modulator of the aging process.,” *Trends Biochem. Sci.*, vol. 37, no. 11, pp. 466–76, Nov. 2012.

[98] J. M. Sedivy, G. Banumathy, and P. D. Adams, “Aging by epigenetics--a consequence of chromatin damage?,” *Exp. Cell Res.*, vol. 314, no. 9, pp. 1909–17, Jun. 2008.

[99] Wellcome Trust, “What are ‘model organisms’?,” 2002. [Online]. Available: http://genome.wellcome.ac.uk/doc_WTD020803.html.

[100] F. C. Ludwig and R. M. Elashoff, “Mortality in syngeneic rat parabionts of different chronological age.,” *Trans. N. Y. Acad. Sci.*, vol. 34, no. 7, pp. 582–7, Nov. 1972.

[101] P. Kamran, K.-I. Sereti, P. Zhao, S. R. Ali, I. L. Weissman, and R. Ardehali, “Parabiosis in mice: a detailed protocol.,” *J. Vis. Exp.*, no. 80, Jan. 2013.

[102] I. M. Conboy, M. J. Conboy, A. J. Wagers, E. R. Girma, I. L. Weissman, and T. A. Rando, “Rejuvenation of aged progenitor cells by exposure to a young systemic environment.,” *Nature*, vol. 433, no. 7027, pp. 760–4, Feb. 2005.

[103] D. Gems and L. Partridge, “Genetics of longevity in model organisms: debates and paradigm shifts.,” *Annu. Rev. Physiol.*, vol. 75, pp. 621–44, Feb. 2013.

[104] B. K. Kennedy and J. K. Pennypacker, “Drugs that modulate aging: the promising yet difficult path ahead.,” *Transl. Res.*, vol. 163, no. 5, pp. 456–65, May 2014.

[105] E. M. Mercken, B. a Carboneau, S. M. Krzysik-Walker, and R. de Cabo, “Of mice and men: The benefits of caloric restriction, exercise, and mimetics.,” *Ageing Res. Rev.*, vol. 11, no. 3, pp. 390–398, Dec. 2011.

[106] R. J. Colman, R. M. Anderson, S. C. Johnson, E. K. Kastman, K. J. Kosmatka, T. M. Beasley, D. B. Allison, C. Cruzen, H. A. Simmons, J. W. Kemnitz, and R. Weindruch, “Caloric restriction delays disease onset and mortality in rhesus monkeys.,” *Science*, vol. 325, no. 5937, pp. 201–4, Jul. 2009.

[107] J. A. Mattison, G. S. Roth, T. M. Beasley, E. M. Tilmont, A. M. Handy, R. L. Herbert, D. L. Longo, D. B. Allison, J. E. Young, M. Bryant, D. Barnard, W. F. Ward, W. Qi, D. K. Ingram, and R. de Cabo, “Impact of caloric restriction on health and survival in rhesus monkeys from the NIA study.,” *Nature*, vol. 489, no. 7415, pp. 318–21, Sep. 2012.

[108] R. J. Colman, T. M. Beasley, J. W. Kemnitz, S. C. Johnson, R. Weindruch, and R. M. Anderson, “Caloric restriction reduces age-related and all-cause mortality in rhesus monkeys,” *Nat. Commun.*, vol. 5, pp. 1–5, Apr. 2014.

[109] E. Cava and L. Fontana, “Will calorie restriction work in humans?,” *Aging (Albany. NY).*, vol. 5, no. 7, pp. 507–14, Jul. 2013.

[110] J. Campisi, “Aging, Cellular Senescence, and Cancer.,” *Annu. Rev. Physiol.*, vol. 75, pp. 17.1–17.21, Nov. 2013.

[111] R. M. Naylor, D. J. Baker, and J. M. van Deursen, “Senescent Cells: A Novel Therapeutic Target for Aging and Age-Related Diseases.,” *Clin. Pharmacol. Ther.*, no. September, Dec. 2012.

[112] D. J. Baker, T. Wijshake, T. Tchkonia, N. K. LeBrasseur, B. G. Childs, B. van de Sluis, J. L. Kirkland, and J. M. van Deursen, “Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders.,” *Nature*, vol. 479, no. 7372, pp. 232–6, Nov. 2011.

[113] D. Ruffell, F. Mourkioti, A. Gambardella, P. Kirstetter, R. G. Lopez, N. Rosenthal, and C. Nerlov, “A CREB-C/EBPbeta cascade induces M2 macrophage-specific gene expression and promotes muscle injury repair.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 41, pp. 17475–80, Oct. 2009.

[114] G. Naert and S. Rivest, “CC chemokine receptor 2 deficiency aggravates cognitive impairments and amyloid pathology in a transgenic mouse model of Alzheimer’s disease.,” *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 31, no. 16, pp. 6208–20, Apr. 2011.

[115] L. Ferrucci, S. Bandinelli, E. Benvenuti, A. Di Iorio, C. Macchi, T. B. Harris, and J. M. Guralnik, “Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study.,” *J. Am. Geriatr. Soc.*, vol. 48, no. 12, pp. 1618–25, Dec. 2000.

[116] Luigi Ferrucci and Stefania Bandinelli, “InCHIANTI - INVECCHIARE NEL CHIANTI.” [Online]. Available: http://inchiantistudy.net/.

[117] S. Debey-Pascher, D. Eggle, and J. L. Schultze, “RNA stabilization of peripheral blood and profiling by bead chip analysis.,” *Methods Mol. Biol.*, vol. 496, pp. 175–210, Jan. 2009.

[118] Quiagen, “PAXgene Blood mRNA kit handbook.” [Online]. Available: http://www.ebiotrade.com/buyf/productsf/qiagen/1017455paxgenebloodrna42001.pdf.

[119] “NIH Genome Institute - sequencing costs.” [Online]. Available: http://www.genome.gov/sequencingcosts/.

[120] Illumina Inc, “Illumina HumanHT-12 v3 Expression BeadChip Kit.” [Online]. Available: http://res.illumina.com/documents/products/datasheets/datasheet_humanht_12.pdf.

[121] Illumina Inc, “Illumina Gene Expression Analysis.” [Online]. Available: http://res.illumina.com/documents/products/datasheets/datasheet_gene_exp_analysis.pdf.

[122] A. R. Wood, D. G. Hernandez, M. A. Nalls, H. Yaghootkar, J. R. Gibbs, L. W. Harries, S. Chong, M. Moore, M. N. Weedon, J. M. Guralnik, S. Bandinelli, A. Murray, L. Ferrucci, A. B. Singleton, D. Melzer, and T. M. Frayling, “Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association.,” *Hum. Mol. Genet.*, vol. 20, no. 20, pp. 4082–92, Oct. 2011.

[123] H. Yaghootkar, “NCBI Gene Expression Omnibus: InCHIANTI follow-up 3 gene expression data.” [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48152.

[124] E. Raitoharju, I. Seppälä, L.-P. Lyytikäinen, M. Levula, N. Oksala, N. Klopp, T. Illig, R. Laaksonen, M. Kähönen, and T. Lehtimäki, “A comparison of the accuracy of Illumina HumanHT-12 v3 Expression BeadChip and TaqMan qRT-PCR gene expression results in patient samples from the Tampere Vascular Study.,” *Atherosclerosis*, vol. 226, no. 1, pp. 149–52, Jan. 2013.

[125] Assay Depot, “Taqman Low Density Array (TLDA) Service Description.” [Online]. Available: https://www.assaydepot.com/service/taqman-low-density-array-tlda.

[126] S. Fleige, V. Walf, S. Huch, C. Prgomet, J. Sehm, and M. W. Pfaffl, “Comparison of relative mRNA quantification models and the impact of RNA integrity in quantitative real-time RT-PCR.,” *Biotechnol. Lett.*, vol. 28, no. 19, pp. 1601–13, Oct. 2006.

[127] BioSistemika, “What is real-time PCR?” [Online]. Available: http://biosistemika.com/workshops/qpcr-basics/.

[128] Applied Biosystems, “TaqMan Probe-Based Gene Expression Analysis.” [Online]. Available: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/real-time-pcr/taqman-probe-based-gene-expression-analysis.html.

[129] P. Beineke, K. Fitch, H. Tao, M. R. Elashoff, S. Rosenberg, W. E. Kraus, and J. A. Wingrove, “A whole blood gene expression-based signature for smoking status.,” *BMC Med. Genomics*, vol. 5, no. 1, p. 58, Jan. 2012.

[130] labstats.net, “Statistics for Experimental Biologists: What exactly is a p-value?” [Online]. Available: http://labstats.net/articles/pvalue.html.

[131] M. Goldman, “Statistics for Bioinformatics,” *Berkeley statistics course*, 2008. [Online]. Available: http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf.

[132] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *J. R. Stat. Soc. Ser. B*, pp. 289–300, 1995.

[133] M. E. Sobel, “Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models,” *Sociol. Methodol.*, vol. 13, p. 290, 1982.

[134] K. Imai, L. Keele, and T. Yamamoto, “Identification, Inference and Sensitivity Analysis for Causal Mediation Effects,” *Stat. Sci.*, vol. 25, no. 1, pp. 51–71, 2010.

[135] M. S. Fritz and D. P. Mackinnon, “Required sample size to detect the mediated effect.,” *Psychol. Sci.*, vol. 18, no. 3, pp. 233–9, Mar. 2007.

[136] T. R. Ten Have and M. M. Joffe, “A review of causal estimation of effects in mediation analyses.,” *Stat. Methods Med. Res.*, vol. 21, no. 1, pp. 77–107, Feb. 2012.

[137] “Scientific Linux.” [Online]. Available: http://www.scientificlinux.org/.

[138] L. Walll, “Perl quote,” 1994. [Online]. Available: http://en.wikiquote.org/wiki/Larry_Wall#1994.

[139] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, “GeneCards: integrating information about genes, proteins and diseases.,” *Trends Genet.*, vol. 13, no. 4, p. 163, Apr. 1997.

[140] B. M. Psaty, C. J. O’Donnell, V. Gudnason, K. L. Lunetta, A. R. Folsom, J. I. Rotter, A. G. Uitterlinden, T. B. Harris, J. C. M. Witteman, and E. Boerwinkle, “Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts.,” *Circ. Cardiovasc. Genet.*, vol. 2, no. 1, pp. 73–80, Mar. 2009.

[141] T. Lang, T. Streeper, P. Cawthon, K. Baldwin, D. R. Taaffe, and T. B. Harris, “Sarcopenia: etiology, clinical consequences, intervention, and assessment.,” *Osteoporos. Int.*, vol. 21, no. 4, pp. 543–59, Apr. 2010.

[142] J. Peake, P. Della Gatta, and D. Cameron-Smith, “Aging and its effects on inflammation in skeletal muscle at rest and following exercise-induced muscle injury.,” *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 298, no. 6, pp. R1485–95, Jun. 2010.

[143] J. A. Faulkner, S. V Brooks, and E. Zerba, “Muscle atrophy and weakness with aging: contraction-induced injury as an underlying mechanism.,” *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 50 Spec No, pp. 124–9, Nov. 1995.

[144] J. G. Tidball, “Inflammatory processes in muscle injury and repair.,” *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 288, no. 2, pp. R345–53, Feb. 2005.

[145] H. X. Nguyen and J. G. Tidball, “Interactions between neutrophils and macrophages promote macrophage killing of rat muscle cells in vitro.,” *J. Physiol.*, vol. 547, no. Pt 1, pp. 125–32, Feb. 2003.

[146] B. A. St Pierre and J. G. Tidball, “Differential response of macrophage subpopulations to soleus muscle reloading after rat hindlimb suspension.,” *J. Appl. Physiol.*, vol. 77, no. 1, pp. 290–7, Jul. 1994.

[147] L. Arnold, A. Henry, F. Poron, Y. Baba-Amer, N. van Rooijen, A. Plonquet, R. K. Gherardi, and B. Chazaud, “Inflammatory monocytes recruited after skeletal muscle injury switch into antiinflammatory macrophages to support myogenesis.,” *J. Exp. Med.*, vol. 204, no. 5, pp. 1057–69, May 2007.

[148] T. Robertson, M. Maley, M. Grounds, and J. Papadimitriou, “The role of macrophages in skeletal muscle regeneration with particular reference to chemotaxis,” *Exp. Cell Res.*, vol. 207, pp. 321–321, 1993.

[149] D. M. Mosser and J. P. Edwards, “Exploring the full spectrum of macrophage activation.,” *Nat. Rev. Immunol.*, vol. 8, no. 12, pp. 958–69, Dec. 2008.

[150] S. A. Villalta, H. X. Nguyen, B. Deng, T. Gotoh, and J. G. Tidball, “Shifts in macrophage phenotypes and macrophage competition for arginine metabolism affect the severity of muscle pathology in muscular dystrophy.,” *Hum. Mol. Genet.*, vol. 18, no. 3, pp. 482–96, Feb. 2009.

[151] M. N. Bradley, L. Zhou, and S. T. Smale, “C/EBP-b Regulation in Lipopolysaccharide-Stimulated Macrophages,” *Society*, vol. 23, no. 14, pp. 4841–4858, 2003.

[152] A. S. Brack, M. J. Conboy, S. Roy, M. Lee, C. J. Kuo, C. Keller, and T. A. Rando, “Increased Wnt signaling during aging alters muscle stem cell fate and increases fibrosis.,” *Science*, vol. 317, no. 5839, pp. 807–10, Aug. 2007.

[153] A. Polesskaya, P. Seale, and M. A. Rudnicki, “Wnt signaling induces the myogenic specification of resident CD45+ adult stem cells during muscle regeneration.,” *Cell*, vol. 113, no. 7, pp. 841–52, Jun. 2003.

[154] V. Dumeaux, K. S. Olsen, G. Nuel, R. H. Paulssen, A.-L. Børresen-Dale, and E. Lund, “Deciphering normal blood gene expression variation--The NOWAC postgenome study.,” *PLoS Genet.*, vol. 6, no. 3, p. e1000873, Mar. 2010.

[155] T. Zeller, P. Wild, S. Szymczak, M. Rotival, A. Schillert, R. Castagne, S. Maouche, M. Germain, K. Lackner, H. Rossmann, M. Eleftheriadis, C. R. Sinning, R. B. Schnabel, E. Lubos, D. Mennerich, W. Rust, C. Perret, C. Proust, V. Nicaud, J. Loscalzo, N. Hübner, D. Tregouet, T. Münzel, A. Ziegler, L. Tiret, S. Blankenberg, and F. Cambien, “Genetics and beyond--the transcriptome of human monocytes and disease susceptibility.,” *PLoS One*, vol. 5, no. 5, p. e10693, Jan. 2010.

[156] S. Vasunilashorn, A. K. Coppin, K. V Patel, F. Lauretani, L. Ferrucci, S. Bandinelli, and J. M. Guralnik, “Use of the Short Physical Performance Battery Score to predict loss of

ability to walk 400 meters: analysis from the InCHIANTI study.," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 64, no. 2, pp. 223–9, Feb. 2009.

[157] B. E. Ainsworth, W. L. Haskell, A. S. Leon, D. R. Jacobs, H. J. Montoye, J. F. Sallis, and R. S. Paffenbarger, "Compendium of physical activities: classification of energy costs of human physical activities.," *Med. Sci. Sports Exerc.*, vol. 25, no. 1, pp. 71–80, Jan. 1993.

[158] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.," *Nat. Genet.*, vol. 34, no. 3, pp. 267–73, Jul. 2003.

[159] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15545–50, Oct. 2005.

[160] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nat. Genet.*, vol. 25, no. 1, pp. 25–9, May 2000.

[161] G. S. Lynch, J. a Faulkner, and S. V Brooks, "Force deficits and breakage rates after single lengthening contractions of single fast fibers from unconditioned and conditioned muscles of young and old rats.," *Am. J. Physiol. Cell Physiol.*, vol. 295, no. 1, pp. C249–56, Jul. 2008.

[162] Y.-W. Liu, H.-P. Tseng, L.-C. Chen, B.-K. Chen, and W.-C. Chang, "Functional cooperation of simian virus 40 promoter factor 1 and CCAAT/enhancer-binding protein beta and delta in lipopolysaccharide-induced gene activation of IL-10 in mouse macrophages.," *J. Immunol.*, vol. 171, no. 2, pp. 821–8, Jul. 2003.

[163] M. J. Gray, M. Poljakovic, D. Kepka-Lenhart, and S. M. Morris, "Induction of arginase I transcription by IL-4 requires a composite DNA response element for STAT6 and C/EBPbeta.," *Gene*, vol. 353, no. 1, pp. 98–106, Jun. 2005.

[164] V. a Fadok, D. L. Bratton, A. Konowal, P. W. Freed, J. Y. Westcott, and P. M. Henson, "Macrophages that have ingested apoptotic cells in vitro inhibit proinflammatory cytokine production through autocrine/paracrine mechanisms involving TGF-beta, PGE2, and PAF.," *J. Clin. Invest.*, vol. 101, no. 4, pp. 890–8, Feb. 1998.

[165] S. Kinoshita, "A member of the C/EBP family, NF-IL6 beta, forms a heterodimer and transcriptionally synergizes with NF-IL6," *Proc. …*, vol. 89, no. February, pp. 1473–1476, Feb. 1992.

[166] L. C. Jones, "Expression of C/EBPbeta from the C/ebpalpha gene locus is sufficient for normal hematopoiesis in vivo," *Blood*, vol. 99, no. 6, pp. 2032–2036, Mar. 2002.

[167] P.-H. Lai, W.-L. Wang, C.-Y. Ko, Y.-C. Lee, W.-M. Yang, T.-W. Shen, W.-C. Chang, and J.-M. Wang, "HDAC1/HDAC3 modulates PPARG2 transcription through the

sumoylated CEBPD in hepatic lipogenesis.," *Biochim. Biophys. Acta*, vol. 1783, no. 10, pp. 1803–14, Oct. 2008.

[168] D. Sag, D. Carling, and R. Stout, "Adenosine 5'-monophosphate-activated protein kinase promotes macrophage polarization to an anti-inflammatory functional phenotype," *J. Immunol.*, 2008.

[169] K. Kang, S. M. Reilly, V. Karabacak, M. R. Gangl, K. Fitzgerald, B. Hatano, and C.-H. Lee, "Adipocyte-derived Th2 cytokines and myeloid PPARdelta regulate macrophage polarization and insulin sensitivity.," *Cell Metab.*, vol. 7, no. 6, pp. 485–95, Jun. 2008.

[170] J. I. Odegaard, R. R. Ricardo-Gonzalez, A. Red Eagle, D. Vats, C. R. Morel, M. H. Goforth, V. Subramanian, L. Mukundan, A. W. Ferrante, and A. Chawla, "Alternative M2 activation of Kupffer cells by PPARdelta ameliorates obesity-induced insulin resistance.," *Cell Metab.*, vol. 7, no. 6, pp. 496–507, Jun. 2008.

[171] T. T. Schug and X. Li, "PPARdelta-mediated macrophage activation: a matter of fat.," *Dis. Model. Mech.*, vol. 2, no. 9–10, pp. 421–2, 2009.

[172] F. Borcherding, M. Nitschke, G. Hundorfean, J. Rupp, D. von Smolinski, K. Bieber, C. van Kooten, H. Lehnert, K. Fellermann, and J. Büning, "The CD40-CD40L pathway contributes to the proinflammatory function of intestinal epithelial cells in inflammatory bowel disease.," *Am. J. Pathol.*, vol. 176, no. 4, pp. 1816–27, Apr. 2010.

[173] L. Lindemann, M. Ebeling, N. a Kratochwil, J. R. Bunzow, D. K. Grandy, and M. C. Hoener, "Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors.," *Genomics*, vol. 85, no. 3, pp. 372–85, Mar. 2005.

[174] G. Talaber, K. Kvell, Z. Varecza, F. Boldizsar, S. M. Parnell, E. J. Jenkinson, G. Anderson, T. Berki, and J. E. Pongracz, "Wnt-4 protects thymic epithelial cells against dexamethasone-induced senescence.," *Rejuvenation Res.*, vol. 14, no. 3, pp. 241–8, Jun. 2011.

[175] T. Rantanen, "Muscle strength, disability and mortality.," *Scand. J. Med. Sci. Sports*, vol. 13, no. 1, pp. 3–8, Feb. 2003.

[176] D. Baylis, G. Ntani, M. H. Edwards, H. E. Syddall, D. B. Bartlett, E. M. Dennison, C. Martin-Ruiz, T. von Zglinicki, D. Kuh, J. M. Lord, A. Aihie Sayer, and C. Cooper, "Inflammation, Telomere Length, and Grip Strength: A 10-year Longitudinal Study.," *Calcif. Tissue Int.*, vol. 95, no. 1, pp. 54–63, May 2014.

[177] L. A. Schaap, S. M. F. Pluijm, D. J. H. Deeg, and M. Visser, "Inflammatory markers and loss of muscle mass (sarcopenia) and strength.," *Am. J. Med.*, vol. 119, no. 6, pp. 526.e9–17, Jun. 2006.

[178] H. P. Patel, N. Al-Shanti, L. C. Davies, S. J. Barton, M. D. Grounds, R. L. Tellam, C. E. Stewart, C. Cooper, and A. A. Sayer, "Lean Mass, Muscle Strength and Gene Expression in Community Dwelling Older Men: Findings from the Hertfordshire Sarcopenia Study (HSS).," *Calcif. Tissue Int.*, Jul. 2014.

[179] S. Melov, M. a Tarnopolsky, K. Beckman, K. Felkey, and A. Hubbard, "Resistance exercise reverses aging in human skeletal muscle.," *PLoS One*, vol. 2, no. 5, p. e465, Jan. 2007.

[180] J. von Maltzahn, N. C. Chang, C. F. Bentzinger, and M. A. Rudnicki, “Wnt signaling in myogenesis.,” *Trends Cell Biol.*, vol. 22, no. 11, pp. 602–9, Nov. 2012.

[181] V. A. Gupta and A. H. Beggs, “Kelch proteins: emerging roles in skeletal muscle development and diseases.,” *Skelet. Muscle*, vol. 4, p. 11, Jan. 2014.

[182] A. P. Sharples, N. Al-Shanti, D. C. Hughes, M. P. Lewis, and C. E. Stewart, “The role of insulin-like-growth factor binding protein 2 (IGFBP2) and phosphatase and tensin homologue (PTEN) in the regulation of myoblast differentiation and hypertrophy.,” *Growth Horm. IGF Res.*, vol. 23, no. 3, pp. 53–61, Jun. 2013.

[183] L. W. Harries, L. C. Pilling, L. D. G. Hernandez, R. Bradley-Smith, W. Henley, A. B. Singleton, J. M. Guralnik, S. Bandinelli, L. Ferrucci, and D. Melzer, “CCAAT-enhancer-binding protein-beta expression in vivo is associated with muscle strength.,” *Aging Cell*, vol. 11, no. 2, pp. 262–268, Apr. 2012.

[184] J. Blackwell, L. W. Harries, L. C. Pilling, L. Ferrucci, A. Jones, and D. Melzer, “Changes in CEBPB expression in circulating leukocytes following eccentric elbow-flexion exercise.,” *J. Physiol. Sci.*, Nov. 2014.

[185] W. B. Kannel, M. Feinleib, P. M. McNamara, R. J. Garrison, and W. P. Castelli, “An investigation of coronary heart disease in families. The Framingham offspring study.,” *Am. J. Epidemiol.*, vol. 110, no. 3, pp. 281–90, Sep. 1979.

[186] A. Hofman, C. M. van Duijn, O. H. Franco, M. A. Ikram, H. L. A. Janssen, C. C. W. Klaver, E. J. Kuipers, T. E. C. Nijsten, B. H. C. Stricker, H. Tiemeier, A. G. Uitterlinden, M. W. Vernooij, and J. C. M. Witteman, “The Rotterdam Study: 2012 objectives and design update.,” *Eur. J. Epidemiol.*, vol. 26, no. 8, pp. 657–86, Aug. 2011.

[187] H. Völzke, D. Alte, C. O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, C. Havemann, T. Ittermann, S. Schipf, R. Haring, S. E. Baumeister, H. Wallaschofski, M. Nauck, S. Frick, A. Arnold, M. Jünger, J. Mayerle, M. Kraft, M. M. Lerch, M. Dörr, T. Reffelmann, K. Empen, S. B. Felix, A. Obst, B. Koch, S. Gläser, R. Ewert, I. Fietze, T. Penzel, M. Dören, W. Rathmann, J. Haerting, M. Hannemann, J. Röpcke, U. Schminke, C. Jürgens, F. Tost, R. Rettig, J. A. Kors, S. Ungerer, K. Hegenscheid, J.-P. J. Kühn, J.-P. J. Kühn, N. Hosten, R. Puls, J. Henke, O. Gloger, A. Teumer, G. Homuth, U. Völker, C. Schwahn, B. Holtfreter, I. Polzer, T. Kohlmann, H. J. Grabe, D. Rosskopf, H. K. Kroemer, T. Kocher, R. Biffar, U. John, and W. Hoffmann, “Cohort profile: the study of health in Pomerania.,” *Int. J. Epidemiol.*, vol. 40, no. 2, pp. 294–307, Apr. 2011.

[188] B. W. J. H. Penninx, A. T. F. Beekman, J. H. Smit, F. G. Zitman, W. A. Nolen, P. Spinhoven, P. Cuijpers, P. J. De Jong, H. W. J. Van Marwijk, W. J. J. Assendelft, K. Van Der Meer, P. Verhaak, M. Wensing, R. De Graaf, W. J. Hoogendijk, J. Ormel, and R. Van Dyck, “The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods.,” *Int. J. Methods Psychiatr. Res.*, vol. 17, no. 3, pp. 121–40, Jan. 2008.

[189] HUGO Gene Nomenclature Committee, “HGNC.” [Online]. Available: http://www.genenames.org/.

[190] R Core Team, “R: A language and environment for statistical computing.” R Foundation for Statistical Computing, Vienna, Austria, 2014.

[191] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models using lme4,” *J. Stat. Softw.*, vol. submitted, p. 51, Jun. 2014.

[192] C. J. Willer, Y. Li, and G. R. Abecasis, “METAL: fast and efficient meta-analysis of genomewide association scans.,” *Bioinformatics*, vol. 26, no. 17, pp. 2190–1, Sep. 2010.

[193] B. Zhang, S. Kirov, and J. Snoddy, “WebGestalt: an integrated system for exploring gene sets in various biological contexts.,” *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W741–8, Jul. 2005.

[194] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013.,” *Nucleic Acids Res.*, vol. 41, no. Web Server issue, pp. W77–83, Jul. 2013.

[195] W. K. Mitchell, J. Williams, P. Atherton, M. Larvin, J. Lund, and M. Narici, “Sarcopenia, dynapenia, and the impact of advancing age on human skeletal muscle size and strength; a quantitative review.,” *Front. Physiol.*, vol. 3, p. 260, Jan. 2012.

[196] A. Siritantikorn, K. Johansson, K. Ahlen, R. Rinaldi, T. Suthiphongchai, P. Wilairat, and R. Morgenstern, “Protection of cells from oxidative stress by microsomal glutathione transferase 1.,” *Biochem. Biophys. Res. Commun.*, vol. 355, no. 2, pp. 592–6, Apr. 2007.

[197] D. Cortez, G. Glick, and S. J. Elledge, “Minichromosome maintenance proteins are direct targets of the ATM and ATR checkpoint kinases.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 27, pp. 10078–83, Jul. 2004.

[198] S. Spaziani, E. Imperlini, A. Mancini, M. Caterino, P. Buono, and S. Orrù, “Insulin-like growth factor 1 receptor signaling induced by supraphysiological doses of IGF-1 in human peripheral blood lymphocytes.,” *Proteomics*, vol. 14, no. 13–14, pp. 1623–9, Jul. 2014.

[199] X. Qiang, W.-L. Yang, R. Wu, M. Zhou, A. Jacob, W. Dong, M. Kuncewitch, Y. Ji, H. Yang, H. Wang, J. Fujita, J. Nicastro, G. F. Coppa, K. J. Tracey, and P. Wang, “Cold-inducible RNA-binding protein (CIRP) triggers inflammatory responses in hemorrhagic shock and sepsis.,” *Nat. Med.*, vol. 19, no. 11, pp. 1489–95, Nov. 2013.

[200] A. A. Khan and J. G. Quigley, “Control of intracellular heme levels: heme transporters and heme oxygenases.,” *Biochim. Biophys. Acta*, vol. 1813, no. 5, pp. 668–82, May 2011.

[201] C. Wu, C. Orozco, J. Boyer, M. Leglise, J. Goodale, S. Batalov, C. L. Hodge, J. Haase, J. Janes, J. W. Huss, and A. I. Su, “BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.,” *Genome Biol.*, vol. 10, no. 11, p. R130, Jan. 2009.

[202] K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O’Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy, and J. M. Ostell, “RefSeq: an update on mammalian reference sequences.,” *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D756–63, Jan. 2014.

[203] J. Thiery, D. Keefe, S. Boulant, E. Boucrot, M. Walch, D. Martinvalet, I. S. Goping, R. C. Bleackley, T. Kirchhausen, and J. Lieberman, “Perforin pores in the endosomal membrane trigger the release of endocytosed granzyme B into the cytosol of target cells.,” *Nat. Immunol.*, vol. 12, no. 8, pp. 770–7, Aug. 2011.

[204] L. C. Pilling, R. Joehanes, D. Melzer, L. W. Harries, W. Henley, J. Dupuis, H. Lin, M. Mitchell, D. Hernandez, S.-X. Ying, K. L. Lunetta, E. J. Benjamin, A. Singleton, D. Levy, P. Munson, J. M. Murabito, and L. Ferrucci, "Gene Expression markers of Age-Related Inflammation in Two Human Cohorts," *Exp. Gerontol.*, p. In submission, 2015.

[205] J. Han, P. Mistriotis, P. Lei, D. Wang, S. Liu, and S. T. Andreadis, "Nanog reverses the effects of organismal aging on mesenchymal stem cell proliferation and myogenic differentiation potential.," *Stem Cells*, vol. 30, no. 12, pp. 2746–59, Dec. 2012.

[206] J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, I. I. Slukvin, and J. A. Thomson, "Induced pluripotent stem cell lines derived from human somatic cells.," *Science*, vol. 318, no. 5858, pp. 1917–20, Dec. 2007.

[207] M. Cesari, B. W. J. H. Penninx, F. Lauretani, C. R. Russo, C. Carter, S. Bandinelli, H. Atkinson, G. Onder, M. Pahor, and L. Ferrucci, "Hemoglobin levels and skeletal muscle: results from the InCHIANTI study.," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 59, no. 3, pp. 249–54, Mar. 2004.

[208] C. Maraldi, A. Ble, G. Zuliani, J. M. Guralnik, C. Mussi, R. Fellin, and S. Volpato, "Association between anemia and physical disability in older patients: role of comorbidity.," *Aging Clin. Exp. Res.*, vol. 18, no. 6, pp. 485–92, Dec. 2006.

[209] I. J. Deary, A. F. Wright, S. E. Harris, L. J. Whalley, and J. M. Starr, "Searching for genetic influences on normal cognitive ageing.," *Trends Cogn. Sci.*, vol. 8, no. 4, pp. 178–84, Apr. 2004.

[210] T. F. Hughes and M. Ganguli, "Modifiable Midlife Risk Factors for Late-Life Cognitive Impairment and Dementia.," *Curr. Psychiatry Rev.*, vol. 5, no. 2, pp. 73–92, May 2009.

[211] P. B. Gorelick, "Role of inflammation in cognitive impairment: results of observational epidemiological studies and clinical trials.," *Ann. N. Y. Acad. Sci.*, vol. 1207, pp. 155–62, Oct. 2010.

[212] O. Philipson, A. Lord, A. Gumucio, P. O'Callaghan, L. Lannfelt, and L. N. G. Nilsson, "Animal models of amyloid-beta-related pathologies in Alzheimer's disease," *FEBS J.*, vol. 277, no. 6, pp. 1389–1409, 2010.

[213] L. Gu, Y. Okada, S. K. Clinton, C. Gerard, G. K. Sukhova, P. Libby, and B. J. Rollins, "Absence of monocyte chemoattractant protein-1 reduces atherosclerosis in low density lipoprotein receptor-deficient mice.," *Mol. Cell*, vol. 2, no. 2, pp. 275–81, Aug. 1998.

[214] J. El Khoury, M. Toft, S. E. Hickman, T. K. Means, K. Terada, C. Geula, and A. D. Luster, "Ccr2 deficiency impairs microglial accumulation and accelerates progression of Alzheimer-like disease.," *Nat. Med.*, vol. 13, no. 4, pp. 432–8, Apr. 2007.

[215] O. C. Maes, S. Xu, B. Yu, H. M. Chertkow, E. Wang, and H. M. Schipper, "Transcriptional profiling of Alzheimer blood mononuclear cells by microarray.," *Neurobiol. Aging*, vol. 28, no. 12, pp. 1795–809, Dec. 2007.

[216] E. Grünblatt, J. Bartl, S. Zehetmayer, T. M. Ringel, P. Bauer, P. Riederer, and C. P. Jacob, "Gene expression as peripheral biomarkers for sporadic Alzheimer's disease.," *J. Alzheimers. Dis.*, vol. 16, no. 3, pp. 627–34, Jan. 2009.

[217] C. A. de Jager, M. M. Budge, and R. Clarke, “Utility of TICS-M for the assessment of cognitive function in older adults.,” *Int. J. Geriatr. Psychiatry*, vol. 18, no. 4, pp. 318–24, Apr. 2003.

[218] G. M. Terrera, F. Matthews, and C. Brayne, “A comparison of parametric models for the investigation of the shape of cognitive change in the older population.,” *BMC Neurol.*, vol. 8, p. 16, Jan. 2008.

[219] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, “Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.,” *Nat. Genet.*, vol. 29, no. 4, pp. 365–71, Dec. 2001.

[220] A. J. Mitchell, “A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment.,” *J. Psychiatr. Res.*, vol. 43, no. 4, pp. 411–31, Jan. 2009.

[221] M. F. Folstein, S. E. Folstein, and P. R. McHugh, “‘Mini-mental state’. A practical method for grading the cognitive state of patients for the clinician.,” *J. Psychiatr. Res.*, vol. 12, no. 3, pp. 189–98, Nov. 1975.

[222] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 16, pp. 9440–5, Aug. 2003.

[223] K. J. Livak and T. D. Schmittgen, “Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method.,” *Methods*, vol. 25, no. 4, pp. 402–8, Dec. 2001.

[224] S. Wood, *Generalized Additive Models: An Introduction with R*. CRC Texts in Statistical Science, 2006.

[225] R. Ihl, L. Frölich, T. Dierks, E. M. Martin, and K. Maurer, “Differential validity of psychometric tests in dementia of the Alzheimer type.,” *Psychiatry Res.*, vol. 44, no. 2, pp. 93–106, Nov. 1992.

[226] F. E. Matthews, B. C. M. Stephan, I. G. McKeith, J. Bond, and C. Brayne, “Two-year progression from mild cognitive impairment to dementia: to what extent do different definitions agree?,” *J. Am. Geriatr. Soc.*, vol. 56, no. 8, pp. 1424–33, Aug. 2008.

[227] R. S. Wilson, S. E. Leurgans, P. A. Boyle, J. A. Schneider, and D. A. Bennett, “Neurodegenerative basis of age-related cognitive decline.,” *Neurology*, vol. 75, no. 12, pp. 1070–8, Sep. 2010.

[228] I. H. G. B. Ramakers, P. J. Visser, P. Aalten, O. Bekers, K. Sleegers, C. L. van Broeckhoven, J. Jolles, and F. R. J. Verhey, “The association between APOE genotype and memory dysfunction in subjects with mild cognitive impairment is related to age and Alzheimer pathology.,” *Dement. Geriatr. Cogn. Disord.*, vol. 26, no. 2, pp. 101–8, Jan. 2008.

[229] N. Terrando, C. Monaco, D. Ma, B. M. J. Foxwell, M. Feldmann, and M. Maze, “Tumor necrosis factor-alpha triggers a cytokine cascade yielding postoperative cognitive decline.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 47, pp. 20518–22, Nov. 2010.

[230] L.-F. Lue, Y.-M. Kuo, T. Beach, and D. G. Walker, “Microglia activation and anti-inflammatory regulation in Alzheimer’s disease.,” *Mol. Neurobiol.*, vol. 41, no. 2–3, pp. 115–28, Jun. 2010.

[231] G. A. Rosenberg, “Inflammation and white matter damage in vascular cognitive impairment.,” *Stroke.*, vol. 40, no. 3 Suppl, pp. S20–3, Mar. 2009.

[232] K. Rezai-Zadeh, D. Gate, C. A. Szekely, and T. Town, “Can peripheral leukocytes be used as Alzheimer’s disease biomarkers?,” *Expert Rev. Neurother.*, vol. 9, no. 11, pp. 1623–33, Nov. 2009.

[233] A. Cedazo-Minguez and B. Winblad, “Biomarkers for Alzheimer’s disease and other forms of dementia: clinical needs, limitations and future aspects.,” *Exp. Gerontol.*, vol. 45, no. 1, pp. 5–14, Jan. 2010.

[234] T. C. Dawson, W. A. Kuziel, T. A. Osahar, and N. Maeda, “Absence of CC chemokine receptor-2 reduces atherosclerosis in apolipoprotein E-deficient mice.,” *Atherosclerosis*, vol. 143, no. 1, pp. 205–11, Mar. 1999.

[235] P.-C. Liao, L. Yu, C.-C. Kuo, C. Lin, and Y.-M. Kuo, “Proteomics analysis of plasma for potential biomarkers in the diagnosis of Alzheimer’s disease.,” *Proteomics. Clin. Appl.*, vol. 1, no. 5, pp. 506–12, May 2007.

[236] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. F. Friedman, D. R. Galasko, M. Jutel, A. Karydas, J. A. Kaye, J. Leszek, B. L. Miller, L. Minthon, J. F. Quinn, G. D. Rabinovici, W. H. Robinson, M. N. Sabbagh, Y. T. So, D. L. Sparks, M. Tabaton, J. Tinklenberg, J. A. Yesavage, R. Tibshirani, and T. Wyss-Coray, “Classification and prediction of clinical Alzheimer’s diagnosis based on plasma signaling proteins.,” *Nat. Med.*, vol. 13, no. 11, pp. 1359–62, Nov. 2007.

[237] F. Licastro, S. Pedrini, L. Caputo, G. Annoni, L. J. Davis, C. Ferri, V. Casadei, and L. M. Grimaldi, “Increased plasma levels of interleukin-1, interleukin-6 and alpha-1-antichymotrypsin in patients with Alzheimer’s disease: peripheral inflammation or signals from the brain?,” *J. Neuroimmunol.*, vol. 103, no. 1, pp. 97–102, Feb. 2000.

[238] K. Ishizuka, T. Kimura, R. Igata-yi, S. Katsuragi, J. Takamatsu, and T. Miyakawa, “Identification of monocyte chemoattractant protein-1 in senile plaques and reactive microglia of Alzheimer’s disease.,” *Psychiatry Clin. Neurosci.*, vol. 51, no. 3, pp. 135–8, Jun. 1997.

[239] T. H. Bak and E. Mioshi, “A cognitive bedside assessment beyond the MMSE: the Addenbrooke’s Cognitive Examination.,” *Pract. Neurol.*, vol. 7, no. 4, pp. 245–9, Aug. 2007.

[240] F. Franco-Marina, J. J. García-González, F. Wagner-Echeagaray, J. Gallo, O. Ugalde, S. Sánchez-García, C. Espinel-Bermúdez, T. Juárez-Cedillo, M. A. V. Rodríguez, and C. García-Peña, “The Mini-mental State Examination revisited: ceiling and floor effects after score adjustment for educational level in an aging Mexican population.,” *Int. Psychogeriatr.*, vol. 22, no. 1, pp. 72–81, Feb. 2010.

[241] T. Singh and A. B. Newman, “Inflammatory markers in population studies of aging.,” *Ageing Res. Rev.*, vol. 10, no. 3, pp. 319–29, Jul. 2011.

[242] L. Ferrucci, A. Corsi, F. Lauretani, S. Bandinelli, B. Bartali, D. D. Taub, J. M. Guralnik, and D. L. Longo, “The origins of age-related proinflammatory state.,” *Blood*, vol. 105, no. 6, pp. 2294–9, Mar. 2005.

[243] L. Ferrucci, T. B. Harris, J. M. Guralnik, R. P. Tracy, M. C. Corti, H. J. Cohen, B. Penninx, M. Pahor, R. Wallace, and R. J. Havlik, “Serum IL-6 level and the development of disability in older persons.,” *J. Am. Geriatr. Soc.*, vol. 47, no. 6, pp. 639–46, Jun. 1999.

[244] P. M. Ridker, N. Rifai, M. J. Stampfer, and C. H. Hennekens, “Plasma concentration of interleukin-6 and the risk of future myocardial infarction among apparently healthy men.,” *Circulation*, vol. 101, no. 15, pp. 1767–72, Apr. 2000.

[245] K. Heikkilä, S. Ebrahim, and D. A. Lawlor, “Systematic review of the association between circulating interleukin-6 (IL-6) and cancer.,” *Eur. J. Cancer*, vol. 44, no. 7, pp. 937–45, May 2008.

[246] A. Economos, C. B. Wright, Y. P. Moon, T. Rundek, L. Rabbani, M. C. Paik, R. L. Sacco, and M. S. V Elkind, “Interleukin 6 plasma concentration associates with cognitive decline: the northern Manhattan study.,” *Neuroepidemiology*, vol. 40, no. 4, pp. 253–9, Jan. 2013.

[247] T. K. Howcroft, J. Campisi, G. B. Louis, M. T. Smith, B. Wise, T. Wyss-Coray, A. D. Augustine, J. E. McElhaney, R. Kohanski, and F. Sierra, “The role of inflammation in age-related disease.,” *Aging (Albany. NY).*, vol. 5, no. 1, pp. 84–93, Jan. 2013.

[248] A. R. J. Young and M. Narita, “SASP reflects senescence.,” *EMBO Rep.*, vol. 10, no. 3, pp. 228–30, Mar. 2009.

[249] A. Bektas, Y. Zhang, W. H. Wood, K. G. Becker, K. Madara, L. Ferrucci, and R. Sen, “Age-associated alterations in inducible gene transcription in human CD4+ T lymphocytes.,” *Aging (Albany. NY).*, vol. 5, no. 1, pp. 18–36, Jan. 2013.

[250] M. Feinleib, W. B. Kannel, R. J. Garrison, P. M. McNamara, and W. P. Castelli, “The Framingham Offspring Study. Design and preliminary data.,” *Prev. Med. (Baltim).*, vol. 4, no. 4, pp. 518–25, Dec. 1975.

[251] R. Joehanes, S. Ying, T. Huan, A. D. Johnson, N. Raghavachari, R. Wang, P. Liu, K. A. Woodhouse, S. K. Sen, K. Tanriverdi, P. Courchesne, J. E. Freedman, C. J. O’Donnell, D. Levy, and P. J. Munson, “Gene expression signatures of coronary heart disease.,” *Arterioscler. Thromb. Vasc. Biol.*, vol. 33, no. 6, pp. 1418–26, Jun. 2013.

[252] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, “Exploration, normalization, and summaries of high density oligonucleotide array probe level data.,” *Biostatistics*, vol. 4, no. 2, pp. 249–64, Apr. 2003.

[253] J. D. Fontes, J. F. Yamamoto, M. G. Larson, N. Wang, D. Dallmeier, M. Rienstra, R. B. Schnabel, R. S. Vasan, J. F. Keaney, and E. J. Benjamin, “Clinical correlates of change in inflammatory biomarkers: The Framingham Heart Study.,” *Atherosclerosis*, vol. 228, no. 1, pp. 217–23, May 2013.

[254] K. Imai, L. Keele, D. Tingley, and T. Yamamoto, “Causal Mediation Analysis Using R,” *Analysis*, pp. 129–154, 2010.

[255] C. E. Bonferroni, “Teoria statistica delle classi e calcolo delle probabilità,” *Pubbl. del R Ist. Super. di Sci. Econ. e Commer. di Firenze*, vol. 8, pp. 3–62, 1936.

[256] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.,” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Jan. 2009.

[257] Y. Liu, S. M. Johnson, Y. Fedoriw, A. B. Rogers, H. Yuan, J. Krishnamurthy, and N. E. Sharpless, “Expression of p16(INK4a) prevents cancer and promotes aging in lymphocytes.,” *Blood*, vol. 117, no. 12, pp. 3257–67, Mar. 2011.

[258] E. S. Didier, C. Sugimoto, L. C. Bowers, I. a Khan, and M. J. Kuroda, “Immune correlates of aging in outdoor-housed captive rhesus macaques (Macaca mulatta).,” *Immun. Ageing*, vol. 9, no. 1, p. 25, Nov. 2012.

[259] D. Rukavina, G. Laskarin, G. Rubesa, N. Strbo, I. Bedenicki, D. Manestar, M. Glavas, S. E. Christmas, and E. R. Podack, “Age-related decline of perforin expression in human cytotoxic T lymphocytes and natural killer cells.,” *Blood*, vol. 92, no. 7, pp. 2410–20, Oct. 1998.

[260] J. E. McElhaney, X. Zhou, H. K. Talbot, E. Soethout, R. C. Bleackley, D. J. Granville, and G. Pawelec, “The unmet need in the elderly: how immunosenescence, CMV infection, co-morbidities and frailty are a challenge for the development of more effective influenza vaccines.,” *Vaccine*, vol. 30, no. 12, pp. 2060–7, Mar. 2012.

[261] G. Del Giudice, B. Weinberger, and B. Grubeck-Loebenstein, “Vaccines for the Elderly.,” *Gerontology*, Nov. 2014.

[262] M. D. Parker and W. F. Boron, “The divergence, actions, roles, and relatives of sodium-coupled bicarbonate transporters.,” *Physiol. Rev.*, vol. 93, no. 2, pp. 803–959, Apr. 2013.

[263] K. Rajamäki, T. Nordström, K. Nurmi, K. E. O. Åkerman, P. T. Kovanen, K. Öörni, and K. K. Eklund, “Extracellular acidosis is a novel danger signal alerting innate immunity via the NLRP3 inflammasome.,” *J. Biol. Chem.*, vol. 288, no. 19, pp. 13410–9, May 2013.

[264] W. R. Farwell and E. N. Taylor, “Serum anion gap, bicarbonate and biomarkers of inflammation in healthy individuals in a national survey.,” *CMAJ*, vol. 182, no. 2, pp. 137–41, Feb. 2010.

[265] M. Rebsamen, L. Pochini, T. Stasyk, M. E. G. de Araújo, M. Galluccio, R. K. Kandasamy, B. Snijder, A. Fauster, E. L. Rudashevskaya, M. Bruckner, S. Scorzoni, P. A. Filipek, K. V. M. Huber, J. W. Bigenzahn, L. X. Heinz, C. Kraft, K. L. Bennett, C. Indiveri, L. A. Huber, and G. Superti-Furga, “SLC38A9 is a component of the lysosomal amino acid sensing machinery that controls mTORC1.,” *Nature*, 2015.

[266] A. Gheldof, P. Hulpiau, F. van Roy, B. De Craene, and G. Berx, “Evolutionary functional analysis and molecular regulation of the ZEB transcription factors.,” *Cell. Mol. Life Sci.*, vol. 69, no. 15, pp. 2527–41, Aug. 2012.

[267] A. Ramasamy, D. Trabzuni, J. R. Gibbs, A. Dillman, D. G. Hernandez, S. Arepalli, R. Walker, C. Smith, G. P. Ilori, A. A. Shabalin, Y. Li, A. B. Singleton, M. R. Cookson, J. Hardy, M. Ryten, and M. E. Weale, “Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies.,” *Nucleic Acids Res.*, vol. 41, no. 7, p. e88, Apr. 2013.

[268] N. Compté, B. Bailly, S. De Breucker, S. Goriely, and T. Pepersack, “Study of the association of total and differential white blood cell counts with geriatric conditions,

cardio-vascular diseases, low-grade inflammation and telomere length.," *Exp. Gerontol.*, Nov. 2014.

[269] J. J. A. Faulkner, S. V. S. Brooks, and J. A. J. Opiteck, "Injury to skeletal muscle fibers during contractions: conditions of occurrence and prevention," *Phys. Ther.*, vol. 73, no. 12, p. 911, 1993.

[270] S. a Studenski, K. W. Peters, D. E. Alley, P. M. Cawthon, R. R. McLean, T. B. Harris, L. Ferrucci, J. M. Guralnik, M. S. Fragala, A. M. Kenny, D. P. Kiel, S. B. Kritchevsky, M. D. Shardell, T.-T. L. Dam, and M. T. Vassileva, "The FNIH sarcopenia project: rationale, study description, conference recommendations, and final estimates.," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 69, no. 5, pp. 547–58, May 2014.

[271] O. Y. a Chan, A. H. van Houwelingen, J. Gussekloo, J. W. Blom, and W. P. J. den Elzen, "Comparison of quadriceps strength and handgrip strength in their association with health outcomes in older adults in primary care.," *Age (Dordr).*, vol. 36, no. 5, p. 9714, Oct. 2014.

[272] D. R. Clemmons, "Involvement of insulin-like growth factor-I in the control of glucose homeostasis.," *Curr. Opin. Pharmacol.*, vol. 6, no. 6, pp. 620–5, Dec. 2006.

[273] G. Naert and S. Rivest, "A deficiency in CCR2+ monocytes: the hidden side of Alzheimer's disease.," *J. Mol. Cell Biol.*, vol. 5, no. 5, pp. 284–93, Oct. 2013.

[274] P. Pezzotti, S. Scalmana, A. Mastromattei, and D. Di Lallo, "The accuracy of the MMSE in detecting cognitive impairment when administered by general practitioners: a prospective observational study.," *BMC Fam. Pract.*, vol. 9, p. 29, Jan. 2008.

[275] S. Jacobs, E. Ruusuvuori, S. T. Sipilä, A. Haapanen, H. H. Damkier, I. Kurth, M. Hentschke, M. Schweizer, Y. Rudhard, L. M. Laatikainen, J. Tyynelä, J. Praetorius, J. Voipio, and C. A. Hübner, "Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 1, pp. 311–6, Jan. 2008.

[276] C. A. Gurnett, R. Veile, J. Zempel, L. Blackburn, M. Lovett, and A. Bowcock, "Disruption of sodium bicarbonate transporter SLC4A10 in a patient with complex partial epilepsy and mental retardation.," *Arch. Neurol.*, vol. 65, no. 4, pp. 550–3, Apr. 2008.

[277] N. Kendrick, "A gene's mRNA level measured by microarrays or RT-PCR does not necessarily predict its protein level: R^2(or Rs) for plots of mRNA versus protein is <=0.4," Madison, WI, 2012.

[278] E. Lozada, J. Yi, J. Luo, and D. K. Orren, "Acetylation of Werner syndrome protein (WRN): relationships with DNA damage, DNA replication and DNA metabolic activities," *Biogerontology*, Jun. 2014.

[279] Z. Fewell, G. Davey Smith, and J. A. C. Sterne, "The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study.," *Am. J. Epidemiol.*, vol. 166, no. 6, pp. 646–55, Sep. 2007.

[280] D. Galasko, I. Abramson, J. Corey-Bloom, and L. J. Thal, "Repeated exposure to the Mini-Mental State Examination and the Information-Memory-Concentration Test results in a practice effect in Alzheimer's disease.," *Neurology*, vol. 43, no. 8, pp. 1559–63, Aug. 1993.

[281] K. Rockwood, P. Stolee, D. Robertson, and E. R. Shillington, “Response bias in a health status survey of elderly people.,” *Age Ageing*, vol. 18, no. 3, pp. 177–82, May 1989.

[282] J. E. Buring, *Epidemiology in Medicine*. Lippincott Williams & Wilkins, 1987.

[283] R. F. Nau, “Testing the assumptions of linear regression,” 2014. [Online]. Available: http://people.duke.edu/~rnau/testing.htm.

[284] P. Khatri, M. Sirota, and A. J. Butte, “Ten years of pathway analysis: current approaches and outstanding challenges.,” *PLoS Comput. Biol.*, vol. 8, no. 2, p. e1002375, Jan. 2012.

[285] R. Doll, “Proof of causality: deduction from epidemiological observation.,” *Perspect. Biol. Med.*, vol. 45, no. 4, pp. 499–515, 2002.

[286] ENCODE Project Consortium, “ENCODE: Encyclopedia of DNA Elements.” [Online]. Available: https://www.encodeproject.org/. [Accessed: 22-Apr-2015].

[287] P. A. C. ’t Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. A. M. Vossen, R. X. de Menezes, J. M. Boer, G.-J. B. van Ommen, and J. T. den Dunnen, “Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.,” *Nucleic Acids Res.*, vol. 36, no. 21, p. e141, Dec. 2008.

[288] V. Chai, A. Vassilakos, Y. Lee, J. A. Wright, and A. H. Young, “Optimization of the PAXgene blood RNA extraction system for gene expression analysis of clinical samples.,” *J. Clin. Lab. Anal.*, vol. 19, no. 5, pp. 182–8, Jan. 2005.

[289] M. Reimers, “Normalizing Expression Arrays.” [Online]. Available: http://www.people.vcu.edu/~mreimers/OGMDA/normalize.expression.html.

[290] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, “A gene atlas of the mouse and human protein-encoding transcriptomes.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 16, pp. 6062–7, Apr. 2004.

[291] J. P. A. Ioannidis, “Why most published research findings are false.,” *PLoS Med.*, vol. 2, no. 8, p. e124, Aug. 2005.

[292] F. Prinz, T. Schlange, and K. Asadullah, “Believe it or not: how much can we rely on published data on potential drug targets?,” *Nat. Rev. Drug Discov.*, vol. 10, no. 9, p. 712, Sep. 2011.

[293] A. C. Holly, L. C. Pilling, D. Hernandez, B. P. Lee, A. Singleton, L. Ferrucci, D. Melzer, and L. W. Harries, “Splicing Factor 3B1 hypomethylation is associated with altered SF3B1 transcript expression in older humans.,” *Mech. Ageing Dev.*, Jan. 2014.

[294] A. C. Holly, S. Grellscheid, P. van de Walle, D. Dolan, L. C. Pilling, D. J. Daniels, T. von Zglinicki, L. Ferrucci, D. Melzer, and L. W. Harries, “Comparison of senescence-associated miRNAs in primary skin and lung fibroblasts,” *Biogerontology*, 2015.

[295] L. W. Harries, A. D. Fellows, L. C. Pilling, D. Hernandez, A. Singleton, S. Bandinelli, J. Guralnik, J. Powell, L. Ferruci, and D. Melzer, “Advancing age is associated with gene expression changes resembling mTOR inhibition: evidence from two human populations.,” *Mech. Ageing Dev.*, vol. 133, no. 8, pp. 556–62, Jul. 2012.

[296] A. D. Fellows, A. C. Holly, L. C. Pilling, D. Melzer, and L. W. Harries, “Age related changes in mTOR-related gene expression in two primary human cell lines,” *Heal. Aging Res.*, pp. 1–8, Dec. 2012.

[297] K. Hek, A. Demirkan, J. Lahti, A. Terracciano, A. Teumer, M. C. Cornelis, N. Amin, E. Bakshis, J. Baumert, J. Ding, Y. Liu, K. Marciante, O. Meirelles, M. a. Nalls, Y. V. Sun, N. Vogelzangs, L. Yu, S. Bandinelli, E. J. Benjamin, D. a. Bennett, D. Boomsma, A. Cannas, L. H. Coker, E. de Geus, P. L. De Jager, A. V. Diez-Roux, S. Purcell, F. B. Hu, E. B. Rimm, D. J. Hunter, M. K. Jensen, G. Curhan, K. Rice, A. D. Penman, J. I. Rotter, N. Sotoodehnia, R. Emeny, J. G. Eriksson, D. a. Evans, L. Ferrucci, M. Fornage, V. Gudnason, A. Hofman, T. Illig, S. Kardia, M. Kelly-Hayes, K. Koenen, P. Kraft, M. Kuningas, J. M. Massaro, D. Melzer, A. Mulas, C. L. Mulder, A. Murray, B. a. Oostra, A. Palotie, B. Penninx, A. Petersmann, L. C. Pilling, B. Psaty, R. Rawal, E. M. Reiman, A. Schulz, J. M. Shulman, A. B. Singleton, A. V. Smith, A. R. Sutin, A. G. Uitterlinden, H. Völzke, E. Widen, K. Yaffe, A. B. Zonderman, F. Cucca, T. Harris, K.-H. Ladwig, D. J. Llewellyn, K. Räikkönen, T. Tanaka, C. M. van Duijn, H. J. Grabe, L. J. Launer, K. L. Lunetta, T. H. Mosley, A. B. Newman, H. Tiemeier, and J. Murabito, “A genome-wide association study of depressive symptoms,” *Biol. Psychiatry*, vol. 73, no. 7, pp. 667–78, Apr. 2013.

[298] T. Huan, T. Esko, M. J. Peters, L. C. Pilling, K. Schramm, C. Schurmann, B. H. Chen, C. Liu, R. Joehanes, A. D. Johnson, C. Yao, S. Ying, P. Courchesne, L. Milani, N. Raghavachari, R. Wang, P. Liu, E. Reinmaa, A. Dehghan, A. Hofman, A. G. Uitterlinden, D. G. Hernandez, S. Bandinelli, A. Singleton, D. Melzer, A. Metspalu, M. Carstensen, H. Grallert, C. Herder, T. Meitinger, A. Peters, M. Roden, M. Waldenberger, M. Dörr, S. B. Felix, T. Zeller, R. Vasan, C. J. O’Donnell, P. J. Munson, X. Yang, H. Prokisch, U. Völker, J. B. J. van Meurs, L. Ferrucci, and D. Levy, “A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension,” *PLOS Genet.*, vol. 11, no. 3, p. e1005035, 2015.

[299] C. A. Ibrahim-verbaas, J. Bressler, S. Debette, M. Schuur, A. V Smith, J. C. Bis, G. Davies, S. Trompet, and J. A. Smith, “GWAS for executive function and processing speed suggests involvement of the CADM2 gene,” *Mol. Psychiatry*, pp. 1–9, 2015.

[300] T. Fletcher, T. S. Galloway, D. Melzer, P. Holcroft, R. Cipelli, L. C. Pilling, D. Mondal, M. Luster, and L. W. Harries, “Associations between PFOA, PFOS and changes in the expression of genes involved in cholesterol metabolism in humans.,” *Environ. Int.*, vol. 57–58C, pp. 2–10, Apr. 2013.

[301] T. S. Galloway, T. Fletcher, O. J. Thomas, B. P. Lee, L. C. Pilling, and L. W. Harries, “PFOA and PFOS are associated with reduced expression of the parathyroid hormone 2 receptor (PTH2R) gene in women,” *Chemosphere*, vol. 120, pp. 555–562, Feb. 2015.

[302] D. Melzer, L. C. Pilling, A. D. Fellows, A. C. Holly, L. W. Harries, and L. Ferrucci, “Gene Expression Biomarkers and Longevity,” in *Annual Review of Gerontology and Geriatrics*, vol. 33, no. 1, Ingenta Connect, 2013, pp. 233–258.

[303] B. K. Winiarski, N. Cope, M. Alexander, L. C. Pilling, S. Warren, N. Acheson, N. J. Gutowski, and J. L. Whatmore, “Clinical Relevance of Increased Endothelial and Mesothelial Expression of Proangiogenic Proteases and VEGFA in the Omentum of Patients with Metastatic Ovarian High-Grade Serous Carcinoma,” *Transl. Oncol.*, vol. 7, no. 2, pp. 267–276.e4, Apr. 2014.