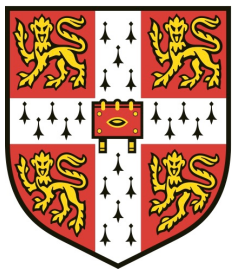
Computational studies on ageing and age-related diseases



Handan Melike Dönertaş

European Bioinformatics Institute (EMBL-EBI)
University of Cambridge

This dissertation is submitted for the degree of
Philosophy of Science

Darwin College

April 2020

Anne ve anneanneme, her zaman yanımdaydılar ve bugün bu tezi yazabiliyor olmayı onlara borçluyum…

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration with others, except when specified in the declarations at the beginning of the chapters. This work is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices and has fewer than 150 figures.

Handan Melike Dönertaş
April 2020

Acknowledgements

I am indebted to many people who made this thesis possible. First and foremost, I would like to express my gratitude to my supervisor, Janet Thornton. I am grateful for her time, guidance, discussions and support, which were indispensable for this thesis and the early steps of my journey to become a scientist. I was also very lucky to be supervised by Janet, as at a very early stage of my career, I had the chance to observe first hand that it is possible to become a great scientist without compromising values and virtues.

I am deeply thankful to Linda Partridge, whose input was vital for the development of this thesis. I also want to thank my thesis advisory committee, Susan Ozanne, Pedro Beltrao, and Wolfgang Huber, for their suggestions and guidance.

I want to thank all members of the spice family for creating such a friendly and supportive environment. My special thanks to Daniel Fabian and Matias Fuentealba, who were not only great colleagues but also comrades in my journey. During my PhD, I had the chance to work closely with visitors in our group. Many thanks, Veronika Kedlian, Lori Kregar, Victor Laigle, and Gozde Turan! I also want to acknowledge the contributions of Ulas Isildak, another great colleague we worked together from distance. It was a privilege to work with all of you! Our discussions were incredibly helpful and your contributions were fundamental for my PhD.

I also want to thank all members of the societies I have been a part of, ISCBSC RSG-Turkey, epiSTEM Turkiye, and EkoEvo. They gave me invaluable chances to contribute to the scientific environment in my home country, even though I was far away. I also want to thank ISCB Student Council for sparing a place for me in their organization and sharing my enthusiasm to communicate research to everyone for free.

A huge thanks to all PhD students at the EBI. Also, I want to thank all my Turkish-speaking friends in Cambridge, especially to Poorya Parvizi and Ozge Gizlenci. My special thanks to Betul Taskoparan, Hamit Izgi, Buse Isbilir, Ekin Saglican, and Gulsah Kilinc, who keep proving that geographic distance has no effect on true friendship. I

have remained sane(-ish) thanks to all of your support and friendship. I also want to thank Mehmet Somel, who motivated me years ago to pursue research on ageing and remains as a trusted mentor.

I want to express my gratitude to the essential components which made my journey to here possible in the first place. I must acknowledge the EMBL-EBI for funding the studies of overseas students, a rare example of true inclusiveness. The only reason I could imagine to come here to pursue my PhD at the EBI and University of Cambridge is the public school system I studied in, which allowed free and high-quality education for everybody, starting from elementary school to post-graduate degrees. I hope that creating equal opportunities for everyone, regardless of the economic background, could also be included in the discussion of diversity in academia.

Last but certainly not least, I want to thank my mother and grandmother for their continuous love and encouragement. This thesis would not exist without their support at every stage of my life.

H. Melike Donertas
April 2020

Abstract

Age is the major risk factor for a variety of non-communicable diseases. As life expectancy increases, ageing poses significant challenges to the individual, society, and healthcare systems. Ageing is a complex process involving multiple interconnected cellular and organismal phenotypes. Thus, understanding the molecular mechanisms and finding potential interventions is challenging and requires systems-level approaches. In this PhD I have addressed three main questions about ageing, using high-throughput data and computational methods.

My first study considers interindividual heterogeneity in gene expression during ageing. Previous studies had suggested that phenotype, epigenome and gene expression become more heterogeneous with age. However, the list of genes and pathways reported as heterogeneous in late age showed differences in the literature and did not resolve whether the increase in heterogeneity is a time-dependent process starting at birth or is restricted to the ageing period (*i.e.* after 20 years of age). Using different data pre-processing steps and heterogeneity measures on the same transcriptome dataset, we have shown that the inconsistency in the literature could reflect technical issues as well as biological variability. Next, applying a meta-analysis scheme that relies on consistent results across multiple datasets to increase reproducibility, we have shown that the increase in inter-individual heterogeneity starts after the age of 20. Moreover, the genes that become more heterogeneous during ageing have a higher number of transcriptional regulators (miRNAs and transcription factors) and are associated with known longevity pathways.

My second study focuses on the link between ageing and age-related diseases. Many diseases show age-dependency, but the molecular nature of this relationship is not fully understood. Using UK Biobank data, I have characterised 116 common diseases based on their age-of-onset profiles and genetic associations. I first showed diseases following the same age-of-onset distribution are genetically more similar, and this similarity could not be explained by disease categories, co-occurrences, or causal relationships. Two groups of diseases showed age-dependent profiles, starting to become

more prevalent after the ages of 20 and 40 respectively. They both showed an association with known ageing-related genes but had different functional and evolutionary profiles. I found support for the two evolutionary genetic theories of ageing, mutation accumulation, and antagonistic pleiotropy, using the variants linked to diseases with different age-of-onsets. I also identified some drugs that could be repurposed to target multiple conditions and potentially decrease the need for polypharmacy in the elderly.

Finally, I followed a systems-level approach to identify drugs that can target ageing in the human brain. Using transcriptome datasets from multiple brain regions, I first identified the gene expression changes that can characterise ageing. Then, compared with the drug-perturbed gene expression profiles in the Connectivity Map, I identified 24 drugs that are significantly associated with the ageing signature. Some of these drugs may function as anti-ageing drugs by reversing the detrimental changes that occur during ageing, others by mimicking the cellular defence mechanisms. The drugs that we identified included a significant number of already identified pro-longevity drugs, indicating that the method can discover de novo drugs that ameliorate ageing.

Table of Contents

List of tables

List of figures

Nomenclature

Abbreviations

AP Antagonistic Pleiotropy Theory of Aging

ARD Age-Related Disease

BY Benjamini-Yekutieli

CORT Temporal Correlation Coefficient

DR Dietary Restriction

eQTL Expression Quantitative Trait Loci

FDA U.S. Food and Drug Administration

FDR False Discovery Rate

GCP Genetic Causality Proportion

GSEA Gene Set Enrichment Analysis

GWAS Genome-Wide Association Study

HES Hospital Episode Statistics

ICD International Classification of Diseases

IQR Interquartile Range

ITP NIA Interventions Testing Program

KS Kolmogorov–Smirnov

LCV Latent Causal Variable

LD Linkage Disequilibrium

MA Mutation Accumulation Theory of Aging

MAF Minor Allele Frequency

MHC Major Histocompatibility Complex

NES Normalised Enrichment Score

NSN Nutrient-Sensing Network

PAM Partitioning Around Medoids

PCA Principal Component Analysis

ROS Reactive Oxygen Species

SASP Senescence-Associated Secretory Phenotype

SNP Single Nucleotide Polymorphisms

SR Self-Reported

SVA Surrogate Variable Analysis

TF Transcription Factor

Databases and Tools

CMap Connectivity Map

DO Disease Ontology

GO BP Gene Ontology Biological Process

GO CC Gene Ontology Cellular Component

GO MF Gene Ontology Molecular Function

GTEx Genotype-Tissue Expression

UKBB UK Biobank

Chapter 1

Introduction

Human life expectancy has been increasing steadily worldwide (Oeppen & Vaupel, 2002) and is currently around 80 years of age in developed countries (Max Roser & Ritchie, 2020). Improvements in the healthcare system, immunisation against infectious diseases, antibiotics, better sanitary measures, housing and lifestyle, resulted in a decline in early life mortality (Waite, 2004). There has also been a significant improvement in late-life mortality, mostly thanks to economic and social developments and ongoing medical advances (Vaupel, Carey, & Christensen, 2003).

The scientific community has long been discussing whether there is a limit to lifespan (Barbi, Lagona, Marsili, Vaupel, & Wachter, 2018; Dong, Milholland, & Vijg, 2016; Oeppen & Vaupel, 2002; Olshansky, 2016; Olshansky, Carnes, & Cassel, 1990; Olshansky, Carnes, & Désesquelles, 2001). For example, using conditional probabilities integrating disease data, in 1990 and 2001, Olshansky et al. had suggested the required changes in disease rates are too much to achieve 35 years of additional life expectancy at the age of 50, and thus it is highly unlikely to achieve (Olshansky et al., 1990, 2001). However, female life expectancy at the age of 50 in Japan surpassed this limit (Oeppen & Vaupel, 2002). Dong et al. (2016) used global demographic data and suggested that the survival rates at later ages slow down and maximum lifespan in humans is fixed. A recent study, on the other hand, suggested the mortality rate decreases and essentially reaches a plateau after the age of 105 (Barbi et al., 2018). Similarly, Kontis et al. (2017) followed a probabilistic Bayesian model averaging (BMA) approach that considers an ensemble of 21 forecasting models and predicted a continued increase in longevity.

The increase in health-span, *i.e.*, healthy years of life, increases at a slower rate and 16 to 20% of life is spent with multiple diseases (Jagger et al., 2008). Moreover, the difference between lifespan and health-span is higher in females, and in individuals with a lower socioeconomic status. The force of natural selection declines with age, and thus variants with deleterious effects in late-life accumulate and can cause diseases in the elderly (Niccoli & Partridge, 2012). Compression of the time spent ill at the end of life by targeting ageing is one of the trends in the ageing research as it becomes more and more clear that it is achievable (Partridge, Deelen, & Slagboom, 2018). There are some 'blue zones' in the world; Okinawa in Japan, Sardinia in Italy, Ikaria in Greece, Nicoya in Costa Rica, and Loma Linda in the US (Poulain, Herm, & Pes, 2013). People living in these regions enjoy increased lifespan and health-span. We have only recently started to live long periods, and thus genetics is not expected to have a significant role in these populations (Partridge et al., 2018). Indeed, these populations are not genetically different from their neighbours, but their lifestyle and social networks can explain the difference (Poulain et al., 2013). Diet, education, and physical activity contributes to a healthier life in human populations (Crimmins, 2015). Another line of evidence for the malleability of life- and health-span comes from the model organism studies. Multiple genetic, environmental, and pharmacological interventions are shown to both increase lifespan and decrease the functional decline in late life (Fontana, Partridge, & Longo, 2010).

Nevertheless, application of the accumulated knowledge to humans to improve health-span requires a better understanding of the molecular biology of ageing, age-related changes, and age-related diseases. In this thesis, I aim to approach this aim by using computational studies.

1.1 What is ageing?

Ageing is a phenomenon that everybody is familiar with through either observation or experience, yet there is no clear definition. While it could be used to only mean an increase in chronological age; to most of the researchers working in the field, ageing is a biological process with specific characteristics. Throughout this thesis, I will also make a distinction between the changes with chronological age and ageing. I will consider ageing as a process that starts after the average age at first reproduction, and the time before then as postnatal development. The reasoning comes from the

evolutionary theories stating that the force of natural selection is the highest before the reproductive period (see Section 1.2).

In the literature, ageing is often defined as the time-dependent functional decline, increased susceptibility to pathologies, and an increased risk of death (López-Otín, Blasco, Partridge, Serrano, & Kroemer, 2013). While I strongly agree with this definition, I want to state that ageing-related changes observed through high-throughput experiments are not necessarily always detrimental. Instead, they may include neutral or beneficial changes as a response to harmful effects seen during ageing (Kowald & Kirkwood, 2016). Unfortunately, the characterisation of these changes is still lacking. Thus, even though, in principle, I would consider only the processes that result in functional decline as contributors of ageing, in practice, all age-related changes after the average age at reproduction will be considered as ageing-related changes in the upcoming chapters. The challenges and results of this ambiguity are further discussed in the relevant chapters.

1.2 Evolution of ageing

If ageing results in a decline in organismal fitness and an increase in the risk of mortality, how could it survive through evolution, and how can it be almost ubiquitous, seen in most animals? There have been more than 300 theories of ageing (Medvedev, 1990), and many attempted to answer this question. Overall, we can classify these theories into two as programmed and non-programmed theories of ageing. Today, the mainstream opinion favours the non-programmed theories (Austad, 2004); however, I will briefly describe both categories.

1.2.1 Programmed theories of ageing

The programmed theories of ageing suggest that functional decline in post-reproductive age and shorter lifespan confers evolutionary benefit and thus is evolved under selection. This could be driven to prevent post-reproductive individuals from competing for resources or to increase genetic diversity and the speed of evolution by increasing the number of generations. There are three main counter-arguments against programmed theories, and these are presented and reviewed multiple times by Tom Kirkwood and his colleagues (Kirkwood, 1977, 2005; Kirkwood & Melov, 2011; Kowald &

Kirkwood, 2016). The first counter-argument is that although ageing is almost universal for many species in captivity (Flatt & Partridge, 2018), it is rarely observed in the wild. In natural conditions, mortality is generally due to extrinsic factors (*e.g.* predation and pathogens), and rarely caused by ageing-related mortality. Thus, ageing-related mortality cannot be under selection. This argument has been challenged by recent field studies documenting that multiple species experience ageing in the wild (Brunet-Rossinni & Austad, 2005; Nussey, Froy, Lemaitre, Gaillard, & Austad, 2013). The second counter-argument is that if ageing were programmed, there would be a limited number of genes controlling ageing. This argument was also challenged by the discovery of genes and pathways that modulate lifespan. However, these genes only postpone the ageing process and do not eliminate it. Last and probably, the most persuasive argument is that these theories are based on group selection, which is much weaker than selection on individuals, except for a few special conditions (Smith, 1976). If there were factors that accelerate ageing because of its benefits for the species, they would not be selected as their disadvantageous effects on individuals would override the selection at the group level.

1.2.2 Non-programmed theories of ageing

Non-programmed theories suggest that ageing is non-adaptive, and is a result of either decreasing natural selection or is a side-effect of the selective pressure on other traits (Flatt & Partridge, 2018). The force of natural selection at different ages can be described as a function of reproduction and extrinsic mortality (Kirkwood & Melov, 2011). As age increases, especially after the age of reproduction, the force of natural selection is predicted to decline because of extrinsic hazards. There are two main evolutionary genetic theories of ageing that are based on population genetics principles, *i.e.* mutation accumulation and antagonistic pleiotropy theories of ageing. Mutation accumulation theory of ageing (MA) predicts an intergenerational accumulation of deleterious variants in the population if they are only expressed (*i.e.* functional) at a late age (Medawar, 1953). Antagonistic pleiotropy theory of ageing (AP) focuses on pleiotropic variants that have opposite effects on multiple traits. If the variant is beneficial for a trait that is important when the force of natural selection is strong, *e.g.* survival or reproductive success, these variants would be favoured to be passed on the future generations despite their detrimental effects in late age (Williams, 1957). A modified version of AP that is expressed in terms of physiological trade-offs is the disposable soma theory. It is based on the principle that cellular resources are limited

and are distributed between growth, reproduction, and maintenance. Thus, greater investment in growth and reproduction would result in reduced investment in maintenance, which would result in the accumulation of cellular damages (Kirkwood, 1977).

The discovery of genes and pathways that regulate lifespan in model organisms had challenged the non-adaptive theories of ageing (Johnson, 2002). However, the fact that these genes are mostly in pathways that sense the environmental resources (*e.g.* nutrient-sensing network (NSN), mTOR / PI3K pathway) or regulate genetic maintenance (*e.g.* DNA repair genes) supports the notion that these genes do not control the ageing process, but instead regulate the resource allocation (Kirkwood & Melov, 2011). Moreover, none of these genes causes complete abolition of ageing, but they rather postpone the process, which supports the predictions of non-programmed theories.

1.3 The hallmarks of ageing

The first signals that genetics may play a role in modulation of lifespan started to emerge around 40 years ago when the first long-lived strains of *Caenorhabditis elegans* were isolated (Klass, 1983). The first genetic element discovered to modulate lifespan was *age-1*. However, the impact of this gene was controversial as it also affects total food intake, possibly causing caloric restriction. Later, Friedman and Johnson showed 65% lifespan extension was due to the mutation itself, not a reduction in the caloric intake (Friedman & Johnson, 1988). Ten years later, Kenyon et al. showed *daf-2* mutant worms also have potential to double their lifespans, and more importantly, without a cost on their reproductive success (Kenyon, Chang, Gensch, Rudner, & Tabtiang, 1993). *daf-2* also codes for a protein in the NSN, a homologue of IGF-1 receptor in humans. Many other discoveries followed these and scientists found different pathways regulating lifespan in model organisms. López-Otín et al. (2013), later reviewed these genetic elements together with the changes that occur in physiological ageing in mammals to compile a set of characteristics that define the ageing phenotype. Authors considered three criteria to consider these characteristics as hallmarks: i) they should manifest themselves in physiological ageing, ii) their experimental exacerbation should worsen the ageing phenotype, and iii) their reduction should improve lifespan or at least delay functional decline. Not all of the molecular changes can be tested extensively in multiple organisms, and some of the hallmarks do not satisfy all three. Another layer of complexity comes from the interconnected-

ness between hallmarks as they are not independent. Nevertheless, the list of nine hallmarks presents the functional characteristics associated with ageing.

1.3.1 Primary hallmarks

These are the hallmarks that explain likely causes of cellular damage. These have adverse effects on the functionality, and their reduction would ameliorate the ageing process.

Genomic instability: DNA integrity and stability are threatened by various intrinsic (*e.g.* DNA replication errors, reactive oxygen species (ROS)), or extrinsic (*e.g.* UV radiation, cancerogenic factors) factors throughout life (Hoeijmakers, 2009). These result in point mutations, translocations, chromosomal abnormalities like aneuploidy that can affect the structure or regulation of essential genes and result in a decrease in the functional capacity of cells (Faggioli, Wang, Vijg, & Montagna, 2012; Forsberg et al., 2012; García-Nieto, Morrison, & Fraser, 2019). Genes functioning in the DNA repair mechanisms have been found important for physiological ageing (Gorbunova, Seluanov, Mao, & Hine, 2007) as well as accelerated ageing syndromes (Kubben & Misteli, 2017).

Telomere attrition: While the accumulation of somatic damage throughout the genome seems to be random, telomeres are the primary sites being affected during ageing (Blackburn, Greider, & Szostak, 2006). DNA polymerase cannot replicate the ends of linear DNA molecules and, in the absence of telomerase, the telomere region gets shortened in each replication (Shay, 2018). Most mammalian species lack telomerase in somatic cells, and thus telomere attrition is one of the characteristics of ageing. This also limits the number of possible replications and leads to replicative senescence (Hayflick & Moorhead, 1961; Olovnikov, 1996). Shorter telomeres are associated with shorter lifespan (Boonekamp, Simons, Hemerik, & Verhulst, 2013), and induced telomerase activity rescues this phenotype (Jaskelioff et al., 2011).

Epigenetic alterations: Just like somatic genetic mutations, epi-mutations can result in disruption of cellular activities (Talens et al., 2012). Characteristic changes in the epigenome involve global DNA hypomethylation, region-specific DNA hypermethylation, and histone modifications (Pal & Tyler, 2016). These changes are also used to develop the epigenetic ageing clock, the most accurate age-predictor today (see Section 1.4.3) (Horvath & Raj, 2018). The epigenome is vital for the regulation of gene expression and disruptions can result in a decrease in functional capacity. While ge-

netic mutations are mostly irreversible, understanding the epigenome-wide changes can offer ways to intervene in ageing-related phenotypes (Freije & López-Otín, 2012; Rando & Chang, 2012).

Loss of proteostasis: Proteostasis, or protein homeostasis, is the sum of all activities to ensure the stability and functionality of the proteome. It involves many regulators, including molecular chaperones, protein posttranslational modification enzymes, proteasome, and lysosome (Hipp, Kasturi, & Hartl, 2019). Accumulation of nonfunctional and misfolded proteins is a characteristic of physiological ageing and age-related diseases, especially neurodegenerative diseases such as Alzheimer's and Parkinson's (Klaips, Jayaraj, & Hartl, 2018). Improvement of proteostasis through experimental interventions delayed ageing in mammals (Zhang & Cuervo, 2008).

1.3.2 Antagonistic hallmarks

This category involves hallmarks that might have positive or negative effects depending on the intensity. They generally occur as a response to the primary hallmarks, and when their amplitude or duration is excess, they result in pathological consequences.

Deregulated nutrient sensing: Genes and proteins functioning in the nutrient-sensing network (NSN) have been shown to modulate life-, and health-span in model organisms (Fontana & Partridge, 2015). In general, anabolism accelerates the ageing process and decreases lifespan (Fontana et al., 2010). In line with this observation, dietary restriction improves lifespan in a wide range of organisms (see Section 1.6) (Colman et al., 2014; Mattison et al., 2012).

Mitochondrial dysfunction: Ageing-related accumulation of ROS, and mutations in the mitochondrial DNA result in the disruption of the energy metabolism and decreased ATP availability (Green, Galluzzi, & Kroemer, 2011; Kauppila, Kauppila, & Larsson, 2017). However, a limited amount of ROS was observed to trigger mitohormesis, a process that prevents further damage. Thus, while a small amount of ROS can activate protective mechanisms, the excess can disrupt energy metabolism (Balaban, Nemoto, & Finkel, 2005).

Cellular senescence: Broadly, cellular senescence can be defined as cell cycle arrest upon various triggers, such as telomere attrition and oncogene activation (Campisi & Fagagna, 2007). Cellular senescence prevents the accumulation of damaged cells by preventing further divisions. However, the senescent cells also secrete vari-

ous proinflammatory molecules (senescence-associated secretory phenotype, SASP), that, in turn, cause neighbouring cells to become senescent (Rodier & Campisi, 2011). Because it prevents nonfunctional cells from dividing and accumulating, inducing cellular senescence in modest levels may improve lifespan (Matheu et al., 2009). However, eliminating the senescent cells that accumulate with age also extends lifespan (Childs et al., 2017).

1.3.3 Integrative hallmarks

The integrative hallmarks are the results of the previously mentioned hallmarks. They induce system-level changes and result in an overall ageing phenotype that causes a decline in the organism's functional capacity.

Stem cell exhaustion: Decline in tissues' proliferative capacity is one of the well-known characteristics of ageing (Goodell & Rando, 2015). As a result of the multi-layered accumulation of damage, tissues lose their regenerative capacity (Janzen et al., 2006; Rossi et al., 2007). Improving the division capacity of the stem cells may improve lifespan at the organismal level (Rando & Chang, 2012).

Altered intracellular communication: Apart from cell-intrinsic changes, hormonal and non-hormonal intercellular communication is also disrupted with age (Russell & Kahn, 2007; Zhang et al., 2013). Inflammaging, which is a continuously inflamed state of aged tissues, is one of the characteristics of ageing and it is triggered by the pro-inflammatory signals, inefficient clearing of pathogens and their hosts, and SASP (Franceschi, Garagnani, Parini, Giuliani, & Santoro, 2018). Coordinated functional decline in multiple organ systems is also observed through communication via gap-junctions, cell-to-cell contacts, and other signalling channels (Durieux, Wolff, & Dillin, 2011; Nelson et al., 2012). Blood-borne interventions, such as parabiosis, offer opportunities to modulate lifespan at the intercellular level (Conboy et al., 2005; Villeda et al., 2011).

1.4 High-throughput 'omics' studies of ageing

The majority of today's knowledge about ageing and regulation of lifespan has been driven by the genetic studies on model organisms, investigating how genetic perturbations influence lifespan. A relatively recent development has been the advances

in the high-throughput technologies, allowing the study of physiological ageing at the molecular level. In this section, I summarise some of the conclusions driven by these high-throughput omics studies focusing on ageing in humans.

1.4.1 Genomics

High-throughput measurements became first available for genomics. Using chips designed to capture specific single nucleotide polymorphisms, or later using DNA sequencing technologies, it became possible to resolve genetic variation at a single nucleotide level.

Germline mutations and longevity

The heritability of longevity estimated to be only 20-25% (Deelen, Beekman, Capri, Franceschi, & Slagboom, 2013; Murabito, Yuan, & Lunetta, 2012); however, many age-related diseases which influence health-span and longevity have rather high heritability estimates. For example, Alzheimer’s show 60-80% heritability (Van Cauwenberghe, Van Broeckhoven, & Sleegers, 2016), and cataract has around 50% (Hammond, Snieder, Spector, & Gilbert, 2000).

Studies comparing centenarians, *i.e.* people reaching 100 years of age, with younger controls found several candidate genes influencing longevity, including *CDKN2A/B* and *APOE* genes and FOXO transcription factors which were found repeatedly in independent studies (Deelen et al., 2013; Partridge et al., 2018; Pilling et al., 2017; Wright et al., 2019). Despite the relatively low heritability of longevity, studies using parental longevity as a proxy to individual longevity also replicated these results and identified other candidates, which were also associated with late-onset diseases (Deelen et al., 2019).

Somatic mutation accumulation with age

Spontaneous stochastic mutations occur in somatic cells throughout life. Most of them are thought to be harmless, but some may occasionally affect a gene or its regulation, leading to functional consequences. The role of somatic mutations and cancer is better understood today (Martincorena & Campbell, 2015). Some non-malignant neurological diseases, such as cerebral cortical malformation, epilepsy, autism spectrum

disorder, are also related to somatic mutations that occur early in life (Rodin & Walsh, 2018). Progressive accumulation of somatic mutations is proposed to contribute to ageing (López-Otín et al., 2013); however, the characterisation of the extent requires improved sequencing technologies.

The somatic mutation rate in human B and T lymphocytes, fibroblasts, retina, and the intestinal epithelium is estimated to be in the order of 2 to 10 mutations per diploid genome per cell division, which is ten times more than germline mutations (Lynch, 2010; Martincorena & Campbell, 2015). Using whole-genome/exome sequencing, single-cell DNA sequencing, or RNA sequencing technologies, studies showed that somatic mutations occur stochastically but are affected by the underlying genetic programs such as biases in the DNA repair mechanisms. They can lead to clonal expansion as in cancer, and can involve cancer driver mutations (Lodato & Walsh, 2020; Risques & Kennedy, 2018). Moreover, mutation landscape is tissue-specific, but genes whose expression is associated with increased mutation load in multiple tissues include DNA repair, immune response, autophagy, and cell adhesion (García-Nieto et al., 2019).

1.4.2 Transcriptomics

Genes are transcribed into RNA molecules, and the total of these transcripts are called transcriptomics. It includes both coding and non-coding RNA molecules. Similar to genomics, both microarrays and sequencing methods are used to measure the abundance of transcripts in the cell. Most of the studies have a cross-sectional design, *i.e.* have multiple samples from different individuals at different ages, to measure how the expression of genes changes with age. Most of the human studies have a limited sample size, and thus, replication of the results is an issue (Valdes, Glass, & Spector, 2013). Although the genes that show an increase or decrease in expression with age do not overlap well, studies agree on overall up- and down-regulated pathways. However, these pathways show a high tissue-specificity. Blood tissue shows significant changes in immune regulated pathways and apoptosis (Magalhães, Curado, & Church, 2009; Peters et al., 2015). Brain tissue is associated with changes in synapse-related genes (Berchtold et al., 2008; Colantuoni et al., 2011; Kang et al., 2011). Down-regulation of mitochondrial genes, on the other hand, is observed across multiple tissues (Valdes et al., 2013). Age-related change in transcriptome is not limited to abundances but also involves expression of different isoforms. Both the most

abundant isoform and the number of co-existing isoforms change with age, especially in the brain tissue (Bhadra, Howell, Dutta, Heintz, & Mair, 2019).

Most of the transcriptome studies for ageing focus on one species at a time, and investigate one of the most widely used laboratory organisms (McCarroll et al., 2004; Stegeman & Weake, 2017), humans (Colantuoni et al., 2011; Glass et al., 2013; Lu et al., 2004), or long-lived species such as naked mole-rat (Kim et al., 2011) or bowhead whale (Seim et al., 2014). However, natural lifespan is highly variable across species and could be investigated to understand what contributes to a longer lifespan. The biggest of these studies analysed gene expression in kidney, brain and liver of 33 mammals (Fushan et al., 2015) and found a parallel evolution of gene expression and lifespans. More specifically, they identified a positive correlation between lifespan and expression of genes involved in DNA repair, immune response, and damage response. They also showed down-regulation of lipid metabolism, TCA cycle, and amino acid degradation with increased lifespan. Another study used primary skin fibroblasts from rodents, bats, and a shrew and again found DNA repair as up- and proteolysis as down-regulated with increased lifespan (Ma et al., 2016). However, it is important to note that these studies compare gene expression in adult organisms and do not consider the regulation of gene expression across the lifespan.

The described studies use bulk tissue, and thus, whether these changes are cell-autonomous or due to the changes in the cellular composition is not clear. Recent developments in single-cell sequencing technology allowed to measure gene expression levels in a single cell. The number of human studies using scRNAseq is minimal. Nevertheless, they suggest that both the changes in cell-type composition and cell-specific changes contribute to the age-related changes observed using bulk tissue (Darmanis et al., 2015; Enge et al., 2017). A recent study sampling multiple tissues at different ages in *Mus musculus* suggested distinct cell types undergo different ageing trajectories, but similar cell types in different tissues show similar changes; thus ageing and cell-type identity are coupled (Kimmel et al., 2019).

Here I summarised the transcriptomics studies focusing on coding transcripts, however, there are several studies suggesting non-coding RNAs such as miRNAs (Danka Mohammed, Park, Nam, & Kim, 2017; Kinser & Pincus, 2019; Victoria, Nunez Lopez, & Masternak, 2017), circRNAs (Knupp & Miura, 2018; Mahmoudi & Cairns, 2019), lncRNAs (Bink, Lozano-Vidal, & Boon, 2019; Marttila, Chatsirisupachai, Palmer, & Magalhães, 2020; Sousa-Franco, Rebelo, Rocha, & Bernardes de Jesus, 2019) also show changes with age and control some of the critical regulators of longevity.

1.4.3 Other omics

Epigenomics is the total of chemical compounds that change the DNA regulation without altering the base sequence. This can be achieved by modifying the DNA or histone proteins. Epigenome undergoes substantial changes with age, including a global DNA hypomethylation, regional DNA hypermethylation, and histone modifications (Pal & Tyler, 2016), and global heterochromatin loss (Tsurumi & Li, 2012). The first studies identified specific CpG sites that show differential methylation between young and old in different tissues and cell types (Bork et al., 2010; Rakyan et al., 2010; Teschendorff et al., 2010). Moreover, one of the characteristic observations in the age-related change in DNA methylation is the increase in inter-individual variability (see Section 1.4.4), which is called 'epigenetic drift' (Teschendorff, West, & Beck, 2013). Despite this increased heterogeneity, DNA methylation offers the most accurate biomarkers of ageing. Bocklandt et al. (2011), for the first time, showed that epigenetic changes in only two CpG sites could predict the chronological age with 5.2 years of error. After this pioneering study, many different epigenetic clocks are designed for different tissues or species. Horvath (2013) and Hannum et al. (2013) epigenetic clocks constitute the most accurate biomarker of human ageing so far with 3.6 and 4.9 years of mean absolute errors. In these models, the deviation from the chronological age is interpreted as acceleration or deceleration of ageing process and is associated with a wide range of health-related conditions, such as cancer, obesity, menopause, Werner syndrome, and Huntington's disease (Horvath & Raj, 2018).

Proteomics is the study of the entirety of the proteins, their alternative isoforms and modifications. Due to alternative splicing and post-translational modifications, the number of proteins is estimated to be two orders of magnitude higher than the coding transcripts. However, due to technical challenges, only a small fraction of proteins or their variants are examined. Although only around 3000 proteins could be examined, sampling more than 4000 individuals, Lehallier et al. (2019) showed the changes in proteome are not linear but show changes in direction, as also reported for transcriptome and proteome studies with much smaller sample size (Anisimova et al., 2020; Colantuoni et al., 2011; Dönertaş et al., 2017; Somel et al., 2010). Although the number of proteins that could be investigated is limited, previous studies also report that the changes in protein abundance mostly reflect transcript abundances (Ori et al., 2015; Somel et al., 2010), however translational efficiency impacts the protein abundance (Ori et al., 2015). Moreover, the composition of key protein complexes, such as nuclear pore complex, polycomb repressive complex, exon junction complex, and

complexes involved in vesicular transport, changes with age; potentially leading to misassembly or change in functionality (Ori et al., 2015).

Metabolome, which encompasses low-molecular-weight molecules in a biological system, also shows substantial changes with age. Studies report age-related changes in two-thirds of the entire metabolomic content measured (Menni et al., 2013; Yu, Wang, Han, & He, 2012). The changes are also associated with a wide range of age-related functional decline, including changes in bone mineral density, cholesterol levels, cancer, and type 2 diabetes (Zierer, Menni, Kastenmüller, & Spector, 2015).

Microbiomics investigates all of the genetic material and their products within the entire collection of microorganisms in a given environment. The majority of ageing research focuses on intracellular molecular changes with age; however, there is growing evidence that the gut microbiome changes with age and in ageing-related diseases (Buford, 2017). The human gut is inhabited by over 1000 microbial species, which have co-evolved with humans to live together with mutual benefit. These microbial communities are involved in various processes, including modulation of the nutrient absorption, insulin signalling pathway, overall metabolic state, and the immune system, which are also implicated in ageing (Seidel & Valenzano, 2018). During ageing, the interaction between host and gut microbiota disrupts, leading to dysbiosis and infections (Aleman & Valenzano, 2019). There is some inconsistency about which microbial families in the human gut are affected the most as changes in lifestyle, background population, and antibiotic usage that accompany ageing can influence the populations as well. Nevertheless, all studies agree that the microbial diversity decreases with age (O’Toole & Jeffery, 2015). Older ages are associated with an increased number of pathogenic species and a reduced number of short-chain fatty acid-producing bacteria (Smith et al., 2017). The permeability of intestinal epithelial barrier also increases in diverse species (with a limited proof in humans), and likely causes chronic inflammation (Camilleri, 2019). The experiments in mice and killifish show microbiota transfer from young to old individuals have effects on both age-related phenotypes and lifespan and suggest that the age-associated changes in gut microbiota play a functional role rather than being just a correlation (Kundu et al., 2019; Smith et al., 2017). Disruption of the stability in gut microbiota contributes to various food allergies and inflammatory bowel disorders. However, in recent years it became evident that the changes in microbial composition are relevant for even organs distant from the intestine. Diseases associated with changes in the microbial composition include colon cancer, Parkinson’s and Alzheimer’s diseases, obesity, diabetes, cardiovascu-

lar conditions, frailty, osteoporosis, gout, and rheumatoid arthritis, majority of which are age-related conditions (Buford, 2017).

1.4.4 Heterogeneity in molecular changes associated with age

Despite the ubiquity of ageing in all living organisms, the molecular mechanisms responsible still require further elucidation. Ageing differs phenotypically among individuals, including monozygotic twins (Herndon et al., 2002; Kirkwood, 2005), and within tissues from the same individuals (Horvath, 2013). Researchers have observed an age-related increase in variability in the epigenome (Cheung et al., 2018; Fraga et al., 2005) and transcriptome (Somel, Khaitovich, Bahn, Pääbo, & Lachmann, 2006) of genetically identical samples, which may underlie the phenotypic differences. Age-related expression variability has been detected in many different cells, and tissue types including mice stem cells, cardiomyocytes and immune cells (Bahar et al., 2006; Hernando-Herraez et al., 2019; Martinez-Jimenez et al., 2017), rat neural retina (Li, Wright, & Royland, 2009), fruit-fly, mice and human brain (Angelidis et al., 2019; Brinkmeyer-Langford, Guan, Ji, & Cai, 2016; Davie et al., 2018; Somel et al., 2006; Ximerakis et al., 2019) as well as human pancreas, lung, blood, skin, fat and human fibroblasts in vitro (Angelidis et al., 2019; Enge et al., 2017; Viñuela et al., 2018; Wiley et al., 2017). Despite these reports, there is no agreement on the underlying mechanisms, extent and functional consequences. Suggested mechanisms include somatic (Bahar et al., 2006; Enge et al., 2017) and germline mutations (Brinkmeyer-Langford et al., 2016; Viñuela et al., 2018), changes in the DNA methylation (Hernando-Herraez et al., 2019; Slieker et al., 2016; Viñuela et al., 2018) and chromatin modifications (Cheung et al., 2018) and resulting chromatin compaction (Davie et al., 2018) as well as global dysregulation, caused by the change in a transcription factor or miRNA expression (Inukai, Pincus, Lencastre, & Slack, 2018). Both genome-wide and hypothesis-driven approaches have been employed to explore the extent of expression variability with age. Among the former, some show a transcriptome-wide increase (Angelidis et al., 2019; Davie et al., 2018; Enge et al., 2017; Hernando-Herraez et al., 2019; Somel et al., 2006), while others focus only on those genes showing significant changes in their variability. Brinkmeyer-Langford et al. (2016) report that an equal number of genes show a significant increase or decrease in their expression variability, whereas a recent study from Viñuela et al. (2018) suggests more genes with a decrease in variability of expression. Hypothesis-driven studies mostly show an increase in variability for the genes measured (Bahar et al., 2006;

Martinez-Jimenez et al., 2017; Wiley et al., 2017), whereas Warren et al. (2007) suggest this might be specific only to the non-renewing tissues. Similarly, Ximerakis et al. (2019) show that change in transcription variability is in different directions in different cell types of the mouse brain. The reports also vary in terms of the functional association of this variability. While some consider that increase in variability is widespread (Davie et al., 2018; Somel et al., 2006), others report that variability is concentrated in various cellular functions (Brinkmeyer-Langford et al., 2016; Li et al., 2009; Slieker et al., 2016) – although these functions also differ between reports. In Chapter 2, I present an analysis of the extent and functional associations of the age-related change in gene expression heterogeneity in the human brain.

1.5 Ageing and diseases

Ageing is the major risk factor for many diseases. Increase in life expectancy poses a threat to both individuals and society, considering that late-life is generally associated with multiple diseases. Thus, it is vital to elucidate the nature of the relationship between ageing and diseases.

Many diverse diseases have an increased risk with age, including cardiovascular, neurodegenerative, metabolic diseases, and cancer (Niccoli & Partridge, 2012; Partridge et al., 2018). However, even before the diagnosis of diseases, structural, mechanic, and endocrine changes occur throughout the body. For example, age-related changes in bone density, muscle mass, strength, cognitive abilities, hormone levels, insulin resistance, and vascular stiffness can result in medically defined diseases at later ages.

Although these diseases involve different organs and pathologies, they all show a strong dependence on age (Niccoli & Partridge, 2012) and could, therefore share common aetiologies based upon the underlying mechanisms of ageing. It is therefore essential to understand if the ageing process itself leads age-related deterioration in common pathways and thus different age-related conditions, or if these diseases instead have independent causes. Previous studies have suggested the presence of common pathways that are associated with different age-related diseases. A computational study, comparing GWAS summary results for different age-related diseases has shown that pathways such as proteostasis and NSN are associated with various age-related diseases at the pathway level (Johnson, Dong, Vijg, & Suh, 2015).

Similarly, independent studies have shown the importance of NSN in cardiovascular diseases (Nishimura, Ocorr, Bodmer, & Cartry, 2011) and neurodegeneration (Bové, Martínez-Vicente, & Vila, 2011). A comprehensive review on the role of ageing-related pathways in age-related diseases suggests that not only NSN but mitochondrial function, DNA damage response and autophagy pathways are also linked to age-related diseases (Niccoli & Partridge, 2012). Components of the proteostasis system have also been repeatedly shown to have a role in many age-related diseases including but not limited to ALS, Alzheimer's, cataracts, cardiomyopathy, cardiovascular disease, frontotemporal dementia, and type II diabetes (Kaushik & Cuervo, 2015).

The unhealthy proportion of life at late ages, often with more than two diseases (*i.e.* multimorbidity) (Kingston, Robinson, Booth, Knapp, & Jagger, 2018; Marengoni et al., 2011; Violan et al., 2014), and the use of multiple drugs for treatment (*i.e.* polypharmacy) (Bushardt, Massey, Simpson, Ariail, & Simpson, 2008; Gu, Dillon, & Burt, 2010; Guthrie, Makubate, Hernandez-Santiago, & Dreischulte, 2015; Parameswaran Nair et al., 2016) poses challenges for the individuals, their carers and social networks, and healthcare systems. Limiting the effect of the unhealthy proportion at the end of life is a high priority for the researchers, national governments, and international health organisations (*World report on Ageing And Health*, 2015). A plausible idea is to target ageing to alleviate the effect of multimorbidity and associated polypharmacy at late ages. However, this assumes that multiple diseases have a common underlying factor that can be attributed to age-related changes. There are a few studies investigating the link between ageing and age-related diseases. However, to our knowledge, as a part of this thesis, I present the most comprehensive and systemic analysis of diseases that are clustered based on their age-relevance using an unbiased data-driven approach in Chapter 3.

1.6 Pharmacological interventions to improve life- and health-span

The genetic studies showed that the lifespan of model organisms could be modulated (Kenyon, 2010). Later studies showed lifespan could be extended not only through genetic modifications but also by environmental perturbations, more specifically through dietary restriction (Fontana et al., 2010). Dietary restriction (DR), *i.e.* restricted food intake without malnutrition, can improve both life- and health-span in diverse organ-

isms (Fontana & Partridge, 2015; Kapahi, Kaeberlein, & Hansen, 2017). DR has been shown to reduce plasma triglycerides, diabetes, cardiovascular disease, neoplasms, and brain atrophy in rhesus macaque (Colman et al., 2014; Mattison et al., 2012). NSN, including insulin/insulin-like growth factor and mTOR signalling pathways, monitor the changes in diet. Genetic perturbations that reduce the activity of NSN have led to improved lifespan in diverse organisms (Fontana et al., 2010; Kenyon, 2010; Pan & Finkel, 2017). This, in turn, directed the attention to finding pharmacological interventions that improve health- and lifespan. *Could it be possible that we take some drugs and live longer and healthier?*

Model organism studies suggest that there are at least several chemical compounds that can improve lifespan in diverse species. Moreover, combinations of drugs acting on separate pathways or branches of pathways can be used to achieve even longer lifespan (Admasu et al., 2018; Castillo-Quan et al., 2019). As of March 2020, DrugAge, which is a literature-based database of chemical compounds that modulate lifespan in model organisms, harbours 567 unique drugs that modulate lifespan of at least one model organism (Barardo et al., 2017). Not all of these compounds have a considerable influence on the lifespan of multiple organisms, but there are a few that are repeatedly shown to extend lifespan in independent studies. Here, I summarise three that has been tested in *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus* and recorded 20% extension in average lifespan in at least one experiment:

Rapamycin: Rapamycin, or sirolimus, inhibits the mTOR Complex 1 (TORC1) activity and is used after tissue transplants to prevent rejection (“Sirolimus,” n.d.). Shortly after the first discovery of the importance of mTOR signalling in ageing (Kapahi et al., 2004; Vellai et al., 2003), lifespan extension by rapamycin was shown in diverse organisms, including yeast (Kaeberlein et al., 2005), worm (Robida-Stubbs et al., 2012), fly (Bjedov et al., 2010), and mice (Harrison et al., 2009). It also alleviates many age-related conditions in ageing mice (Kennedy & Lamming, 2016). All seven studies that are recorded in DrugAge database and use the three model organisms mentioned above extended lifespan (Barardo et al., 2017).

Metformin: Metformin is a drug used to treat type 2 diabetes (“Metformin,” n.d.). There are multiple ageing-related mechanisms targeted by metformin. In essence, it decreases insulin levels, decreasing IGF-1 signalling (Liu et al., 2011), inhibits mTOR (Kickstein et al., 2010), inhibits mitochondrial complex 1, reducing endogenous ROS production (Batandier et al., 2006), activates AMPK (Foretz, Guigas, Bertrand, Pollak, & Viollet, 2014), and reduces DNA damage (Algire et al., 2012). However, it is not yet

clear if the effect of metformin on lifespan is through multiple independent pathways or the effects are downstream of one primary action (Barzilai, Crandall, Kritchevsky, & Espeland, 2016). Moreover, metformin reduces all-cause mortality in people with diabetes compared to the ones receiving non-metformin therapies (Campbell, Bellman, Stephenson, & Lisy., 2017) and is under clinical trials for protection against the effects of ageing (Barzilai et al., 2016; Espeland et al., 2016). 62% of 13 studies on the three model organisms that are recorded in DrugAge extended lifespan (Barardo et al., 2017).

Resveratrol: Resveratrol is a compound found in red wine. It is not an approved drug, though it is being investigated to treat cold sores ("Resveratrol," n.d.). Since 1992, there have been several reports showing cardioprotective, anti-cancer, stress-resistance effects of resveratrol, as well as its capacity to extend lifespan (Baur & Sinclair, 2006). The mechanism of action of resveratrol remains controversial. Early studies suggested it acts directly by activating *SIRT1*. However, recent studies suggest that *SIRT1* is activated by the inhibition of cAMP phosphodiesterase (Novelle, Wahl, Diéguez, Bernier, & Cabo, 2015). Nevertheless, direct or indirect, resveratrol seems to act through activation of sirtuins, which are NAD^+-dependent deacetylases, and promote survival and stress resistance (Baur & Sinclair, 2006). Only 50% of 32 studies using the three model organisms that are recorded in DrugAge database extended lifespan (Barardo et al., 2017).

Recent computational developments and high-throughput testing technology also allowed a new area of research: computational drug repurposing. Instead of designing new compounds that target key contributors of a disease / phenotype, drug repurposing strategies search for drugs that are already in use or designed for other conditions but could be used for a new purpose, ageing in our case. There are already a few studies in this area, and as part of this thesis, I present a new approach in Chapter 4, together with an overview of other computational methods developed to achieve this aim.

1.6.1 Challenges involved in drug studies

Several drugs have been shown to extend lifespan in model organisms; however, a significant challenge remains to be the reproducibility of the effects. As noted in the previous section, although there are many candidates, only three had been found to extend lifespan in multiple organisms, and even they do not have a very high repro-

ducibility (50%, 62%, and 100% of the studies extended the average lifespan of tested organisms). NIA Interventions Testing Program (ITP) is a critical multi-institutional program aiming to test the effects of chemical or environmental perturbations on the lifespan of genetically heterogeneous mice (Nadon, Strong, Miller, & Harrison, 2016). The program aims to increase reproducibility and find candidates to modulate ageing.

Although some of the candidates indicate potential use for lifespan extension, their effects in humans are difficult to assess experimentally. We need access to human data in old ages, longitudinal data enabling advanced computational modelling, and reliable biomarkers of ageing. The need for biomarkers of ageing is becoming ever more critical to allow a rapid assessment of the effects of a drug on an organism other than by conducting lengthy, expensive lifespan experiments. The emergence of several 'epigenetic ageing clocks' (Hannum et al., 2013; Horvath, 2013; Levine et al., 2018) provides opportunities to monitor ageing in both cells and organisms actively. Combining molecular, clinical, and lifestyle data is crucial to model the effects of drugs in humans (Dönertaş, Fuentealba, Partridge, & Thornton, 2019).

Moreover, even though the targets of the candidates are highly conserved, we see that their effects on lifespan change drastically between organisms. Drugs have much more notable effects on short-lived species such as *C. elegans* and *D. melanogaster* than in mice. *Homo sapiens* already lives much longer than its closely related species and whether the effects seen in model organisms are reliable predictions for their impact for humans remains an open question. Moreover, the challenge today is not to increase the lifetime of humans. Instead, we should aim to improve healthy years of life by reducing the multimorbidity associated with old age. This challenge has not been adequately addressed, either in model organism research or in clinical studies.

1.7 Thesis overview and objectives

In this chapter, I introduced some of the key concepts of ageing biology, with a focus on the current gap in understanding and the challenging aspects. In the rest of this thesis, I will introduce the computational approaches aiming to address some of these challenges.

In Section 1.4.4, I introduced the inter-individual heterogeneity in gene expression as a challenge in ageing research. Although multiple studies have investigated how heterogeneity changes with age, several questions remain unanswered. In Chapter

2, I will summarise our analyses aiming to understand i) how the inter-individual heterogeneity in gene expression changes with age, at the transcriptome-, pathway-, and gene-level?, ii) what is the influence of different pre-processing steps and measures of heterogeneity on the conclusions?, iii) if the change in heterogeneity is linear throughout lifespan or is different between postnatal development (0 to 20 years of age) and ageing (20+ years of age)?

In Section 1.5, I introduced ageing as a major risk factor for diverse diseases. The exact nature of this relationship, however, is not yet known. In Chapter 3, I will present an analysis of 116 diseases in the UK Biobank, aiming to characterise the common underlying genetic factor between diseases with similar age profiles concerning their relevance to ageing, functional, and evolutionary characteristics. I will present a list of drugs targeting multiple late-onset diseases.

In Section 1.6, I summarised the pharmacological interventions that extend lifespan in model organisms. In Chapter 4, I will first give an overview of the published computational approaches to find potential lifespan-extending candidates. Then, I will introduce a drug repurposing approach to find drugs that target the characteristic age-related gene expression changes in the human brain. Lastly, I will present a comparison of the results of all published studies, including our approach.

In Chapter 5, I will summarise our contributions and future directions of research.

Chapter 2

Age-related changes in gene expression heterogeneity in the human brain

Declaration

This work includes projects I have conducted in collaboration with Veronika Kedlian, a visiting student from the Taras Shevchenko National University of Kyiv, and Ulas Isildak, an undergraduate student from the Department of Biological Sciences, Middle East Technical University (Işıldak, Somel, Thornton, & Dönertaş, 2020; Kedlian, Donertas, & Thornton, 2019). My contributions mostly involved the design and supervision of the study, as well as some computational analyses. I also acknowledge the contributions of my supervisor, Janet Thornton, and Mehmet Somel who contributed through fruitful discussions.

Kedlian, V. R.*, Donertas, H. M.*, & Thornton, J. M. (2019). The widespread increase in inter-individual variability of gene expression in the human brain with age. Aging. * denotes an equal contribution

Işıldak, U., Somel, M., Thornton, J. M., & Dönertaş, H. M. (2020). Temporal changes in the gene expression heterogeneity during brain development and aging. Scientific Reports, 10(1), 4080.

Data Availability

Supplementary tables and all summary statistics are available in the BioStudies database under accession number S-BSST273. Supplementary tables are referred with the corresponding file names throughout the text.

2.1 Introduction

Ageing is a complex process characterised by a gradual decline in maintenance and repair mechanisms, accompanied by an increase in genetic and epigenetic mutations, and oxidative damage to nucleic acids, protein and lipids (Gorbunova et al., 2007; Lu et al., 2004). The human brain experiences dramatic structural and functional changes in the course of ageing. These include decline in grey matter and white matter volumes (Sowell, Thompson, & Toga, 2004), increase in axonal bouton dynamics (Grillo et al., 2013) and reduced synaptic plasticity, all processes that may be associated with decline in cognitive functions (Dorszewska, 2013). Changes during brain ageing are suggested to be a result of stochastic processes, unlike changes associated with postnatal neuronal development that are known to be primarily controlled by adaptive regulatory processes (Polleux, Ince-Dunn, & Ghosh, 2007; Schratt, 2009; Stefani & Slack, 2008). The molecular mechanisms underlying age-related alteration of regulatory processes and eventually leading to ageing-related phenotypes, however, are little understood.

Over the past decade, a number of transcriptome studies focusing on age-related changes in human brain gene expression profiles were published (Kang et al., 2011; Lu et al., 2004; Miller et al., 2014; Somel et al., 2010; Tebbenkamp, Willsey, State, & Šestan, 2014). These studies report ageing-related differential expression patterns in many functions, including synaptic functions, energy metabolism, inflammation, stress response, and DNA repair. By analysing age-related change in gene expression profiles in diverse brain regions, we previously showed that for many genes, gene expression changes occur in opposite directions during postnatal development (pre-20 years of age) and ageing (post-20 years of age), which may be associated with ageing-related phenotypes in healthy brain ageing (Dönertaş et al., 2017). While different brain regions are associated with specific, and often independent, gene expression profiles (Kang et al., 2011; Miller et al., 2014; Tebbenkamp et al., 2014), these studies also show that age-related alteration of gene expression profiles during ageing is a widespread effect across different brain regions.

One of the suggested effects of ageing is increased heterogeneity between individuals and somatic cells, which has been previously reported by several studies. Some of these studies find an increase in age-related heterogeneity in heart, lung and white blood cells of mice (Angelidis et al., 2019; Bahar et al., 2006; Martinez-Jimenez et al., 2017), *Caenorhabditis elegans* (Herndon et al., 2002), and human twins (Fraga

et al., 2005). A study analysing microarray datasets from different tissues of humans and rats also reported an increase in age-related heterogeneity in expression as a general trend (Somel et al., 2006). However, this study found no significant consistency across datasets, nor any significant enrichment in functional gene groups. That said, the generality of increase in expression heterogeneity remains unresolved. For instance, Viñuela et al. (2018) find more decrease than an increase in heterogeneity in human twins and Ximerakis et al. (2018) show the direction of the heterogeneity change depends on cell type in ageing mice brain. Using GTEx data covering different brain regions (20 to 70 years of age), Brinkmeyer-Langford et al. (2016) identify a set of differentially variable genes between age groups, but they do not observe increased heterogeneity at old age. Meanwhile, another study performing single-cell RNA sequencing of human pancreatic cells, identifies an increase in transcriptional heterogeneity and somatic mutations with age (Enge et al., 2017). My previous research also suggested more shared expression patterns during development than in ageing, implying an increase in inter-individual heterogeneity (Dönertaş et al., 2017).

2.1.1 Research objectives

Whether age-related increase in heterogeneity is a universal phenomenon remains contentious. The studies analysed different organisms, organs, and age ranges, making the comparison difficult. Thus, we first employ different methodologies to assess the same dataset, to understand the extent of technical influence on these different results. Furthermore, where it can be detected, whether the increase in heterogeneity is a time-dependent process that starts at the beginning of life or whether this increase and its functional consequences are only seen after developmental processes are completed, have not yet been explored. In this study, we retrieved transcriptome data from three independent studies covering the whole lifespan, including data from diverse brain regions, and conducted a comprehensive analysis to identify the prevalence of age-related heterogeneity changes in human brain ageing compared with those observed during postnatal development. We further investigate the pathways and biological functions associated with the increased heterogeneity.

2.2 Comparison of different methods to measure the change in gene expression heterogeneity with age

In order to study the effect of different preprocessing methods and heterogeneity measures on gene expression heterogeneity during ageing, we used one of the biggest published human brain transcriptome datasets, generated using microarray technology (Colantuoni et al., 2011). We limited the age range to between 20 and 80 years, resulting in RNA expression data for 147 prefrontal cortex samples. We excluded prenatal, infant and childhood samples (up to 20 years old) because their expression levels are inherently coupled with developmental processes in the brain.

2.2.1 Methods to correct for batch effects

We applied two different batch correction strategies to account for technical and biological confounders:

1. Regression: Quantile normalisation followed by linear regression to correct for known co-variates,
2. SVA: Quantile normalisation followed by Surrogate Variable Analysis

We choose these two approaches as representatives of one supervised and one unsupervised approach. Linear regression requires the known confounders to be added as factors to the model, whereas SVA is unsupervised and deduces the factors from data. We also applied only quantile normalisation and ComBat to correct for batch effects and the results are available in the paper (Kedlian et al., 2019). For the brevity here I just summarise the results with these two methods. **Quantile Normalisation** was performed using the 'normalize.quantiles' function from the 'preprocessCore' R library. **Linear Regression** was applied considering: technical batches (N = 19), sex (N=2), race (N=4), post-mortem interval, RNA integrity number, and pH. **Surrogate Variable Analysis (SVA)** was applied using 20 inferred surrogate variables using the 'sva' package in R (Leek et al., 2019). Details of the methods are available in Section 2.2.6.

2.2.2 Methods to measure age-related changes in gene expression heterogeneity

We also applied two different strategies to measure the change in the gene expression heterogeneity with age, namely *continuous* and *grouped* approaches. The continuous approach detects continuous monotonic change in variation from 20 to 80 years of age. The grouped approach compares the gene expression variation between two age groups: young (20 - 40 years old, N = 53) and old (60 - 80 years old, N = 22).

In the *continuous approach*, we first fit a linear model to explain age-dependent change in expression, and then use the residuals from this model to represent the heterogeneity. To measure change in the expression heterogeneity with age, we calculate the Spearman correlation coefficient ($\Delta var(\rho)$) between the absolute value of residuals and age.

In the *grouped approach*, we first generated a distribution of expected heterogeneity in gene expression for the young individuals and treated it as a null distribution to compare with the heterogeneity from the old individuals. We used interquartile range (IQR) as a measure of heterogeneity because it is robust to outliers. In order to calculate a distribution of expected heterogeneity in the young group, we randomly selected a subsample of 22 individuals (the number of samples in the old group) from the 53 individuals in the young group for 10,000 times and calculated IQR. The change in heterogeneity, $\Delta var(IQR)$, was measured as a fractional change in the IQR between old and young groups. The p-value was determined by calculating how many times we observed a value as extreme as IQR_{old}.

Details of the methods are available in Section 2.2.6.

2.2.3 Changes in heterogeneity at the transcriptome-level

The change in heterogeneity calculated using continuous approach, $\Delta var(\rho)$, ranged between -0.32 and 0.36 and were normally distributed (Shapiro-Wilk test, p > 0.05) (Figure 2.1A). The distributions were significantly shifted towards positive values for both correction methods (Wilcoxon test, $p < 2.2e^{-16}$, median values range between 0.01 and 0.03). Although the shift in the distribution was small, 57 and 63% of the genes showed increase in heterogeneity with age, for SVA and Regression corrections, respectively. However, we noted that the changes in heterogeneity calculated

for each gene, using regression- and SVA-corrected data were only weakly correlated ($\rho = 0.35$, Figure 2.1B).

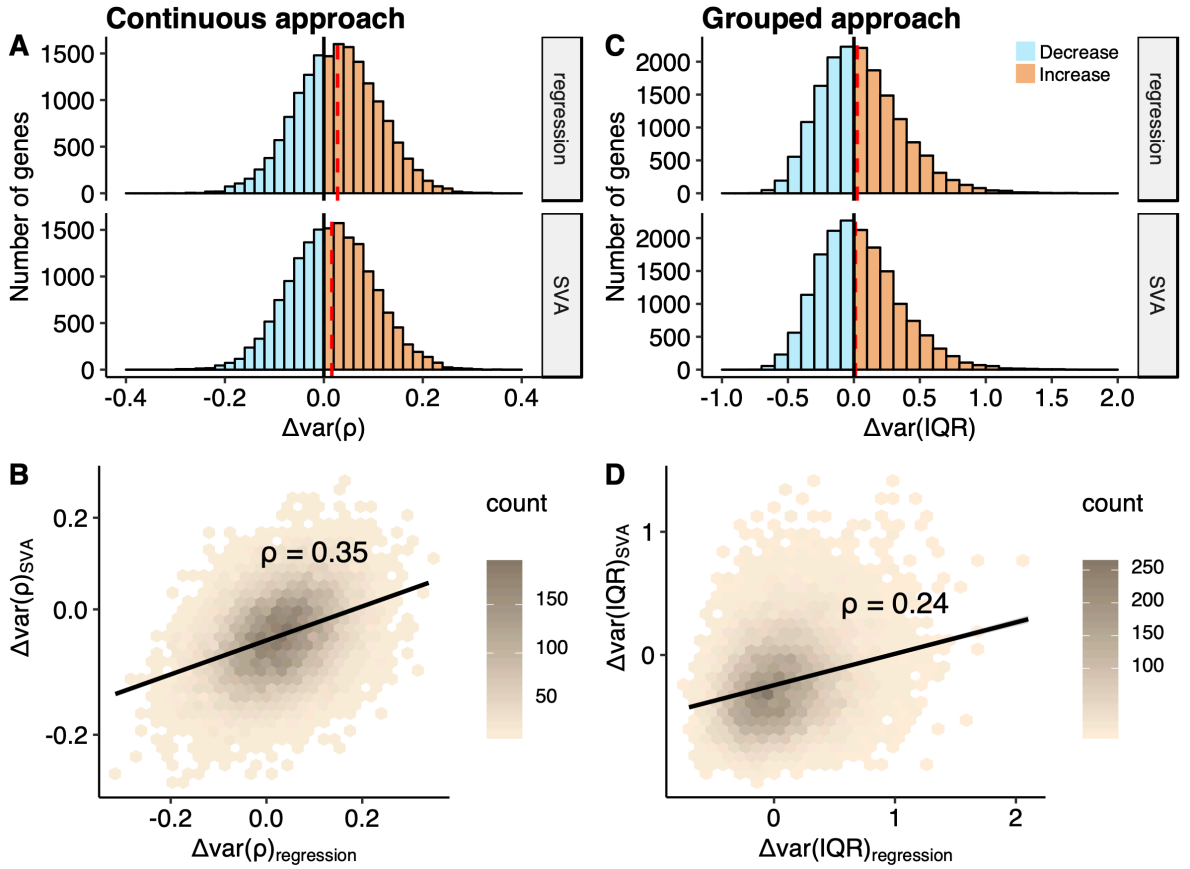


Figure 2.1 *Distributions of the Δvar-measures for all the genes (n = 16675) obtained in the continuous (A) and grouped (C) approaches. Increase in the heterogeneity with age, $\Delta var > 0$, is coloured in orange, while decrease in heterogeneity, $\Delta var < 0$, is marked in blue. The red dashed line depicts median of the distribution. The $\Delta var(\rho)$ distributions are normal with their mean and median values equal to 0.03 and 0.02 for regression and SVA, respectively; The $\Delta var(IQR)$ distributions are moderately skewed: skewness values are 0.66 and 0.68 for regression and SVA, respectively. The mean and median values of the $\Delta var(IQR)$ distribution are 0.05 and 0.02 for regression and 0.04 and 0.01 for SVA, respectively. Hexagonal heat maps illustrate relationship between regression and SVA-corrected measures of the heterogeneity for each gene, obtained in continuous - $\Delta var(\rho)$ (B) and grouped - $\Delta var(IQR)$ (D) approaches. The colour gradient represents the density of the data. The linear regression line and the Spearman correlation estimate, ρ, for the corresponding variables are shown on each graph.*

The distributions of the change in heterogeneity calculated using grouped approach, $\Delta var(IQR)$, were moderately skewed to the right and ranged from -0.7 to 2.1 for the regression corrected data and from -0.78 to 1.71 for the SVA corrected data (Fig-

ure 2.1C). The skew to the right was expected given that we calculate heterogeneity change as a fraction and thus, it was more sensitive to increase in heterogeneity. In both cases, the distributions demonstrated a significant deviation from zero (Wilcoxon test, $p < 2.2e^{-16}$). 6% and 2% more genes showed higher heterogeneity in the old group, for regression- and SVA-corrected data, respectively. Similar to the continuous approach, the effect sizes calculated using different batch correction methods weakly correlated ($\rho = 0.24$, Figure 2.1D).

2.2.4 Changes in heterogeneity at the gene-level

We then asked if we could detect any genes with a significant change in heterogeneity. Using the continuous approach, we did not detect any significant change in heterogeneity with age after multiple testing correction. The grouped approach leads to 741 and 746 differentially variable (DV) genes (FDR corrected p ≤ 0.05) using the regression and SVA correction, respectively. However, the two sets of DV genes had only 83 genes in common, one of which shows an opposite direction of change in the two sets. The correlation between $\Delta var(IQR)$ for regression and SVA corrected data is weak (ρ = 0.24), but correlation increases when we select only the common DV genes (ρ = 0.44). In agreement with our overview analysis above, we find twice as many DV genes with an increase in heterogeneity as those that decrease heterogeneity, using both correction methods: i) 533 genes increase and 208 decrease their heterogeneity after the regression correction, ii) 505 genes increase and 241 decrease their heterogeneity after the SVA correction.

2.2.5 Changes in heterogeneity at the pathway-level

Following the individual gene analysis, we explored whether genes that tend to increase or decrease heterogeneity with age are localised in particular functional groups. We performed multiple gene set enrichment analyses (GSEA) using the change in the heterogeneity with age (Δvar) measures obtained in the continuous and grouped approaches on the gene sets from KEGG and Biological Process GO categories. We observed no genome-level significant enrichment in particular functional groups on the data either from the continuous (SVA correction), or the grouped approach (Regression and SVA corrections). However, we found that 4 pathways, namely beta-Alanine metabolism, Ras signaling pathway, Phosphatidylinositol signaling system, Bacterial

invasion of epithelial cells (FDR corrected p ≤ 0.05) were enriched among the genes showing more heterogeneity of expression in the continuous approach (Regression correction). These pathways had positive normalised enrichment scores (NES), *i.e.* enrichment for the genes that increase heterogeneity with age. Moreover, these pathways also had positive NES for other approaches, even though they were not significant.

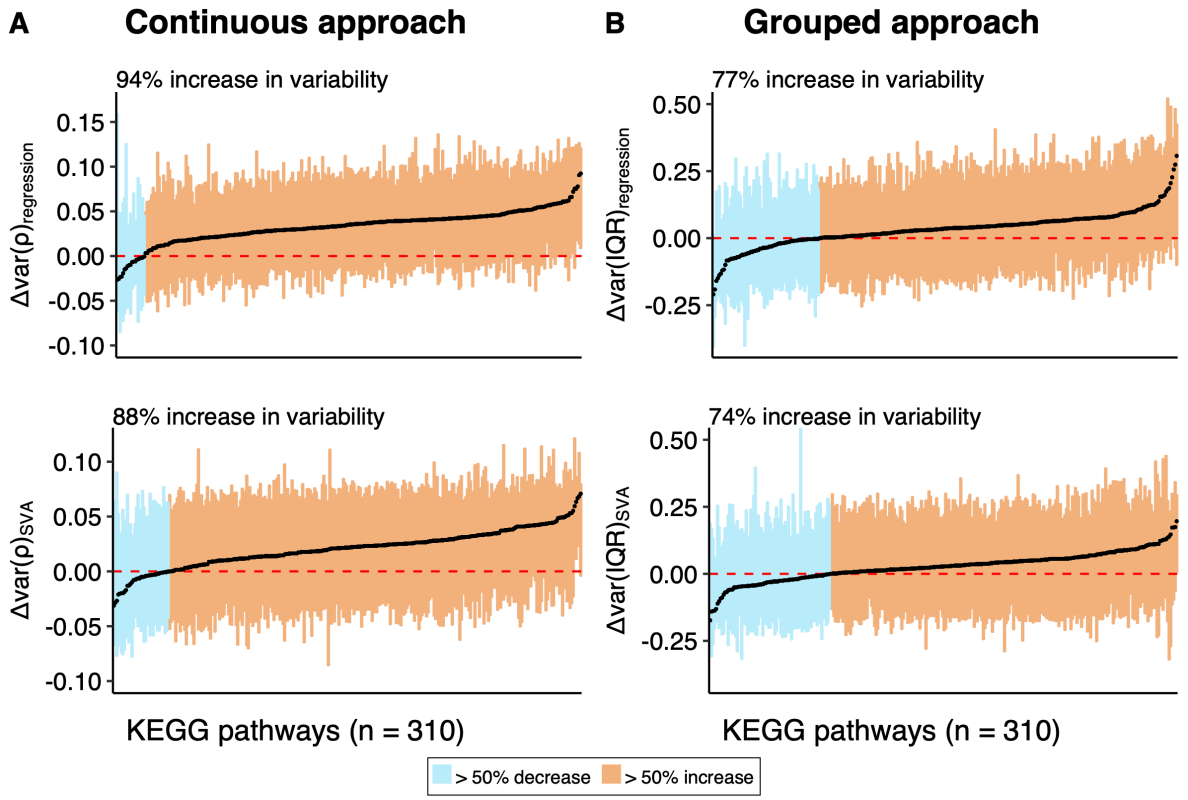


Figure 2.2 *The distribution of the heterogeneity measures (Δvar) for the genes within each pathway is represented as a box, encapsulating part of the distribution between 1st and 3rd quantile, median of the box is coloured in black. (A) represents heterogeneity measure distributions for the continuous, while (B) for the grouped approaches. Pathways are ordered by increasing median. Boxes are coloured in orange if the corresponding pathways have median $\Delta var > 0$, and in blue, if median $\Delta var < 0$. Text label on the plot shows percentage of pathways with median $\Delta var > 0$. Red dashed line marks $\Delta var = 0$. The mean values for the median across all pathways are 0.033 for $\Delta var(\rho)_{regression}$, 0.021 for $\Delta var(\rho)_{SVA}$, 0.033 for $\Delta var(IQR)_{regression}$, and 0.027 for $\Delta var(IQR)_{SVA}$.*

The gene set enrichment analysis shows if there are particular gene sets that include the genes with the highest increase or decrease. Failing to detect such functional categories, we asked how the heterogeneity measures for the genes were distributed in the different functional groups of genes. For each of 310 KEGG pathways, en-

compassing 5922 unique genes, we analysed the distributions of Δvar measures, focusing on the median value for the change in heterogeneity (Figure 2.2A, B). In line with the overall tendencies we observed (Figure 2.1A, C), the majority of pathways contained a larger number of genes that become more variable with age, irrespective of the approach or correction method used. Although the increase in heterogeneity is ubiquitous and is observed across the majority of the pathways (74-94%), the increase is small (the mean value for the shift in distributions range between 0.021 and 0.033) – in accordance with the small, but significant increase observed in the distribution for all genes. Since the pathways are not mutually exclusive, we checked if there are particular genes that are present in many different pathways and cause the shift. However, no significant correlation between the pathway membership of gene and its heterogeneity measure (Δvar) was detected.

2.2.6 Methods

Data processing steps

Dataset Selection: We utilised one of the largest age-series human brain expression datasets, featuring 269 prefrontal cortex samples from healthy individuals and spanning the whole lifespan from development (prenatal samples) through ageing (80 years) (Colantuoni et al., 2011). These data were collected using microarray technology from people of both sexes and 4 races, namely African American (AA), Caucasian (CAUC), Hispanic (HISP) and Asian (AS). In the current analysis, we excluded foetal, childhood and early adulthood samples before the age of 20, thus limiting our sample size to 147. This was to exclude developmental processes taking place in the brain until the end of early adulthood, which exhibit discontinuous expression changes between early adulthood and ageing (Dönertaş et al., 2017). Our main motivation was to study changes in gene expression variability during ageing, considering 20 years old as a starting point.

Data Characterisation: The pre-processed data (loess normalisation was applied on the background corrected log2 intensity ratios (sample/reference)(Colantuoni et al., 2011)); sample and gene (probe set to Entrez gene mapping) annotations were obtained from the NCBI Gene Expression Omnibus (GEO) at accession number GSE30272. Samples were processed in 19 batches, had different quality measurements, namely pH and RNA integrity number (RIN), and differed in the time of collection after death (post-mortem interval (PMI)). Using a PCA, we found no sample outliers as judged

by visual inspection of the first two principal components. However, the relationship analysis between the above-mentioned factors (*i.e.* batch, RIN, PMI and others) and age yielded significant correlations for sex, post-mortem interval and RNA integrity, pointing to potential confounders in the data. We further checked the overlap between significantly differentially variable genes in our analysis and previously reported genes that are affected by PMI and detected only a limited overlap.

Probe set to gene summarisation: If one probe-set was mapped to several genes, it was deleted to avoid duplication. Conversely, when one gene had several probe-set expression values, they were averaged to obtain a unique gene expression value. In total 16675 genes were measured on the array. *Batch correction:* To compensate for technical variation between samples, quantile normalisation (QN) was performed using the 'normalize.quantiles' function from the 'preprocessCore' R library. To differentiate between the age effect and the effect of the unwanted technical and biological variability, we have applied different expression correction strategies: linear regression of the known covariates, unsupervised estimation of covariates using surrogate variable analysis (SVA) (Leek & Storey, 2007). As a result, we analysed the same data two times, corrected using QN+regression and QN+SVA. Different corrections work by adjusting for the different covariates in the linear model that explains the gene expression, namely: i) QN – no covariates were added; ii) QN+regression – considering technical batches (N = 19), sex (N=2), race (N=4), post-mortem interval, RNA integrity number, pH; iii) QN + SVA – 20 surrogate variables (SV) were inferred from the expression data using the 'sva' function from 'SVA' R library (Leek et al., 2019).

Differential variability

The continuous approach: First, a linear model to fit gene expression during ageing, using $age^{0.25}$ and potential confounders, was constructed. Next, the Spearman correlation was calculated between the absolute values of the residuals, $|\epsilon_i|$ from the linear model and age. Consequently, Spearman correlation estimates were used as a measure of the change in variability, referred as $\Delta var_i(\rho)$. p values for the Spearman correlation estimates were corrected for multiple testing using FDR. FDR adjusted p ≤ 0.05 was used as a threshold to define significantly DV genes.

$$\Delta var_i(\rho) = \rho(|\epsilon_i, age|)$$

The grouped approach: First, a corrected expression matrix was obtained by removing the effect of covariates (see data processing steps) from the data using the residuals from a linear regression model. The 'grouped approach' is a custom resampling-based test designed to compare gene expression variability between young (20 – 40 years old) and old (60-80 years old) groups using an interquartile range (IQR). IQR corresponds to the difference between the 75th and 25th percentiles of the distribution and is considered to be a robust measure of variability, meaning it is not susceptible to outliers and departure from normality in the data. In order to adjust for the unequal sample size of the young (N = 53) and old (N = 22) groups, we, first, calculated a null distribution of the IQR values for the young group by resampling it 10,000 times with the size of the old group. Next, we calculated significance as a percentage of samples where IQR_{old} was more extreme than IQR_{young} and corrected it for multiple testing using FDR correction, q ≤ 0.05. The 'grouped' measure of change in the variability, $\Delta var_i(IQR)$, for the gene i, corresponds to the difference between IQR value for the old, $IQR_{i,old}$, and $\overline{IQR_{i,young}}$ (*i.e.* mean IQR value from the young distribution), which is then divided by the latter, see formula:

$$\Delta var_i(IQR) = \frac{IQR_{i,old} - \overline{IQR_{i,young}}}{\overline{IQR_{i,young}}}$$

Gene Set Enrichment Analysis for KEGG pathways

Δvar measures from the differential variability analyses were used to perform gene set enrichment analysis, GSEA (Subramanian et al., 2005) using the 'clusterProfiler' R library (Yu et al., 2012). 315 KEGG pathways with the size of between 10 and 500 genes were considered as gene sets for the GSEA.

Heterogeneity Distributions in Pathways

KEGG pathway to gene mapping was obtained from 'KEGGREST' R library and pathways were pre-filtered to contain between 5 and 500 genes. As a result, 310 KEGG pathways that comprise 5922 unique genes were used for the subsequent analysis. The boxplots illustrated distributions of the Δvar measure for genes in each pathway. Pathways were sorted according to their median Δvar measure in ascending order. The percentage of pathways that have their median Δvar above zero was calculated.

Distribution tests

Distributions of the Δvar - measures for all the genes were tested for normality using the Shapiro-Wilk test in R ('Shapiro.test' function) on the multiple subsamples, consisting of 5000 measures. Skewness of the distributions was calculated using the 'fBasics' function from 'BasicStatistics' R library.

2.3 Temporal landscape of the changes in gene expression heterogeneity during brain development and ageing

To investigate how heterogeneity in gene expression changes with age, we used 19 published microarray datasets from three independent studies. Datasets included 1,010 samples from 17 different brain regions of 298 individuals whose ages ranged from 0 to 98 years (Supplementary Table S1, Figure A.1). In order to analyse the age-related change in gene expression heterogeneity during ageing compared to the change in development, we divided datasets into two subsets as development (0 to 20 years of age, n = 441) and ageing (20 to 98 years of age, n = 569). We used the age of 20 to separate pre-adulthood and adulthood based on commonly used age intervals in earlier studies (see Section 2.3.6). For the analysis, we focused only on the genes for which we have a measurement across all datasets (n = 11,137).

2.3.1 Age-related change in gene expression levels

To quantify age-related changes in gene expression, we used a linear model between gene expression levels and age (see Section 2.3.6, Figure A.2). We transformed the ages to the fourth root scale before fitting the model as it provides relatively uniform distribution of sample ages across the lifespan, but we also confirmed that different age scales yield quantitatively similar results (Figure A.3). We quantified expression change of each gene in ageing and development periods separately and considered regression coefficients from the linear model (β values) as a measure of age-related expression change (Figure A.4).

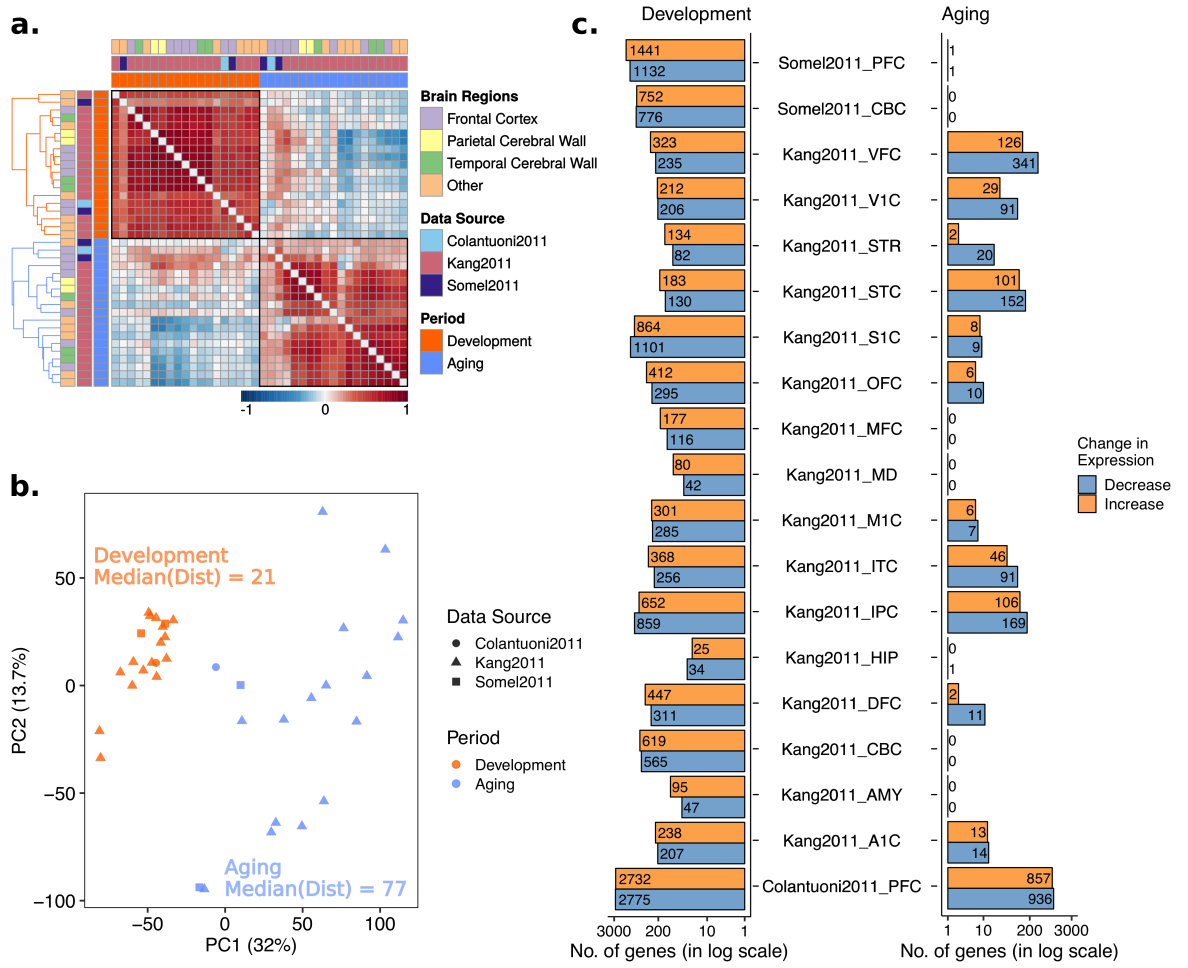


Figure 2.3 *Age-related change in gene expression during postnatal development and ageing. (a) Spearman correlations among age-related expression changes (β values) across datasets. The colour of the squares indicates if the correlation between the corresponding pair of datasets (across β values of 11,137 common genes) is positive (red) or negative (blue), while darker colour specifies a stronger correlation. Diagonal values were removed in order to enhance visuality. Annotation rows and columns indicate data source, brain region and period of each dataset. Hierarchical clustering was performed for each period separately (colour of the dendrogram indicates periods) to determine the order of datasets. (b) Principal component analysis (PCA) of age-related expression changes during ageing and development. The analysis was performed on age-related expression change values of 11,137 common genes among all 38 datasets. The values of the first principal component on the x-axis and second principal component on the y-axis were drawn, where the values in the parenthesis indicate the variation explained by the corresponding principal component. Median Euclidean pairwise distances among development and ageing datasets calculated using PC1 and PC2 were annotated on the figure. Different shapes show different data sources and colours show development (dark orange) and ageing (blue) (c) Number of significant (FDR corrected $p < 0.05$) gene expression changes in development (left panel) and ageing (right panel). The x-axis shows the number of genes in the log scale. The colour of the bars shows the direction of change, decrease (steel grey), and increase (orange). The exact number of genes are also displayed on the plot.*

We first analysed similarity in age-related expression changes across datasets by calculating pairwise Spearman's correlation coefficients among the β values (Figure 2.3a). Both development (median correlation coefficient = 0.56, permutation test p < 0.001, Figure A.5) and ageing datasets (median correlation coefficient = 0.43, permutation test p = 0.003, Figure A.5) showed moderate correlation with the datasets within the same period. Although the difference between dataset correlations within development and ageing datasets was not significant (permutation test p = 0.1, Figure A.6), weaker consistency during ageing may reflect the stochastic nature of ageing, causing increased heterogeneity between ageing datasets.

The principal component analysis (PCA) of age-related expression changes (β) revealed distinct clusters of development and ageing datasets (Figure 2.3b). Moreover, ageing datasets were more dispersed than development datasets (median pairwise Euclidean distances between PC1 and PC2 were 77 for ageing and 21 for development), which may again reflect stochasticity in gene expression change during ageing and can indicate more heterogeneity among different brain regions or datasets during ageing than in development.

We next identified genes showing significant age-related expression change (FDR-corrected p < 0.05), for development and ageing datasets separately (Figure 2.3c). Development datasets showed more significant changes compared to ageing (permutation test p = 0.003, Figure A.6), which may again indicate higher expression variability among individuals during ageing. The direction of change in development was mostly positive (14 datasets with more positive and 5 with more negative), whereas in ageing datasets, we observed more genes with a decrease in expression level (13 datasets with more genes decreasing expression and 5 with no significant change, and 1 with an equal number of positive and negative changes).

2.3.2 Age-related change in gene expression heterogeneity

To assess age-related change in heterogeneity, we obtained the unexplained variance (residuals) from the linear models used to calculate the change in gene expression level. For each gene in each dataset, we separately calculated Spearman's correlation coefficient (ρ) between the absolute value of residuals and age, irrespective of whether the gene shows a significant change in expression (Figure A.2). We considered ρ values as a measure of heterogeneity change, where positive values mean an increase in heterogeneity with age. We also repeated this approach using loess

regression instead of a linear model between expression level and age, and found high correspondence between ρ values based on linear and loess regression models (Figure A.7). Still, loess regression was more sensitive to the changes in sample sizes and parameters and we therefore continued downstream analyses with the ρ estimates based on the residuals from the linear model.

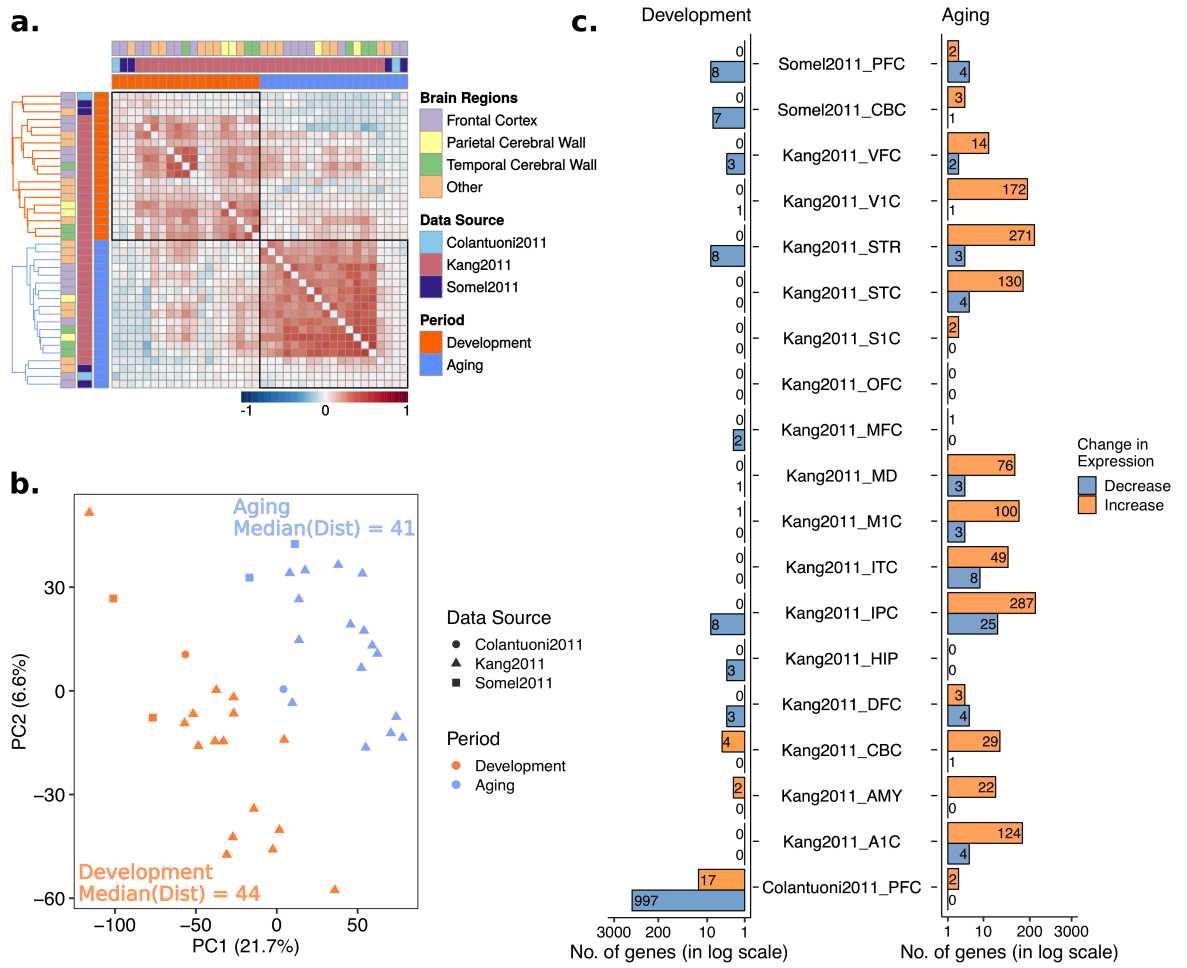


Figure 2.4 *Age-related change in gene expression heterogeneity during development and ageing. The procedures are similar to those in the previous figure, except, age-related heterogeneity changes (ρ values) were used instead of expression changes (β values). (a) Spearman correlations among age-related heterogeneity changes (ρ values) across datasets. (b) Principal component analysis (PCA) of heterogeneity change with age. (c) The number of genes showing significant heterogeneity change in ageing and development.*

We next asked if datasets show similar ρ, *i.e.* age-related changes in heterogeneity, by calculating pairwise Spearman's correlation between pairs of datasets, across shared genes (Figure 2.4a). Unlike the correlations among expression level changes,

ρ values did not show a higher consistency during development. In fact, although the difference is not significant (permutation test p = 0.2, Figure A.6), the median value of the correlation coefficients was higher in ageing (median correlation coefficient = 0.21, permutation test p = 0.24, Figure A.5), than in development (median correlation coefficient = 0.11, permutation test p = 0.25, Figure A.5).

A principal component analysis (PCA) showed that, like expression change, heterogeneity change with age can also differentiate ageing datasets from development (Figure 2.4b). Similar to the pairwise correlations (Figure 2.4a), ageing datasets clustered more closely than development datasets, though the effect is small compared to the changes in expression level (median pairwise Euclidean distances are 41 and 44 for ageing and development, respectively). Both observations imply more similar changes in heterogeneity during ageing.

Using the p-values from Spearman's correlation between age and the absolute value of residuals for each gene, we then investigated the genes showing a significant change in heterogeneity during ageing and development (FDR corrected p-value < 0.05). We found almost no significant change in heterogeneity during development, except for the Colantuoni2011 dataset, for which we have high statistical power due to its large sample size. In ageing datasets, on the other hand, we observed more genes with significant changes in heterogeneity (permutation test p = 0.06, Figure A.6) and the majority of the genes with significant changes in heterogeneity tended to increase in heterogeneity (Figure 2.4c). However, the genes showing a significant change did not overlap across ageing datasets (Figure A.8).

Nevertheless, our analyses indicated relatively more consistent heterogeneity changes among datasets in ageing compared to development, implying that heterogeneity change could be a characteristic linked to ageing.

2.3.3 Consistent increase in heterogeneity during ageing

As our previous analyses suggested age-related changes in heterogeneity can differentiate development from ageing and show more similarity during ageing, we sought to characterise the genes displaying such changes. Since the significance of the changes is highly dependent on the sample size, instead of focusing on significant genes identified within individual datasets, we leveraged upon the availability of multiple datasets and focused on their shared trends, capturing weak but reproducible

trends across multiple datasets. Consequently, we used the level of consistency in age-related heterogeneity change across datasets to sort genes.

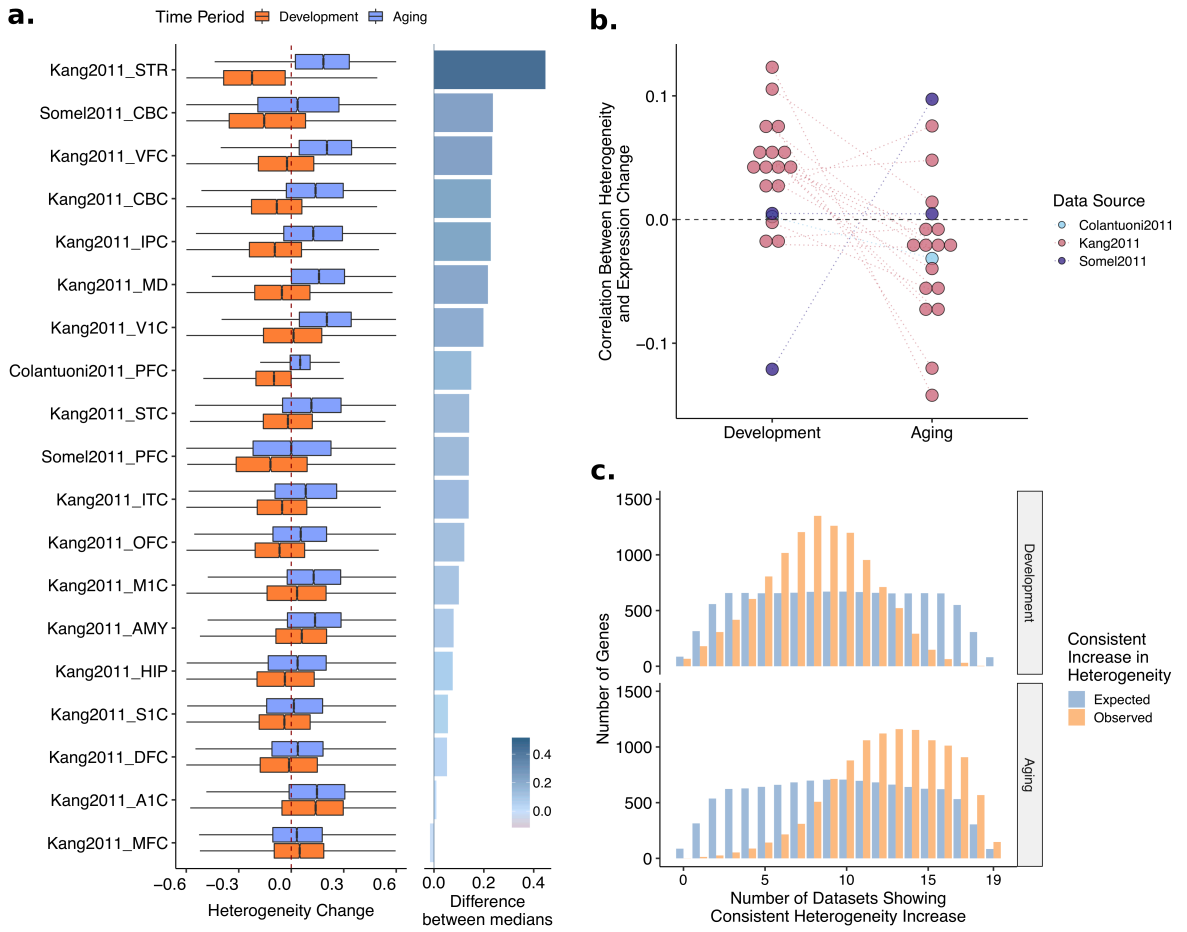


Figure 2.5 *(a) Boxplots, showing distributions of age-related heterogeneity changes (ρ values) of 11,1137 common genes for each dataset and period separately. The dotted red line (vertical line at x = 0) reflects no change in heterogeneity. The difference between median heterogeneity change in ageing and development is given as a bar plot on the right panel. Datasets are ordered by the differences in median heterogeneity changes in ageing and development. (b) The relationship between expression and heterogeneity change with age. Spearman correlation analysis was performed between age-related expression changes (β values) and age-related heterogeneity changes (ρ values) of 11,137 common genes, separately for each dataset. The dotted grey line at y = 0 reflects no correlation between expression and heterogeneity. (c) Expected and observed consistency in the heterogeneity change across datasets in development and ageing. There is a significant shift toward heterogeneity increase in ageing (permutation test p<10-7) (lower panel), while there is no significant consistency in either direction in development (upper panel). The expected distribution is constructed using a permutation scheme that accounts for the dependence among datasets and is more stringent than random permutations*

We first examined profiles of age-related heterogeneity change in ageing and development. Among ageing datasets 18/19 showed more increase than decrease in heterogeneity with age (median $\rho > 0$, *i.e.* higher numbers of genes with increase), while the median heterogeneity change in one dataset was zero. In development, on the other hand, only 5/19 datasets showed more increase in heterogeneity, while the remaining 14/19 datasets showed more decrease with age (median $\rho < 0$) (Figure 2.5a). The age-related change in heterogeneity during ageing was significantly higher than during development (permutation test $p < 0.001$, Figure A.6). We also checked if there is a relationship between changes in heterogeneity during development and during ageing (*e.g.* if those genes that decrease in heterogeneity tend to increase in heterogeneity during ageing) but did not find any significant trend (Figure A.9).

A potential explanation why we see different patterns of heterogeneity change with age in development and ageing could be the accompanying changes in the expression levels, as it is challenging to remove dependence between the mean and variance. To address this possibility, we first calculated Spearman's correlation coefficient between the changes in heterogeneity (ρ values) and expression (β values), for each dataset. Overall, all datasets had values close to zero, suggesting the association is not strong. Moreover, we saw an opposing profile for development and ageing; while the change in heterogeneity and expression were positively correlated in development, they showed a negative correlation in ageing (Figure 2.5b).

Having observed both a more consistent heterogeneity change among ageing datasets (Figure 2.4a-b) and an increased heterogeneity in the ageing datasets compared to development (Figure 2.5a), we asked which genes show consistent increase in heterogeneity across datasets in ageing and development. We therefore calculated the number of datasets with an increase in heterogeneity during development and ageing for each gene (Figure 2.5c, A.10c). To calculate significance and expected consistency, while controlling for dataset dependence, we performed 1,000 random permutations of individuals' ages and re-calculated the heterogeneity changes. In development, there was no significant consistency in heterogeneity change in either increase or decrease. During ageing, however, there was a significant signal of consistent heterogeneity increase, *i.e.* more genes showed consistent heterogeneity increase across ageing datasets than randomly expected (Figure 2.5c, lower panel). We identified 147 common genes with a significant increase in heterogeneity across all ageing datasets (permutation test $p < 0.001$, Supplementary Table S2), whereas only one gene (*GPR137B*) showed a consistent decrease in heterogeneity during ageing. Based on our permutations, we estimated that 84/147 genes could be expected to

have consistent increase just by chance, suggesting only ~40% true positives. In development, in contrast, there was no significant consistency in heterogeneity change in either direction (increase or decrease). Nevertheless, comparing the consistency in ageing and development, there was an apparent shift towards a consistent increase in ageing – even if we cannot confidently report the genes that become significantly more heterogeneous with age across multiple datasets.

2.3.4 Heterogeneity trajectories

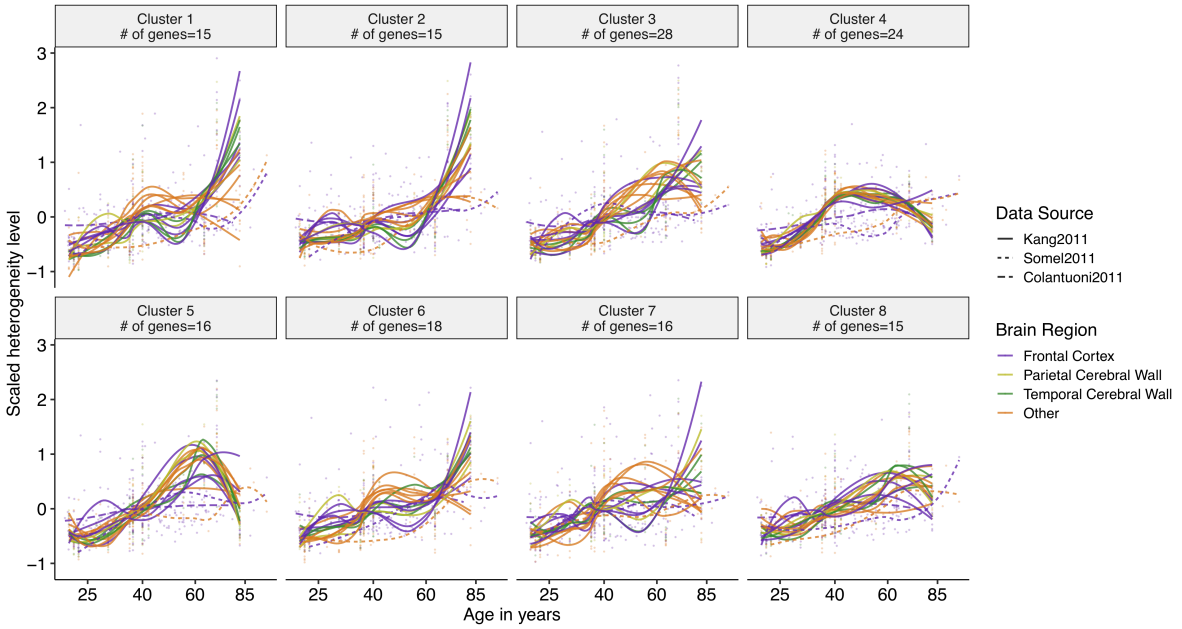


Figure 2.6 *Clusters of genes showing a consistent heterogeneity increase in ageing (n = 147). Clustering was performed based on patterns of the change in heterogeneity, using the k-means clustering method. The x- and y-axes show age and heterogeneity levels, respectively. Mean heterogeneity change for the genes in each cluster was drawn by spline curves. The colours and line-types of curves specify different brain regions and data sources, respectively.*

We next asked if there are specific patterns of heterogeneity change, *e.g.* increase only after a certain age. We used the genes with a consistent increase in heterogeneity with age during ageing (n = 147) to explore the trajectories of heterogeneity change (Figure 2.6). Genes grouped with k-means clustering revealed three main patterns of heterogeneity increase (Supplementary Table S2): i) genes in clusters 3 and 7 show noisy but a steady increase throughout ageing, ii) genes in clusters 4, 5 and 8 show increase in early ageing but a later slight decrease, revealing a reversal (up-down)

pattern, and iii) the remaining genes in cluster 1, 2 and 6 increase in heterogeneity dramatically after the age of 60. We next asked if these genes have any consistent heterogeneity change pattern in development (Figure A.11), but most of the clusters showed no or only weak age-related changes. We also analysed the accompanying changes in mean expression levels for these clusters. Except for cluster 1, which shows a decrease in expression level at around the age of 60 and then shows a dramatic increase, all clusters show a steady scaled mean expression level at around zero, *i.e.* different genes in a cluster show different patterns (Figure A.12).

We further tested the genes showing dramatic heterogeneity increase after the age of 60 (clusters 1, 2 and 6) for association with Alzheimer's Disease, as the disease incidence increases after 60 as well (Evans et al., 1989); however, we found no evidence for such an association (Figure A.13).

2.3.5 Functional analysis

To examine the functional associations of heterogeneity changes with age, we performed gene set enrichment analysis (GSEA) using KEGG pathways (Kanehisa, Sato, Furumichi, Morishima, & Tanabe, 2019), Gene Ontology (GO) categories (Ashburner et al., 2000; The Gene Ontology Consortium, 2019), Disease Ontology (DO) categories (Kibbe et al., 2015), Reactome pathways (Fabregat et al., 2018), transcription factor (TF) targets (TRANSFAC) (Matys et al., 2003), and miRNA targets (MiRTarBase) (Chou et al., 2016). Specifically, we rank-ordered genes based on the number of datasets that show a consistent increase in heterogeneity and asked if the extremes of this distribution are associated with the gene sets that we analysed. There was no significant enrichment for any of the functional categories and pathways for the consistent changes in development. The significantly enriched KEGG pathways for the genes that become consistently heterogeneous during ageing included multiple KEGG pathways known to be relevant for ageing, including the longevity regulating pathway, autophagy (Rubinsztein, Mariño, & Kroemer, 2011), mTOR signaling (Johnson, Rabinovitch, & Kaeberlein, 2013) and FoxO signaling (Martins, Lithgow, & Link, 2016) (Figure 2.7a). Among the pathways with a significant association (listed in Figure 2.7a), only protein digestion and absorption, primary immunodeficiency, linoleic acid metabolism, and fat digestion and absorption pathways had negative enrichment scores, meaning these pathways were significantly associated with the genes having the least number of datasets showing an increase. However, it is important to note

that this does not mean these pathways have a decrease in heterogeneity as the distribution of consistent heterogeneity levels is skewed (Figure 2.5c, lower panel). We also calculated if the KEGG pathways that we identified are particularly enriched in any of the heterogeneity trajectories we identified. Although we lack the necessary power to test the associations statistically due to small number of genes, we saw that i) group 1, which showed a stable increase in heterogeneity, is associated more with the metabolic pathways and mRNA surveillance pathway, ii) group 2, which showed first an increase and a slight decrease at later ages, is associated with axon guidance, mTOR signaling, and phospholipase D signaling pathways, and iii) group 3, which showed a dramatic increase after age of 60, is associated with autophagy, longevity regulating pathway and FoxO signaling pathways. The full list is available as Figure A.14.

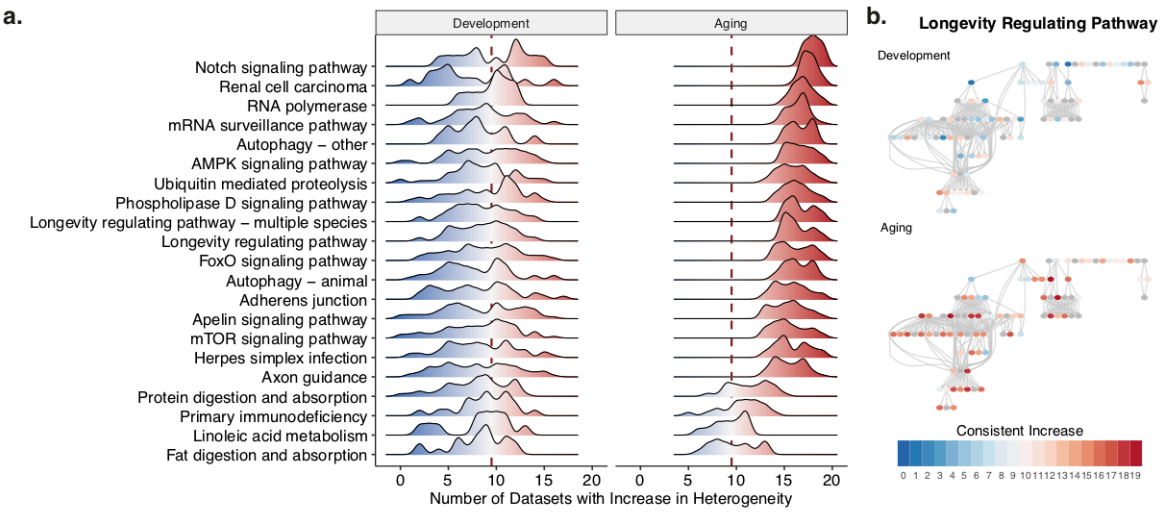


Figure 2.7 *Functional analysis of consistent heterogeneity changes. (a) Distribution of consistent heterogeneity increase for the significantly enriched KEGG pathways, in development and ageing. x- and y-axes show the number of datasets with a consistent increase and the density for each significant pathway, respectively. The dashed red line shows x = 9.5, which is the middle point for 19 datasets, representing no tendency to increase or decrease. Values higher than 9.5, shown with red colour, indicate an increase in heterogeneity while values lower than 9.5, shown with blue colour, indicate a decrease in heterogeneity and the darkness shows the consistency in change across datasets. b) The longevity regulating pathway (KEGG Pathway ID: hsa04211), exemplifying the distribution of the genes (circles), their heterogeneity across datasets (colour – the same colour scheme as panel (a)), and their relationship in the pathway (edges).*

The distribution of consistent heterogeneity in development and ageing also showed a clear difference. The pathway scheme for the longevity regulating pathway (Figure

2.7b), coloured based on the number of datasets with a consistent increase, shows how particular genes compare between development and ageing. Other significantly enriched gene sets, including GO, Reactome, TF and miRNA sets, are included as Supplementary Tables S3-10. In general, while the consistent heterogeneity changes in development did not show any enrichment (except for miRNAs, see Supplementary Table S10), we detected a significant enrichment for the genes that become more heterogeneous during ageing, with the exception that Disease Ontology terms were not significantly associated with consistent changes in either development or ageing. The gene sets included specific categories such as autophagy and synaptic functions as well as broad functional categories such as regulation of transcription and translation processes, cytoskeleton or histone modifications. We also performed GSEA for each dataset separately and confirmed that these pathways show consistent patterns in ageing . There were 30 significantly enriched transcription factors, including EGR and FOXO, and 99 miRNAs (see Supplementary Table S8-9 for the full list). We also asked if the genes that become more heterogenous consistently across datasets are known ageing-related genes, using the GenAge Human gene set (Tacutu et al., 2018), but did not find a significant association (Figure A.15).

It has been reported that the total number of distinct regulators of a gene (apart from its specific regulators) is correlated with gene expression noise (Barroso, Puzovic, & Dutheil, 2018). Accordingly, we asked if the total number of transcription factors (TFs) or miRNAs regulating a gene might be related to the heterogeneity change with age (Figure 2.8). We calculated the correlations between the total number of regulators and the heterogeneity changes and found a mostly positive (18 / 19 for miRNA and 15 / 19 for TFs), and higher correlation between change in heterogeneity and the number of regulators in ageing ($p = 0.007$ for miRNA and $p = 0.045$ TFs). We further tested the association while controlling for the expression changes in development and ageing since regulation of expression changes during development could confound the relationship. However, we found that the pattern is mainly associated with the genes that show a decrease in expression during ageing, irrespective of their expression during development (Figure A.16).

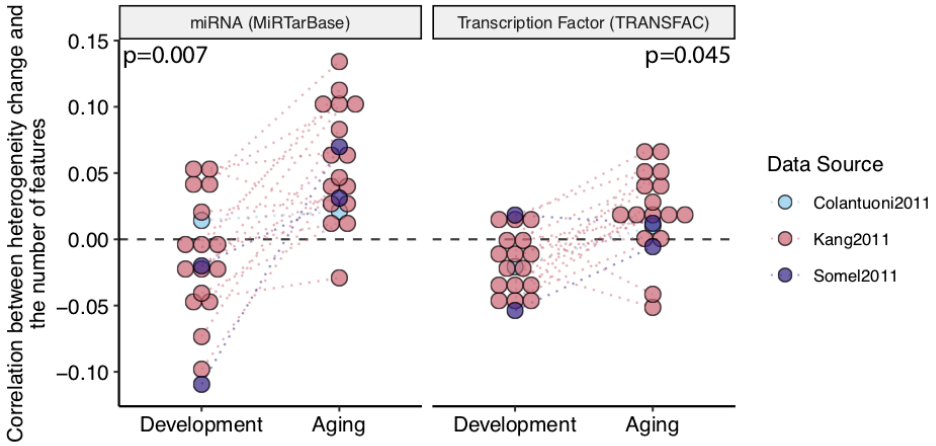


Figure 2.8 *Correlation between the change in heterogeneity and number of transcriptional regulators, i.e. miRNA and transcription factors. Each point represents a dataset, and the colour shows the data source. p-values are calculated using a permutation test. The dashed line at y = 0 shows zero correlation.*

We further tested if genes with a consistent heterogeneity increase in ageing are more central in the protein interaction network using STRING database (Mering et al., 2005). Using multiple cutoffs and repeating the analysis, we observed a higher degree of interactions for the genes with increasing heterogeneity (Figure A.17).

Johnson et al. (2015) previously compiled a list of traits that are age-related and have been sufficiently tested for genome-wide associations (n = 39). Using the genetic associations for GWAS Catalog traits, we tested if there are significantly enriched traits for the consistent changes in heterogeneity during ageing (Supplementary Table S11). Although there was no significant enrichment, all these age-related terms had positive enrichment scores, *i.e.* they all tended to include genes that consistently become more heterogeneous with age during ageing.

Using cell-type specific transcriptome data generated from FACS-sorted cells in mouse brain (Cahoy et al., 2008), we also analysed if there is an association between genes that become heterogeneous with age and cell-type specific genes, which could be expected if brain cell-type composition progressively varied among individuals with age. Although there was an overlap with oligodendrocytes and myelinated oligodendrocytes, there was no significant enrichment (which could be attributed to low power due to small overlap – n= 9 genes - between ageing and cell-type specific expression datasets) (Figure A.18). We further analysed if the cell-type proportions change or become heterogeneous with age, using a deconvolution method. Overall, we see

an increase in myelinated oligodendrocytes in development and astrocytes in ageing (Figure A.19a). Focusing specifically on the 147 genes that become consistently heterogeneous with age, we do not see any consistent pattern in ageing (Figure A.19b), suggesting these genes are not more relevant for certain cell-types. Moreover, we checked if the cell-type proportions, calculated for all genes and only for the 147 genes, become heterogeneous with age during ageing, and found that all cell types seem to become heterogeneous when the heterogeneous gene set analysed, again suggesting their effects are more likely to be broad and not cell-type specific (Figure A.19d).

2.3.6 Methods

Dataset collection

In this study, we performed re-analysis of publicly available transcriptome datasets to test age-related change in gene expression heterogeneity. All data collection in these previous studies were performed in accordance with relevant guidelines, regulations and approved experimental protocols, including informed consents for the use of samples for research from all donors or their next of kin.

Microarray datasets: Raw data used in this study were retrieved from the NCBI Gene Expression Omnibus (GEO) from three different sources (Supplementary Table S1). All three datasets consist of human brain gene expression data generated on microarray platforms. In total, we obtained 1017 samples from 298 individuals, spanning the whole lifespan with ages ranging from 0 to 98 years (Figure A.1).

RNA sequencing dataset: We used the transcriptome data generated by the GTEx Consortium (v6p) (Ardlie et al., 2015). We only used the samples with a death circumstance of 1 (violent and fast deaths due to an accident) and 2 (fast death of natural causes) on the Hardy Scale excluding individuals who died of illnesses. As we focus only on the brain, we used all 13 brain tissue data in GTEx. We thus analysed 623 samples obtained from 99 individuals.

Separating datasets into development and ageing datasets: To differentiate changes in gene expression heterogeneity during ageing from those during development, we used the age of 20 to separate pre-adulthood from adulthood. It was shown that the age of 20 corresponds to the average first age of reproduction in human societies (Walker et al., 2006). Structural changes after the age of 20 in the human brain were

previously linked to age-related phenotypes, specifically neuronal shrinkage and a decline in total length of myelinated fibers (Sowell et al., 2004). Earlier studies examining age-related gene expression changes in different brain regions also showed a global change in gene expression patterns after the age of 20 (Colantuoni et al., 2011; Dönertaş et al., 2017; Somel et al., 2010). Thus, consistent with these studies, we separated datasets using the age of 20 into development (0 to 20 years of age, n = 441) and ageing (20 to 98 years of age, n = 569).

Preprocessing

Microarray datasets: RMA correction (using the 'oligo' library in R) (Carvalho & Irizarry, 2010) and log2 transformation were applied to Somel2011 and Kang2011 datasets. For the Colantuoni2011 dataset, as there was no public R package to analyse the raw data, we used the preprocessed data deposited in GEO, which had been loess normalised by the authors. We quantile normalised all datasets using the 'preprocessCore' library in R (Bolstad, 2019). Notably, our analysis focused on consistent patterns across datasets, instead of considering significant changes within individual datasets. Since we don't expect random confounding factors to be shared among datasets, we used quantile normalisation to minimise the effects of confounders, and we treated consistent results as potentially a biological signal. We also applied an additional correction procedure for Somel2011 datasets, in which there was a batch effect influencing the expression levels, as follows: for each probeset (1) calculate mean expression (M), (2) scale each batch separately (to mean = 0, standard deviation = 1), (3) add M to each value. We excluded outliers given in Supplementary Table S1, through a visual inspection of the first two principal components for the probeset expression levels as in Dönertaş et al. (2017) and Dönertaş, Fuentealba Valenzuela, Partridge, & Thornton (2018). We mapped probeset IDs to Ensembl gene IDs 1) using the Ensembl database, through the 'biomaRt' library (Durinck, Spellman, Birney, & Huber, 2009) in R for the Somel2011 dataset, 2) using the GPL file deposited in GEO for Kang2011, as probeset IDs for this dataset were not complete in Ensembl, and 3) using the Entrez gene IDs in the GPL file deposited in GEO for the Colantuoni2011 dataset and converting them into Ensembl gene IDs using the Ensembl database, through the 'biomaRt' library in R. Lastly, we scaled expression levels for genes (to mean = 0, standard deviation = 1) using the 'scale' function in R. Age values of individuals in each dataset were converted to the fourth root of age (in days) to have a linear relationship between age and expression both in development and ageing.

However, we repeated the analysis using different age scales and confirmed that the results were quantitatively similar (Figure A.3).

RNA sequencing dataset: The genes with median RPKM value of 0 were excluded from the dataset. The RPKM values provided in the GTEx data were log2 transformed and quantile-normalised. Similar to the microarray data, we excluded the outliers based on the visual inspection of the first and second principal components (Supplementary Table S1). In GTEx, ages are given as 10 year intervals. We therefore used the middle point of these age intervals to represent that individual's age.

Age-related expression change

We used linear regression to assess the relationship between age and gene expression. The model used in the analysis is:

$$Y_i = \beta_{i_0} + \beta_{i_1} * Age^{1/4} + \epsilon_i \tag{2.1}$$

where Y_i is the scaled log2 expression level for the i^{th} gene, β_{i_0} is the intercept, β_{i_1} is the slope, and ϵ_i is the residual. We performed the analysis for each dataset (development and ageing datasets separately) and considered the β_1 value as a measure of change in expression. p-values obtained from the model were corrected for multiple testing according to Benjamini and Hochberg procedure (Benjamini & Hochberg, 1995) by using 'p.adjust' function in R.

Age-related heterogeneity change

In order to quantify the age-related change in gene expression heterogeneity, we calculated Spearman's correlation coefficient (ρ). The correlations were calculated between the absolute values of residuals obtained from equation (2.1) and the fourth root of individual age. We regarded the absolute values of residuals as a measure of heterogeneity. Therefore, high positive correlation coefficients suggest that heterogeneity increases with age, whereas strong negative correlation implies heterogeneity decreases with age. p-values were calculated from the correlation analysis and corrected for multiple testing with the Benjamini and Hochberg method using the 'p.adjust' function in R. We further confirmed our results using a different measure of hetero-

geneity, Breusch-Pagan test, implemented as 'ncvTest' function in the 'car' package in R (Fox & Weisberg, 2019) (Figure A.27).

Principal Component Analysis

We conducted principal component analysis on both age-related changes in expression (β) and heterogeneity (ρ). We followed a similar procedure for both analyses, in which we used the 'prcomp' function in R. The analysis was performed on a matrix containing β values (for the change in expression level) and ρ values (for the change in heterogeneity), for 11,137 commonly expressed genes for all 38 development and ageing datasets. In each dataset, the estimates of expression change (β) or heterogeneity change (ρ) values were scaled for each dataset before calculating principal components.

Permutation test

We used a permutation scheme that we developed earlier (Dönertaş et al., 2018, 2017), taking into account that samples across Somel2011 and Kang2011 datasets are not independent (*i.e.* these datasets include multiple samples from the same individuals for different brain regions). Specifically, we first randomly permuted ages among individuals, not samples, for 1,000 times in each data source, using the 'sample' function in R. Next, we assigned ages of individuals to their corresponding samples, making sure that multiple samples from the same individual annotated with the same age, retaining the dependency between datasets. Then, we calculated age-related expression and heterogeneity change for each dataset, using permuted ages. For the tests related to the changes in gene expression with age, we used a linear model between gene expression levels and the randomised ages. In contrast, for the tests related to the changes in heterogeneity with age, we measured the correlation between the randomised ages and the absolute value of residuals from the linear model that we obtained from equation (2.1) using non-randomised ages for each gene. In this way, we preserved the relationship between age and expression, and we were able to ensure that our regression model was appropriate for calculating age-related heterogeneity change (Somel et al., 2006). Using expression and heterogeneity change estimates calculated using permuted ages, we tested (a) if the correlation of expression (and heterogeneity) change in ageing is higher than in development datasets; (b) if the correlations of expression (and heterogeneity) change in

development and in ageing datasets are significantly higher than null expectation; (c) if the number of genes showing significant change in expression (and heterogeneity) is significantly higher in ageing than in development datasets; (d) if the overall increase in age-related heterogeneity during ageing is significantly higher than development; (e) if the observed consistency in heterogeneity increase is significantly different from expected. All tests using permuted ages were performed one-tailed. We also demonstrate that our permutation strategy is more stringent than random permutations in Figure A.10, giving the distributions calculated using both dependent permutations and random permutations.

To test the overall correlation within development or ageing datasets for the changes in expression (β) and heterogeneity (ρ), we calculated median correlations among independent three subsets of datasets (one Kang2011, one Somel2011 and the Colantuoni2011 dataset), taking the median value calculated for each possible combination of independent subsets (16 x 2 x 1 = 32 combinations). Using 1,000 permutations of individuals' ages, we generated an expected distribution for the median correlation coefficient for triples and compared these with the observed values, asking how many times we observe a higher value. We used this approach to calculate expected median correlation among development (and ageing) datasets, because the number of independent pairwise comparisons are outnumbered by the number of dependent pairwise comparisons, causing low statistical power.

To further test the significance of the difference between correlations among development and ageing datasets, we calculated the median difference in correlations between ageing and development datasets for each permutation. We next constructed the null distribution of 1,000 median differences and calculated empirical p-values comparing the observed differences with these null distributions. Next, to test the significance of the difference in the number of significantly changing genes between development and ageing, we calculated the difference in the number of genes showing significant change between development and ageing datasets for each permutation. Empirical p-values were computed according to observed differences. Likewise, to test if the overall increase in age-related heterogeneity during ageing is significant compared to development, we computed median differences between median heterogeneity change values of each ageing and development dataset, for each permutation, followed by an empirical p-value calculation to answer if the ageing datasets have a higher increase in age-related heterogeneity.

Expected heterogeneity consistency

Expected consistency in heterogeneity change was calculated from heterogeneity change values (ρ) measured using permuted ages. For each permutation, we first calculated the total number of genes showing consistent heterogeneity increase for N number of datasets (N = 0, …, 19). To test if observed consistency significantly differed from the expected, we compared observed consistency values to the distribution of expected numbers, by performing a one-sided test for the consistency in N number of datasets, N = 1, …, 19.

Clustering

We used the k-means algorithm ('kmeans' function in R) to cluster genes showing consistent heterogeneity change (n=147) according to their heterogeneity profiles. We first took the subset of the heterogeneity levels (absolute value of the residuals from equation (2.1)) to include only the genes that show a consistent increase with age and then scaled the heterogeneity levels to the same mean and standard deviation. Since the number of samples in each dataset is different, just running k-means on the combined dataset would not equally represent all datasets. Thus, we first calculated the spline curves for scaled heterogeneity levels for each gene in each dataset (using the 'smooth.spline' function in R, with three degrees of freedom). We interpolated at 11 (the smallest sample size) equally distant age points within each dataset. Then we used the combined interpolated values to run the k-means algorithm with k = 8, a liberal choice, given the total number of genes being 147.

To test association of the clusters with Alzheimer's Disease, we retrieved overall AD association scores of the 147 consistent genes (n = 40) from the Open Targets Platform (Carvalho-Silva et al., 2019).

Functional Analysis

We used the "clusterProfiler" package in R to run Gene Set Enrichment Analysis (GSEA), using Gene Ontology (GO) Biological Process (BP), GO Molecular Function (MF), GO Cellular Compartment (CC), Reactome, Disease Ontology (DO), and KEGG Pathways. We used Normalised Enrichment Scores (NES) from ClusterProfiler package, which gives a weighted Kolmogorov-Smirnov (KS) test statistic divided by the

mean of KS statistics obtained from random permutations. We performed GSEA on all gene sets with a size between 5 and 500, and we corrected the resulting p-values with the Benjamini and Hochberg correction method. To test if the genes with a consistent increase or decrease in their expression are associated with specific functions, we used the number of datasets with a consistent increase to run GSEA. Since we are running GSEA using number of datasets showing consistency, our data includes many ties, potentially making the ranking ambiguous and non-robust. In order to assess how robust our results are, we ran GSEA 1,000 times on the same data and counted how many times we observed the same set of KEGG pathways as significant (Supplementary Table S3). The lowest number among the pathways with a significant positive enrichment score was 962 out of 1,000 (Phospholipase D signaling pathway). Moreover, we repeated the same analysis using the heterogeneity change levels (ρ), instead of using the number of datasets with a consistent change, for each dataset to confirm the gene sets are indeed associated with the increase or decrease in heterogeneity . We visualised the KEGG pathways using 'KEGGgraph' library in R and coloured the genes by the number of datasets that show an increase.

We also performed an enrichment analysis of the transcription factors and miRNA to test if specific TFs or miRNAs regulate the genes that become more heterogeneous consistently. We collected gene-regulator association information using the Harmonizome database (Rouillard et al., 2016), 'MiRTarBase microRNA Targets' (12086 genes, 596 miRNAs) and 'TRANSFAC Curated Transcription Factor Targets' (13216 genes, 201 TFs) sets. We used the 'fgsea' package in R, which allows GSEA on a custom gene set. We tested the association for each regulator with at least 10 and at most 500 targets. Moreover, we tested if the number of regulators is associated with the change in heterogeneity. We first calculated the correlation between heterogeneity change with age (or the number of datasets with an increase in expression heterogeneity) and the number of TFs or miRNAs regulating that gene, for ageing and development separately. We repeated the analysis while accounting for the direction of expression changes in these periods (*i.e.* separating genes into down-down, down-up, up-down, and up-up categories based on their expression in development and ageing, Supplementary Figure A.16). To test the difference in the correlations between ageing and development, we used 1,000 random permutations of the number of TFs. For each permutation, we randomised the number of TFs and calculated the correlation between heterogeneity change (or the number of datasets with an increase in heterogeneity) and the randomised numbers. We then calculated the percentage

of datasets where ageing has a higher correlation than development. Using the distribution of percentages, we tested if the observed value is expected by chance.

Protein-protein interaction network analysis

We downloaded all human protein interaction data from the STRING database (v11) (Mering et al., 2005). Ensembl Peptide IDs are mapped to Ensembl Gene IDs using the 'biomaRt' package in R. Here we aimed to test whether genes showing consistent increase in heterogeneity have a different number of interactors than other genes. For this we calculated the degree distributions for the genes that become consistently more heterogeneous with age and all remaining genes using different cutoffs for interaction confidence scores. In order to calculate the significance of the difference, we i) calculated the number of interactors (degree) for each gene, ii) for 10,000 times, randomly sampled k genes from all interactome data (k = number of genes that become heterogeneous with age across all datasets and have interaction information in STRING database, after filtering for a cutoff), iii) calculated the median of degree for each sample. We then calculated an empirical p-value by asking how many of these 10,000 samples we see a median degree that is equivalent to or higher than our original value. The number of genes and interactions after each cutoff are given in Figure A.17.

Cell-type specificity analysis

Using FACS-sorted cell-type specific transcriptome data from the mouse brain (Cahoy et al., 2008), we checked if there is any overlap between genes that become heterogeneous with age and cell-type specific genes. We downloaded the dataset from the GEO database (GSE9566) and preprocessed it by performing: i) RMA correction using the 'affy' package in R (Gautier, Cope, Bolstad, & Irizarry, 2004), ii) log2 transformation, iii) quantile normalisation using the 'preprocessCore' package in R (Bolstad, 2019), iv) mapping probeset IDs to first mouse genes, and then human genes. We only included genes that have one to one orthologs in humans, after filtering out probesets that map to multiple genes.

We first defined the cell-type specific genes by calculating an effect size (Cohen's D) for each gene and cell type and identifying genes that have an effect size higher than or equal to 2 as specific to that cell type. At this cutoff, there was no overlap between

cell type specific gene lists. To test for association between heterogeneity and cell type specificity, we used the Fisher's exact test using the R 'fisher.test' function.

Following another approach, we analysed the relative cell-type contribution to the expression profiles and calculated both how the level and heterogeneity of the relative contributions change with age. This analysis requires positive expression values and since the Colantuoni2011 dataset is loess-normalised, we only analysed Kang2011 and Somel2011 datasets. To determine the relative contribution of different cell-types we used a linear regression based approach, where the contributions are determined based on the following formula:

$$Expr = \alpha + \beta_A \zeta_A + \beta_{Oli} \zeta_{Oli} + \beta_{MOli} \zeta_{MOli} + \beta_{OPC} \zeta_{OPC} + \beta_N \zeta_N + \epsilon \tag{2.2}$$

Where; A: Astrocytes, Oli: Oligodendrocytes, M_Oli: Myelinated Oligodendrocytes, OPC: Oligodendrocyte Precursor Cells, N: Neurons. β's represent the regression coefficients estimate the relative contributions of each cell type, ζ's represent expression level of each cell type (averaged across replicates), and ϵ represents residuals.

Effect of sex-specific gene expression on heterogeneity

To test the effect of sex on increased heterogeneity during ageing, we first obtained residuals from equation (2.1) for 147 consistent genes. Then, we performed two sample Wilcoxon test, using 'wilcox.test' function in R, to test if the distribution of residuals differs significantly between sexes. As we used the real values of residuals here (not absolute values), we would expect a significant difference in their distribution if the expression change is in different directions between sexes. p-values obtained from Wilcoxon test for each gene in each dataset separately were corrected for multiple testing with the Benjamini and Hochberg method using the 'p.adjust' function in R. The number of genes showing a significant difference, and their consistency among ageing datasets are shown in Figure A.23.

Code Availability

All analysis was performed using R and the code to calculate heterogeneity changes with age is available as an R package 'hetAge', documented at https://mdonertas.

github.io/hetAge/. ‘ggplot2’ (Wickham, 2017), ‘ggpubr’ (Kassambara, 2018), and ‘pheatmap’ (Kolde, 2019), R libraries were used for the visualisation.

2.4 Discussion

Ageing is characterised by a gradual decrease in the ability to maintain homeostatic processes, which leads to functional decline, age-related diseases, and eventually to death. This age-related deterioration, however, is thought as not a result of expression changes in a few individual genes, but rather as a consequence of an age-related alteration of the whole genome, which could be a result of an accumulation of both epigenetic and genetic errors in a stochastic manner (Enge et al., 2017; Vijg, 2004). This stochastic nature of ageing impedes the identification of conserved age-related changes in gene expression from a single dataset with a limited number of samples.

In this study, we examined 19 gene expression datasets compiled from three independent studies to identify the changes in gene expression heterogeneity with age. While all datasets have samples representing the whole lifespan, we separated postnatal development (0 to 20 years of age) and ageing (20 to 98 years of age) by the age of 20, as this age is considered to be a turning-point in gene expression trajectories (Dönertaş et al., 2017). We implemented a regression-based method and identified genes showing a consistent change in heterogeneity with age, during development and ageing separately. At the single gene level, we did not observe significant age-related heterogeneity change in most of the datasets, possibly due to insufficient statistical power due to small sample sizes and the subtle nature of the phenomenon. We hence took advantage of a meta-analysis approach and focused on consistent signals among datasets, irrespective of their effect sizes and significance. Although this approach fails to capture patterns that are specific to individual brain regions, it identifies genes that would otherwise not pass the significance threshold due to insufficient power. Furthermore, we demonstrated that our method is robust to noise and confounding effects within individual datasets.

By analysing age-related gene expression changes, we first observed that there are more significant and more similar changes during development than ageing. Additionally, genes showing significant change during ageing tended to decrease in expression (Figure 2.3). These results can be explained by the accumulation of stochastic detrimental effects during ageing, leading to a decrease in expression levels (Lu et

al., 2004). Our initial analysis of gene expression changes suggested a higher heterogeneity between ageing datasets.

We next focused on age-related heterogeneity change between individuals and found a significant increase in age-related heterogeneity during ageing, compared to development. Notably, increased heterogeneity is not limited to individual brain regions, but a consistent pattern across different regions during ageing. We found that age-related heterogeneity change is more consistent among ageing datasets, which may reflect an underlying systemic mechanism. Further, a larger number of genes showed more significant heterogeneity changes during ageing than in development, and the majority of these genes tended to have more heterogeneous expression.

It was previously proposed that somatic mutation accumulations (Lodato et al., 2018; Lombard et al., 2005; Lu et al., 2004; Vijg, 2004) and epigenetic regulations (Cheung et al., 2018) might be associated with transcriptome instability. While Enge et al. (2017) and Lodato et al. (2018) suggested that genome-wide substitutions in single cells are not so common as to influence genome stability and cause transcriptional heterogeneity at the cellular level, epigenetic mechanisms may be relevant. Although we cannot test age-related somatic mutation accumulation and epigenetic regulation in this study, an alternative mechanism might be related to transcriptional regulation, which is considered to be inherently stochastic (Maheshri & O'Shea, 2007). Several studies demonstrated that variation in gene expression is positively correlated with the number of TFs controlling gene's regulation (Barroso et al., 2018). We also found that genes with a higher number of regulators and a decrease in expression during ageing become more heterogeneous. Further, significantly enriched TFs include early growth response (EGF), known to be regulating the expression of many genes involved in synaptic homeostasis and plasticity, and FOXO TFs, which regulate stress resistance, metabolism, cell cycle arrest and apoptosis. Together with these studies, our results support that transcriptional regulation may be associated with age-related heterogeneity increase during ageing and may have important functional consequences in brain ageing.

We next confirmed that observed increase in heterogeneity was not a result of low statistical power (Figure A.1) or a technical artefact (Figure 2.5b, A.19, A.20). Specifically, we tested whether increased heterogeneity during ageing can be a result of the mean-variance relationship, but we found no significant effect that can confound our results. In fact, the mean-variance relationship in development and ageing showed opposing profiles. We further analysed this by grouping genes based on their expres-

sion in development and ageing (Figure A.19). The genes that decrease in expression both in development and ageing showed the most opposing profiles in terms of the mean-variance relationship, which could suggest that the decrease in development are more coordinated and well-regulated whereas the decrease in ageing occurs due to stochastic errors. Another potential confounder is the post-mortem interval (PMI), which is the time between death and sample collection. Since we do not have this data for all datasets we analysed, we could not account for PMI in our model. However, using the list of genes previously suggested as associated with PMI (Zhu, Wang, Yin, & Yang, 2017), we checked if the consistency among ageing datasets could be driven by PMI. Only 2 PMI-associated genes were among the 147 that become consistently heterogeneous, and the distribution also suggested there is no significant relationship (Figure A.20). We also confirmed that the increase in heterogeneity is not caused by outlier samples in datasets (Figure A.21) or by the confound of sex with age (Figure A.22). Moreover, we asked if differential expression trajectories between different sexes may contribute to the observed heterogeneity. Specifically, we tested if the distribution of residuals differs between males and females using 147 consistent genes. We observed that only 15 of 147 genes show a significant difference in at least one dataset (Figure A.23).

Gene set enrichment analysis of the genes with increased heterogeneity with age revealed a set of significantly enriched pathways that are known to influence ageing, including longevity regulating pathway, autophagy, mTOR signaling pathway (Figure 2.7a). Furthermore, GO terms shared among these genes include some previously identified common pathways in ageing and age-related diseases . We have also tested if these genes are associated with age-related diseases through GWAS, and although not significant, we found a positive association with all age-related traits defined in Johnson et al. (2015). Overall, these results indicate the effect of heterogeneity on pathways that modulate ageing and may reflect the significance of increased heterogeneity in ageing. Importantly, we identified genes that are enriched in terms related to neuronal and synaptic functions, such as axon guidance, neuron to neuron synapse, postsynaptic specialisation, which may reflect the role of increased heterogeneity in synaptic dysfunction observed in the mammalian brain, which is considered to be a major factor in age-related cognitive decline (Morrison & Baxter, 2012). We also observed genes that become more heterogeneous with age consistently across datasets are more central (*i.e.* have a higher number of interactions) in a protein-protein interaction network (Figure A.17). Although this could mean the effect of het-

erogeneity could be even more critical because it affects hub genes, another explanation is research bias that these genes are studied more than others.

2.4.1 Limitations

One important limitation of our study is that we analyse microarray-based data. Since gene expression levels measured by microarray do not reflect an absolute abundance of mRNAs, but rather are relative expression levels, we were only able to examine relative changes in gene expression. A recent study analysing single-cell RNA sequencing data from the ageing Drosophila brain identified an age-related decline in total mRNA abundance (Davie et al., 2018). It is also suggested that, in microarray studies, genes with lower expression levels tend to have higher variance (Aris et al., 2004). In this context, whether the change in heterogeneity is a result of the total mRNA decay is an important question. As an attempt to see if the age-related increase in heterogeneity is dependent on the technology used to generate data, we repeated the initial analysis using RNA sequencing data for the human brain, generated by GTEx Consortium (Ardlie et al., 2015) (Figures A.24, A.25, A.26). Nine out of thirteen datasets displayed more increase than decrease in heterogeneity during ageing, consistent with 18/19 microarray datasets, while the remaining four datasets showed the opposite pattern (BA24, cerebellar hemisphere, cerebellum and substantia nigra). Unlike what we observed for the microarray datasets, the change in expression levels and heterogeneity were strongly positively correlated (Figure A.26). Unfortunately, average expression levels and variation levels in RNA sequencing is challenging to disentangle. Thus, the biological relevance of the relationship between the age-related change in expression levels and expression heterogeneity still awaits to be studied through novel experimental and computational approaches. Nevertheless, RNA sequencing analysis also suggests an overall increase in age-related heterogeneity.

It should also be noted that these results may be biased toward the cortex region, as the independent data sources mostly included samples from the cortex region.

Another limitation is related to use of bulk RNA expression datasets, where each value is an average for the tissue. While it is important to note that our results indicate increased heterogeneity between individuals rather than cells, the fact that the brain is composed of different cell types raises the question if increased heterogeneity may be a result of changes in brain cell-type proportions. To explore the association between

heterogeneity and cell-type specific genes, we used FACS-sorted cell type specific transcriptome dataset from mouse brain (Cahoy et al., 2008). We only had nine genes that have consistent heterogeneity increase and are specific to one cell-type. Eight out of nine were highly expressed in oligodendrocytes. However, we did not observe any significant association between cell-type specific genes and heterogeneity (Figure A.18, A.28).

2.4.2 Conclusion

In the first study (Section 2.2), our comparison showed that the correction strategy plays a pivotal role in identifying the specific set of differentially variable (DV) genes. However, irrespective of the approach and correction method used, a transcriptome-wide increase in the gene expression heterogeneity was observed, *i.e.* more genes showed a tendency to increase than to decrease expression heterogeneity with age. We also showed that most of the functional processes were susceptible to the ageing-related increase in the expression heterogeneity.

In our second study (Section 2.3), by performing a meta-analysis of transcriptome data from diverse brain regions we found a significant increase in gene expression heterogeneity during ageing, compared to development. Increased heterogeneity was a consistent pattern among diverse brain regions in ageing, while no significant consistency was observed across development datasets. Our results support the view of ageing as a result of stochastic molecular alterations, whilst development has a higher degree of gene expression regulation. Similar to our first study, we observed a widespread heterogeneity in genes and functional categories. However, the genes showing a *consistent* increase in heterogeneity during ageing are involved in pathways important for ageing and neuronal function. Therefore, our results demonstrate that increased heterogeneity is one of the characteristics of brain ageing and is unlikely to be only driven by the passage of time starting from developmental stages.

Chapter 3

The link between ageing and age-related diseases

Data Availability

*Supplementary Tables are available to download **here**. At the moment the data is not publicly available but requires a permission which is granted case by case. All supplementary tables and summary statistics will be publicly available in the BioStudies database under accession number S-BSST377 upon publication of this work. Data tables are referred with the corresponding file names throughout the text. This research has been conducted using the UK Biobank Resource (application no. 30688).*

3.1 Introduction

Risk of many age-related diseases is influenced by genetic variation. Genome-wide association studies (GWAS) have identified genetic variants that alter complex traits. Pleiotropy, where variants or genes influence multiple traits, is more prevalent than previously thought (Bulik-Sullivan et al., 2015; Cortes, Albers, Dendrou, Fugger, & McVean, 2020; Cross-Disorder Group of the Psychiatric Genomics Consortium et al., 2013; Pickrell et al., 2016), and it indicates that different traits share common causal pathways (Solovieff, Cotsapas, Lee, Purcell, & Smoller, 2013). Pleiotropy within the disease classification system (Cortes et al., 2020) and in certain disease classes, such as immune-related diseases (Ellinghaus et al., 2016; Parkes, Cortes, Heel, & Brown, 2013) and cancer (Bien & Peters, 2019), have been studied, but understanding of pleiotropy in age-related diseases (ARD) more broadly is limited. Johnson et al.

(2015) manually curated a list of 39 ageing-related biomarkers and diseases, collected their genetic associations from previous studies, and asked if they share commonalities. Although the number of genes implicated in diverse traits was low, they found that traits shared common pathways, such as NSN (*i.e.* nutrient-sensing network) and proteostasis (Johnson et al., 2015). Other studies aimed to understand if ageing and age-related diseases share common genes. Specifically testing if known ageing-related genes and their interactors are implicated in late-onset diseases, they showed significant but limited number of genes (Fernandes et al., 2016; Wang, Zhang, Wang, Chen, & Zhang, 2009). Despite the challenges of combining data from published studies, these studies provided the first hints that at least some late-onset diseases share common pathways (Johnson et al., 2015), and they are related to some ageing-related genes (Fernandes et al., 2016).

Genome-wide germline variations that increase the risk of diseases in old age may not be pruned by natural selection, or may be associated with beneficial phenotypes earlier in life. Indeed, comparing genetic associations in the GWAS-Catalog (Buniello et al., 2019), Rodríguez et al. (2017) found support for both theories.

3.1.1 Research objectives

In this chapter, I present the investigation of causal connections between different ARDs by clustering them according to their age-of-onset, and determining whether the resulting clusters share common genetic risk variants. The UK Biobank (UKBB) (Bycroft et al., 2018) includes genetic and health-related data for almost half a million participants. We extracted age-of-onset profiles for 116 diseases and identified unbiased clusters to define the relationship between disease incidence and age. We identified variants associated with each disease and compared the genetic associations between diseases based on the clusters. In particular, we aimed to answer the following questions:

- Do diseases follow similar age-of-onset profiles?
- Are diseases with the same age-of-onset profile genetically similar? Can this be explained by co-morbidity and disease cause-effect relationships, or does it suggest a common aetiology?
- Does the genetic component shared in late-onset diseases overlap with known ageing-related genes and pathways?
- Is the shared component druggable?

- Do variants associated with different age-of-onset profiles support evolutionary theories of ageing?

3.2 Exploratory analysis of the UKBB for ageing and age-related disease research

In this study, we used self-reported diseases and age-at-diagnoses in the UKBB (Bycroft et al., 2018). Using the samples that pass quality control (n = 484,598), we first performed an exploratory analysis using the basic demographics, disease data, and ageing-related data fields.

There were more females (n=262,758) than males (n = 221,840) (Figure 3.1a). The age range of participants during the first visit was between 37 (minimum age of males = 37, females = 39) to 73 (maximum age of males = 73, females = 71) with a median value of 58 (median age of males = 58, females = 57) (Figure 3.1b). There were 13,697 participants who died after participating in the study and the death rate was higher in males (Figure 3.1c). As expected, height, weight, and BMI also differed in females and males (Figure 3.1d-f).

Participants were also asked how they rate their health, how satisfied they are with their health, smoking status, alcohol drinker status, if other people generally say they look i) younger than they are, ii) about their age, or iii) older than they are, and if they had a close relative who had non-accidental sudden death. Overall, more people rated their health high and were happy with their health (Figure A.29a-b). Most of the UKBB participants either never smoked or were previous smokers (Figure A.29c) and are current alcohol drinkers (Figure A.29d). Most of the participants also reported that people generally think they are either younger than their age or about the same age (Figure A.29e). Most of the participants did not have any close relatives who died suddenly from non-accidental causes (Figure A.29f).

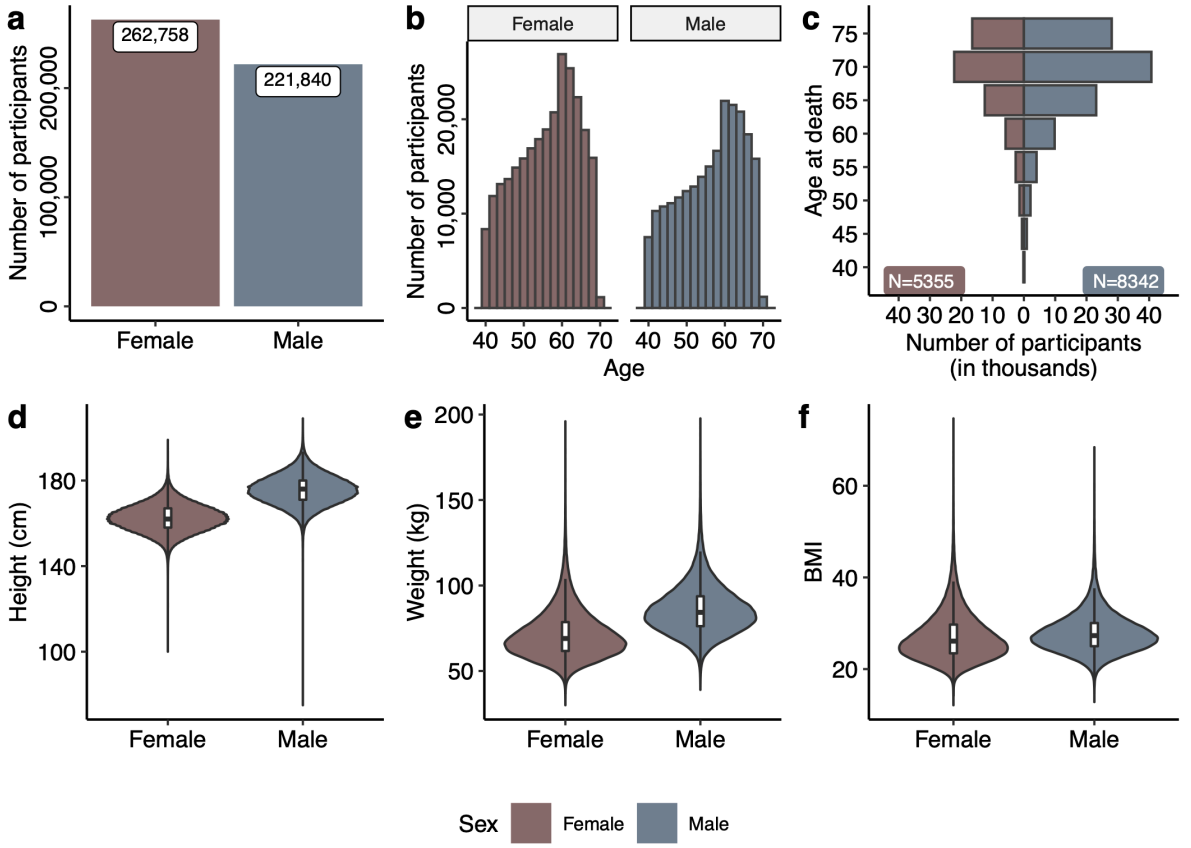


Figure 3.1 *Participant data in the UKBB after quality control steps. a) The number of female and male participants, b) Age distribution when participants first attended the UKBB assessment centre and answered self-reported questions, c) Age at death (every 5 years are binned together) for the participants who died after attending the UKBB assessment centre. The values are corrected for the number of female and male participants who passed the ages specified in the y-axis, d) Distribution of 'standing height' field in the UKBB, e) Distribution of 'weight' field in the UKBB, f) Distribution of BMI field calculated using 'standing height' and 'weight' fields in the UKBB.*

We also checked the distribution of other ageing-related fields, namely parents' age at death, age at menarche, and age at menopause. There were 391,842 participants whose at least one parent is dead. The distribution of age at death for parents was wide (10 to 117), but the majority of the data (between the first and third quantiles) lie between 65 and 79.5 (average age at death) (Figure A.30a). The age at menarche differed between 5 and 25, with a median of 13 (Figure A.30b). The age at menopause differed between 18 and 68, with a median of 50 (Figure A.30c).

The number of self-reported operations ranged between 1 to 32, with a median of 1 and the number of self-reported medications ranged between 0 to 48, with a median

of 2 (Figure A.31a). Among 39,910 participants with cancer, most of them had only one cancer, while there was also a participant with 6 cancers (Figure A.31b).

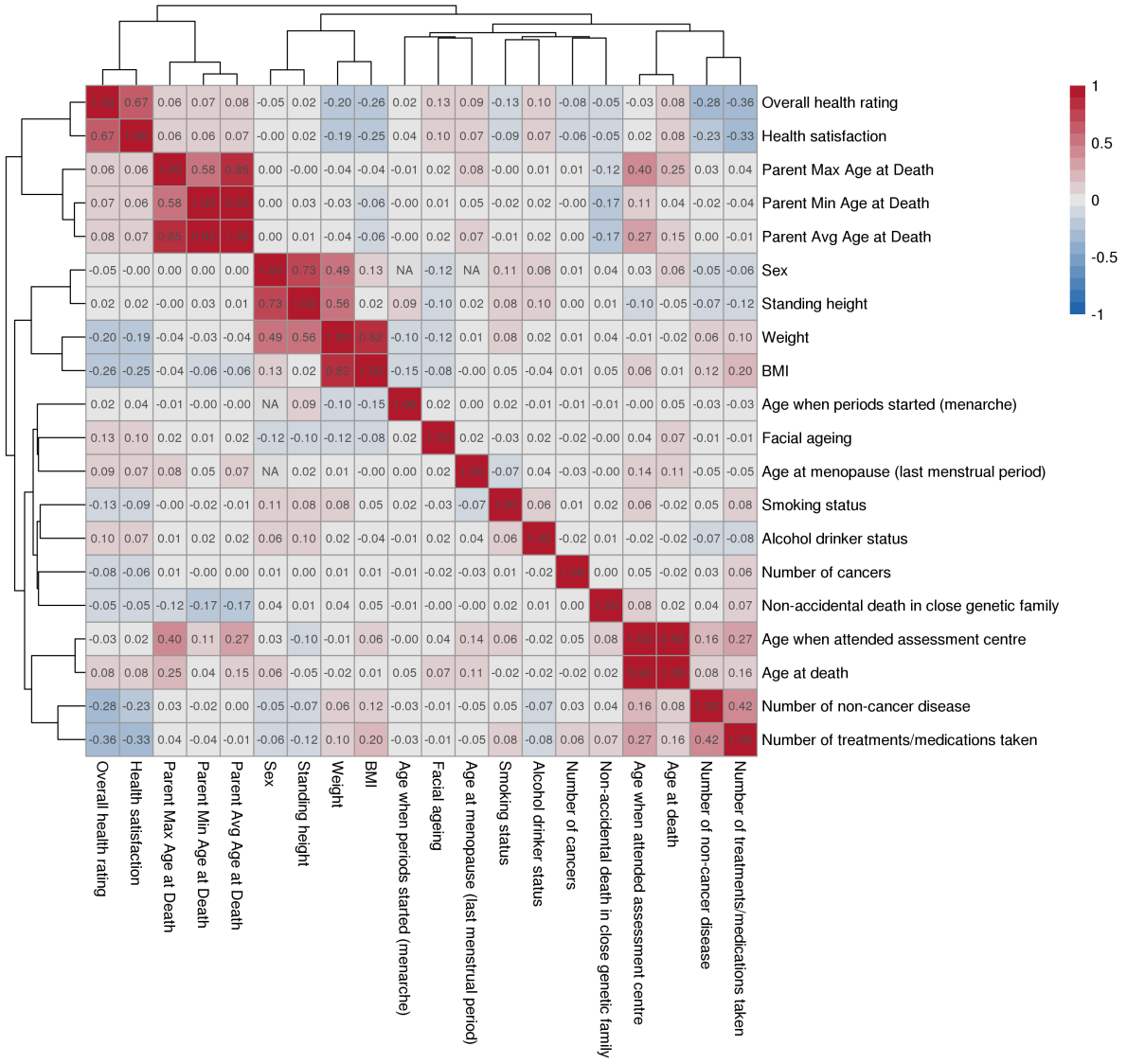


Figure 3.2 *Pairwise correlations between traits. Each row and column show a trait in the UKBB, and the colour shows the pairwise Spearman correlation coefficients between traits. Dark red denotes a strong positive correlation, while darker blue indicates strong negative values. Traits are ordered based on the hierarchical clustering of the correlation coefficients.*

We then checked the correlations between these traits (Figure 3.2). As expected, age when attended assessment centre was very strongly correlated with age at death. It also showed a correlation with parental age at death, the number of non-cancer diseases, and the number of medications taken. 'Overall health rating' and 'health sat-

isfaction' were also correlated with the number of diseases and medications. Moreover, these values also show a correlation to BMI and weight. While 'sex' and 'standing height' are correlated with 'weight', they are not correlated with 'BMI' and 'overall health rating' which are both correlated with 'weight'. BMI is also correlated with 'number of medications taken'.

3.3 Self-reported diseases in the UKBB

The UKBB includes disease information from two sources: i) disease ICD10/9 codes based on hospital episode statistics (HES) and ii) the self-reported (SR) diseases. We analysed the SR diseases as the participants also inform the age at diagnosis. And since they report the diseases they had at earlier ages as well, this data is less biased by the age distribution of the UKBB participants. Moreover, a previous study using the UKBB suggested that GWAS using self-reported diseases and ICD-10 codes were sufficiently similar (Cortes, Dendrou, Fugger, & McVean, 2018).

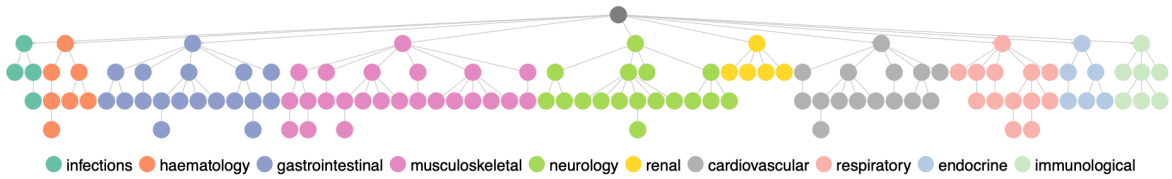


Figure 3.3 *Disease hierarchy for the 116 diseases included in the analysis. The nodes are coloured by the disease categories as indicated in the legend.*

Like ICD10 codes, the UKBB SR Diseases are defined in a hierarchical structure (Figure 3.3). This tree is constructed by the UKBB nurses and it mostly reflects the system or the tissue that disease is most symptomatic in. Participants enter SR disease data with a trained nurse, who guides them. However, we observed that some participants did not consider the disease hierarchy while some did. For example, some patients having 'essential hypertension' also reported having 'hypertension' which is the parent node, while some did not. In order not to bias data, we propagated disease data towards upper levels, so that a participant with a disease at a lower level is always annotated with the connected nodes at upper levels.

Importantly, we only considered 116 non-cancer diseases with at least 2,000 cases and that are not sex-specific. Although we exclude single-sex diseases, we included

the ones that are more prevalent in females (thyroid problem, hypothyroidism, bone disorder, osteoporosis) or in males (abdominal hernia, gout, heart attack) (Figure 3.4).

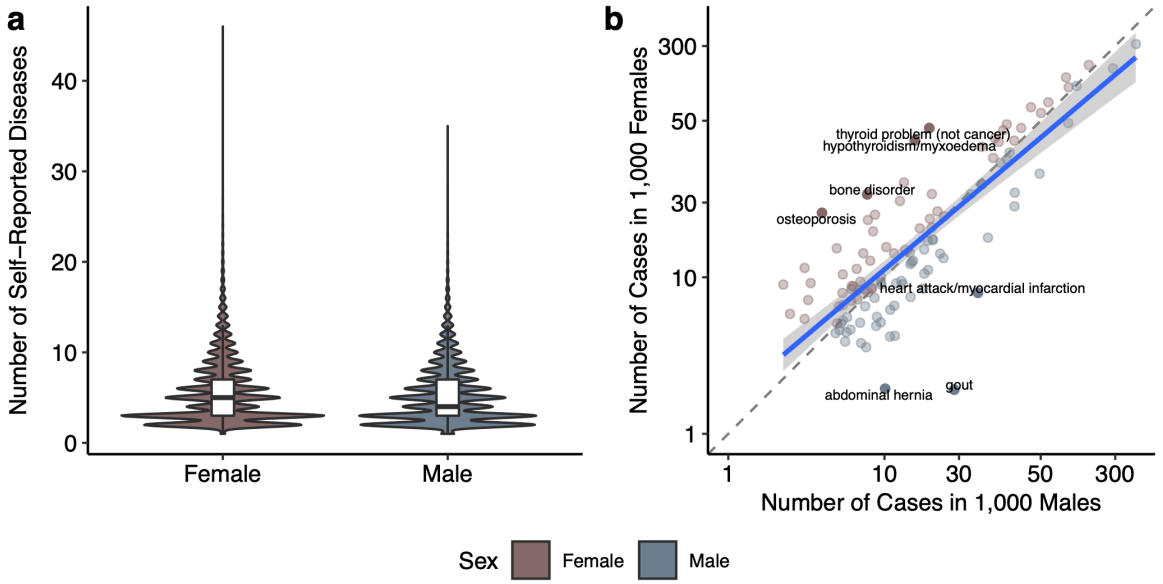


Figure 3.4 *Sex-stratified statistics for 116 selected diseases. (a) The distribution of the number of self-reported diseases (y-axis) stratified by sex (x-axis). (b) The distribution of disease prevalence in males and females. The x- and y-axes show the number of cases in 1,000 males and females (in log scale), respectively. The colour of each point denotes if a disease has a higher prevalence in females (rosy brown, above the dashed line) or males (slate grey, below the dashed line). The linear regression line is depicted as blue. Diseases having a residual value bigger than 3 standard deviations are labeled but not excluded as they are also common in the other sex.*

3.4 Disease co-occurrences

We next calculated disease co-occurrences, using relative risk score to calculate associations and ϕ values as a measure of robustness (Gutiérrez-Sacristán et al., 2018; Jiang, Ma, Shia, & Lee, 2018; Park, Lee, Christakis, & Barabási, 2009; Sanchez-Valle et al., 2018).

Figure 3.5 *Disease association matrix summarizing relative risk scores and correlations. Each row and column denote diseases, ordered by hierarchical clustering of risk scores. The colour is defined by relative risk scores while the size is determined by ϕ value, indicating the robustness. The diagonal tiles are coloured by the UKBB's disease hierarchy to visualise if diseases from the same category cluster together. Associations for the 62 diseases that have at least one relative risk ratio higher than four or lower than minus four ($|log_2 RR| \leq 2$) are plotted.*

There were five major clusters with high relative risk scores and robustness and they seemingly cluster by disease categories: i) musculoskeletal/trauma diseases and early-onset gastrointestinal diseases such as appendicitis, ii) other musculoskele-

tal/trauma diseases such as sciatica and disc problems, iii) respiratory/ENT diseases such as bronchitis and pneumonia, iv) cardiovascular diseases and diabetes, v) retinal problem, glaucoma, and cataract (Figure 3.5). While most of these clusters are biologically plausible, some could be explained by reporting bias, *e.g.* bone fractures and appendicitis. Moreover, we saw a strong negative correlation between osteoarthritis and arthritis (nos). The disease 'arthritis (nos)' does not include osteoarthritis by definition (nos = not osteoarthritis) and seeing this association suggests that we can detect co-occurrences reliably.

3.5 Age-of-onset clusters

In order to characterise diseases based on their age-relevance, we used age at diagnosis as a proxy to disease onset and derived disease age-of-onset profiles (Figure A.32-A.41). On average, cardiovascular and endocrine diseases had a high median age-of-onset, while infections had the lowest (Figure 3.6).

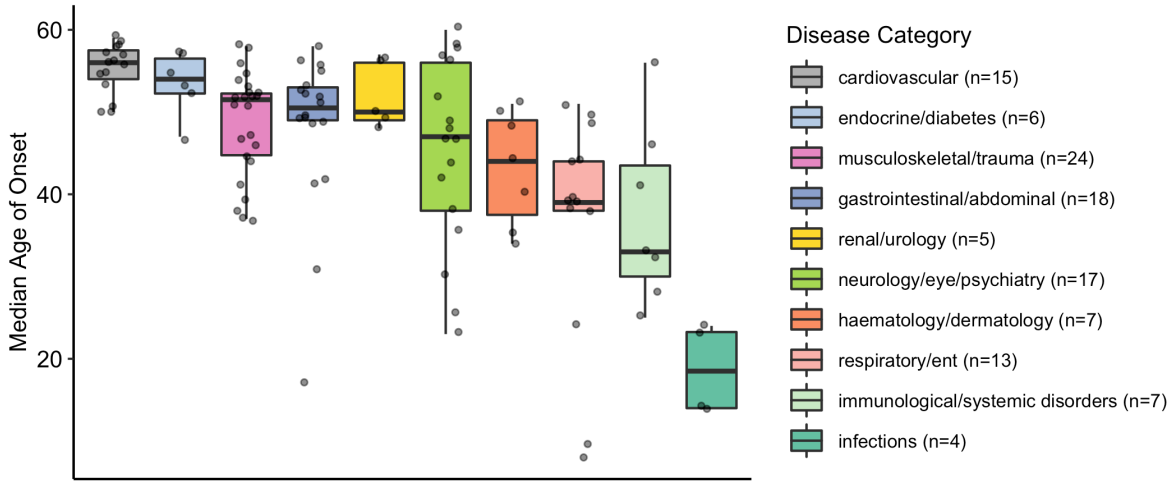


Figure 3.6 *Distribution of median age-of-onset (y-axis) across categories (x-axis). Points show diseases, grouped by the categories (individual boxplots). Categories are ordered by the median value of the median age-of-onset.*

We then clustered diseases into 4 clusters (the optimum number determined by gap statistic) using the PAM algorithm and disease dissimilarities calculated using CORT (temporal correlation measure) distance (Chouakria & Nagabhushan, 2007) (Figure 3.7, Supplementary Table S1).

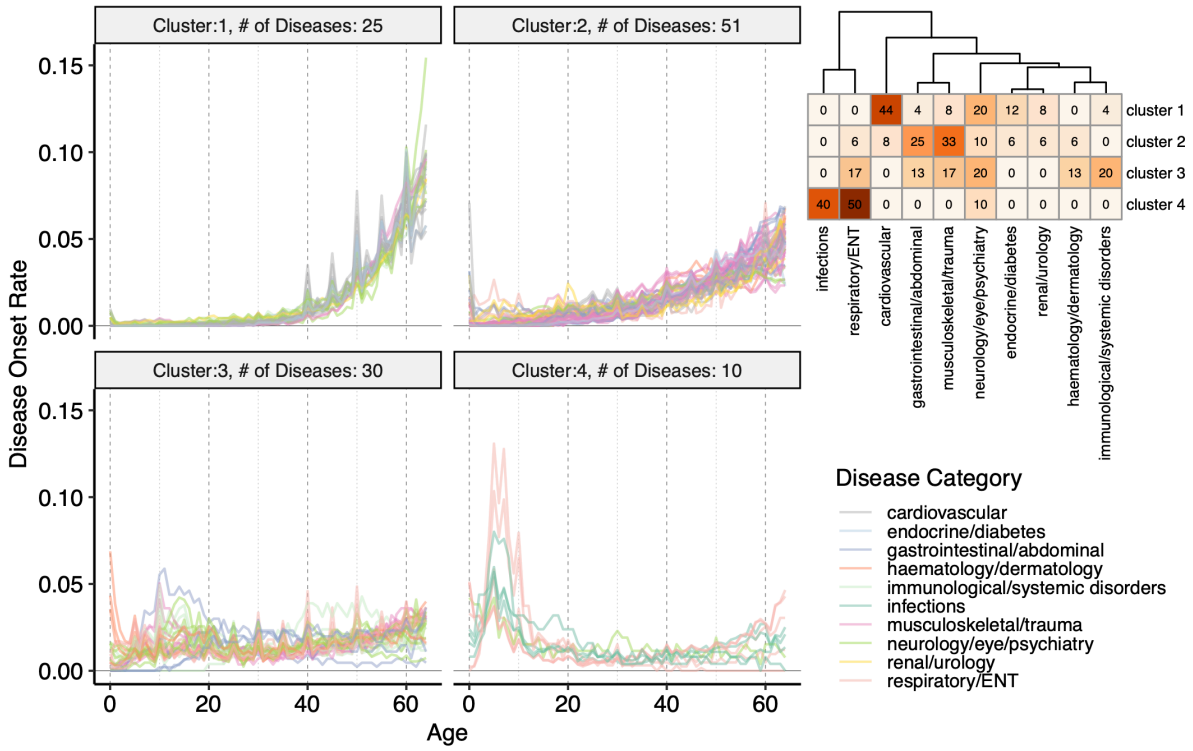


Figure 3.7 *Age-of-onset profiles clustered by the PAM algorithm, using dissimilarities calculated using (CORT). The y-axis shows the number of individuals who are diagnosed with the disease at a certain age, divided by the total number of people having that disease. Values are calculated using the median values for 100 permutations of 10,000 people in the UKBB. The x-axis shows the age-of-onset in years. Each line shows one disease and is coloured by disease categories. The heatmap on the right upper corner shows the percent overlap between categories and clusters. Numbers give the % of an age-of-onset cluster belonging to each category.*

Cluster 1 diseases (n = 25) showed a rapid increase with age after the age of 40; 11 were cardiovascular diseases, but this cluster also included other diseases such as diabetes, osteoporosis, and cataract. Cluster 2 (n = 51) diseases started to increase in the population at an earlier age of 20, but had a slower rate of increase with age; diseases in this cluster were the most diverse, including 17 musculoskeletal, 13 gastrointestinal diseases, as well as others such as anaemia, DVT, thyroid problems, depression. Cluster 3 diseases (n = 30) showed a low age dependency with a mostly uniform distribution across ages, but with slight increases around the ages of 10 and 60 years. This category included similar numbers of immunological, neurological, musculoskeletal, gastrointestinal and respiratory diseases but all have an 'immune' component even if not classified in this way by the UKBB (*e.g.*, inflammatory bowel disease (gastrointestinal), asthma (respiratory), psoriasis (dermatology)). Cluster 4 (n

= 10) had a peak at around 0-10 years of age and included respiratory diseases (n = 5) and infections (n = 4). Notably, all infectious diseases were in this cluster.

3.6 Genetic similarities between diseases

Using linear mixed models implemented in BOLT-LMM (Loh et al., 2015), we performed GWAS on each disease separately and included approximately 10 million common variants that pass quality control. Considering associations with the literature standard p-value lower than $5 * 10^{-8}$ as significant (Panagiotou, Ioannidis, & Genome-Wide Significance Project, 2012; Pe'er, Yelensky, Altshuler, & Daly, 2007), we next quantified the associations for each disease, category, and age-of-onset cluster (Figure 3.8). The major histocompatibility complex (MHC) region is excluded from all analyses, as in the literature, because of its unusually high effect sizes and LD patterns (chr6: 28,477,797 - 33,448,354) (Bulik-Sullivan et al., 2015; "MHC region of the human genome - genome reference consortium," n.d.). Out of 116 diseases, 36 had no significant association and the total number of polymorphic regions with at least one significant association was 93,817. The maximum number of significant associations was 35,001 (hypertension) and the median and mean were 13.5 and 1389.3, respectively. We also checked if diseases from different age-of-onset clusters vary in the number of associations. Cluster 4 had almost no significant associations (the disease with the maximum number of associations had only 3 significant variants). Although cluster 1 had the highest number of significant associations on average, the values across clusters 1, 2, and 3 were not significantly different (Figure 3.8b). Moreover, endocrine, immunological, cardiovascular diseases had the highest number of associations and infections had the lowest (Figure 3.8c). Only 1% of the significant polymorphisms (n=932) were in coding regions, and of these 49% (n=452) were missense and only 1% (n=10) were nonsense. We further found that 47% of significant variants (n=43,810) were associated with multiple diseases, but only ~9% were associated with multiple diseases from different categories (n=8,048) and again ~9% with different age-of-onset clusters (n=8,801) (Figure A.42).

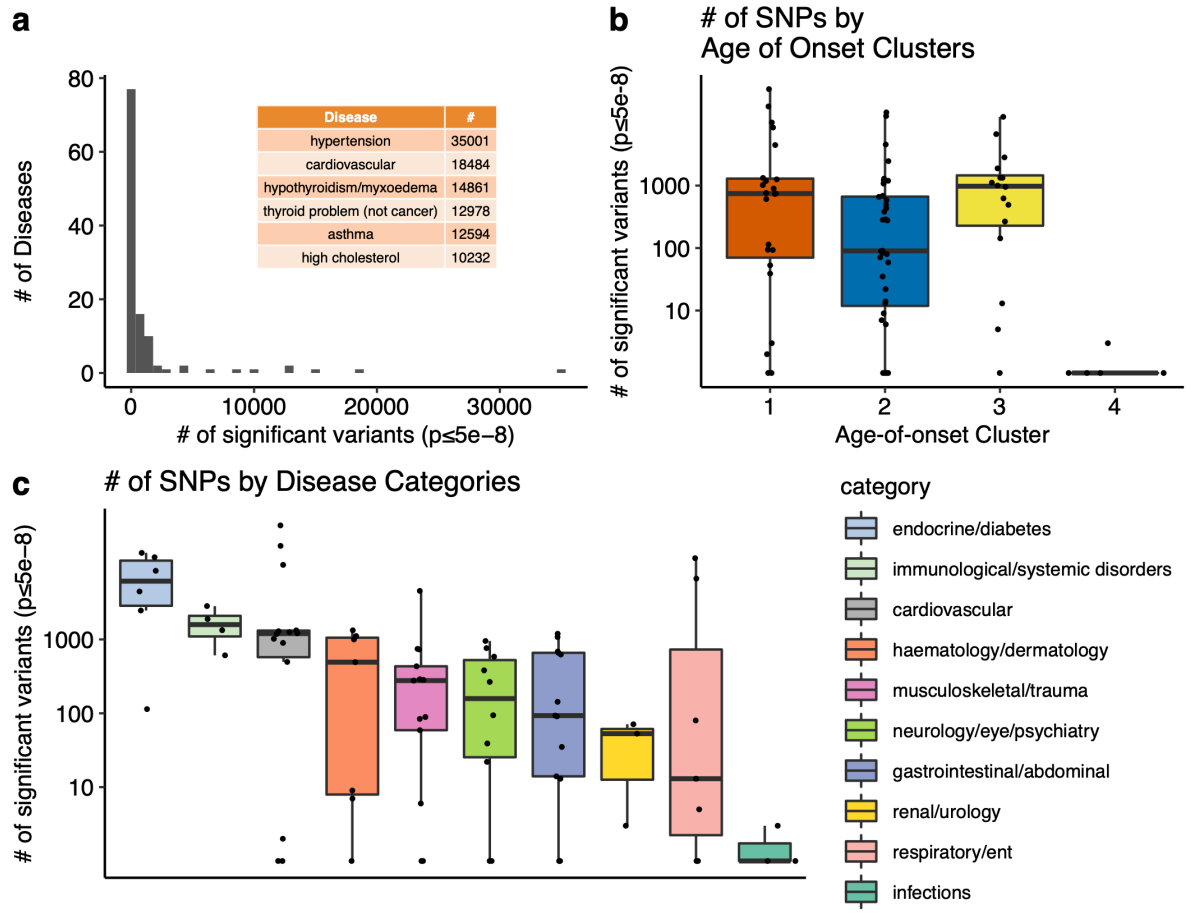


Figure 3.8 *(a) Number of diseases for different number of significant variants (p≤5e-8). Diseases with the highest number of associations (N≥10,000) are given as an inset table. (b) Comparison of the number of significant associations (y-axis, in log scale) across age-of-onset clusters (x-axis) (ANOVA after excluding cluster 4, p = 0.06). Since the y-axis is in log scale, diseases with zero significant associations are not shown on the graph. (c) The same as (b) but for disease categories. Categories are ordered by the median number of significant SNPs.*

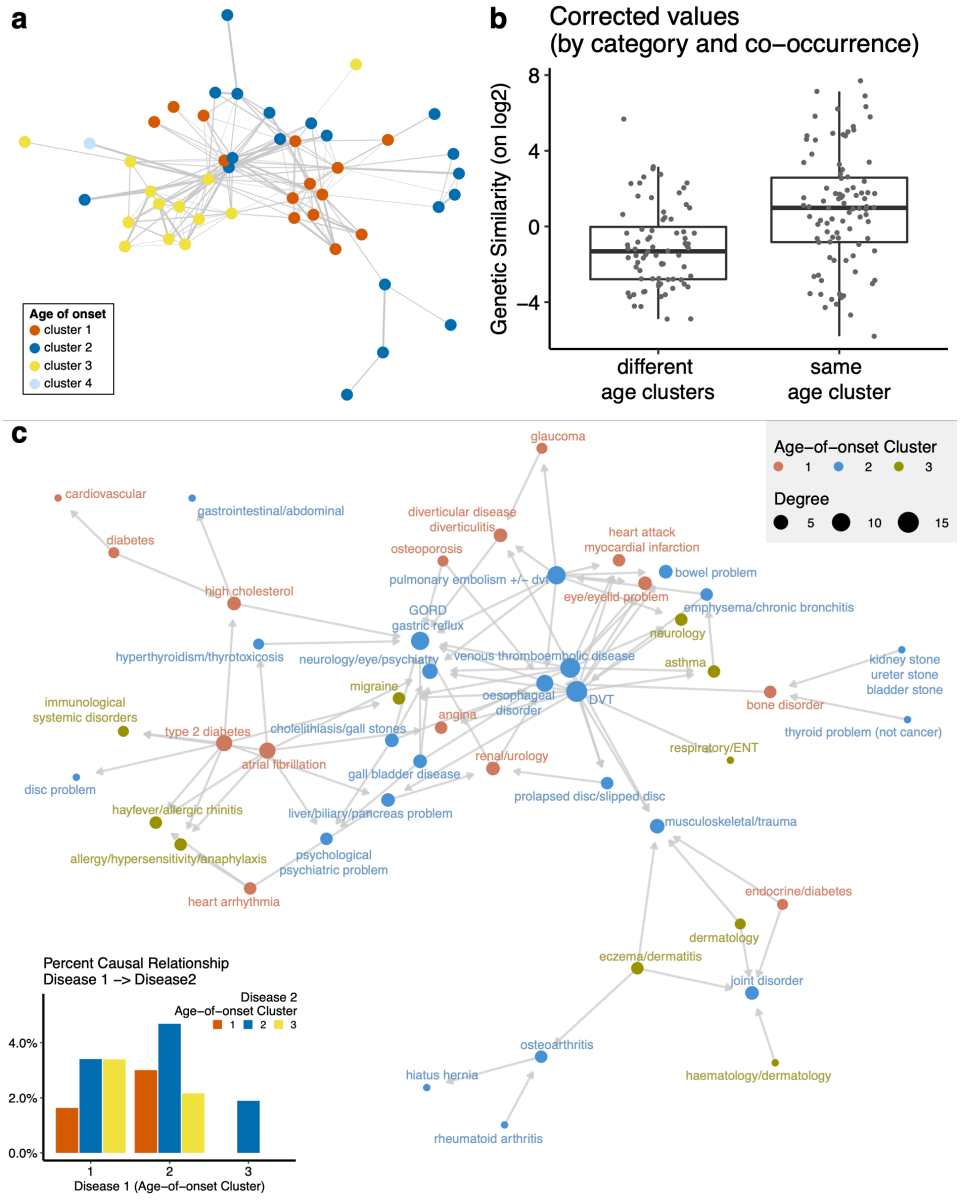


Figure 3.9 *(a) Network representation of the genetic similarities. Nodes (n=47) show diseases with at least one significant genetic similarity and are coloured by the age-of-onset cluster. Edges (n=167) are weighted by the genetic similarity corrected by disease categories and co-occurrences. (b) The difference between genetic similarity within and across the age-of-onset clusters. The y-axis shows genetic similarity corrected by category and co-occurrence on the log2 scale and x-axis groups similarities into different or same age-of-onset clusters. c) Network representation of the cause and effect relationships between diseases. Each node (n=48) shows a disease, coloured by the age-of-onset cluster. Size of the nodes represent the number of significant causal relationships. Arrows show the cause and effect relationship between pairs with FDR corrected p≤0.01 and GCP>0.6. The inset bar plot shows the percent significant causal relationships (y-axis) between disease 1 (x-axis) and disease 2 (bars coloured by the age-of-onset).*

We next sought to characterise the genetic similarities between diseases using a score that shows the excess of overlapping associations between diseases, given the number of significant associations for each disease. Importantly, we calculated genetic similarities between 80 diseases that have at least one significant association, excluding the pairs that are vertically connected (*i.e.* ancestors to child) in the disease hierarchy (*e.g.* essential hypertension and hypertension). We found 47 significant overlaps and diseases with similar age-of-onset profiles showed a higher genetic similarity, even when controlled for disease categories and co-occurrences (F-test p = 1.19 x 10-8, Figure 3.9a-b,A.43). Moreover, this trend was reproducible when each cluster was analysed separately (Figure A.44). We repeated the analysis using 1,703 previously defined LD blocks (Berisa & Pickrell, 2016) instead of taking all SNPs as independent, and found significant similarity only between diseases of the same age-of-onset cluster (Figure A.45). Importantly the disease categories (p=0.89) and co-occurrences (p=0.15) did not explain the similarity.

3.7 Cause-effect relationships between diseases

Using a recent methodology developed by O'Connor & Price (2018), we tested for partial or fully causal relationship between diseases. In particular, the method identifies if a latent causal variable (LCV) mediates the genetic correlation between diseases. Using a genetic causality proportion, it assigns a causal relationship if one of the diseases is more strongly correlated with the LCV. The authors report that, unlike mendelian randomisation, this method can distinguish between the correlation due to common aetiology and causation. We tested for potential causation between 60 diseases, excluding the ones with less than 10 significant genetic variants and low heritability estimates ($Z_h < 7$) (O'Connor & Price, 2018). Also, similar to genetic similarities, we did not calculate the causation between diseases that are vertically connected in the disease hierarchy. Following the same significance criteria proposed in the methods article (FDR corrected p≤0.01 and mean $GCP > 0.6$ - *i.e.* genetic causality proportion), we found significant evidence for fully or partial genetic causality in 91 disease pairs between 48 diseases (Figure 3.9c, Supplementary Table S2). DVT, venous thromboembolic disease, and pulmonary embolism had the highest number of out degrees, meaning they were found as causal for multiple diseases, including all 3 age-of-onset clusters and 5 different disease categories. Gastro-oesophageal reflux (GORD)/gastric reflux and oesophageal disorder, on the other hand, had the high-

est number of in degrees, meaning there are multiple diseases detected as causal. A likely explanation is that there are multiple types of drugs, used for different conditions, that are associated with increased risk or exacerbated symptoms of GORD (Mungan & Pınarbaşı Şimşek, 2017). These causal diseases spanned 5 disease categories and age-of-onset clusters 1 and 2.

Using Fisher's exact test, we also tested if the causal relationships were more common between diseases in certain age-of-onset clusters but did not find any significant difference (FDR corrected p>0.1 for all comparisons, inset bar plot in Figure 3.9c). Thus, we concluded that although there are some causal links, the high genetic similarity within clusters (Figure 3.9a-b) cannot be explained by causation but more likely to include common aetiologies.

3.8 Known ageing-related genes and genes associated with different age-of-onset clusters

We next mapped all variants to genes based on proximity or known eQTLs using the GTEx data (Gamazon et al., 2018). To assess the reproducibility of the genes identified from this study, we compared the significant hits we found with all those reported in the GWAS Catalog. We verified that most of the diseases had significant overlaps with the same or associated traits in the GWAS-Catalog (*e.g.* osteoporosis and bone density), confirming our results are reproducible with independent data (Supplementary Table S3). We next compiled the genes associated with multiple diseases and multiple categories and grouped them based on the age-of-onset cluster of the associated diseases (Supplementary Table S4). In particular we created two sets of genes, 'multidisease' and 'multicategory', for clusters 1, 2, and 3. We exclude cluster 4 from downstream analysis, because the number of variants significantly associated with this cluster was low (n=7 associated with 5 diseases), mapping to only 2 genes (*ZPBP2*, *NPC1L1*). We also compiled genes associated with multiple diseases or categories in combinations of different age-of-onset clusters. Importantly, genes associated with multiple clusters are not in the gene sets for individual clusters as they only include genes specific to those clusters, *e.g.* cluster 1, cluster 2 and cluster 1 & 2 sets all include mutually exclusive sets.

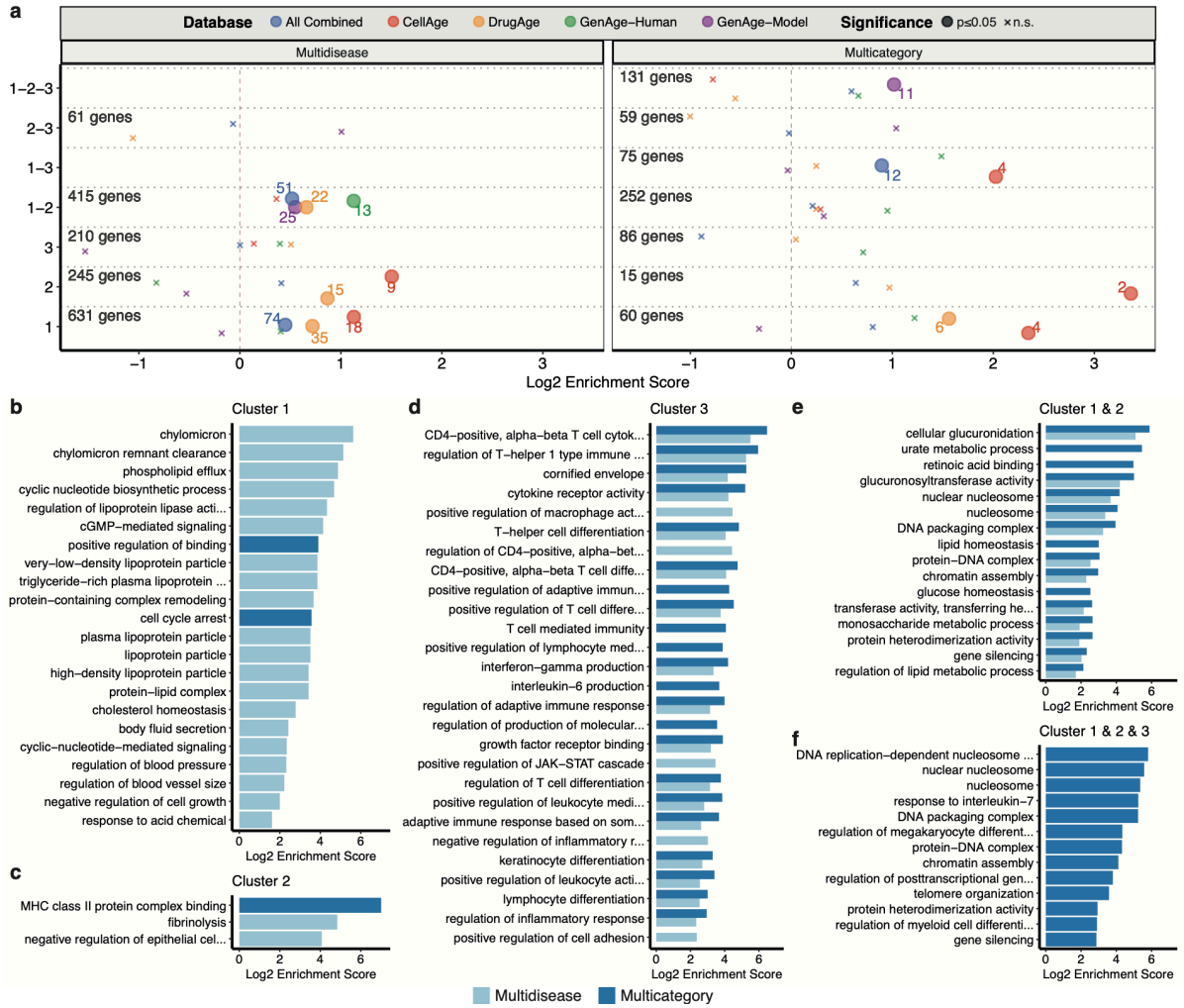


Figure 3.10 *a) Overlap between known ageing-related genes in databases and genes associated with diseases in different clusters. x-axis shows log2 enrichment score, and y-axis shows the age-of-onset clusters. The numbers of genes in each cluster (for both Multidisease and Multicategory genes) are given. The size of the points shows the statistical significance (larger shows marginal p-value≤0.05) and the colour shows different databases. The coloured numbers near the points show the numbers of overlapping genes. b-f) Gene Ontology (GO) Enrichment results for genes associated with diseases in b) Cluster 1, c) Cluster 2, d) Cluster 3, e) Cluster 1 and 2, f) Cluster 1, 2 and 3. Representative GO categories for significantly enriched categories (BY-adjusted p-value ≤ 0.05) are listed on the y-axis. Log2 enrichment scores are given on the x-axis. The colour of the bar shows the result for multidisease and multicategory genes. There was no significant enrichment for cluster 1 & 3 and 2 & 3.*

We next sought to understand if the genes we identified were previously associated with ageing, by comparing the gene lists we compiled from the literature-based ageing databases: GenAge human (genes associated lifespan in humans or closely related species), human orthologs in GenAge model organism (genes modulating lifespan in

model organisms), CellAge (genes regulating cellular senescence), DrugAge targets (drugs modulating lifespan in model organisms), and all databases combined (Avelar et al., 2019; Barardo et al., 2017; Tacutu et al., 2018) (Figure 3.10a). In general, genes associated with clusters 1 and 2 showed significant enrichment with known ageing-related genes, but not cluster 3. The list of overlapping genes is given in Supplementary Table S5. CellAge database showed the largest number of significant overlaps, with genes associated with clusters 1, 2, and both 1 and 3. DrugAge targets had a significant overlap with clusters 1, 2, and both 1 and 2. GenAge Human only had significant association with genes associated with both cluster 1 and 2. GenAge model organism data significantly overlapped with genes associated with both cluster 1 and 2, and all clusters (1 & 2 & 3). The results showed that, although the number of overlaps is low, the clusters 1 and 2, constituting age-associated diseases, share a significant genetic component with known ageing-related genes, while cluster 3 did not.

3.9 Biological functions of the genes associated to different age-of-onset clusters

We next did a Gene Ontology (GO) enrichment analyses using Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) categories. Cluster 1 was associated with many lipoprotein-related categories, cellular signaling, cellular response, cell cycle arrest, and blood pressure (Figure 3.10b). Cluster 2 showed association to MHC class II binding, fibrinolysis, and negative regulation of epithelial cell (Figure 3.10c). Cluster 3 had associations to many immune-related categories and cell adhesion (Figure 3.10d). Genes associated with both cluster 1 and cluster 2 were related to nucleosome complex, glucose homeostasis, retinoic acid binding (Figure 3.10e). Genes in clusters 1 and 3, and in 2 and 3 did not have any significant associations. Genes associated with at least one disease in all clusters were related to nucleosome complex, interleukin-7 response, differentiation, telomere, and gene silencing (Figure 3.10f). Here we listed the categories that are representative to all other significant associations. The full list is given in Supplementary Table S6, and the procedure of selecting representatives is described in Section 3.12.

3.10 Drug repurposing to improve health-span

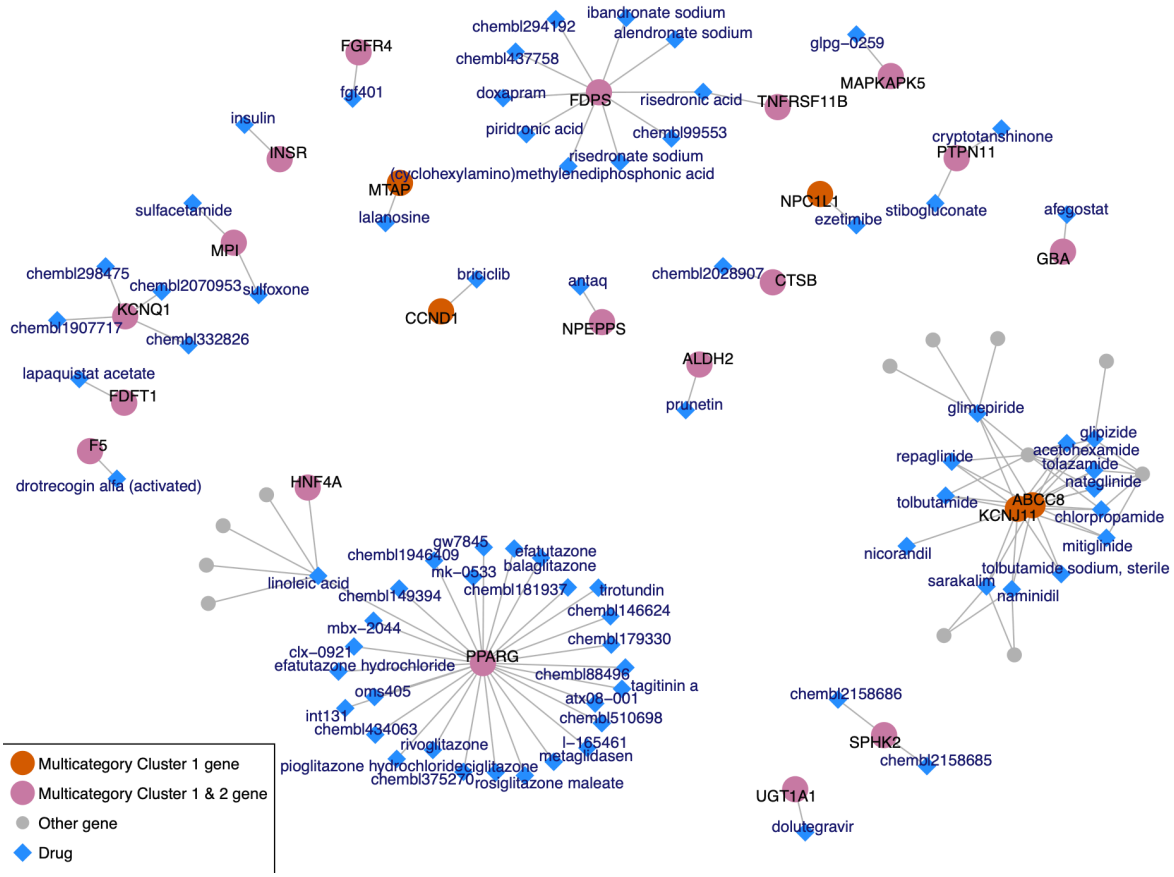


Figure 3.11 *Drug-target gene interaction network for the drugs that specifically target multicategory cluster 1, cluster 2 or cluster 1 and 2 genes as determined by Fisher's exact test. Blue diamonds show the drugs with significant association or targeting only one gene in these gene groups. Circles represent the genes targeted by the significant hits, coloured by their age-of-onset cluster. Grey circles show the genes targeted by these drugs but are not among the gene set of interest. The interaction data is collected from DGIdb.*

Identification of drugs that can target the multicategory genes in clusters 1 or 2 poses a possibility to treat many diseases simultaneously and thus extend health-span in the elderly. Thus, we investigated if there are drugs that target these genes specifically ($p \leq 0.01$ or having only one specific target, Figure 3.11). We found drugs targeting multicategory cluster 1 genes i) *ABCC8* and *KCNJ11*, which code for parts of K-ATP channels, ii) *CCND1*, iii) *MTAP*, iv) *NPC1L1.* There were also several drugs targeting multicategory genes associated with both cluster 1 and 2 diseases, such as *PPARG, INSR, FGFR4, MAPKAPK5, ALDH2, PTPN11*. Interestingly, 23 of the

significant drugs were approved drugs for 14 conditions, including diabetes, hyperlipidaemia, neoplasms, osteoporosis, cardiovascular diseases (list of all drugs and indications available in Supplementary Table S7). Moreover, one of the drugs we identified, prunetin (targeting *ALDH2*), was previously shown to improve the lifespan of male Drosophila melanogaster (Piegholdt, Rimbach, & Wagner, 2016).

3.11 Evolution of ageing and age-related diseases

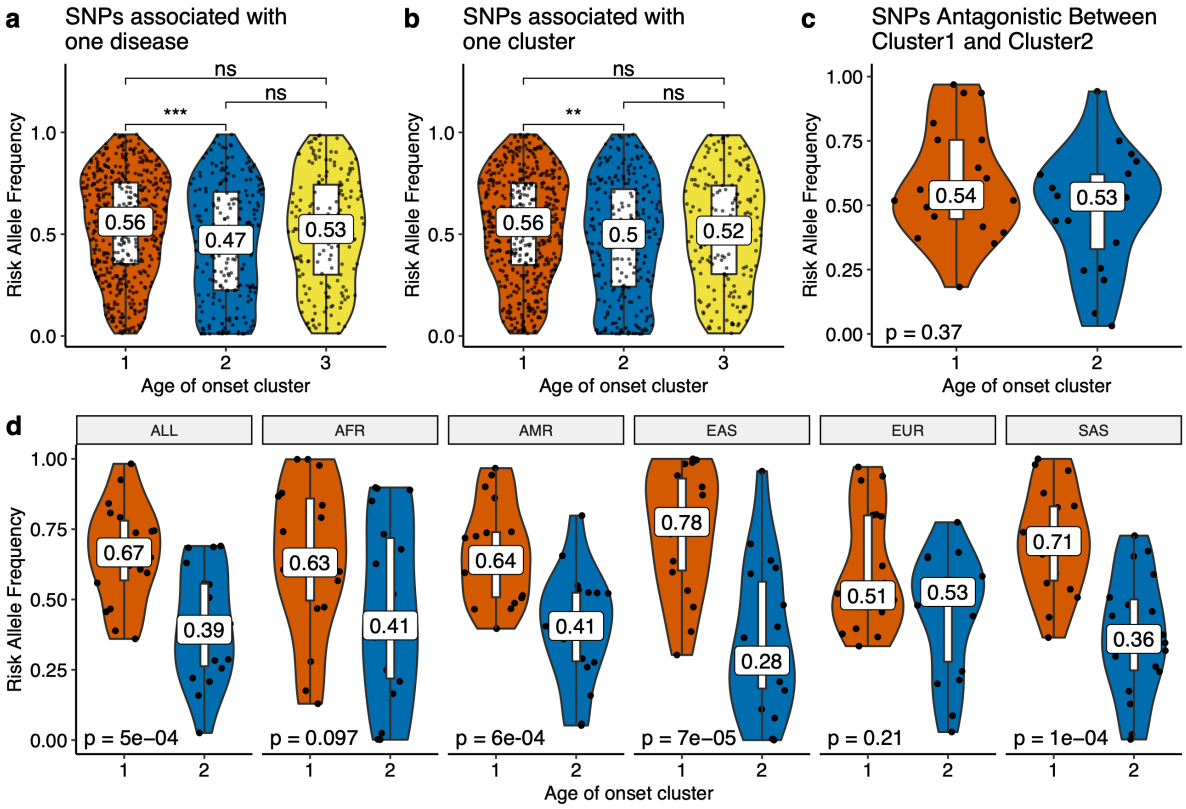


Figure 3.12 *Risk allele frequency distributions (y-axis) of different age-of-onset clusters (x-axis) in the UKBB for a) SNPs associated with one disease (excluding antagonistic associations), b) SNPs specific to one cluster (excluding antagonistic associations) (ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001), and c) SNPs that have antagonistic association with cluster 1 and 2 (excluding agonists between cluster 1 and 2). d) The same as panel c but for different 1000 Genomes super-populations (ALL: complete 1000 Genomes cohort, AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian). The p-values specified on the plots are marginal p-values calculated for two-sided tests and are not corrected for multiple testing.*

We next sought to understand the abundance of disease-associated variants in the population and their relationship with the evolutionary theories of ageing. We first hypothesized that according to the mutation accumulation theory of ageing (MA), SNPs associated with later-onset diseases (Cluster 1) would have a higher frequency than the SNPs associated with diseases that occur at earlier ages (Clusters 2 & 3). We therefore compared the risk allele frequencies of diseases in different age-of-onset clusters. As SNPs which are close together in the genome are expected to have similar allele frequencies due to linkage, we calculated the median risk allele frequency for SNPs within previously defined LD blocks (Berisa & Pickrell, 2016). Diseases of cluster 1 had significantly higher risk allele frequencies than cluster 2, both for the SNPs associated with one disease (Figure 3.12a, Wilcoxon test p = 0.00033) or with one cluster (Figure 3.12b, Wilcoxon test p = 0.0068, also confirmed by bootstrapping n = 100 loci for B = 1,000; Figure A.46). We also analysed the risk allele frequencies in populations of the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) data and confirmed that the same trend is observed in all super-populations (Figure A.47-A.48).

To test the antagonistic pleiotropy theory (AP), we first asked if the diseases with different onsets have an excess of antagonistic SNPs. Here, following the literature (Rodríguez et al., 2017), we defined a pleiotropic biallelic SNP as agonistic if the risk allele is the same for different diseases, and as antagonistic if opposite alleles are associated with increased risk for different diseases. If one of these diseases are under a stronger negative selection, then the risk allele of the other disease could increase with time. Comparing the proportion of agonist and antagonist SNPs within and between the age-of-onset clusters, we found that there is an excess of antagonistic pleiotropy between diseases with different age-of-onset profiles (Fisher's exact test p<0.001, Supplementary Table S8). Next, we tested the differences in risk allele frequencies between the clusters as AP predicts a higher risk allele frequency for late-onset diseases. Interestingly, the difference between the risk allele frequencies for cluster 1 and cluster 2 was not significant for the UKBB population (Figure 3.12c). However, all 1000 Genome super-populations except for Europeans had higher risk allele frequencies for cluster 1 diseases (Figure 3.12d). A potential explanation is that it is easier to detect antagonistic associations for SNPs with a minor allele frequency closer to 0.5 because of the increased power for both alleles. Indeed, associations with a larger effect size showed the expected differences in allele frequencies, although the number of independent loci was limited (Figure A.49). We also examined the type of diseases and genes associated with antagonistic pleiotropy. The main

driver of the pattern was the loci with *ABCG8* and *ABCG5* genes, showing antagonistic relationship for high cholesterol (cluster 1) and other lipid-related diseases that are in cluster 2, such as gallbladder disease and cholelithiasis. Another locus included variants that show antagonistic relationship with cardiovascular disease (cluster 1) and cluster 2 diseases gout (*ADH1B*), osteoarthritis and joint disorder (*SLC39A8*), and osteoarthritis (*BANK1*). Another potential candidate was a locus associated with hypertension (cluster 1) and musculoskeletal diseases (cluster 2), but this locus included multiple candidate genes (Supplementary Table S9).

3.12 Methods

UK Biobank Data

Data is downloaded using bash and following the guidelines provided by the UK Biobank.

Sample quality control

After excluding all samples from individuals who have withdrawn their data from UK Biobank, we first filtered out all samples without genotypes (N = 14,248). Then, we used the following criteria for the remaining 488,295 samples.

Discordant sex: Data includes two entries for sex: 1) self-reported and 2) genetic sex determined using the call intensities on sex chromosomes. There are multiple reasons why these two entries may not correspond, such as sample mishandling, errors in data input, transgender individuals, and sex chromosome aneuploidies (Anderson et al., 2010; Bycroft et al., 2018). Since we used sex as a covariate in our GWAS model, we preferred to be cautious about this issue and excluded all cases where the genetic sex and self-reported sex do not correspond and all cases where sex chromosome aneuploidy is detected. Specifically, we used the fields '31-0.0' (Sex) and '22001-0.0' (Genetic sex) to compile discordant information. There were 235 self-reported males being identified as female by the genetics, and 143 self-reported females being identified as males by the genetics. We excluded these 378 cases, making only 0.077% of the data. Moreover, field '22019-0.0' (Sex chromosome aneuploidy) is used to exclude cases with sex chromosome aneuploidy. There were 651 cases of aneuploidy, making 0.133% of all data. 181 of these cases (27.80% of aneuploidy cases) are also detected as discordant information in the first step. This corresponds to 47.88% of

discordant sex cases. Overall, we identified 848 samples to be excluded based on this criterion.

Genotype call rate & Heterozygosity: For exclusions based on missingness and heterozygosity we only used the suggested exclusions by UK Biobank. Specifically, we used the field '22010-0.0' (Recommended genomic analysis exclusions) and determined the cases with 'poor heterozygosity/missingness' (N = 469). We next used the field '22018-0.0' (Genetic relatedness exclusions) and noted down the cases with 'Participant self-declared as having a mixed ancestral background' (N = 692), and the cases with 'High heterozygosity rate (after correcting for ancestry) or high missing rate' (N = 840). Lastly, there were 968 cases that are suggested as outliers for heterozygosity or missing rate, field '22027-0.0' (Outliers for heterozygosity or missing rate). Genotype missingness and heterozygosity are widely used as a measure of DNA sample quality. We then checked the scatter plots for logit(Missingness) vs. Heterozygosity for each Ethnic Background, in accordance with the identification of samples to exclude by the UK Biobank (Bycroft et al., 2018) (Figure A.51). Logit transformation is used to linearize sigmoidal distribution of missingness. Investigation of heterozygosity can detect DNA sample contamination, inbreeding, or mixed ethnicity (Anderson et al., 2010). This quality check enables seeing that people with a mixed ethnicity tend to have a higher heterozygosity, even after correcting for PCs. We confirmed these are in accordance with the original article and excluded the samples suggested by the UK Biobank.

Overall, there were 3,697 samples excluded based on these two criteria. Please note that the numbers presented above may not add up to this number. That is because there were some samples excluded because of multiple criteria. The percent overlap across multiple criteria is given in Figure A.52.

After the exclusions, we had 484,598 samples.

Preparing the trait data

Using the samples that pass the quality control (N=484,598), we subset the data so that it includes only the baseline visit. Apart from the data that is already available in UK Biobank, we calculated some other values:

1. BMI: Using the columns for 'Weight' and 'Standing height' we calculated BMI as: Weight / (Standing Height / 100)^2,
2. Parent Age at Death - Minimum: The youngest age at which either parent died.

3. Parent Age at Death - Maximum: The age of death for the parent who lived longest.
4. Parent Age at Death - Average: The average age of death for the two parents. If neither of the parents died, or if the data is unavailable, these values (2-4) are set to be NA. If only one parent died, we use the corresponding age as both the minimum, maximum, and average.
5. The number of self-reported non-cancer diseases: The number of unique self-reported non-cancer illnesses each participant recorded in the baseline recruitment.
6. The number of self-reported cancers: The number of unique self-reported cancers each participant recorded in the baseline recruitment.
7. Self-reported diseases after taking the disease hierarchy into consideration (Propagated disease data): The self-reported diseases in UK Biobank are not independent, but rather are organised in a hierarchical manner. Using the relationship information between diseases, we propagate disease-participant associations, upwards, including terms higher up the tree. For example, if a person reports having “essential hypertension”, we also annotate that person with “hypertension”, and “cardiovascular disease”.
8. Age at diagnosis for the self-reported diseases after taking the disease hierarchy into consideration (Propagated age at diagnosis data): We re-defined age at diagnosis using the minimum age at diagnosis for all the diseases that are child term for a particular disease.
9. The number of self-reported non-cancer diseases after taking the disease hierarchy into consideration (Propagated number of non-cancer diseases): The number of unique self-reported diseases each participant records after taking into account the data propagation.
10. Age when the last deceased person died: We calculated the age of each person when the last death entry in the UKBB happened. This value is used to calculate the proportion of people who died at a certain age interval in Figure 3.1c.

Selecting diseases to analyse

We calculated the disease occurrences for all self-reported diseases in UK Biobank. Specifically, among the cohort we will use, we calculated how many participants and what proportion of males and females reported each disease. Since we analyse the same set of SNPs that have MAF>=0.01 across multiple diseases, to decrease the false positive rate in GWAS, we limited the diseases to a subset with at least 2,000 cases (n = 129 out of 472). Moreover, we only focused on diseases that are common

and not sex-specific, *i.e.* we only considered diseases that are seen in 1 in every 1,000 males and females (n = 189 out of 472). The intersection of these two conditions was 116 diseases and we excluded all others.

In this study, we only analysed self-reported non-cancer diseases (field '20002') and did not combine self-reported cancers (field '20001'), mainly because i) the number of cases is low (45,224 compared to 384,906 for other diseases), ii) cancer is thought as a result of a complex interaction between germline and somatic mutations (Kanchi et al., 2014; Khurana et al., 2016), whereas the evidence for the effect of somatic mutations in other diseases is limited to rare and neurological disorders (Poduri, Evrony, Cai, & Walsh, 2013; Zhang & Vijg, 2018), iii) the relationship between cancer and ageing is complex, *e.g.* while telomere attrition and cellular senescence, are thought to be evolved as a tumour suppressor mechanisms; ageing-related changes in epigenomic landscape and genomic instability contribute cancer occurrence (Finkel, Serrano, & Blasco, 2007). Thus, although a similar analysis using cancers would be interesting, we only focused on non-cancer self-reported diseases in this study.

Disease co-occurrences

Table 3.1 *Contingency table for disease co-occurrence calculations.*

.	Disease_B	Not_Disease_B	Total
Disease A	Nab	Nanb	Ta
Not Disease A	Nnab	Nnanb	Tna

Relative risk (RR) score

Relative risk is an estimate of having the disease A, when already affected by disease B. Overall it measures if disease A co-occurs with disease B more frequently than expected if these diseases were independent in the population. It is calculated as a fraction between the number of patients diagnosed with both diseases and a random expectation based on disease prevalence (Sanchez-Valle et al., 2018). Mathematically it can be expressed as follows, using the values from Table 3.1:

$$P_{exposed} = \frac{N_{ab}}{T_a}, P_{notexposed} = \frac{N_{nab}}{T_n a}$$

$$RR = \frac{P_{exposed}}{P_{notexposed}} \tag{3.1}$$

$$CI = lnRR \pm 1.96\sqrt{\frac{\frac{T_a - N_{ab}}{N_{ab}}}{T_a} + \frac{\frac{T_{na} - N_{nab}}{N_{nab}}}{T_{na}}}$$

ϕ value (Pearson correlation for binary variables)

It measures the robustness of the association between diseases based on co-occurrences (Gutiérrez-Sacristán et al., 2018). Mathematically, it can be expressed as:

$$\phi_{AB} = \frac{C_{AB}N - P_A P_B}{\sqrt{P_A P_B (N - P_A)(N - P_B)}} \tag{3.2}$$

N: the total number of individuals

P_A: Prevalence of disease A

C_{AB}: Number of patients with both diseases

ϕ ranges between -1 and 1, where the sign indicates the type of association.

Disease age-of-onset

Disease dissimilarity measure

Temporal correlation: In order to calculate dissimilarities among diseases, we use CORT (Chouakria & Nagabhushan, 2007) distance as included in R package TSclust (Montero & Vilar, 2014). Euclidean distance and dynamic time warping (Berndt & Clifford, 1994) are the two most widely used proximity measures for time series proximity. However, they are both calculated based on the closeness of the values and disregard the growth behaviour. Correlation-based measures are also used to calculate the similarity between time series. However, Pearson correlation overestimates the similarity because of the underlying temporal dependency and Spearman correlation fails to consider the growth rate as it is based on ranks. Chouakria et al., on the other hand, suggests a measure that also considers the proximity-based on growth behaviour,

CORT. Temporal correlation between two time series objects S1=(u1,u2,…,up) and S2=(v1,v2,…,vp) is calculated as follows:

$$CORT(S_1, S_2) = \frac{\sum_{i=1}^{p-1}(u_{i+1} - u_i)(v_{i+1} - v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{i+1} - u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{i+1} - v_i)^2}} \tag{3.3}$$

CORT ranges between -1 and 1. A value of CORT = 1 implies that two time series increase or decrease simultaneously with the same growth rate, whereas a value of -1 shows the same growth rate but in the opposite direction. If the value is 0, it means there is no temporal correlation between the series.

Dissimilarity Index: The dissimilarity index suggested by Chouakria et al, is calculated based on an automatic adaptive tuning function and considers similarity based on both values and behaviour, *i.e.* the strength of monotonicity and closeness of the growth rates as calculated by CORT measure introduced in the previous section. They suggest a dissimilarity index D as follows:

$$D(S_1, S_2) = f(CORT(S_1, S_2)).\delta_{conv}(S_1, S_2) \tag{3.4}$$

Where $f(x)$ is an exponential adaptive tuning function:

$$f(x) = \frac{2}{1 + exp(kx)}, k \geq 0 \tag{3.5}$$

As k increases, the contribution of behaviour increases. We use k = 2 and as a result behaviour (CORT) contributes 76.2% to D and values (δ_{conv}) contribute 23.8%. For δ_{conv} we used conventional Euclidean distance.

Clustering diseases by age-of-onset

We clustered data using 'partition around medoids (PAM)' algorithm (Kaufman & Rousseeuw, 1990) based on the distance measure calculated using the previous step. The aim of this algorithm is to minimise the average distance (based on any dissimilarity measure) between the objects and their closest selected medoid object. It works very similarly to k-means, except instead of defining arbitrary points as the means, it defines medoids among the objects. Thus, it can incorporate any distance measure instead of just using the mean distance between points (*i.e.*, euclidean distances). The algorithm first searches for k number of objects that represent the structure of the data (Here the number k is assumed to be known a priori but see the next section for

the determination of k). After finding a set of k medoids, k clusters are constructed by assigning each observation to the nearest medoid. Overall, the goal is to find k representative objects such that the sum of dissimilarities of the observations to their closest representative is as small as possible. After each assignment, medoid and non-medoid data points are swapped and a cost (sum of distances of points to the new medoid) is calculated. If the total cost of configuration is decreased, then the new configuration is maintained, otherwise, it is reversed. We used 'pam' function in the 'cluster' package (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2019) in R to apply this algorithm.

Choosing the optimum number of clusters

The clustering algorithm we used, PAM, clusters data into k clusters, which is determined by the user. So, even if there is no real structure in data, as we increase the number of clusters, we can get more and more clusters. A potential way to decide on the number of clusters is using the gap statistic (Tibshirani, Walther, & Hastie, 2001). This value is calculated by comparing logarithm of the within-sum-of-squares and comparing it to averages from simulated data where there is no structure.

$$WSS_k = \sum_{l=1}^{k} \sum_{x_i \in C_l} d^2(x_i, \overline{x_l}) \tag{3.6}$$

k: number of clusters

Cl: objects in the l-th cluster

$\overline{x}$: the average point.

Calculating only WSS, however, is not enough as it would be minimised when each point has its own cluster. Thus, we use the gap statistic which suggests calculating $log(WSS_k)$ for a range of values of k and compare it to that obtained by WSS calculated based on simulated data. So, after WSS is calculated for various values of k, the algorithm involves generating B (we choose B=1,000) reference datasets, using Monte Carlo sampling from a homogeneous distribution and re-calculate WSS for all k values. Using these values gap(k) statistic is calculated:

$$\begin{aligned} gap(k) &= \overline{l_k} - log(WSS_k) \\ \overline{l_k} &= \frac{1}{B} \sum_{b=1}^{B} log(W_{kb}^*) \end{aligned} \tag{3.7}$$

If the clustering is good (WSS is small) we expect lk to be higher than $log(WSS)$. Thus gap statistic is mostly positive and we are interested in the highest value. Tibshirani et al. (2001) suggests using the smallest k such that,

$$gap(k) \geq gap(k+1) - s'_{k+1} \tag{3.8}$$

where

$$\begin{aligned} s'_{k+1} &= sd_{k+1}\sqrt{1+\frac{1}{B}} \\ sd_k^2 &= \frac{1}{B-1}\sum_{b=1}^{B}(log(W_{kb}^*) - \overline{l_k})^2 \end{aligned} \tag{3.9}$$

Using this approach, we determined k = 4.

Genome wide association study

Preparing the files required for GWAS

Fixing FAM files: In UK Biobank FAM files, the column for 'phenotype' includes batch that is coded with characters. In order to use BOLT-LMM (Loh et al., 2015), we updated all the entries in this column to numeric values (Loh, 2017).

'Remove' files for BOLT-LMM: BOLT-LMM accepts a list of individuals to be removed from the analysis as an input. These files are called 'remove' files and are in the FAM format. We prepared these files for i) withdrawn samples (n = 51), ii) samples that failed the quality control (n = 3,779), iii) samples that have information in PLINK files but lack BGEN files (n = 968).

Calculating the SNP statistics: In order to apply a quality filter for SNPs, using PLINK (Anderson et al., 2010), we calculated i) p-values for each SNP showing whether it deviates from Hardy-Weinberg equilibrium, and ii) Minor allele frequencies (MAF).

SNP Quality Control: We excluded SNPs that deviate from Hardy-Weinberg equilibrium (p ≤ 1e-6, n = 202,473) or with a minor allele frequency (MAF) smaller than 0.01 (n = 127,969). In total, we discarded 314,697 SNPs (Note that the numbers do not add up as these SNPs can overlap).

Phenotype File: We created a phenotype file that can be used as an input for BOLT-LMM, including the following fields: sex, age when attended assessment centre, cal-

culated BMI, assessment centre, ethnicity, batch, first 20 PCs, and self-reported diseases (one column per disease).

GWAS run using BOLT-LMM

For each disease, we run GWAS separately using BOLT-LMM with the following inputs:

- We remove the samples that are in plink files but now in bgen; samples that did not pass our QC; samples from the individuals who have withdrawn their data from the UKBB
- We excluded the SNPs that deviate from Hardy-Weinberg equilibrium, and have minor allele frequency lower than 0.01.
- We used Sex, Age, BMI, assessment centre, ethnicity, batch, and the first 20 PCs as covariates.
- To run the mixed-model, a reference LD score table is required. We used LD scores generated using 1000G European-ancestry samples, which is provided with the BOLT-LMM download.
- Genetic map for hg19 file provided in the BOLT-LMM website.
- We set 'bgenMinMAF' argument to $10^{-}2$ and 'bgenMinINFO' parameter to 0.5 to only include SNPs that pass these criteria.

GWAS Results

We removed MHC region (chr6: 28,477,797 - 33,448,354) from the analysis and considered positions with a p-value lower than $5 * 10^{-8}$ as a significant association.

Coding Variants

We used VarMap (Stephenson, Laskowski, Nightingale, Hurles, & Thornton, 2019) to map variants to proteins and domains. VarMap provides detailed information about coding variants, including annotations for the missense, synonymous, and nonsense variations. In our analysis, if a variant is not annotated as a coding variant in VarMap output, we assumed it is non-coding.

Genetic similarities between diseases

In order to calculate the overlap between diseases we used the number of SNPs that are significantly associated with both diseases, but corrected by the number that is expected by chance, if two diseases are independent:

$$GeneticSimilarity = \frac{N_{common}}{N_{d1} * N_{d2}} * N_{total} \tag{3.10}$$

N_{common}: Number of SNPs in common.

N_{dx}: Number of SNPs associated with disease X.

N_{total}: Total number of SNPs analysed in the study.

The statistical significance of these genetic similarities is calculated using binomial test, and the similarity is only considered for downstream analysis if p<=0.01. Moreover, the value is only calculated if two diseases do not have any hierarchical relationships in the disease hierarchy.

In order to assess the genetic similarity within age-of-onset clusters, we further used linear regression to correct this value by disease co-occurrences (risk ratios) and disease categories (binary data showing whether two diseases are of the same category). The 'corrected genetic similarity' is the residuals from this linear model.

LD Blocks

In order to assess the similarity between different diseases we use overlaps across significant associations and thus preferred not to do fine mapping. However, a significant challenge is that genomic variations are not independent but instead linked in the genome. To understand the effect of linkage or overcome it, we made use of linkage disequilibrium blocks previously defined for human genome (Berisa & Pickrell, 2016). We repeated the analysis for genetic similarity after collapsing all positions within an LD block and thus creating independent genomic loci (n = 1,703). We use binary information for LD blocks, *i.e.* blocks with at least one significant association are considered as a hit, and the rest are not.

Analysis of causal relationships between diseases

Using the LCV method developed by O'Connor & Price (2018), we tested the causal relationships between diseases. We used the R functions developed by the authors

and provided on GitHub (github.com/lukejoconnor/LCV/). We calculated the genetic causality proportion (GCP) between each disease pair, if the diseases have at least 10 significant variants and a significant heritability estimate as suggested by the developers (Zh≥7). We only calculated GCP if the diseases are not vertically connected on the disease hierarchy. Following the criteria applied by the developers, we considered pairs with FDR corrected p≤0.01 and mean GCP>0.6 as significant.

SNP to gene mapping

We map all SNPs analysed in GWAS to genes based on proximity and eQTL results.

Using proximity

Using VariantAnnotation (Obenchain et al., 2014), GenomicRanges (Lawrence et al., 2013), and TxDb.Hsapiens.UCSC.hg19.knownGene (Carlson & Maintainer, 2015) packages in R, we mapped the genomic coordinates for each SNP to genes. Specifically, if a gene is within the coding region, intron, 5' or 3' UTR, or 1kb down- or up-stream of the transcription start site, we annotated that SNP to the gene. As a result, we had 4,443,872 SNP-gene associations for 4,236,176 SNPs and 22,228 Entrez gene IDs. We used the Ensembl biomaRt (Durinck et al., 2009) package in R to retrieve HGNC symbols (17,994), Ensembl Gene IDs (20,507), and gene descriptions for the Entrez gene IDs obtained from TxDb.Hsapiens.UCSC.hg19.knownGene database.

Using GTEx eQTL data

Using SNP-gene associations based on GTEx v7 eQTL data (accessed on 04.09.2018) (Gamazon et al., 2018), we associated SNPs with the genes they could potentially regulate. We generated a combined tissue list, which associates SNP to the gene if there is at least one tissue in which there is a significant (p <= 5e-8) association. As a result, there are 2,166,300 unique SNPs associated with 15,312 Ensembl Gene IDs. We used the biomaRt (Durinck et al., 2009) package in R to retrieve HGNC Symbols (12,292), Entrez IDs (10,163), and gene descriptions.

Comparison of proximity and eQTL based mapping

Instead of only focusing on disease-associated SNPs, we first mapped all SNPs that we analysed to discover if there is a bias for certain genes (*e.g.* some genes could have many more SNPs because they are longer, or because they are already associated with certain traits and the chip is designed in that way). There were as much as 19,195 SNPs mapped to one gene (CSMD1) by proximity, whereas there were 82

SNPs per gene on average (median). The number of SNPs per gene was on average, higher for the mappings by eQTL (Figure A.53a). The maximum was 8473 SNPs for HLA-C gene and the median number of SNPs per gene was 218. However, we did not consider MHC region in our downstream analysis and thus this region is also excluded. The correlation between the number of SNPs per gene was low (rho = 0.13, Figure A.53b). Since the proximity based mapping is by definition dependent on the gene length, we also tested if there is a significant correlation between the number of SNPs per gene and gene length. While the correlation is low for gene mappings by eQTL (Spearman's correlation rho = 0.03, p = 1.073e-4), mappings by proximity show a high correlation as expected (Spearman's correlation rho = 0.87, p < 2.2e-16). This also explains the low correlation between eQTL and proximity-based mappings. We next checked the correlation between the number of SNPs per gene mapped by proximity but only to promoter region. The correlation between the number of SNPs and gene length decreased (rho = 0.21), and the correlation with the number of SNPs by eQTL slightly increased but was still low (rho = 0.08). Overall, we concluded that both eQTL data and proximity-based mapping could capture different information and decided to use both for the downstream analyses.

GWAS Catalog analysis

We accessed the GWAS Catalog on 30-07-2019 and used v1.0.2 e96 dataset (Buniello et al., 2019). We excluded all studies which used UK Biobank dataset (n = 190, data courtesy of GWAS Catalog team). Using the associations with a p-value lower than $5 * 10^{-8}$, we compiled significant associations between MAPPED_GENEs and MAPPED_TRAITs. We use GWAS catalog analysis to check if our GWAS hits are supported by previous studies and applied a Fisher test between all traits in GWAS catalog and the diseases in our study. p-values are corrected for multiple testing using FDR correction.

Analysis of the relevance with ageing

We downloaded GenAge human, GenAge model organism (Tacutu et al., 2018) and DrugAge (Barardo et al., 2017) data on Aug 13, 2019 and CellAge (Avelar et al., 2019) data on Oct 02, 2019 (CellAge data is kindly provided by Avelar et al.). We used HGNC Symbols for GenAge and CellAge genes. In order to compile genes that are targeted by the drugs in DrugAge database, using the drug names in DrugAge

data, we first compiled PubChem IDs using PubChem REST API (Kim et al., 2019). Using UniChem (Chambers et al., 2013), we mapped PubChem IDs to ChEMBL IDs (Gaulton et al., 2017). Next, using DGIdb (Cotto et al., 2018), we compiled the genes targeted by these ChEMBL IDs. As a result, we had 307 genes from GenAge human database, 902 genes from GenAge model organism database, 279 genes from CellAge database, and 714 genes targeted by DrugAge drugs. We next calculated the overlaps between these databases and the genes associated with multiple diseases or multiple categories in different age-of-onset clusters. To calculate the expected values and statistical significance, we used 10,000 permutations calculating the overlap for the same number of random genes among genes that can be detected by GWAS. Then, an odds ratio is calculated by dividing the observed value to the mean of expected values.

Functional Enrichment Test

Using the goseq package in R (Young, Wakefield, Smyth, & Oshlack, 2010), which takes the gene length bias into account, we performed a functional analysis of the genes associated with different age-of-onset clusters. Using GO categories with more than 10 and less than 500 annotated genes, we applied an enrichment test for the Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium & The Gene Ontology Consortium, 2019) Biological Process (BP), Molecular Function (MF), and Cellular Compartment (CC) categories. BY correction (Benjamini & Yekutieli, 2001) is applied to the p-values for all tests for all clusters and 3 GO Categories (BP, MF, and CC) combined. We considered associations with a BY-corrected p-value lower than 0.05 as significant. For the ease of visualisation and comprehension we selected representative categories for significant associations as follows: For each cluster and GO Ontology (*i.e.* BP, MF, CC) separately; i) Jaccard similarity index (*i.e.* number of genes in common divided by the number of unique genes in each category combined) is calculated between all significantly associated GO Categories; ii) Jaccard indices are hierarchically clustered and cut to k number of groups, where k is the minimum number of clusters which ensures median Jaccard similarity within a cluster is above 0.5; iii) The category with the highest average similarity to other categories in the same cluster is assigned as the representative.

Drug Repurposing

We searched for the drugs that specifically target multicategory genes in cluster 1, cluster 2, or cluster 1 and 2. Using the Fisher's exact test, we compiled the drugs in DGIdb (Cotto et al., 2018) specifically targets these genes ($p \leq 0.01$) and drugs that target only one gene in one of these cluster. The interaction data is compiled from DGIdb, and the names of the drugs are obtained from ChEMBL REST API (Gaulton et al., 2017).

Evolutionary Analysis

In order to test the mutation accumulation and antagonistic pleiotropy theories of ageing we used the risk allele frequencies in UK Biobank and 1000 Genomes super-populations (1000 Genomes Project Consortium et al., 2015). Risk allele is an allele that shows positive association with a disease. Since the SNPs are not independent and have similar allele frequencies in a given LD block, we analysed LD blocks instead of individual SNPs and used the median risk allele frequency for a given LD block. We used only the biallelic SNPs for this analysis. Allele frequencies for UK Biobank are calculated using BOLT-LMM and the allele frequencies for 1000 Genome super-populations are obtained from the vcf file provided on the 1000 Genomes project website. To test the antagonistic pleiotropy excess, we calculated the proportion of antagonistic vs. agonist SNPs within the same vs. different age-of-onset clusters using Fisher's exact test. We considered pleiotropic SNPs as agonist if the risk allele for two or more diseases are the same, and antagonist if the risk alleles are opposite. We only tested the risk allele frequency differences between cluster 1 and cluster 2. Also, we excluded any SNPs that are antagonistic within an age-of-onset cluster and agonist between clusters.

3.13 Discussion

The number and the incidence of diseases increase with age. *Is this just a result of the accumulation of time-dependent stochastic damage, or is there a common genetic component among ageing-related diseases, which can also be attributed to the ageing process?* In this study, we aimed to answer this question by comparing genetic associations and age-of-onset distributions of 116 self-reported diseases in the UKBB.

We first used an unsupervised, data-driven approach to classify diseases based on their age-of-onset profiles and found 4 main clusters (Figure 3.12); diseases that i) rapidly increase after 40 years of age, ii) increase after 20 years of age, iii) do not show any age-related pattern, and iv) peak at around 10 years of age. Notably, unlike previous studies (Fernandes et al., 2016; Wolfson, Budovsky, Tacutu, & Fraifeld, 2009), using this unsupervised approach, we detect a distinction between cluster 1 and cluster 2, which show age-dependency but distinct age-of-onset distributions.

We next found that the diseases in the same age-of-onset cluster show significantly higher genetic similarity, even when corrected for disease categories and co-occurrences (Figure 3.9a-b). While correcting for the disease categories and co-occurrences, we may remove some true positive signals from the analysis. However, this correction is particularly necessary for our study, as we use the same cohort for multiple diseases, and thus, diseases vertically connected on disease hierarchy and those that co-occur uses the same set of samples. Nevertheless, we retain a significant signal even after this correction, suggesting diseases on average are more similar to others that have a similar age-of-onset profile. We next asked if cause and effect relationships, rather than a common aetiology, may explain this observation (Figure 3.9c). We found 91 disease pairs with a significant causal association; however, there was no bias towards any specific age-of-onset cluster. Overall, our results suggest that even though each disease may have separate contributors, a genetic component may influence the age-dependency of diseases and the average age-of-onset.

By mapping the significantly associated variants to genes, we defined pleiotropic genes that are specific to each cluster or combinations of clusters. For example, *MTAP* and *CDKN2B* genes were associated with the highest number of diseases in cluster 1, spanning 3 disease categories; cardiovascular, endocrine, and eye diseases. *MTAP*, coding for an enzyme in polyamine biosynthesis, is found deleted in a number of cancers (Piñero et al., 2019). Being adjacent to a tumour suppressor gene, *CDKN2A*, its loss in cancer was attributed to being a bystander or regulator of *CDKN2A* expression (Schmid et al., 1998). Recent studies suggest *MTAP* itself is a tumour suppressor (Kadariya et al., 2013, 2009) and controls IGF1R activity (Xu et al., 2019), which has been repeatedly shown to have evolutionary conserved relevance for ageing (López-Otín et al., 2013). *CDKN2B* also codes for a tumour suppressor and is associated with multiple ageing-related diseases and parental longevity (Pilling et al., 2017). Notably, we did not analyse cancer associations in our dataset, and the diseases we analysed did not show a wide-spread co-occurrence with cancers that can bias the results (Figure A.50). Although cancer associations are not included, finding

these tumour suppressor genes is interesting; but potentially could be explained by the involvement of these genes in cell cycle regulation and cellular senescence.

Genes associated with multiple diseases or multiple categories in clusters 1 and 2 were enriched among genes previously associated with ageing (GenAge, CellAge, DrugAge databases), but not cluster 3 alone (Figure 3.10a). However, the association is established through a small subset of genes, as also previously described in the literature (Fernandes et al., 2016; Johnson et al., 2015). Especially the genes in CellAge database, which include genes regulating cellular senescence in human cell lines, and genes targeted by the drugs in DrugAge, which include drugs modulating lifespan in model organisms, showed strong associations. Eleven genes associated with multiple categories and diseases spanning all three clusters had a significant association with the GenAge model organism genes, which modulates longevity in model organisms. Overall, both cluster 1 and cluster 2 showed significant association with known lifespan modulators. This suggests the age-associated diseases in clusters 1 and 2 share a genetic component with longevity determinants.

Our analysis so far suggested a potential role of regulation of cellular senescence in age-related diseases, but we also found several functional categories associated with different clusters. Cluster 1 genes were involved in lipoprotein-related functions, cellular response, and cell cycle arrest. Cluster 2, on the other hand, had a different functional profile, including epithelial cell apoptosis, blood coagulation and fibrinolysis, and MHC II protein binding. Cluster 3 had a strong immune-related profile, as we expect from the diseases in this category. The genes that are associated with both clusters 1 and 2 were related to the nucleosome, chromatin structure, homeostasis, metabolic process, and gene silencing. Interestingly, multicategory genes in Cluster 1 & 2 and 3, also had associations with the nucleosome, chromatin structure, and gene silencing. Since cluster 3 did not have an age-dependent profile, these categories could represent pleiotropic genes in general. Genes in this group were also associated with the interleukin-7 signaling pathway. Overall, these results suggest, although both cluster 1 and 2 genes were linked to previously identified ageing-related genes, they have distinct functional profiles. Moreover, the genes associated with diseases spanning these two categories were related to homeostasis and metabolic process, which were previously reported to change with age and modulate lifespan (López-Otín et al., 2013).

Therapies targeting the multicategory genes in cluster 1 and 2 diseases may alleviate the effects of multimorbidity at late ages and extend health-span. We found

many drugs that are currently in use for different conditions, including age-related diseases we analysed, might be repurposed for other age-related conditions. Some of these drugs are already considered for multiple diseases from different categories. For example, acetohexamide, which targets the K-ATP channel, is in use for diabetes mellitus and is undergoing clinical trials for cataracts ("Compound," n.d.-a).

Can we use SNP-disease associations to understand the evolution of ageing? Extending the work of Rodríguez et al. (2017), we tested the two most widely appreciated evolutionary genetic theories of ageing (see Section 1.2.2). In line with the MA theory, allele frequencies of the variants associated with ageing-related diseases were higher than the frequencies of cluster 2 disease variants (*i.e.*, diseases that start as early as the age of 20). Variants associated with cluster 3, which includes immune-related diseases, were not significantly different than cluster 1, although cluster 3 diseases can occur even at an earlier age. The genes in this cluster were also enriched in immune response categories, which were previously suggested to be under long-term balancing selection in humans (Bitarello et al., 2018), explaining higher minor allele frequencies. We then showed that AP between different age-of-onset clusters is more common than expected. Comparing the allele frequencies of antagonistic variants between cluster 1 and 2 diseases in the UKBB and 1000 Genomes super-populations, we found support for the AP theory (*i.e.*, higher risk allele frequency for cluster 1 variants) only in non-European populations but not in the UKBB or European population. A potential explanation why we do not see the same trend in the UKBB and European population is that the probability of detecting significant association for two diseases associated with different alleles is higher for SNPs with a minor allele frequency closer to 0.5. Indeed, repeating the analysis with different effect size cutoffs (Figure A.49), we saw that stronger associations had a clear risk allele frequency difference in these populations as well. Importantly, the number of independent loci is minimal, and thus, although there is a signal supporting the AP theory, it is unlikely that the effect is genome-wide.

Overall, we showed that diseases clustered by their age-of-onset profiles show higher genetic similarity and are associated with specific genes, functions, and evolutionary characteristics. *How may these genes and functions affect whether a disease is an adult- or late-onset disease?* Here, we analysed the associations of the germline mutations, and thus, to understand their effects on the age-of-onset, we should consider their dynamic interaction with the environment, time-dependent somatic mutations, and the age-related changes in their expression. Somatic mutation load increases with age across almost all tissues, but the mutation profiles are tissue-specific (García-

Nieto et al., 2019). Thus, if somatic genetic mutations lead to the increased risk of disease, one may expect to see a clustering by tissue or disease category but not necessarily by the age-of-onset profiles. Genes with consistently increased mutational load across multiple tissues were found to be associated with autophagy, DNA damage repair, and some immune response categories. Although the functional associations we identified for late-onset diseases do not overlap with these, the interaction between the identified functions and those found to be associated with mutational load could still play a role. Another potential contributor is the change in gene or protein expression. Age-related changes in the transcriptome and proteome are not linear and around 20 and 60 years of age, include changes in the direction or rate (Anisimova et al., 2020; Colantuoni et al., 2011; Dönertaş et al., 2017). Similarly, a recent study investigating the changes in proteome after the age of 20, suggested three groups of proteins with changes at 34, 60, and 78 years of age (Lehallier et al., 2019). They further compared these groups to genes associated with Alzheimer's and cardiovascular disease. The authors showed Alzheimer's had a significant association with only the proteins that change at later ages (60 and 78). However, cardiovascular disease was also associated with proteins that change at 34, and this could potentially explain an earlier age-of-onset. Similarly, the genes defining cluster 1 and cluster 2 genes may have different gene or protein expression trajectories due to various environmental or intrinsic factors. .

3.13.1 Limitations

We employed one of the largest comprehensive available datasets, including medical and genotype data for almost half a million participants, to perform GWAS on human diseases. However, our study is not without limitations.

- We used the same cohort to perform multiple GWAS and thus, data is not independent. In our genetic analysis, we correct for disease co-occurrence values which consider if the same people have the diseases, however, it is still possible that the effect is not fully neutralised and we detect more commonality between diseases that co-occur together. This, however, is not likely to cause misinterpretations as most of the late-onset diseases do co-occur and the results still could have biological relevance.
- In this study, we only used self-reported diseases. Throughout the study, there were several instances that revealed reporting bias. For example, seemingly

unrelated diseases such as fractures and appendicitis had high co-occurrence values. Another example is the peaks in age-of-onset distributions of diseases at every 5 years. These both suggest the reporting is not perfectly accurate.

- We used the UKBB dataset, which has samples mostly from British descendants and thus our results should be validated in independent cohorts.
- In order to discover the common genetic component between diseases, we followed a GWAS approach and thus were limited to common variants. Another related outcome is that we only analysed common and complex diseases. These limit our ability to detect truly causal variants and instead we detect incremental contributions to risk of disease development. In general, considering that ageing is a gradual and subtle phenotype and most of the late-onset diseases are common, this approach may not prevent the analysis of many interesting diseases for ageing research. However, especially the analysis for evolutionary theories of ageing could benefit inclusion of rare variants and diseases. Moreover, it is important to note that complex diseases generally have strong interactions with environmental factors. However, we only included BMI as a covariate in our GWAS model. Especially studies aiming to find more causal genetic factors influencing multiple diseases should model environmental factors.
- In this study, we had a limited age range, covering individuals up to 65 years old and thus, could not analyse diseases of later ages, such as Alzheimer's Disease.
- In order to calculate the genetic similarities between diseases, we used a simple overlap-based measure. The main reason was to specifically test the overlapping variants so that we can further analyse that specific set of variants for ageing-relevance, functional associations, and evolutionary characteristics. However, this method is limited as it only considers 'significant' variants and ignores potential association between diseases that have limited sample size or more complex nature due to power issues. Future studies aiming to discover the associations between diseases may use probabilistic approaches to calculate disease-disease similarities that can also consider the factors we considered in this study, such as disease categories and co-occurrences.
- Lastly, the UKBB only includes genotypes and lack other 'omics data types. Also, although participants provide self-reported data and some of the biomarkers multiple times, they are genotyped for only once and the genetic profile is assumed to represent germline variants. As most of the age-related diseases are influenced by environment, especially to drive conclusions about the causal

nature of the identified variants, we would need data spanning somatic mutations, transcriptome and epigenome.

Future cohorts with a broader age range and data spanning different 'omics data, somatic mutations, health outcomes, and lifestyle information, can enable a better understanding of the genetic mechanisms of age-of-onset determination and establishing the causal link with candidate genes.

3.13.2 Conclusion

In this study, we have taken the first steps in understanding the relationship between ageing and diseases through a comprehensive analysis of disease age-of-onset profiles and genetic associations in the UK population. Diseases with more similar age-of-onset profile were genetically more similar, and this similarity was not explained by disease categories, co-occurrences, or cause-effect relation between diseases, suggesting biological pleiotropy and common aetiology. The shared genetic component between ARDs overlapped with some ageing-related genes, but diseases with different age-of-onset profiles had different functional associations. We also identified drugs targeting the common genetic component between ARDs, which could target multiple diseases at once and thus alleviate the effects of polypharmacy in the elderly. Lastly, variants associated with ARDs that start to occur at different ages had different evolutionary characteristics, supporting the mutation accumulation and antagonistic pleiotropy theories of ageing.

Chapter 4

Drug repurposing for ageing

Declaration

This work includes results that I have published as a research article in Aging Cell (Dönertaş et al., 2018). All the results presented in this chapter are products of my own work. I have conceived the study and performed all the analysis presented here. I must acknowledge the contributions of Matias Fuentealba, Linda Partridge, and my supervisor, Janet M Thornton. Their suggestions helped design this work better and improved it significantly. I also benefited from another article we have published as a review in Trends in Endocrinology and Metabolism (Dönertaş et al., 2019). The summary of existing drug repurposing studies in the introduction of this chapter stems from this article, and Matias Fuentealba had a significant contribution in compiling and reviewing these existing studies. The analysis of different studies, investigating their coverage and comparison is a product of my own work, but again benefits from our discussions with Matias. This work was also supervised by Linda Partridge and Janet M Thornton.

Dönertaş, H. M., Fuentealba Valenzuela, M., Partridge, L., & Thornton, J. M. (2018). Gene expression-based drug repurposing to target aging. Aging Cell, 17(5), e12819.

Dönertaş, H. M.*, Fuentealba, M.*, Partridge, L., & Thornton, J. M. (2019). Identifying Potential Ageing-Modulating Drugs In Silico. Trends in Endocrinology and Metabolism: TEM, 30(2), 118–131. * denotes an equal contribution

Data Availability

Supplementary tables are available as a BioStudies entry S-BSST330. Supplementary tables are referred using the corresponding file names throughout the text.

4.1 Introduction

4.1.1 The malleability of ageing

Since ageing is the major risk factor for poor functioning and disease, intervening to ameliorate its effects could also prevent multiple age-related conditions simultaneously. There is growing evidence for the feasibility of this approach. People who die when they are very old (100, 105, 110) show progressively less multimorbidity at the end of their lives (Andersen, Sebastiani, Dworkis, Feldman, & Perls, 2012; Christensen, McGue, Petersen, Jeune, & Vaupel, 2008). Thus, a healthy ageing phenotype in humans can be achieved, and if we could understand the mechanisms leading to it, we might be able to extend it to the general population. Moreover, work over the past 20 years has shown that lifespan of laboratory model organisms can be greatly extended by genetic and environmental interventions, which also improve health and function during ageing (Clancy et al., 2001; Lucanic, Lithgow, & Alavez, 2013; Xiao et al., 2013).

Dietary restriction (DR), a reduction in food intake that avoids malnutrition, can extend lifespan and induce a marked improvement in health during ageing in diverse organisms, including rodents (Fontana & Partridge, 2015; Kapahi et al., 2017). Two studies of rhesus monkeys subjected to DR found that the animals had lowered plasma triglycerides, diabetes, cardiovascular disease, sarcopenia, incidence of neoplasms, and brain atrophy, all features of ageing in humans (Colman et al., 2014; Mattison et al., 2012; Vaughan et al., 2017). However, compliance with DR regimes in humans is low, and for this reason, it is not a practical public health intervention.

Changes in diet are monitored by many nutrient-sensing systems, including the insulin / insulin-like growth factor and target of rapamycin (mTOR) signalling network. Many of the interventions further target components of the nutrient-sensing network (NSN), and decrease the activity of IGF / Insulin / TOR signalling (Fontana et al., 2010). This highly conserved network senses nutrients, growth factors, and stress status, and modulates the costly activities of the organism, such as metabolism, growth, and reproduction, accordingly. Genetic interventions that reduce the activity of the network have proved to extend lifespan in nematode worms, fruit flies, and mice (Fontana et al., 2010; Kenyon, 2010; Pan & Finkel, 2017). These long-lived mutants are protected against many natural pathologies of old age and also those associated with genetic models of age-related diseases. Mechanisms of ageing are highly conserved during

evolution, and the process shows a set of characteristic hallmarks of ageing, which are also present in the aetiology of age-related diseases (López-Otín et al., 2013; Partridge et al., 2018). Interventions that improve health during ageing and increase lifespan in laboratory animals do so by reducing the impact of one or more of these hallmarks.

Pharmacological intervention can also extend animal lifespan. The DrugAge database reports drug-induced lifespan extensions up to 1.5-fold for *C. elegans*, 1.1-fold for *D. melanogaster*, and 31% for *M. musculus* (Barardo et al., 2017). Some of these chemicals may mimic the effects of DR (Fontana et al., 2010). For example, resveratrol, which induces a similar gene expression profile to dietary restriction (Pearson et al., 2008), can increase lifespan of mice on a high-calorie diet, although not in mice on a standard diet (Strong et al., 2013). Rapamycin, directly targets the mTORC1 complex, which plays a central role in nutrient sensing network (NSN) and has an important role in lifespan extension by DR (Mair & Dillin, 2008). Rapamycin extends lifespan by affecting autophagy and the activity of the S6 kinase in flies. However, it can further extend the fly lifespan beyond the maximum achieved by DR, suggesting that different mechanisms might be involved (Bjedov et al., 2010). Nevertheless, the mechanisms of action for most of the drugs are not well known.

4.1.2 Previous in silico studies to discover anti-ageing drugs

Here we review 11 recent in-silico studies aiming to identify and prioritise pro-longevity drugs for animal models and humans. All such studies have been enabled by the development of powerful databases for the annotation and curation of genes/proteins (Ensembl (Herrero et al., 2016), UniProt (Consortium, 2017)), their associated functions and pathways (Gene Ontology (Ashburner et al., 2000), KEGG (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016), Reactome (Fabregat et al., 2018)) and chemical ligands and drugs interacting with them (ChEMBL (Gaulton et al., 2017), DrugBank (Law et al., 2014), STITCH (Szklarczyk et al., 2016), DGIdb (Griffith et al., 2013), PDB (Berman et al., 2000)) or affecting their expression (Connectivity Map (Lamb et al., 2006), CREEDS (Wang et al., 2016)), as well as drugs (DrugAge (Barardo et al., 2017), Geroprotectors.org (Moskalev et al., 2015)) and targets (GenAge (Tacutu et al., 2018), Ageing Clusters (Blankenburg, Pramstaller, & Domingues, 2018)) implicated in ageing and age-related disease mechanisms.

Although the published studies of drug repurposing to target ageing use different strategies and sources of data, they can be classified into two main categories: methods employing the structural information to predict drugs potentially interacting with proteins already identified as being involved in ageing, and methods based on the similarity between ageing-related drugs or genes based on molecular structure, interactions, pathways or networks.

Methods using structural information to find drugs targeting known regulators of lifespan in model organisms

The concept here is to find drugs which are known to target those genes that have been implicated in ageing. Two studies adopted methods based on the hypothesis that proteins or ligands with similar structures are likely to bind similar ligands or proteins, respectively, to predict drug-target interactions. The first of these (Snell et al., 2016) aimed to identify novel drugs targeting 3 specific temperature sensing proteins implicated in ageing in the rotifer Brachionus manjavacas *(TRP7, S6P, FhBC)*. The authors used a virtual screening software called FINDSITEcomb that combines protein modelling with sophisticated threading approaches to model the target. The pockets in the model are then compared against the pockets in experimentally determined structures of proteins with ligands or modelled structures with known binders. The ligands of the top 100 ranked pockets are then compared against a library of screened ligands and ranked by ligand similarity. The authors screened 1,347 FDA approved drugs in silico and tested four drugs for each target experimentally in the rotifers for their effects on lifespan and health-span. Five out of the 12 compounds tested significantly increased the rotifers' lifespan. Changes in health-span, approximated by swimming speed, reproduction and mitochondrial activity, were also observed. In a subsequent study by the same authors (Snell et al., 2018), the number of proteins analysed was expanded to a set of ageing-related genes found in other animal models with orthologues genes in rotifers. This time a total of 94 targets were screened in silico using the FINDSITEcomb software. The top 1% binding compounds for each target were further ranked by their cumulative lifespan extension achieved by genetic interventions into their targets as taken from experimental model organism data, and filtered according to availability and previously predicted side effects (Zhou, Gao, & Skolnick, 2015). From the 31 drugs experimentally tested in rotifers by two ten-days survival screens, seven drugs were further tested in two whole life survival analyses, two of them resulting in a median lifespan extension of 13-42%. The pro-longevity

effect of these drugs was observed even when drug treatment was initiated in middle age.

Another in silico screening study was restricted to a single gene, AMP-activated protein kinase (*AMPK*), whose activation partially mediates the effects on ageing of dietary restriction (DR) (Mofidifar, Sohraby, Bagheri, & Aryapour, 2018). To find new molecules to activate AMPK and theoretically mimic DR, Mofidifar et al. (2018) performed virtual screening using molecular docking of 1,908 FDA approved drugs. The interaction between the top-ranked compounds and their targets was then further checked by more detailed molecular dynamics. The study reported 4 compounds with predicted high affinity for AMPK, but these were not tested experimentally.

Similarity-based methods to discover drugs targeting ageing

Using a priori information on known ageing-related genes, pro-longevity drugs or gene expression profiles, several studies have implemented a series of similarity-based approaches to identify novel anti-ageing drugs.

Finding drugs that target known ageing-related genes. Given that drugs targeting ageing-related gene products are expected to affect the ageing process, Fernandes et al. (2016) focused on finding drugs that target human genes whose orthologues in animal models are associated with longevity. The drugs were ranked by the likelihood of targeting ageing-related genes among all targets. For this calculation, only inhibitory drugs interacting with anti-longevity genes and activators targeting pro-longevity genes were considered. In total, 376 drugs were obtained of which 20 were considered to be statistically significant. Thirteen targeted histone deacetylases, and three were previously associated with lifespan extensions in animal models. Recently, a study I have contributed used a composite set of ageing-related genes with direct evidence for influencing human ageing, together with physical and functional drug-protein interactions, to implement a similar gene-set overlap analysis (Fuentealba et al., 2019). This study also considered other levels of biological actions including pathways, functions and protein-protein interactions. Three of the top 10 compounds ranked highest on an aggregate score were previously shown to increase lifespan in animal models, and seven had been proposed to affect longevity by other drug-repurposing methods. The pro-longevity effects of the top-ranked compound (tanespimycin) was experimentally validated in *Caenorhabditis elegans*.

Finding drugs similar to known pro-longevity drugs. An alternative approach is to find drugs similar to known pro-longevity drugs using machine learning, which is a strategy well-suited for prediction tasks. Liu et al. (2016) attempted to predict new pro-longevity drugs for *C. elegans*. They adopted a semi-supervised algorithm trained with high-confidence pro-longevity drugs derived from an experimental screen for *C.elegans* (Ye, Linton, Schork, Buck, & Petrascheck, 2014), together with their associated ageing-related genes curated from the literature and the GenAge database (Tacutu et al., 2018). They produced a rank-ordered list of 785 drugs with a potential to increase lifespan in worms, with experimental validation for one drug in their list, using a lifespan assay. A separate machine learning approach was trained with chemical descriptors of known pro-longevity drugs, and functional annotation of their targets (Barardo et al., 2017). Using a supervised algorithm (*i.e.* random forest), they generated a ranked list of drugs predicted as lifespan extending compounds, although no validation was performed.

Comparing transcriptome signatures from ageing and drugs. The Connectivity Map (CMap) Resource provides drug-induced expression profiles for 1,309 compounds (version 2). Comparing these profiles with ageing-related gene expression signatures using a gene-set enrichment analysis can reveal drugs that generate changes in expression correlated (positively or negatively) to those seen in ageing (or any other biological process or disease). This approach requires no a priori list of 'ageing genes' and can therefore potentially identify new targets, based solely on expression profile similarities. Calvert et al. (2016) used dietary restriction (DR) expression profiles in rats and rhesus monkeys to find DR mimetics. They identified 11 drugs that could potentially increase lifespan by mimicking DR. They experimentally tested several of the drugs in *C.elegans* and most extended lifespan. I also follow a similar methodology and instead aim to find drugs that target genes that show the most reproducible age-related changes in gene expression in the human brain. The results of this study are presented under Section 4.2. Importantly using this data-based approach I aim to identify novel drugs and genes, not previously associated with ageing. More recently, after the publication of the work I present in this Chapter, Yang et al. (2020) used a network-based methodology, called ANDRU (ageing network-based drug discovery). Instead of relying on model organisms, this approach was also driven by human transcriptome data (GTEx) from young and old adipose and artery tissues and signatures from the CREEDS database (Wang et al., 2016) to identify differentially expressed genes within the ageing-related networks and drugs reversing these changes. They report three distinct drugs ranking as the first five. Although none is

previously reported as a life-span modulator, these drugs target pathways that change in expression with age, such as metabolic enzymes and lipid metabolism.

Approaches to prioritise drugs for testing

One of the major challenges in developing anti-ageing drugs is experimental validation. Since clinical trials involve many ethical considerations and are very expensive, such drugs are pre-tested in model organisms. In this spirit, Aliper et al. (2016), aimed to predict which pro-longevity drugs previously tested in *C.elegans* could work in humans. Using young and old human stem cell expression profiles and an algorithm called Geroscope that maps the gene expression changes with age to ageing-related signalling pathways, they ranked a set of candidate drugs by their likelihood of targeting these pathways. To do this they calculate the pathway activation strength (PAS) for each drug. They shortlisted ten compounds with pro-longevity effects in *C. elegans*, and tested six of them for geroprotective effects in senescent human fibroblast cultures. While the majority of tested drugs improved senescence-associated phenotypes, one drug (PD-98059), a highly selective MEK1 inhibitor, also showed life-prolonging and rejuvenating effects.

Comparably, to assess which 'human' drugs and chemicals are likely to modulate the *C. elegans* and *Drosophila melanogaster* orthologue of the target, Ziehm et al. (2017) developed a method to rank chemicals binding to genes implicated in human ageing. They generated an empirical scoring function that considers the conservation of the domain and binding site at the sequence level between the animal and the human protein, predicted binding energy for the compounds for the human targets and experimental bioavailability, in addition to scores for drug-likeness, promiscuity, purchasability and development status. Although the authors provided no experimental validation, they conducted a comprehensive literature mining and molecular docking procedure to validate their results.

4.1.3 Research objectives

Although previous studies tried to discover drugs that can affect ageing, they all focus on genes or drugs related to lifespan regulation. The role of these drugs in promoting healthy ageing in humans is still an open question. In this study, using gene expression data for human brain ageing, we aimed to discover not only new pro-longevity

drugs but also those that can improve health during ageing. Human brain undergoes substantial structural changes with age, including changes in brain weight, white and grey matter volumes. Accompanied by the altered intercellular communication and synaptic loss, these changes bring about cognitive decline, neurodegeneration and memory loss (Salthouse, 2009). The biological processes showing a change in expression include pathways related to synaptic and cognitive functions as well as proteostasis (Lu et al., 2004), suggesting gene expression changes in the ageing brain could be used as a surrogate to find drugs to target detrimental effects.

Here, we extended the previous approaches to identification of new anti-ageing drugs for humans, by focusing directly on human ageing data. We used a framework that does not require any prior knowledge and is thus robust to biases in the literature and databases on ageing. Moreover, using human age-series data, this methodology has the potential to discover drugs affecting both life- and health-span. Through a meta-analysis of multiple gene expression datasets, we aimed to first compile a robust signature that can characterise ageing in the human brain. We then aimed to identify a list of potential drug candidates that could influence human brain ageing, using drug-induced RNA expression profiles deposited in the Connectivity Map (CMap) (Lamb et al., 2006). We also provide an analysis of the results combined with the results from other *in silico* studies.

4.2 Gene expression-based drug repurposing to target ageing

4.2.1 Analysis of age-related changes in RNA expression in human brains

We analysed data from seven, published, microarray-based studies of age-related changes in RNA expression (Barnes et al., 2011; Berchtold et al., 2008; Colantuoni et al., 2011; Kang et al., 2011; Lu et al., 2004; Maycox et al., 2009; Somel et al., 2010, 2011). The data came from 22 different brain regions, and the ages of the donors ranged from 20 to 106 years (Figures 4.1a, A.54). The data for each brain region in each study were analysed separately, resulting in 26 datasets.

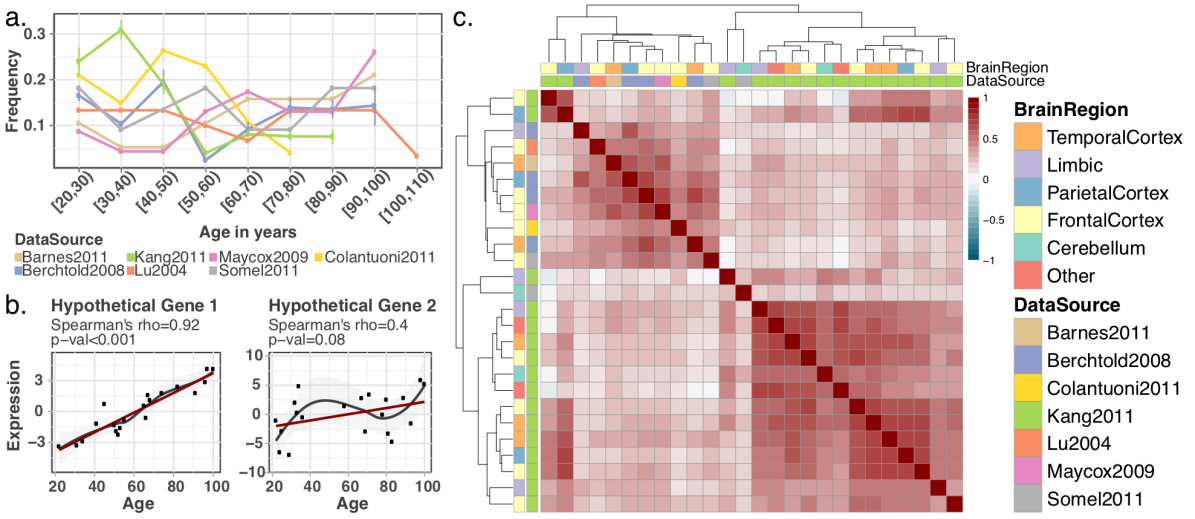


Figure 4.1 *a) Age distribution of the brains from which the data sets used in the study were derived. The error bars show the standard deviation of the sample frequency for different brain regions in data sources with multiple brain regions. b) Hypothetical gene expression plots, demonstrating how Spearman's correlation coefficient and p-value behave when the association is weak or nonmonotonic. c) Pairwise Spearman's rank correlation coefficients across data sets. The intensity of the colours on the heatmap shows the magnitude of the correlation coefficient*

To characterise the association between the gene expression and age, we calculated the Spearman's correlation between the expression level and age, for each gene, in each dataset separately. We first calculated the number of significant changes (FDR corrected p<0.05) in each dataset (Figure A.55). While there were two datasets with a large number of significant changes, most of the datasets did not show substantial significant change. This can be explained by several factors, most importantly i) most of the datasets had a small sample size, providing insufficient power to detect changes in most of the cases, and ii) Spearman's correlation test calculates significant monotonic changes, whereas it is likely that many of the changes are not exclusively monotonic throughout ageing. Thus, we applied another approach, using the correlation coefficient to capture significant trends across datasets, instead of within a dataset (see Section 4.4). While the p-value is affected by the number of the samples and the strength of the monotonic relationship (Figure 4.1b), the sign of the correlation coefficient can be used to capture consistent trends of up- or down-regulation once coupled with an appropriate testing scheme. This strategy requires the datasets to be concordant and reflect genuine age-related changes. We first investigated if this assumption was valid. To assess the concordance among datasets, we used Spearman's corre-

lation coefficients and calculated the correlation between expression-age correlations between datasets (Figure 4.1c). We observed a weak correlation with a median pairwise correlation coefficient of 0.29. To calculate the significance of this correlation, we developed a stringent permutation scheme specifically designed to account for the dependence between genes as well as the datasets (see Section 4.4). We concluded that a median correlation coefficient of 0.09 would be expected by chance and that our observation (median ρ=0.29), is statistically significant (p<0.001). Based on these correlations, datasets clustered according to the data source rather than to the brain region. This observation is in line with the previous studies suggesting that ageing-related changes are small and heterogeneous, making them difficult to detect (Somel et al., 2006). We therefore tested for significant correlations across datasets from different studies. When we excluded the correlation coefficients among the datasets generated by the same studies, we still observed a significant correlation coefficient of 0.22 (permutation test p<0.001, ρ=-0.002 would be expected by chance), showing that we have significant correlations among different data sources as well. Using these correlations, we proceeded to compile the ageing-signatures, reflecting consistent trends.

4.2.2 Defining the ageing signature

To construct a robust ageing signature, we identified the age-related changes that were observed across all datasets, irrespective of the effect size. We thus focused on global age-related changes in the brain, rather than region-specific changes, and the set of genes that showed gene expression changes in the same direction across all datasets (Figure 4.2a).

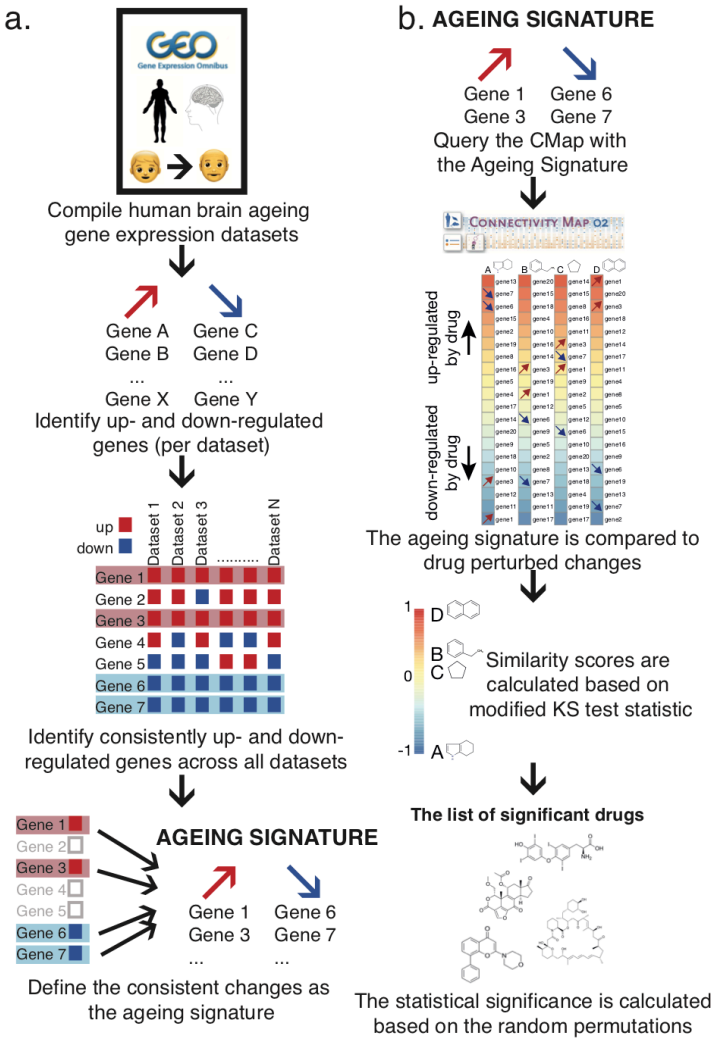


Figure 4.2 *Method summary for (a) compiling the ageing signature and (b) theCMap algorithm.*

This profile consisted of only 100 up- and 117 down-regulated genes (Supplementary Table S2, Figures A.56, A.57), 'the ageing signature'. To establish the robustness of the ageing signature, we calculated the statistical significance of the number of consistent changes with the same permutation scheme used to test the correlations among datasets. This methodology randomises the age of each individual, making it possible to test the null hypothesis where there is no association between expression and age while retaining the dependence between genes and datasets (see Section 4.4 for details). The number of consistent expression changes across brain regions was significant ($p<0.001$, Figure A.58a,b), establishing that the ageing signature indeed has biological meaning.

To further test the robustness of the ageing signature, we used an independent data set, consisting of gene expression in human brain generated by the GTEx Consortium (GTEx Consortium, 2015), consisting of data from 99 individuals, 13 brain regions and ages between 20-79 (Figure A.56b, Supplementary Table S1). These data were generated using RNA-Seq, allowing us to assess the robustness of the ageing signature to different technology platforms. We used the pipeline previously applied to the microarray data to calculated age-related expression changes for each gene in each brain region separately. The pairwise correlations between the GTEx datasets were higher than with the other dataset, and they tended to cluster together (Figure A.59). We found 1189 up- and 1352 down-regulated genes that showed the same direction of change across all GTEx brain regions (Supplementary Table S2), compared with only 100 and 117 in the microarray ageing signature. A likely explanation is that samples from different brain regions from the same individuals were used in GTEx, while the microarray ageing signature combined seven independent studies and different microarray platforms. The numbers of shared expression changes based on permutations were 127 and 131.5, for down- and up-regulated genes, suggesting a higher false positive rate in the GTEx dataset. Nevertheless, the numbers of consistent up- and down-regulated genes in the GTEx dataset were also significant (p=0.001, Figure A.58c-d). The numbers of common up- and down-regulated genes across the GTEx and microarray signatures were 50 and 48, respectively, both statistically significant (binomial test $p < 2.2e-16$ for both), demonstrating that the ageing signature was reproducible.

4.2.3 Biological processes associated with the ageing signature

We next investigated the biological processes associated with the microarray ageing signature. Using the genes that were consistently expressed in all data sources as background, we did Gene Ontology enrichment tests for consistently up- and down-regulated genes, separately (Figure 4.3, Supplementary Tables S3-4). Down-regulated genes were enriched in synaptic functions and biosynthetic processes (FDR corrected $p<0.05$), while differentiation and proliferation-related categories showed enrichment for the up-regulated genes (FDR corrected $p<0.05$). These results are consistent with the findings of earlier brain ageing transcriptome studies (Lu et al., 2004; Naumova et al., 2012; Xue et al., 2007). Oddly, ossification-related biological processes also showed significant enrichment for the up-regulated genes. However, except for one gene, these ossification-related categories shared all genes with the more generic

development-related categories. Thus, this result could be interpreted as a general up-regulation of the development-related processes rather than ossification-related categories.

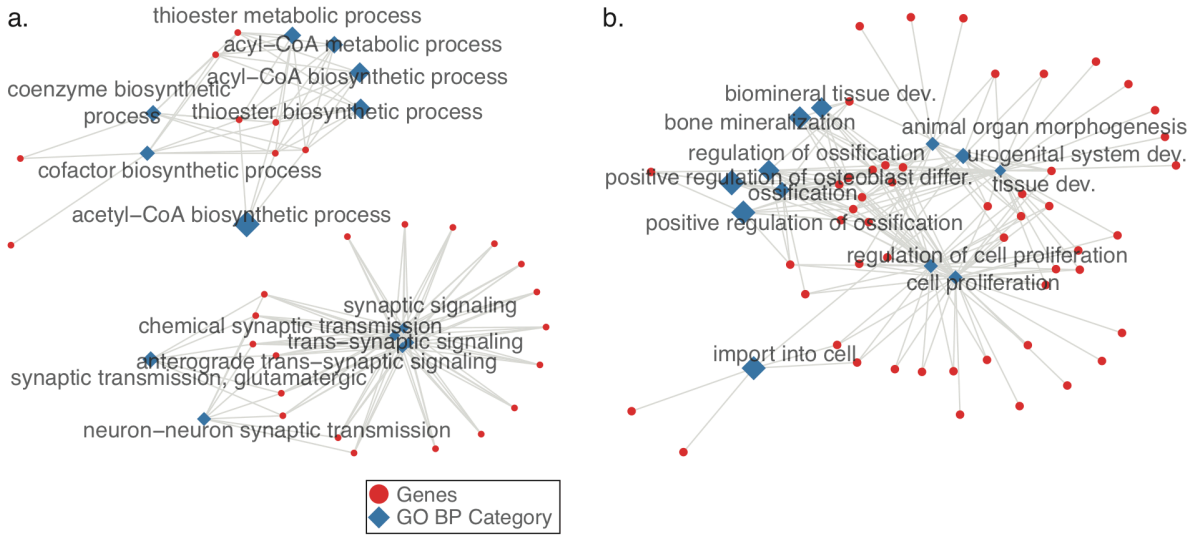


Figure 4.3 *Gene Ontology Biological Process Categories significantly enriched in (a) down- and (b) upregulated genes in the microarray ageing signature. Red circles represent the genes, and diamonds show the significantly associated GO categories, where FDR adjusted $p < 0.05$. The size of the diamonds represents the effect size (odds ratio).*

We repeated the enrichment analysis using the GTEx ageing signature and found 194 and 256 GO BP categories as significantly associated with down- and up-regulated genes, respectively (Supplementary Tables S7-8). Since the number of genes in the GTEx signature is higher, we had more power to detect smaller changes and thus had a higher number of significant associations. However, the effect sizes (odds ratios) for each GO BP category calculated for microarray and the GTEx ageing signature were correlated (Figure A.60). Correlations between the odds ratios calculated for all of the GO categories calculated in both methods were 0.46 and 0.37, for the enrichment in the down- and up-regulated genes, respectively. Correlations increase when we considered only the GO categories that are significantly associated with at least one of the ageing signatures; 0.55 and 0.60, for the enrichment in the down- and up-regulated genes, respectively. This further shows that the ageing signatures are robust. The categories enriched in down-regulated genes included biological processes related to neuronal and synaptic functions, autophagy, post-translational modifications, and translation (see Supplementary Table S7 for the full list). Processes related to response pathways, immune response, macromolecule organisation and lipid

metabolism showed enrichment in up-regulated genes (see Supplementary Table S8 for the full list). Interestingly, categories related to ossification were also among the GO categories significantly associated with up-regulation, based on GTEx data.

4.2.4 Mapping the ageing signature onto drug-perturbed expression profiles

The Connectivity Map (CMap) is a database of drug perturbed gene expression profiles (Lamb et al., 2006). It consists of 6100 gene expression profiles for 1309 drug perturbation experiments performed on five different cell lines. The CMap algorithm uses a modified Kolmogorov-Smirnov test statistic to calculate the similarity of a drug-perturbed expression profile to the gene expression profile used to query the database. A positive similarity score means that the drug-perturbed expression profile is similar to the query, whereas a negative score indicates a negative correlation (Figure 4.2b). Based on the random permutations, the statistical significance of the similarity score for each drug is calculated. Thus, the p-value shows the probability of finding the same association when a random signature is supplied. We queried the CMap database and identified drugs that showed significant associations in either direction with the ageing signatures. To determine the robustness of this procedure, we queried the CMap data using the microarray ageing signature, and the top 500 up- and 500 down-regulated genes from the GTEx ageing signature (see Section 4.4). The correlation was significant (r=0.52, p<2.2e-16, Figure A.61a) showing that the two ageing signatures produce reproducible overlaps with the CMap database. In order to test the reproducibility and not bias the results due to the technology used to generate the data, we preferred not to combine ageing signatures but report the resulting drug hits from the two signatures separately. Nevertheless, it is noteworthy that the drug similarity scores, generated using the overlap between signatures, show significant correlation with the lists generated using both microarray and GTEx signatures (Figure A.63).

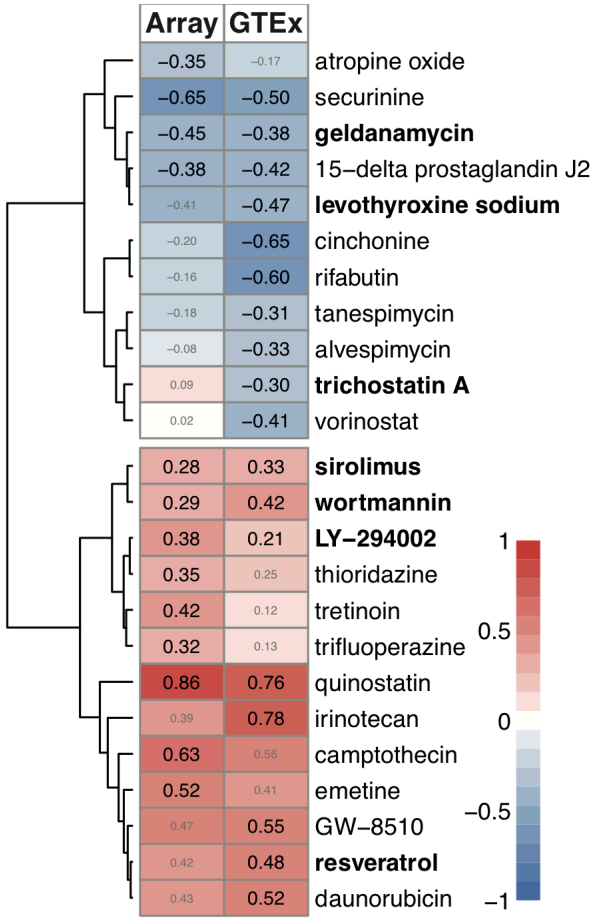


Figure 4.4 *Similarity score table for the drugs having at least one significant association with the ageing signatures. Each row corresponds to a drug and columns correspond to two independent ageing signatures—using the microarray and the GTEx data sets. The size of score labels indicates the significance of the results (FDR-corrected p < 0.05). The row labels written in bold indicates the drugs in the DrugAge database.*

Querying the CMap database, we identified 13 drugs significantly associated (FDR corrected p<0.05) with the microarray ageing signature (Table 4.1, Figure 4.4). Four of these drugs were previously shown to extend lifespan in worms or flies in at least one experiment (Supplementary Table S9). The number of pro-longevity drugs rediscovered using this methodology was statistically significant (p=0.004), and only one drug would be expected based on 10,000 random permutations of drugs. Repeating the same analysis with the GTEx ageing signature, we identified 18 drugs, seven of which were in common with the microarray results, including the four known pro-longevity drugs. In total, 24 drugs were significantly associated with at least one of the ageing signatures. The correlation between the drug similarity scores for these 24 drugs calculated based on the microarray and GTEx data was 0.88 (p<9.44e-09,

Figure A.61b), indicating high concordance. Since the similarity scores show high correlation, the rest of the results will be presented for the 24 drugs that are associated with at least one of the ageing signatures.

Table 4.1 *The drugs that are significantly associated (FDR corrected p<0.05) with at least one of the ageing signatures. Drug names denoted with (*) shows the drugs in DrugAge database. "Score" is the mean similarity score given in the CMap output, based on the KS test. The similarity scores denoted with (*) show the significant associations. The list is ordered by the mean of the similarity scores from negative to positive. Target or mechanism of action is manually curated from literature or extracted from CHEMBL, DrugBank, and PubChem databases. The targets denoted with (*) are found in the GenAge model organism or GenAge human databases.*

Drug name	Array	GTEx	Target or Mechanism of action
Securinine	-0.65	-0.50	GABRA1-5, GABRB1-3
Levothyroxine sodium*	-0.41	-0.47	THRA, THRB*
Cinchonine	-0.20	-0.65	CYP2D6*
Geldanamycin*	-0.45	-0.38	HSP90AA1*
15-delta prostaglandin J2	-0.38	-0.42	PPARG*
Rifabutin	-0.16	-0.60	BCL6
Atropine oxide	-0.35	-0.17	-
Tanespimycin	-0.18	-0.31	HSP90AA1*
Alvespimycin	-0.08	-0.33	HSP90AA1*
Vorinostat	0.02	-0.41	HDAC1*, HDAC2*, HDAC3*, HDAC6
Trichostatin A*	0.09	-0.30	HDAC6, HDAC7, HDAC8
Trifluoperazine	0.32	0.13	DRD2, DRD3, DRD4, HTR2A, HTR2C
Tretinoin	0.42	0.12	RARA*, RARB*, RARC*
LY-294002*	0.38	0.21	PI3KCG*
Thioridazine	0.35	0.25	DRD2, DRD3, DRD4, HTR2A, HTR2C
Sirolimus*	0.28	0.33	mTOR*
Wortmannin*	0.29	0.42	PI3KR1*, PI3KCA*, PI3KCG*

Resveratrol*	0.42	0.48	SULT1B1, YARS, LTA4H, TTR, NQO2, PTGS2*, PTGS1, MAT2B, CSNK2A1, CYP3A4*, ESR1*, PPARG*, SIRT1*, SIRT5, CYP1A2, CYP1A1, CYP1B1, NCOA2, TNNC1
Emetine	0.52	0.41	Protein Synthesis Inhibition
Daunorubicin	0.43	0.52	TOP2A*, TOP2B*
GW-8510	0.47	0.55	CDK2, CDK5*
Irinotecan	0.39	0.78	TOP1*
Camptothecin	0.63	0.56	TOP1*
Quinostatin	0.86	0.76	PI3KCA*

Overall, the method re-discovered seven known pro-longevity drugs in DrugAge database (p=0.00023, based on 100,000 random permutations); resveratrol, LY-294002, wortmannin, sirolimus (also known as rapamycin), trichostatin A, levothyroxine sodium, and geldanamycin (Supplementary Table S9).

4.2.5 Targets of the drugs

Next, we investigated the targets of these 24 drugs, using the ChEMBL, PubChem and DrugBank databases as well as through manual curation of the literature (Table 4.1), and whether these targets were previously implicated in ageing, using GenAge human and model organism databases (Figure 4.5). Except for four (rifabutin, securinine, thioridazine, trifluoperazine); all drugs or their target genes had been previously implicated in ageing. Moreover, the drug-target association network showed several clusters with multiple drugs sharing the same targets: i) quinostatin was in the same cluster with two known pro-longevity drugs, wortmannin and LY-294002, targeting PI3K subunits, ii) tanespimycin and alvespimycin shared the same target with another DrugAge drug, geldanamycin, targeting HSP90, iii) vorinostat shared one of its targets, HDAC6, with trichostatin A, another DrugAge drug, iv) thioridazine and trifluoperazine had dopamine and serotonin receptors as targets and v) irinotecan and camptothecin shared TOP1 as their target. The fact that drugs targeting the same proteins / acting through the same mechanism had similar CMap similarity scores (Figure

4.4) further shows that our results are biologically relevant and reflects potential mechanisms to target ageing.

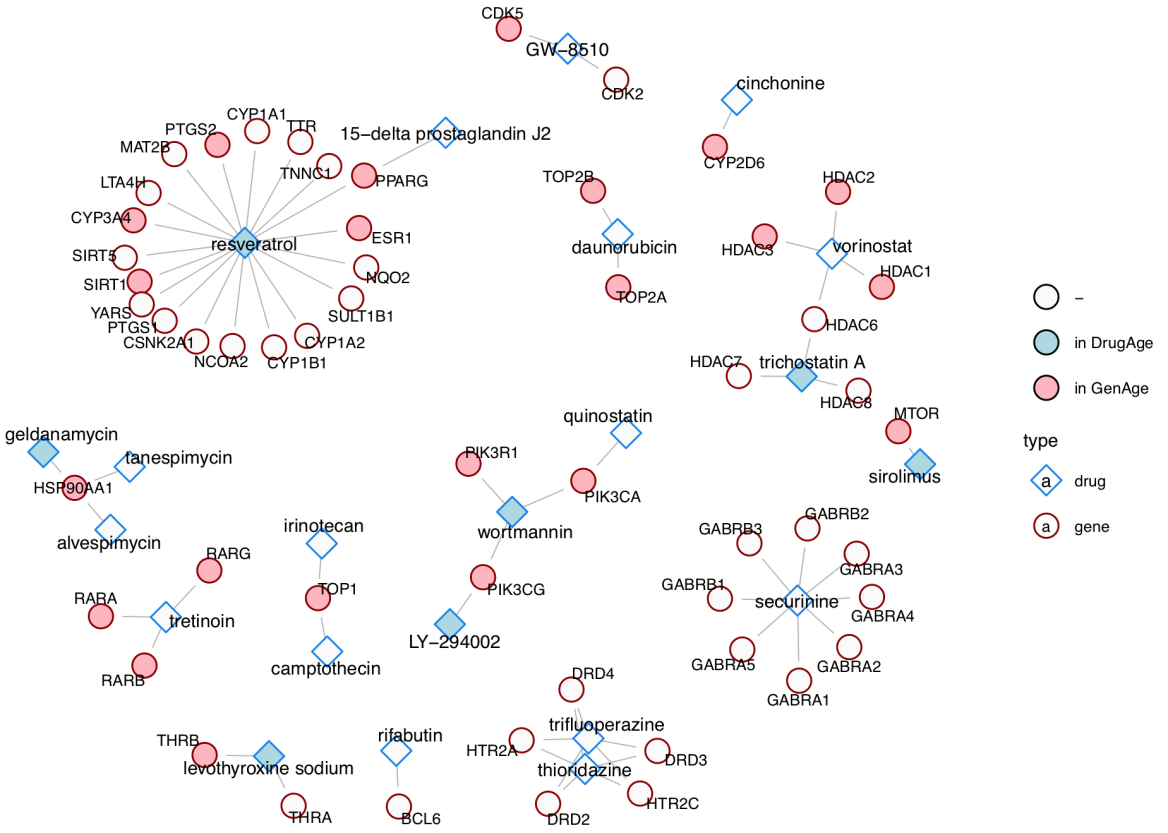


Figure 4.5 *Schematic representation of the drug–target associations as a network. Blue and red nodes show drugs and targets, respectively. The drugs with a light blue background are present in DrugAge database and the targets with a pink background are in either GenAge model organism or GenAge human databases.*

4.2.6 Drugs can act both by reversing ageing effects and mimicking responses

The general expectation from an 'omics-based drug repurposing study is the identification of drugs that can reverse the abnormalities detected in the disease state *i.e.* identification of drugs with negative similarity scores (Duran-Frigola, Mateo, & Aloy, 2017). Following the same logic, one might expect drugs with anti-ageing potential to have negative scores. Interestingly, some of the known pro-longevity drugs had positive similarity scores to the ageing signatures, meaning that the drug-induced profile

was similar to the ageing signature. A plausible explanation for this observation is that ageing signatures may partly reflect cellular defense responses, helping to alleviate the damaging effects of ageing.

4.2.7 Characterising the biological functions associated with pro-longevity drugs

In order to identify the biological processes associated with the changes that were reversed or mimicked by the pro-longevity drugs, we used the drugs documented in DrugAge, that were re-discovered in our analysis. We grouped the microarray ageing signature into five categories, based on the expression changes in ageing (up or down), and the pro-longevity drug-induced profile (up, down or inconsistent) (Supplementary Table S5). To compile the pro-longevity drug profile, for each probe-set in the microarray ageing signature, we asked if the seven DrugAge drugs induced similar changes. If the same direction of change was induced by more than half of these DrugAge drugs, then we included these changes in the pro-longevity drug profile (see Section 4.4 for the details). We then analysed the biological processes associated with the genes in these categories. The number of genes is small, with no significant changes after multiple test correction. We therefore report the associations based on the highest odds ratios only. For genes down-regulated in ageing, the changes mimicked by the drugs were associated with autophagy and metabolic processes (Supplementary Table S6), while for up-regulated genes, pro-longevity drugs tended to mimic the changes in protein complex / cellular complex assembly-related functions and to reverse the changes observed in protein localisation and immune-related functions (Supplementary Table S6). These findings are consistent with the mechanism of action for the most well-known pro-longevity drugs. For example, sirolimus (rapamycin) is an immunosuppressant approved for human use, and similar drugs can enhance the response of elderly humans to immunisation against influenza (Mannick et al., 2014).

4.2.8 Similarity among significant drugs based on the expression changes at the functional level

In order to analyse the similarities among drugs based on expression level changes, we performed a gene-set enrichment analysis (GSEA) for the drug-induced expres-

sion profiles, including all genes irrespective of whether a given gene is in the ageing signature (see Section 4.4). To measure the similarity between drugs, we calculate the Spearman's rank correlation coefficients between the enrichment scores and then cluster drugs based on these correlation coefficients. Notably, drugs targeting the same proteins or pathways, *e.g.* PI3K inhibitors LY294002, wortmannin and quinostatin, clustered together. Using this functional level approach, we grouped drugs into four groups: i) known pro-longevity drugs, ii) drugs clustering together with at least one pro-longevity drug, iii) drugs which clustered together but did not cluster with any known pro-longevity drugs, and iv) drugs which did not cluster with any other drugs (Figure A.64).

4.2.9 Ageing signature in other tissues

Since our analysis is based on an ageing signature compiled using only the brain tissue, we also explored if this signature is representative of the other tissues. A plausible way to approach this question is repeating the same analysis using other tissues. However, it is not straightforward because i) the number of datasets available for the other tissues limits the capacity of our approach to compile consistent signatures, increasing false positives, and ii) we find that the ageing-related changes in other tissues are not as consistent as in brain (Figure A.65a). Thus, we choose another approach and asked if the direction of change for the ageing signature we compiled is similar to the direction of change in other tissues (Figure A.65c). We also tested the significance of the similarity in the direction of change based on random permutations. As expected, GTEx brain data showed the highest percent similarity to the array signature. 8/35 datasets showed more dissimilarity for the down-regulated genes (*i.e.* percent similarity was lower than 50%), while only two were statistically significant, namely, liver and atrial appendage. Similarly, only 6/35 datasets showed more dissimilarity for the up-regulated genes, while none was significant. We repeated the analysis with the GTEx signature and observed similar results with only exception that there were five datasets with significant dissimilarity for the down-regulated genes (Figure A.65e). Thus, it is possible that brain signature includes some brain-specific changes but based on significant similarity, we can say it is also representative of other tissues.

4.3 Comparing the results with other eleven in silico studies

My study and all the other studies described in the Section 4.1.2 had different aims, methods, and data sources. To facilitate their comparison, we have summarised each study in terms of i) the drugs identified, ii) the genes targeted by these drugs, and iii) biological pathways (KEGG) in which these genes are involved (Figure 4.6). Additionally, we compared the results with the manually curated databases of ageing-related genes (GenAge) and drugs (DrugAge).

Drugs: Overall only 12% of all DrugAge drugs are prioritised by at least one study (41 of 346 drugs in DrugAge), with one in every four drugs discovered already present in DrugAge, reflecting the prioritisation process and the low number of drugs reported as significant by each study (15 drugs on average). In addition, the 163 drugs identified usually differ between studies with 91% (149 drugs) of them being specific to one study. From the remaining 14 drugs present in more than one study, trichostatin, geldanamycin, tanespimycin and vorinostat were identified by three studies (Figure 4.6a) and, while only the first two are present in the DrugAge database, the remaining two have been experimentally validated for pro-longevity effects in animal models (Fuentealba et al., 2019; McDonald, Maizi, & Arking, 2013). Most studies resulted in a list of drugs containing mainly novel candidates not present in DrugAge (122 drugs were classified as novel discovery), the only exception being Aliper et al. (2016), which focused only on a set of known pro-longevity drugs. We also note that 66% of the 122 drugs (*i.e.* 81 drugs) known to target ageing-related proteins were prioritised by the computational studies reviewed above, as expected considering that these drugs are included in some of the databases used by some of the methods during the prioritisation process.

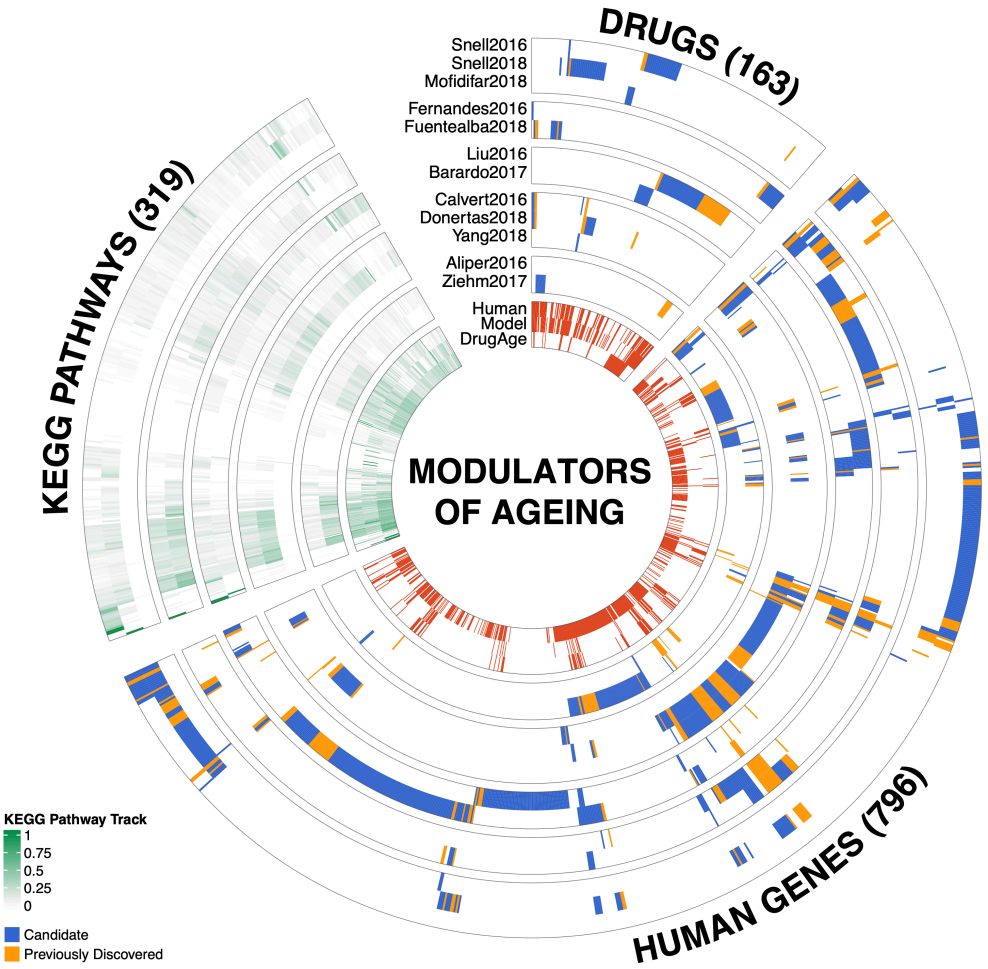


Figure 4.6 *Drugs, human genes and KEGG pathways discovered in the 12 studies. Circular heatmap of the drugs discovered by each of the 12 studies (drugs sector), genes targeted by these drugs (human genes sector), and the pathways including these genes (KEGG pathways sector). Drugs, genes and pathways are clustered independently to reflect discovery patterns from the studies. For the drugs and human genes sectors, the inner circle shows whether drugs or genes were previously associated with ageing, based on the DrugAge or GenAge database, respectively. If a drug was not present in DrugAge, it was classified as 'candidate', and the cell was coloured blue, whereas if the drug was already in DrugAge, it was classified as 'previously discovered' and the cell coloured in orange. An equivalent strategy using the GenAge databases instead of DrugAge was used for the human gene sector. In the inner wheel we present the overlap with drugs targeting ageing-related genes (drug sector – GenAge Human/Model tracks) and for the human gene sector the overlap with genes targeted by the drugs in DrugAge (Human genes sector – DrugAge track). The KEGG pathways sector shows the proportion of genes on each pathway targeted by the drugs discovered by each study. The cells representing KEGG pathways were coloured using a continuous gradient from white to green, where white means that none of the genes in that pathway were targeted by the drugs identified. In the section closer to the centre of the heatmap, we also showed the proportion of ageing-related genes in these pathways as well as the coverage of genes targeted by drugs in the DrugAge database. Data and code for generating this plot are provided in Github: https://github.com/mdonertas/ageing_drug_review.*

Genes: Overall, 34% of the GenAge human genes (103 genes) and 10% of the GenAge model organism genes (94 genes) were identified in at least one study, reflecting at least in part the different sizes of the datasets, with more than three times the number of Model Organism genes in GenAge. For clarification, the computational methods identified candidate drugs, which are predicted to modulate ageing, amongst the known drugs, most of which are currently used as therapy for a specific disease.

Based on the DGIdb database (Griffith et al., 2013), 27% of the druggable genome (*i.e.* 796 genes) is targeted by at least one of the drugs identified in the computational studies (Figure 4.6b) and, while few genes were identified in multiple studies, some of them were present in the GenAge database (Tacutu et al., 2018). Two of these genes *DDIT3* (DNA Damage Inducible Transcript 3) and *ERBB2* (Erb-B2 Receptor Tyrosine Kinase 2) were targeted by the drugs prioritised in eight studies. However, nine studies also identified drugs targeting *BIRC5* (Baculoviral IAP Repeat Containing 5) and *KRAS* (KRAS Proto-Oncogene, GTPase), and ten studies predicted drugs modulating *ABCB1* (ATP Binding Cassette Subfamily B Member 1), which have not previously been related to human ageing. Despite this, genes discovered by multiple studies do not necessarily suggest higher relevance to ageing, and may instead reflect research bias (*e.g.* genes targeted by many drugs because of a role in prevalent disease such as cancer). We also observed that 80% of known pro-longevity drugs (*i.e.* 122 of 152 drugs with known targets) target at least one gene targeted by the candidate geroprotective drugs identified by these twelve computational studies.

Pathways: Intriguingly, among all 319 KEGG pathways, 92% include at least one gene targeted by the drugs identified in the twelve studies. The same tendency was observed for genes in GenAge (83% Model GenAge & 74% Human GenAge), or genes targeted by the DrugAge drugs (88%). While this may suggest ageing is ubiquitous and affects all pathways, another possibility is that genes present in many pathways could be discovered repeatedly because they play a central role in diseases and regulatory mechanisms. Although this may not conclusively prove that ageing is ubiquitous, the prioritised candidate drugs clearly have a genome-wide effect.

4.4 Methods

Datasets

In order to define the gene expression changes during ageing, we only included datasets with samples across different ages. In this way, we calculated the changes that occur monotonically throughout the ageing process, rather than looking at differences in the young and old group. Datasets used in this study are all published datasets and include both microarray and RNA-seq data. The pre-processing steps for each are described below.

Microarray datasets. We used seven microarray-based RNA expression studies with samples from 22 brain regions, that are not mutually exclusive (Supplementary Table S1). Data from different brain regions are processed and analysed separately, resulting in 26 datasets. The number of individuals in each dataset ranges between 11 and 148. The total number of individuals is 304, and the total number of samples is 805 (after removing the outliers). Some studies include samples covering the whole lifespan. However, in this study, we only considered samples above 20 years of age, which corresponds to the age at first reproduction in human societies (Walker et al., 2006). Previous human brain ageing studies using transcriptome data have also suggested gene expression patterns before and after the age of 20 are discontinuous (Colantuoni et al., 2011; Dönertaş et al., 2017). Since we are interested in finding consistent tendencies in terms of the direction of change, which can characterise ageing, we only included samples above 20 years of age. As a result, the samples included in the analysis had ages between 20-106. The microarray data were downloaded from NCBI GEO (Barrett et al., 2013) using the accession numbers in Supplementary Table S1. Using “affy” (Gautier et al., 2004) or “oligo” (Carvalho & Irizarry, 2010) libraries in R, RMA background correction is applied to the expression data. The data is then log2 transformed, and quantile normalised (using “preprocessCore” library in R). By visual inspection of the first and second principal components of the probe-set expression levels, outliers were excluded from the further analysis (Supplementary Table S1). The age distributions for the datasets after outlier removal are given in Figure A.54a. Gene annotations for the probe-sets are obtained from the Ensembl database using the ‘biomaRt’ library (Durinck et al., 2009) in R. Because the annotations for the probe-sets used in Kang2011 and Colantuoni2011 are not available in Ensembl, we used the GPL files deposited in GEO. If Ensembl gene IDs are not provided in the GPL files, Entrez gene IDs were extracted and converted to Ensembl Gene IDs using the

'biomaRt' package. Probe-set level expression information is then mapped to gene IDs. In order not to duplicate expression values, we excluded the probe-sets corresponding to multiple genes. Expression values for the genes with multiple probe-sets were summarised using the mean expression levels. The PCA plots for the samples using gene expression levels are given in Figure A.54b.

RNA-seq dataset: We analysed transcriptome data generated by GTEx project (v6p) (GTEx Consortium, 2015). Samples are filtered based on the cause of death circumstances (4-point Hardy Scale). Only the cases with a death circumstance of 1 (violent and fast deaths due to an accident) and 2 (fast death of natural causes) are used for the downstream analysis and the samples with illnesses are excluded. Among all tissues, only the ones having at least 20 samples are considered. We also excluded 'Cells - Transformed Fibroblasts' category to include only the samples from tissues. As a result, 35 datasets (17 major tissue type) are used for the downstream analysis, 13 of which were from the brain. The final set that we analysed includes 2152 (623 for the brain) samples from 120 (99 for the brain) individuals. The genes with median RPKM value of 0 are also excluded from data. The RPKM values provided in the GTEx database are log2 transformed and quantile normalised for the downstream analysis. Similar to the microarray data, we excluded the outliers based on the visual inspection of the first and second principal components (Supplementary Table S1). Distribution of the ages and the PCA plots after outlier exclusion are given in Figure A.54.

Batch correction. In this study, each dataset is analysed separately, and only the gene expression changes that are consistent across all datasets are considered for the downstream analysis. Since multiple datasets are not combined, and datasets generated at different labs using different platforms unlikely to have the same confounders, we did not apply a correction method other than quantile normalisation and outlier removal based on the PCA (using probe-set level expression data for microarrays and gene-level expression data for RNA-seq as described above). Moreover, most of the datasets have a homogenous sample set as the number of samples is low and for the datasets with a large number of samples, we do not detect any clustering.

Age-related expression changes and the ageing signature

The Spearman's rank correlation coefficients between age and gene expression levels are used to measure age-related expression changes. Instead of combining the

datasets, we calculated the Spearman's correlation for each gene, for each dataset separately. As a result, each gene had two measures to assess its age-related expression: 1) a correlation coefficient (rho), indicating the strength and the direction of change with age and 2) a p-value, showing the significance of the association. The p-values are corrected for multiple testing using p.adjust function in R, with method="FDR" argument. As the power to detect significant changes in each dataset is different and the sample size is small for most of the datasets, for the downstream analysis we only used the correlation coefficients (rho) and assessed the significant gene expression change tendencies that are consistent in all datasets. When a gene is up-regulated by age throughout the lifespan, then it would have a positive Spearman's correlation coefficient that is close to one. In contrast, a gene would have negative correlation coefficient if it is down-regulated. When the association is not strong, the magnitude of the correlation coefficient decreases, but the sign still reflects the direction of change that is observed in most of the time-points. We used the sign of correlation coefficient, *i.e.* the direction of change, to compile the set of genes that show consistent changes across all datasets. This set of genes are referred to as the 'ageing signature'. The ageing signature, thus, does not reflect the dramatic changes in gene expression but captures consistent trends that are observed across all datasets. The statistical significance of the ageing signature is calculated using a permutation scheme, testing the significance of the consistency.

Permutation test

We used a permutation scheme that we developed earlier (Dönertaş et al., 2017), to simulate the null hypothesis that there is no association between age and the gene expression, while retaining the dependence between genes and the datasets. Particularly, the ages of individuals in each study are permuted (randomised) 1,000 times and if that individual donated multiple samples for different brain regions, each sample is annotated with the same age. Then, the Spearman's correlation coefficient between these randomised ages and the gene expression value for all genes are calculated. In this way, we retain the dependence between genes (*e.g.* those regulated by the same transcription factor) and the samples (*e.g.* donated by the same individuals). Permutations are performed using 'sample' function in base R.

Using the correlation coefficients calculated through permutations performed as explained above, we tested i) significance of the correlations among datasets, ii) significance of the finding the same or a higher number of consistently up- or down-regulated

genes, *i.e.* the ageing signature. In order to test the significance of the correlations among datasets, we calculated the correlations between the expression-age correlation coefficients calculated using the permutations. We constructed the distribution for the median correlation coefficient among datasets (distribution of the 1,000 values), and calculated how many times the randomised values have higher correlation than the value we calculate using the real ages. In this way, we calculate an empirical p-value. The median of the permuted values reflects the value that would be expected by chance. Similarly, in order to test the significance of the ageing signature, we compiled permuted ageing signatures, for 1,000 times, and asked how many times we have the same or higher value than the calculated number of genes in the microarray or GTEx ageing signatures. In this way, we calculate the empirical p-value and median of the number of shared tendencies based on permutations, reflecting what would be expected by chance.

Gene Ontology Enrichment

Using "topGO" (Alexa & Rahnenfuhrer, 2019) and "org.Hs.eg.db" (Carlson, 2019) libraries in R, we performed a functional analysis of the ageing signature. Using GO categories with more than 10 annotated genes, we applied an enrichment test for the Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium & The Gene Ontology Consortium, 2019) Biological Process (BP) categories.

Connectivity Map Analysis

A list of genes showing a consistent change in ageing (the ageing signature) is used to query the Connectivity Map (Lamb et al., 2006). Since the Connectivity Map input requires probe-set ids, the "biomaRt" package in R is used to convert the gene list to the probe-set ids that are compatible with the CMap data. The probe-sets that are in both up- and down-regulated probe-set lists are excluded from both lists. The final lists are used to query CMap database to associate the ageing signature with the drug perturbed expression profiles in the database. The resulting p-values are FDR corrected to account for multiple testing and adjusted $p<0.05$ is used as the significance threshold.

The ageing signature compiled using the GTEx data had more than 500 probe-sets in both up and down lists. Since the algorithm requires an input with less than 500

entries, we used the ones with the higher magnitude of expression change (median Spearman's rank correlation coefficients across 13 brain regions). In order to show that this does not bias the results, we repeated this step for 1,000 times by randomly selecting 500 of the probe-sets in the GTEx ageing signature. In order to automatize this process, we re-implemented CMap algorithm in R and calculated the drug similarity scores using the 'rankMatrix.txt' data provided on the CMap website. Drug similarity scores generated using the top 500 and randomly selected 500 of the GTEx ageing signature showed a significant correlation (median rho = 0.81, range = (0.80,0.82)), suggesting that this approach does not bias the results.

Searching the drug databases for CMap drugs

Entries in the Connectivity Map are composed of the drug names, which are generally the catalogue names for the drugs from chemical vendors. Similarly, DrugAge drugs also do not have an ID that is possible to map across different databases. The DrugAge database was retrieved on 11th May 2017, from the DrugAge website. In order to compare the drugs in the Connectivity Map and the DrugAge, we first used the PubChem database (Kim et al., 2016) to make a transition across different sources. We obtained PubChem compound IDs for each drug in the Connectivity Map and DrugAge using PubChem API accessed through R programming environment and 'RCurl' and 'jsonlite' libraries.

Targets of the drugs that are significantly associated with ageing

We compiled the drug-target associations for the drugs significantly associated with ageing mostly through literature research. For the cases where the database entries are manually curated and consistent, we used CHEMBL (Bento et al., 2014), DrugBank (Law et al., 2014), and PubChem (Kim et al., 2016). We downloaded GenAge model organism and human datasets (Tacutu et al., 2018) on 10th October 2017 using GenAge website. Using the human orthologues for the model organisms (genage_models_orthologs_export.tsv) and the human dataset, we asked if any of the drug targets were previously shown to be implicated in ageing. In order to construct the drug – target network, we used 'ggnetwork' package in R.

The Pro-Longevity Drug Expression Profile

In order to compile a set of gene expression changes that can be associated with the known pro-longevity drug profile, we first downloaded the pre-processed data matrix with the drug-induced expression changes ('amplitudeMatrix.txt' from CMap FTP server ftp://ftp.broadinstitute.org/distribution/cmap). Using this matrix, for the seven pro-longevity drugs in DrugAge that are among the significant associations according to our analysis, we generated a pro-longevity drug profile. We first identified the drug-induced gene expression changes for each of these seven drugs and each of the probe-sets that are in the microarray ageing signature. For each drug – probe-set pair, we take the direction of change that is observed in at least 60% of the experiments (using different doses or different cell lines) as the effect of that drug on the expression of that probe-set. After deciding on the individual drug effects, we took the type of change observed in at least four of seven drugs as the pro-longevity drug profile. The reason why we do not seek a perfect overlap among different drugs is to allow potentially different mechanism of actions to be included in the pro-longevity drug profile. As a result, we got five categories: 1) increase in ageing, increased by the drugs; 2) increase in ageing, decreased by the drugs; 3) decrease in ageing, increased by the drugs; 4) decrease in ageing, decreased by the drugs; and 5) the ones that are not affected consistently by the drugs. The full list of genes in the first four categories is given as Supplementary Table S5. We also asked if any of the GO Biological Processes is enriched in any of the first four categories and thus did an enrichment analysis. We calculated the odds ratio for each GO category by keeping the type of change in ageing the same. For example, we asked if a GO category is enriched in genes that increase in ageing and also increased by the drugs, compared to the genes that increase in ageing but decreased by the drugs. Because the number of genes is small, it is not possible to detect significant associations after correcting for multiple testing and thus we only report the odd's ratios for the categories (Supplementary Table S6). We also compared the known pro-longevity drug profile we compiled with the profile induced by the 24 drugs identified in the study (Figure A.62). We calculated the percentage of probe-sets that show the same type of change as the pro-longevity drug profile. For this, we again only considered probe-sets that show the same type of change in at least 60% of the experiments per drug.

Gene-set enrichment analysis for drug-induced changes

Using the 'amplitudeMatrix.txt' downloaded from the CMap website, we determined the expression changes at the gene level for each drug. We first subset the matrix to include only the experiments for the 24 significant drugs we found. We then mapped the probe-set ids (total number of probe-sets = 22,283) to Entrez gene ids using the Ensembl biomaRt package in R. We map 19,222 probe-sets to genes, excluding examples where the same probe-set id maps to multiple genes (628 multi-gene probe-set ids in total). The genes with more than one probe-set id are represented by taking the median expression change induced for the probe-sets (number of genes = 12064). When the experiments for each drug are treated separately, we noticed that the results were confounded by cell-line. Thus, we then summarised multiple experiments for each drug by taking the median of the change they induce. In this way, we trimmed the cell-line specific effects. Then the expression changes (for 12064 genes) for each drug (24 drugs) are rank ordered. Using clusterProfiler package and 'gseKEGG' and 'gseGO' functions, we performed GSEA for the gene expression changes induced by each drug separately. For the KEGG pathway analysis, we only considered the pathways with at least 50 genes (188 pathways), and for GO analysis, we only considered Biological Process categories with at least 50 and maximum of 200 genes (1589 categories).

Comparing Brain Ageing Signature to Other Tissues

We calculated the proportion of genes that show a change in the same direction with the ageing signature compiled using brain data. The proportions are calculated for ageing signatures compiled using the array and GTEx brain data, separately. We also analysed up-regulated and down-regulated genes separately to observe any differential pattern. In order to calculate the significance of similarity or dissimilarity we performed 10,000 permutations as follows: i) N number of genes, where N is the number of genes in a particular group (array / GTEx and up- / down-regulated), were selected randomly from a given GTEx dataset, ii) the proportion of changes in a given direction is calculated, and iii) using the distribution of these proportions, we asked how many times we obtain a value as extreme as the proportion calculated for that tissue and assign empirical p value.

Side Effects

Using compound PubChem IDs, we subset the Side Effect Resource (SIDER 4.1) (Kuhn, Letunic, Jensen, & Bork, 2016), a database of adverse drugs reactions for marketed medicines. The latest version of SIDER code the side effects by using the Medical Dictionary for Regulatory Activities (MedDRA), an adverse event classification dictionary. To obtain term at the system level, we mapped the lowest-level MedDRA terms in SIDER (LLT codes) to MedDRA System Organ Class terms (SOC codes) using hierarchical files downloadable from the MedDRA web-based browser (https://tools.meddra.org/wbb/). A total of 8 drugs among the 24 had labelled side effects.

4.5 Discussion

In this study, using gene expression data, we identified a set of drugs that are likely to modulate ageing in the human brain. Using a meta-analysis approach, we generated a reproducible ageing signature that represents multiple brain regions and is independent of the platform used for the detection of expression. Using the Connectivity Map, we identified drugs highly associated with this ageing signature.

Seven of the 24 drugs were previously shown to modulate lifespan in model organisms. Based on the DrugAge database, seven of these drugs were previously tested on model organisms and prolonged lifespan in at least one experiment. The fact that we successfully re-discovered a statistically significant number of known lifespan modulators, without using any prior drug ageing information, suggests that the other drugs that we identified also have a high potential to be modulators of the ageing process / lifespan.

Eleven of the drugs have targets previously associated with ageing. Vorinostat is a histone deacetylase (HDAC) inhibitor used for the treatment of cutaneous T-cell lymphoma ("Vorinostat," n.d.). Although not reported in the DrugAge database, vorinostat has already been tested on Drosophila for lifespan extension and shown to increase both mortality rate and survival when the drug is given during 'mid- to late-life' (McDonald et al., 2013). This drug had the most pro-longevity drug-like profile (Figure A.62, cluster 6) based on our analysis, suggesting that the methodology, as well as the interpretation, yields biologically relevant results. Quinostatin, targeting the catalytic subunit of PI3K, had the highest CMap score with percent similarity above 50% to the pro-longevity drug profile for all four categories (Figure A.62). Considering that

drugs targeting PI3K, such as LY-294002 and wortmannin, extend lifespan in worms and flies (Barardo et al., 2017), quinostatin is a strong anti-ageing drug candidate. Alvespimycin and tanespimycin inhibit the heat shock protein HSP90, which is also inhibited by geldanamycin. Heat shock proteins are implicated in ageing based on both experiments on worms and flies (Tacutu et al., 2018) and human expression studies. Protein aggregation and disrupted proteostasis are a hallmark of ageing (López-Otín et al., 2013). It is thus plausible that increased activity of HSP90 would reverse the effects of ageing by restoring proteostasis, although its downstream effects might result in reverse (Fuhrmann-Stroissnigg et al., 2017; McClellan et al., 2007). Tretinoin is a retinoic acid receptor (RAR) agonist widely studied for skin (Mukherjee et al., 2006) and brain ageing (Enderlin et al., 1997). RAR genes in are implicated in synaptic plasticity, learning, memory, and pathological conditions such as Alzheimer's disease (Lane & Bailey, 2005). GW-8510 is a cyclin-dependent kinase 2 / 5 inhibitor and was suggested to be neuroprotective (Johnson et al., 2005). 15-d prostaglandin J2 activates PPARG, which shows decreased expression with age, which can be restored by DR (Tacutu et al., 2018). Camptothecin and irinotecan both target TOP1, which alters the topological state of DNA during transcription and can inhibit Warner syndrome protein (WRN), which functions in DNA repair (Shamanna et al., 2016). These two drugs, as well as daunorubicin which targets TOP2A, may therefore worsen health status. Cinchonine, which targets CYP2D6 was similar to pro-longevity drug profile in terms of the genes up-regulated by the pro-longevity drugs but show opposite profile for the down-regulated genes. The functions that are associated with the genes down-regulated by the drugs are autophagy or immune function related categories. Most of the known pro-longevity drugs are suggested to function through inhibition of PI3K / mTOR pathways, favouring autophagy. It appears that cinchonine would not function in the same way. However, considering that it targets CYP2D6, which was shown to have a role in lifespan regulation in *C. elegans*, it is possible that this drug has a distinct mechanism to modulate ageing. Mann et al. previously suggested that expression of CYP2D6 increases with age in the human brain and is lower in Parkinson's disease (Mann et al., 2012). Considering that they suggest this protein might be important to inactivate neurotoxins, inhibiting this protein using cinchonine might function in the same direction and exacerbate ageing by down-regulating one of the cellular responses.

New targets and mechanisms to modulate ageing. Thioridazine and trifluoperazine are serotonin and dopamine receptor antagonists used for the management of psychoses, including schizophrenia. Thioridazine is withdrawn from the market due

to its side effects related to cardiac arrhythmias. Ye et al. (2014) screened a library of compounds for lifespan extension in *C. elegans* and identified a couple of drugs targeting serotonin and dopamine receptor antagonists including thioridazine hydrochloride, which extends lifespan by 31% in *C. elegans*. Thus, it is likely that these drugs also have anti-ageing effects. Emetine is the principal alkaloid of the ipecac root. It is a eukaryotic protein translation inhibitor. A recent study investigated the effect of protein translation inhibition on cellular senescence. They suggest that cytoplasmic protein accumulation is an important cause of the cellular senescence and mild protein translation inhibition can prevent senescence induction in normal and tumour-derived human cells (Takauji et al., 2016). Although both this information and our results suggest that emetine can help alleviating the ageing, Takauji et al. (2016) did test the effect of emetine on senescence and could not detect any significant result. Atropine oxide is predicted to target muscarinic acetylcholine receptors ("Compound," n.d.-b), which are suggested to be important for various brain functions as well as pathologies such as Alzheimer's and Parkinson's diseases (Langmead, Watson, & Reavill, 2008). The information regarding the effect of atropine oxide on different muscarinic acetylcholine receptors, however, is limited to make a conclusion whether this drug could be beneficial or damaging for the human brain ageing. Securinine is a GABA(A) receptor antagonist. GABA receptors are started to gain attention as potential targets for neurodegenerative diseases (Rissman, De Blas, & Armstrong, 2007). The drugs tested and shown to have an impact on cognitive abilities so far, however, are mainly GABA(A) agonists or GABA(B) antagonists (Li et al., 2016). Considering that GABA(A) subunit expression levels show a decrease with age (Supplementary Table S2), and securinine is an antagonist, it is possible that it acts in the same direction as ageing and exacerbates it. Rifabutin shows high similarity to the pro-longevity drug profile and clusters together with levothyroxine sodium and geldanamycin, which are known pro-longevity drugs (Figure A.62, cluster 2). Rifabutin is an antibiotic but it is reported to also target BCL6 (Evans et al., 2014). BCL6 gene is not in GenAge databases, however, there are studies linking this gene to ageing using human gene expression data (Glass et al., 2013) and through its role in cell proliferation and senescence, regulated by miR-127 (Chen, Wang, Guo, Xie, & Cong, 2013). Thus, it is possible that rifabutin helps to reduce the effect of damaging changes induced by ageing, through targeting BCL6.

Challenges. 'Omics-based drug repurposing studies, such as the CMap, aim to identify drugs reversing the profile induced by a biological state of interest. Ageing is a time-dependent, complex phenomenon, which induces subtler changes compared

to development (Dönertaş et al., 2017), or to a disease state such as Alzheimer's (Avramopoulos, Szymanski, Wang, & Bassett, 2011). The 'omics profile reflects two potentially distinct contributions: - the detrimental effects which occur with age (*e.g.* accumulation of mutations) and the potentially beneficial responses to those changes (*e.g.* the immune response). As a result, CMap similarity score is not conclusive on its own. In order to characterise the potential effects of drugs on ageing (anti- or pro-ageing drugs), we use three different approaches: i) comparison of the drug-induced expression profiles with the known pro-longevity drug profile (Figure A.62), ii) functional analysis of the drug-induced gene expression changes (Figure A.64), and iii) compilation of literature on the drugs and targets. Based on these analyses we suggest that eight of seventeen drugs (quinostatin, trifluoperazine, thioridazine, vorinostat, alvespimycin, tanespimycin, rifabutin, and 15-d prostaglandin J2), which are not in DrugAge, are likely to have positive effects, whereas, topoisomerase inhibitors (camptothecin, irinotecan, and daunorubicin) can be detrimental and could act as pro-ageing drugs. Four of the remaining drugs, which are cinchonine, securinine, emetine and tretinoin, do not cluster closely with any known pro-longevity drugs in Figure A.64. Literature, however, suggests cinchonine and securinine are likely to have negative effects, whereas emetine and tretinoin could act as anti-ageing drugs. GW-8510 and atropine oxide could not be classified because neither the clustering results nor literature evidence are conclusive.

4.5.1 Limitations

In this study, we applied a data-driven methodology to overcome research bias and increase the chance of discovering novel drugs to target ageing. However, there are several aspects that could limit our study. It is important to note that none of the cell lines used to generate the CMap data originates from the brain. The assumption for using the CMap algorithm is that the effect we see in diverse cell-lines reflects the global profile of the drug perturbation and thus should be also transferable to the brain. However, it is possible that drugs have cell or tissue-specific effects. Even if the drugs induce the same expression changes in brain cells, an important question is: Can they cross the blood-brain barrier to target the brain? If some of these drugs have side effects on the CNS, it might be an indication that these drugs can affect the brain and can be re-purposed to target brain ageing. Only eight of the 24 compounds have reported side effects and all of them has at least one reported effect on the nervous system, based on MedDRA system organ classes (Supplementary Table S10). This

implies that these drugs can affect CNS, although we do not have information on their ability to cross the barrier. The rest may or may not cross the barrier to influence the expression in the brain, but they may also improve health by targeting generic changes throughout the body. The ageing signatures from brain tissue show a modest but significant similarity to expression profiles from non-brain tissues (Figure A.64). Thus, it is possible that we identified not only drugs specifically targeting ageing in the brain but also drugs targeting ageing in other tissues. It is also possible that there are drugs which can target brain ageing with more potency, but we cannot identify them because we do not have drug-induced expression profiles for brain cells. Another important technical drawback is that the data we used to generate the ageing signature are bulk RNA expression datasets, where the expression profile is an average of all the cell types in the human brain. Focusing on the changes that are observed ubiquitously across all brain regions, we aimed to focus on global changes which are unlikely to be driven by cell type differences. However, future datasets generated using single-cell expression profiling can greatly improve the understanding of both the ageing process itself and how the interventions work.

4.5.2 Conclusion

To summarise, this study provides an unbiased identification of drugs that can target human brain ageing. We first compiled a set of gene expression changes that can characterise human brain ageing and asked if there are drugs which alter the expression of the same genes. We identified 24 drugs, seven of which were among known pro-longevity drugs. Our analysis suggests that anti-ageing drugs may act by mimicking the response while it is also possible that they can reverse the detrimental changes in ageing. Based on the literature research, we concluded that some of the drugs we identified can directly modulate the lifespan, whereas some are more likely to function by improving the cognitive functions and promoting the healthy ageing. We further combined the results with other drug repurposing studies to summarise the state of the field. We found 27% of all druggable genome and 92% of all KEGG pathways are targeted by the drugs suggested by these studies, showing the complex and multifaceted nature of ageing.

Chapter 5

Concluding remarks

In this thesis, the aim was to approach a better understanding of ageing through computational studies. In particular, age-related change in heterogeneity, the link between ageing and age-related diseases, and drug repurposing for ageing were discussed. In this chapter, I aim to both summarise the major findings and also give a future perspective focusing on each chapter separately. Finally, I will summarise future directions by which computational studies can facilitate research on ageing.

5.1 Overview and future perspective

5.1.1 Age-related changes in gene expression heterogeneity

In Chapter 2, the changes in gene expression heterogeneity during postnatal development and ageing were explored. The main conclusions of this study were:

- Age-related increase in heterogeneity is widespread at the gene-, functional category-, pathway-, and transcriptome-levels.
- Previous studies focusing on gene expression variability in ageing reported different genes as differentially variable in ageing. This may stem from the differences in the tissues, organisms, and ages that were analysed or could be due to the stochastic nature of increased variability. Using different methodologies on the same dataset, we showed that the preprocessing steps and heterogeneity measures used in different studies also make a significant difference in detecting the ‘differentially variable’ set of genes. Despite this difference, a widespread

increase in heterogeneity was observed with all combinations of preprocessing steps and measures of heterogeneity.
- We also showed that the age-related increase in heterogeneity is a characteristic of post-adult years (20+) but not pre-adulthood (0-20 years of age).
- Although the increase in heterogeneity is not specific to certain functional categories, or pathways; we found that the genes that become more heterogeneous across all datasets were enriched in many of the longevity-regulating pathways. However, they did not overlap with the exact genes that increase/decrease lifespan in laboratory organisms.
- Finally, we showed that not only specific transcription regulators but also the number of regulators is associated with the increase in heterogeneity during ageing and this is uncoupled from the change in expression during development.

However, some questions remain unanswered and require further experimental and computational studies.

- In this study, we analysed 19 datasets covering 17 brain regions. The brain is a post-mitotic tissue and might include different sources of cellular damage compared to mitotic tissues (e.g. liver) or tissues that are more open to environmental influence (e.g. skin). Thus, future studies comparing different types of tissues can help understand the source of heterogeneity in gene expression.
- We showed an association between the number of regulators and increased heterogeneity. However, both the molecular nature of this association and also whether there are other contributing factors need further exploration. A new study design allowing reliable detection of somatic mutations, and protein levels coupled with the gene expression levels is important to understand the cause of increased heterogeneity during ageing.
- We showed that the reproducible increase in heterogeneity is limited to post-adult ages; however, the reason why there is a difference with postnatal development could not be explored. Likely explanations include active protection from damage or higher tolerance to damage. Moreover, damaging effects may need to accumulate to a certain level to create phenotypic consequences. Understanding the reason would require both a better comprehension of the causes of heterogeneity and a balanced study design with more samples from both development and ageing.
- In this study, we focused on inter-individual heterogeneity with the analysis of bulk RNA expression levels. However, there are already multiple reports sug-

gesting intra-individual heterogeneity also increases with age (Davie et al., 2018; Enge et al., 2017; Hernando-Herraez et al., 2019; Martinez-Jimenez et al., 2017). Moreover, different genes also show a difference in their inter-individual variability at an adult age. A comparative analysis of these two phenomena together with our findings could help understand the mechanisms of heterogeneity.

5.1.2 The link between ageing and age-related diseases

In Chapter 3, we aimed to understand the link between ageing and age-related diseases through a genome-wide association study of 116 diseases in the UK Biobank. The main conclusions were as follows:

- Diseases with similar age-of-onset profiles show a higher genetic similarity which is not explained by disease categories, co-occurrences, and disease cause-effect relationships.
- Common genetic factors shared between age-dependent diseases significantly overlap with regulators of cellular senescence and the targets of drugs known to modulate lifespan in model organisms. However, the intersection is small.
- We identified biological functions associated with different age-of-onset profiles. Although the two age-dependent clusters both had associations with known ageing-related genes, their functional categories were different. Diseases seen after 40 were associated with lipid metabolism, cell cycle, and cellular response, whereas the diseases seen after 20 were associated with MHC 2 protein binding, fibrinolysis, and regulation of apoptosis. Genes associated with both clusters were enriched among regulators of glucose metabolism, chromatin structure, and transcription. However, in general, genes that are highly pleiotropic and associated with all disease clusters were also regulators of chromatin structure and transcription.
- We checked the risk allele frequencies (AF) of the variants associated with each cluster and found diseases with later age-of-onset had significantly higher AF, which supports both mutation accumulation and antagonistic pleiotropy theories of ageing.
- We searched for drugs that specifically target age-dependent clusters and found several significant hits. 23 of these drugs were already in use for 14 different conditions such as diabetes, hyperlipidaemia, neoplasms, osteoporosis, cardiovascular diseases. This means we can more easily repurpose these drugs to

use for the diseases we analysed, which generally co-occur in the elderly. This has the potential to improve health-span and may also alleviate the effects of polypharmacy.

This study, to my knowledge, provides the most comprehensive analysis of the link between ageing and age-related diseases. However, there are multiple aspects that require further investigation:

- In this study, we suggest that there is a common ageing-related genetic factor associated with age-dependent diseases. However, the mechanism by which this factor may influence the age-of-onset profile of diseases could not be explored. Likely explanations include somatic mutations, age-related changes in gene and protein expression levels and epigenome, and gene-environment interactions. We need a significantly improved longitudinal study design including different 'omics and lifestyle data in order to address this question.
- There is a significant association between the targets of lifespan-modulating drugs and the regulators of cellular senescence. However, we did not detect significant overlap with genes that modulate lifespan in model organisms. One plausible explanation is that these genes are deeply evolutionarily conserved and we do not have common variations that we can detect in this study. However, a detailed analysis of this lack of overlap could improve future study design and also may facilitate knowledge transfer from model organism to human studies.
- In this study, we could only include common diseases because of the sample size limitations. However, a better comparison to analyse evolutionary theories of ageing would include variants that are known to be causal for developmental diseases.
- We suggest some drugs that may improve health-span. The next step would be to test these drugs in model organisms, particularly in mice and rat, and disease-models to measure their effect on health- and lifespan.

5.1.3 Drug repurposing for ageing

In Chapter 4, we employed an applied study and aimed to identify pharmacological interventions that can improve lifespan in humans. We followed a system-level approach to find drugs that may mimic or reverse the age-related changes in gene expression.

- Through a meta-analyses approach, we first compiled a robust and reproducible list of gene expression changes that characterise ageing in the human brain.
- We compared this signature with the drug-perturbed gene expression profiles in the Connectivity Map database to find drugs that can mimic or reverse these changes.
- We identified 24 candidates, 7 of which were pro-longevity drugs that were previously tested in model organisms. 18 of the 24 had at least one target that was associated with ageing according to GenAge database. We identified 4 completely novel drugs which may provide another angle in ageing interventions.
- Importantly, we found that some of the pro-longevity drugs act by mimicking the changes in ageing. That means that some of the captured gene expression changes might be beneficial and include responses to the damaging effects.
- We also combined our results with 11 other published studies that perform in-silico drug repurposing to target ageing. Almost 30% of all druggable genome was found to be targeted by at least one drug suggested in these 12 studies. Moreover, 92% of all KEGG pathways had at least one gene targeted by the suggested drugs. These analyses reveal further that ageing is a complex, system-level phenotype.

Rediscovery of a significant number of known pro-longevity drugs suggests our methodology is promising. We also experimentally validated one of the candidate drugs, tanespimycin, in *C.elegans*, as part of another study that I contributed (Fuentealba et al., 2019). Furthermore, since the input data is from humans it is plausible to suggest that these drugs may provide benefits in humans as well. However, there are several study-specific aspects that may further be explored:

- In this study, mostly because of the limitations in the number of available datasets, we restricted our analysis only to the human brain. However, the whole organism is affected during the ageing process. We show that the ageing signature can also represent changes seen in other tissues. Nevertheless, there might be other drugs that can function in other tissues better, and thus, this study could be expanded to other tissues.
- We found that some of the known pro-longevity drugs we identified mimic the changes observed during human brain ageing. This raises a question about their applicability to humans. Humans live longer than their closely related species and may have adapted some response mechanisms that are not observed in the model organisms, which are short-lived. A comparison of the ageing signature

with the age-related changes observed in the model organisms can provide a better understanding.

- Another aspect that we have not addressed in this study is the real-world application of the identified drugs, including factors such as the drug-drug interactions, the age at which drugs could be supplied, the potential effects of long-term use, and bioavailability in humans. Unfortunately, computational studies require more comprehensive experimental datasets that can allow predictive modelling of these factors.
- Lastly, predictive modelling of the overall effect of these drugs on health, healthspan and life-span can help prioritisation of the candidates. Today, it is not possible to perform this task for humans as the human clinical trials for ageing interventions are just beginning. However, thanks to DrugAge database, which records the drug, organism, and outcome of the lifespan assays, future studies employing machine learning approaches can prioritise drugs that may have the highest effect on lifespan.

5.2 Future directions in computational biology of ageing

Our understanding of the biology of ageing is improving every day. Theoretical, experimental, clinical, and applied research on ageing has provided significant insight into the mechanisms and interventions to achieve better health in old age. However, so far, these realms of research have been disconnected. The studies presented in this thesis exemplify how computational biology could help to build bridges between different disciplines. However, the potential of computational biology is not limited to the examples I presented. In this section, I summarise the improvements in computational research that can help us improve our understanding of the biology of ageing and provide the toolset to tackle the challenges of age-related decline in health.

5.2.1 Single-cell genomics of ageing

Ageing has been studied at the population, organismal, tissue, and cell-type levels. With the advancement of single-cell genomic technology, it is now possible to study ageing even at a single-cell resolution. Studying the molecular changes in single cells

within different environments and interactors will provide a more detailed understanding of the mechanisms. This advancement is especially important with regards to the damage-based theories of ageing. Understanding the changes at a single-cell level can suggest better ways of targeting age-related deterioration and help us model age-related changes in specific pathways.

Our understanding of heterogeneity in ageing can also benefit from single-cell technologies. At the moment, studies approach heterogeneity as a stochastic noise; however, Chapter 2 demonstrates that there is some consistency in increased heterogeneity across datasets. This suggests, there might be a mechanism contributing to the age-related increase in heterogeneity. Studying genetic and epigenetic mutations at the single-cell level can help us understand the source of heterogeneity.

5.2.2 Integrative biology of ageing

All the studies I have presented in this thesis further emphasised the importance of having data from multiple layers of information flow (e.g. genome, epigenome, transcriptome, proteome, metabolome). Ageing is a complex phenotype that is far from being driven by just a limited number of genes. By understanding the regulatory changes at multiple levels, we can approach a better intervention strategy. Moreover, it is plausible to argue that the information flow is disrupted at multiple stages during ageing. So far, studies showed an age-related decline in epigenetic stability, which interrupts information flow at a very top level. There are multiple studies focusing on restoring the youthful epigenetic state using drugs targeting epigenetic machinery. However, at the moment we do not know whether the flow of information is disrupted at multiple levels or not. By studying the interactions between different layers of information, we can get a more detailed insight for the source of increased entropy during ageing.

Integration of phenotype level information such as diseases, frailty, and age-related cognitive decline can further provide a better understanding of the effect of molecular changes on the decline in functional integrity during ageing. Furthermore, combining different types of data, even from different organisms, and cultivating the mechanistic insight provided by the model organisms, can help us approach a molecular and cellular model of ageing.

5.2.3 Comparative biology of ageing

So far, most of our functional knowledge about ageing originates in model organism studies. However, these model organisms are short-lived and are under captivity for a long time. Whether the mechanisms and interventions found using these organisms are applicable to humans is an important question that awaits exploration. Comparative studies using both genomics and structural data can help the better transfer of knowledge from model organisms to humans.

Appendix A

Supplementary figures

A.1 Age-related changes in gene expression heterogeneity in the human brain

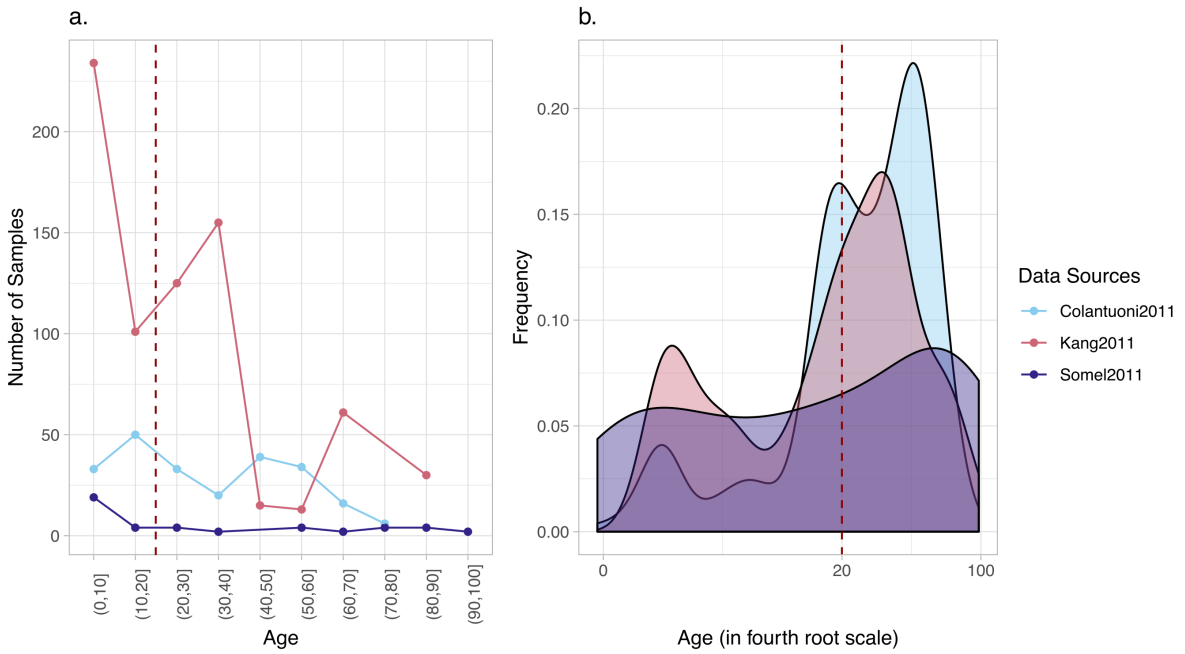


Figure A.1 *The distribution of ages for the samples used in the analysis. (a) Scatter plot showing number of samples (x-axis) included in age intervals (y-axis). (b) Distribution of samples in fourth root scale of ages. The horizontal dotted lines reflect the separation point of development and ageing (age of 20). Data sources are indicated with different colours.*

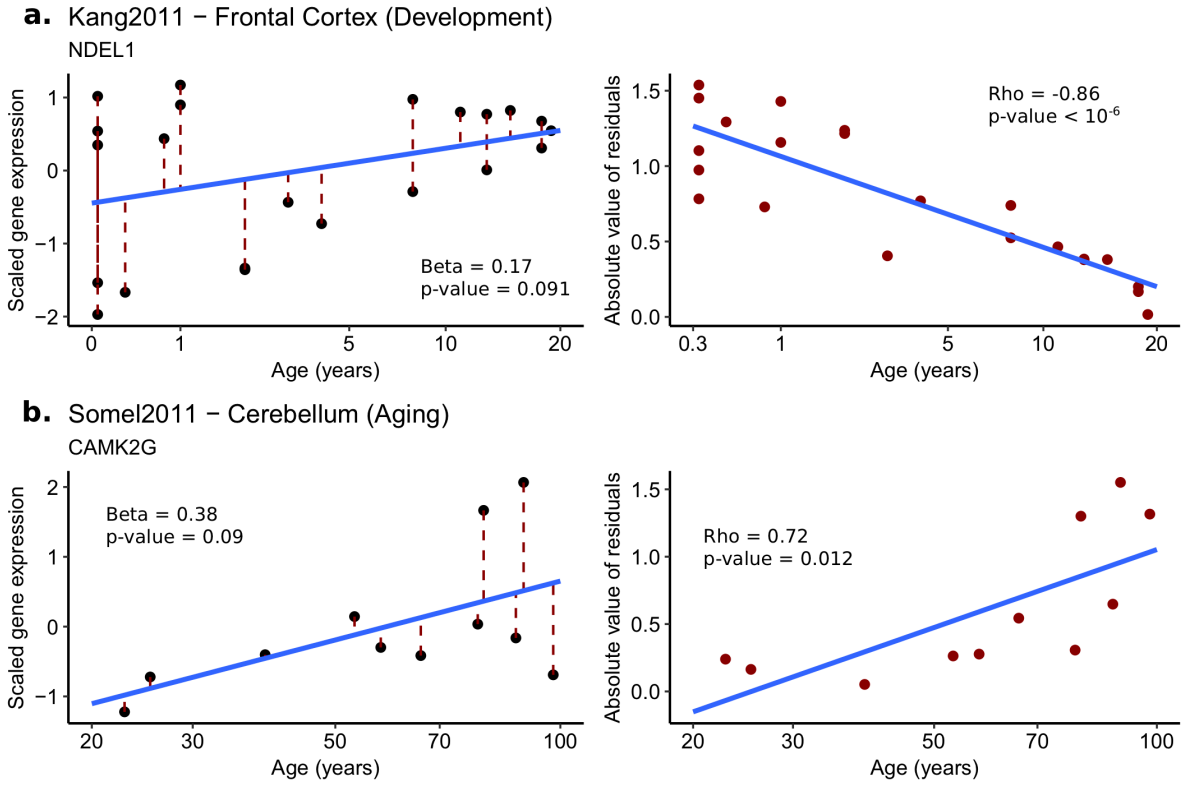


Figure A.2 *Demonstration of the method used to calculate age-related changes in expression level and heterogeneity of two genes during development (a) and ageing (b). Left panels: Age- related expression change is quantified by a linear model using scaled gene expression values (y-axis) and the fourth root of ages (x-axis). Beta values and raw p-values are indicated on the figure. The blue line is drawn based on the linear model, and the red dotted lines represent residuals. Right panels: corresponding scatterplots for the absolute values of residuals (y-axis) and fourth root of age (x-axis). Rho and raw p-values are calculated based on Spearman's correlation test. The blue line is drawn based on a linear model between the absolute value of residuals and the fourth root of age, only for demonstration purposes.*

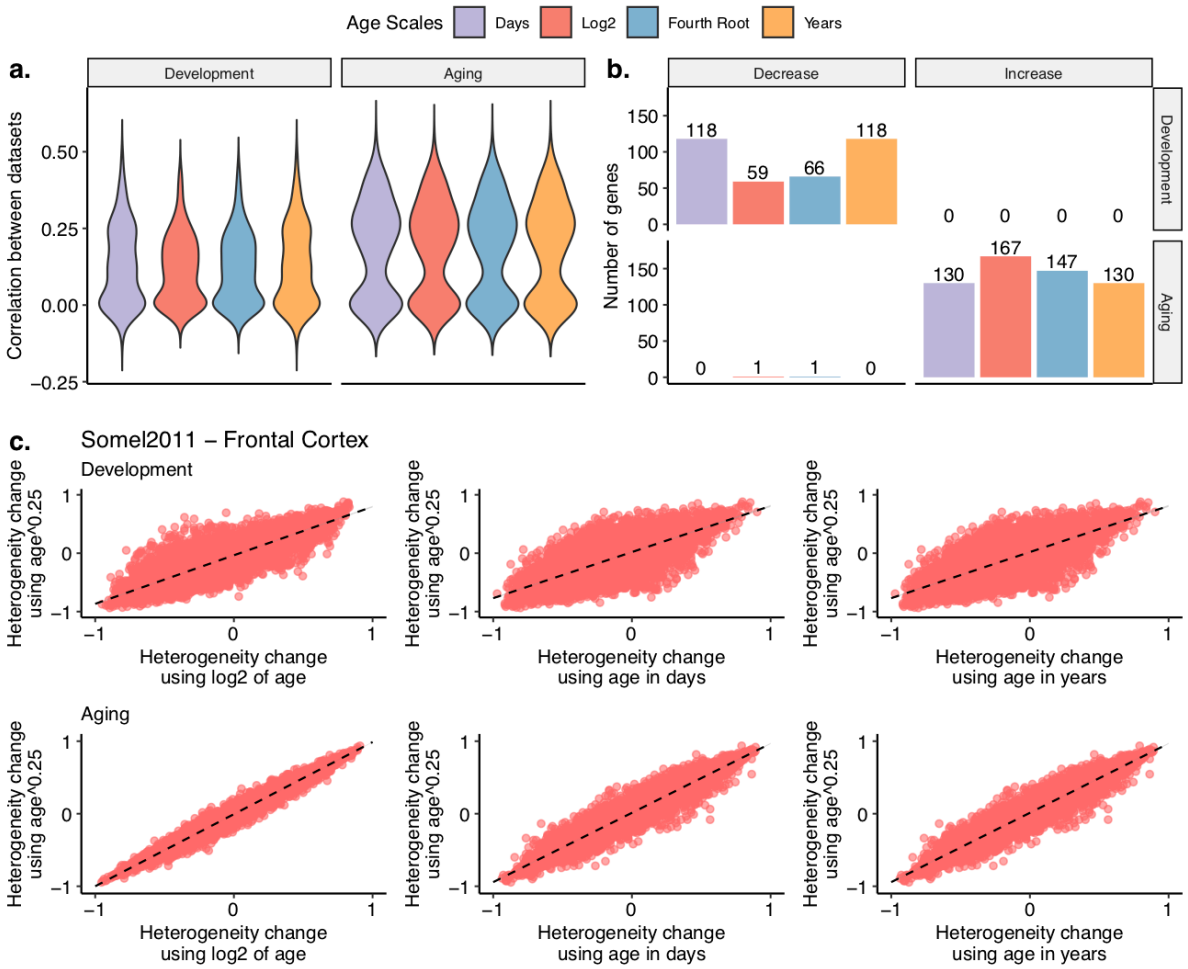


Figure A.3 *Different age scales yields consistent results. We performed downstream analyses using age in days, age in years, log2 of age in days, and fourth root of age in days. (a) Distribution of correlation coefficients between age-related heterogeneity changes across development and ageing datasets (y-axis), using different age scales (x-axis). (b) The number of genes showing consistent heterogeneity change across all 19 datasets (y-axis), using different age scales. (c) Scatterplots of age- related heterogeneity change values of 11,137 genes from one example dataset (Somel2011_PFC) calculated using the fourth root of age in days (y-axis) and different scales including log2 of age in days, age in days and age in years (x-axis) during development (upper panel) and ageing (lower panel).*

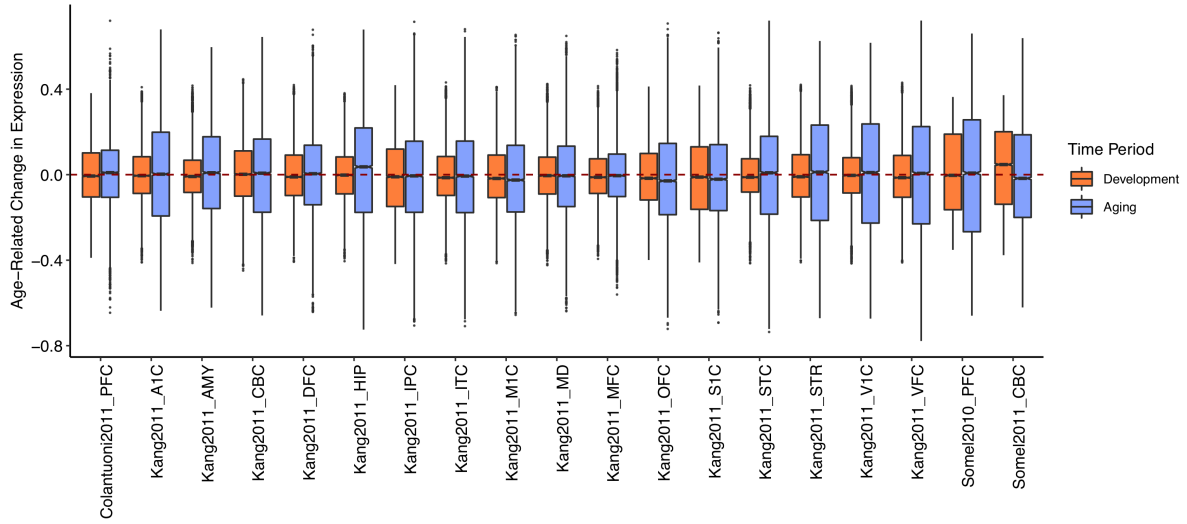


Figure A.4 *Distribution of age-related change in expression values (Beta values) for each dataset (x- axis) during development and ageing. Beta values (y-axis) were computed by linear models using scaled expression values and the fourth root of age for each gene in each dataset during development (orange) and ageing (blue) separately.*

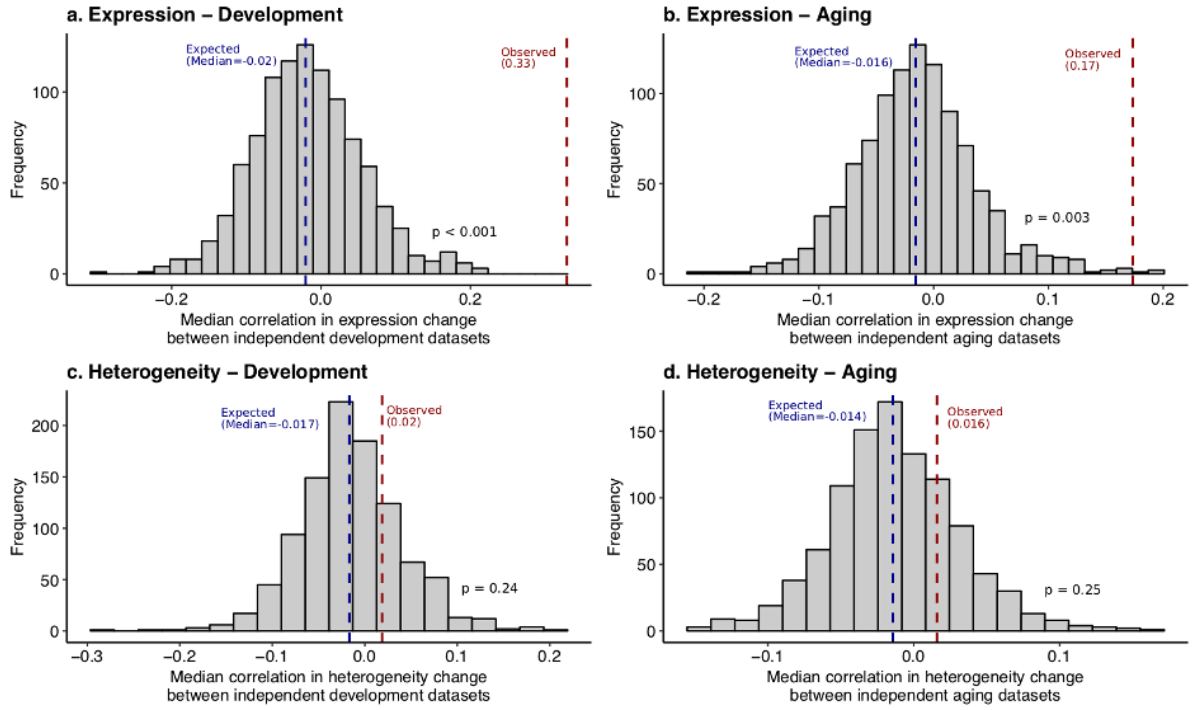


Figure A.5 *Results of the permutation tests for dataset correlations of expression and heterogeneity change during development and ageing. We constructed the distributions by calculating median correlation between all possible three independent datasets from three data sources for all permutations. Similarly, the observed values were calculated as a median value for all three independent dataset combinations from three data sources (i.e. the median value of: Median Correlation(Kang2011_A1C, Somel2011_PFC, Colantuoni2011_PFC), Median Correlation(Kang2011_AMY , Somel2011_CBC, Colantuoni2011_PFC), ...). (a, b) Permutation test results for the significance of observed correlation in gene expression change during development (a) and ageing (b). (c, d) Permutation test results for the significance of observed correlation in heterogeneity change during development (c) and ageing (d).*

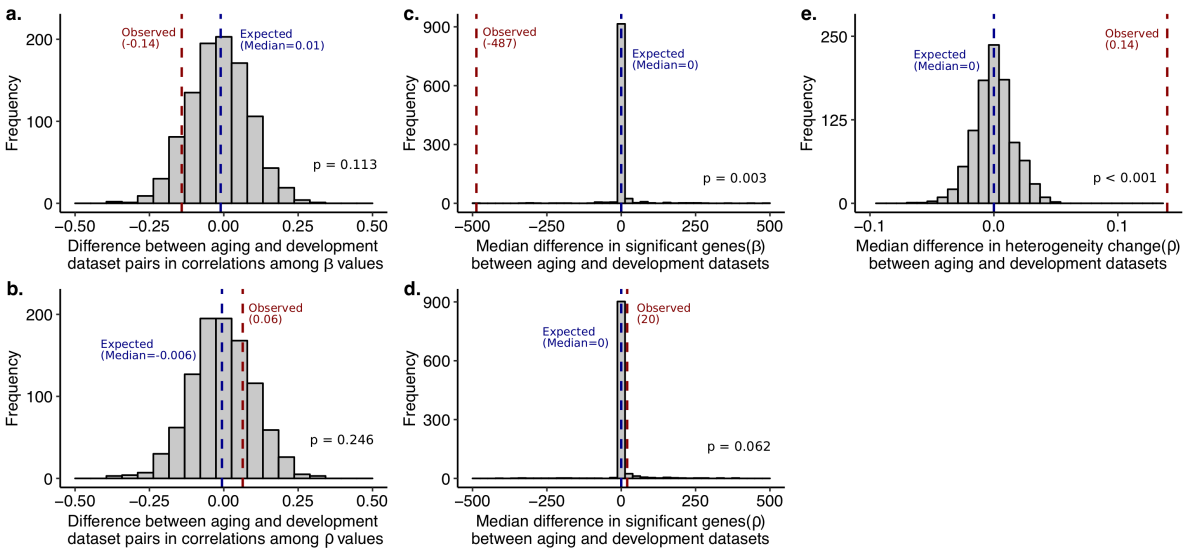


Figure A.6 *(a, b) Results of the permutation tests for differences between dataset correlations during development and ageing. Distributions were constructed using the median differences between correlations among ageing datasets and development datasets for 1000 permutations, i.e. negative values imply higher correlation among development datasets compared to ageing datasets, while positive values indicate higher correlation during ageing. (a) Distribution of median differences between correlations among age-related expression changes (β values) among ageing datasets and development datasets. (b) The same distribution for heterogeneity changes (ρ values). (c, d) Permutation test results for the median difference in the number of genes showing significant change between ageing and development datasets, in terms of (c) expression change and (d) heterogeneity change. For each permutation, p-values were corrected for multiple testing using FDR method. (e) Permutation test result for the median heterogeneity change differences between ageing and development datasets. Associated p-values are from one-tailed tests.*

Figure A.7 *Hexagonal density maps for the changes in heterogeneity based on the residuals from a linear model (x-axis) and loess regressions (y-axis). The colour intensity shows the number of genes in that hexagonal bin, where darker colour shows more genes. The blue line is drawn using linear regression, for demonstration purposes. The correlations between the values are calculated using Spearman's correlation test.*

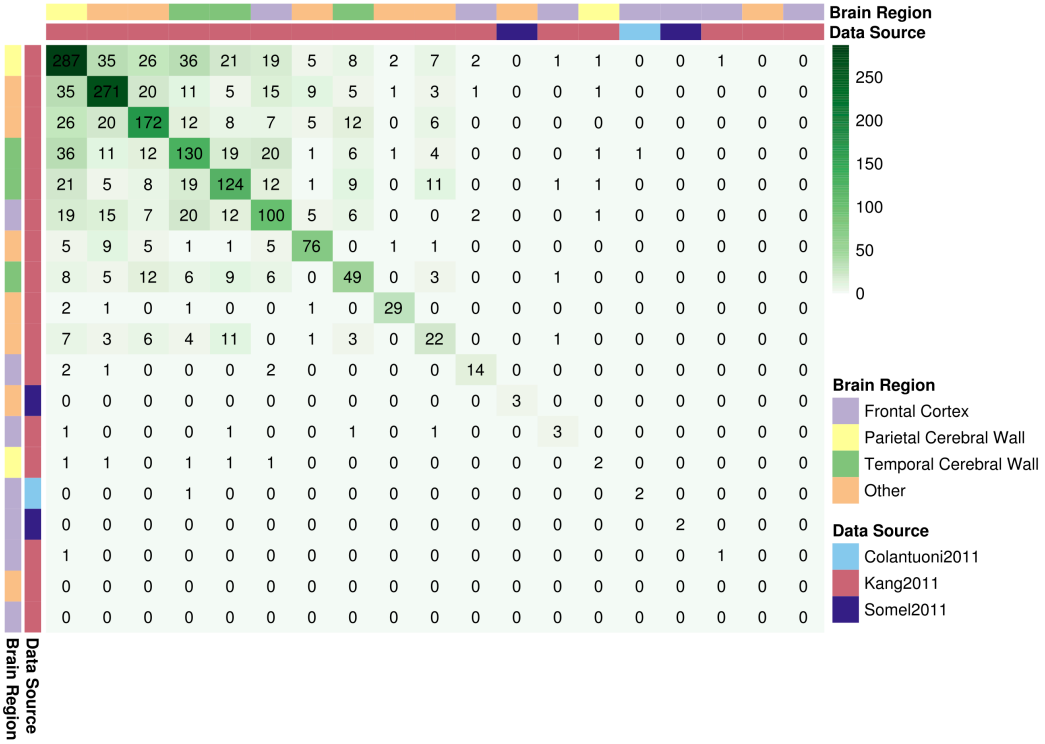


Figure A.8 *Overlaps between datasets for the genes showing a significant increase in heterogeneity. p-values are computed using Spearman correlation test between the absolute value of residuals and fourth root of ages, followed by FDR correction. The colour intensity of cells reflects an increased number of overlapping genes among two corresponding datasets, while the numbers show the exact numbers.*

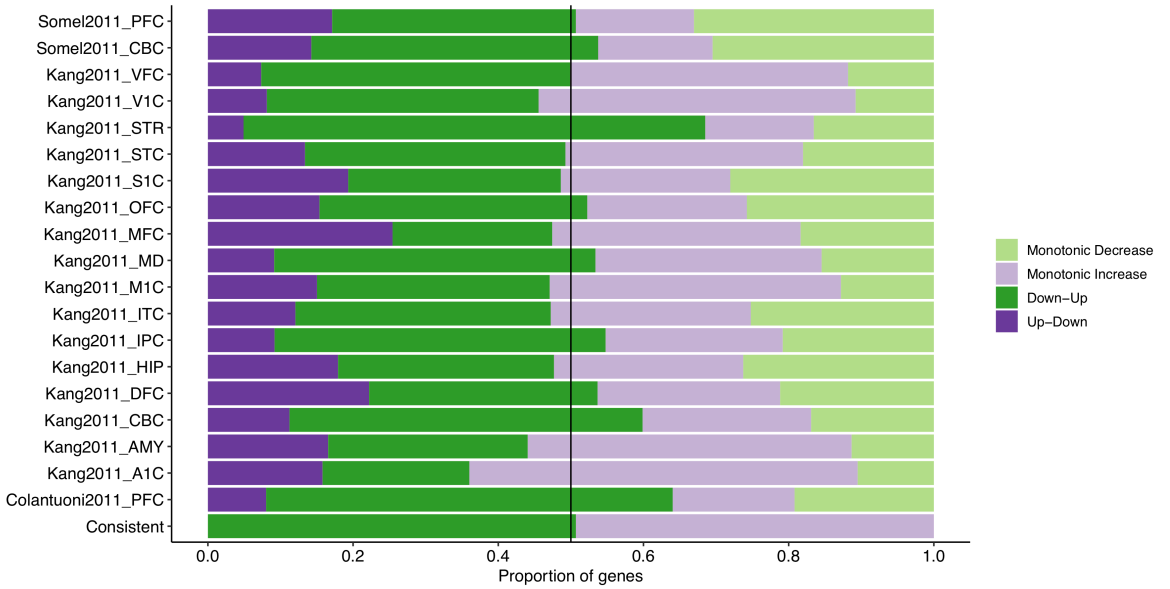


Figure A.9 *The proportion of different trends in age-related heterogeneity change in each dataset and among the genes showing a consistent increase across ageing datasets (n = 147). No effect size or significance cutoff was used. Up-down: increase in development & decrease in ageing; down-up: decrease in development & increase in ageing; monotonic increase: increase in development and ageing; monotonic decrease: decrease in development and ageing.*

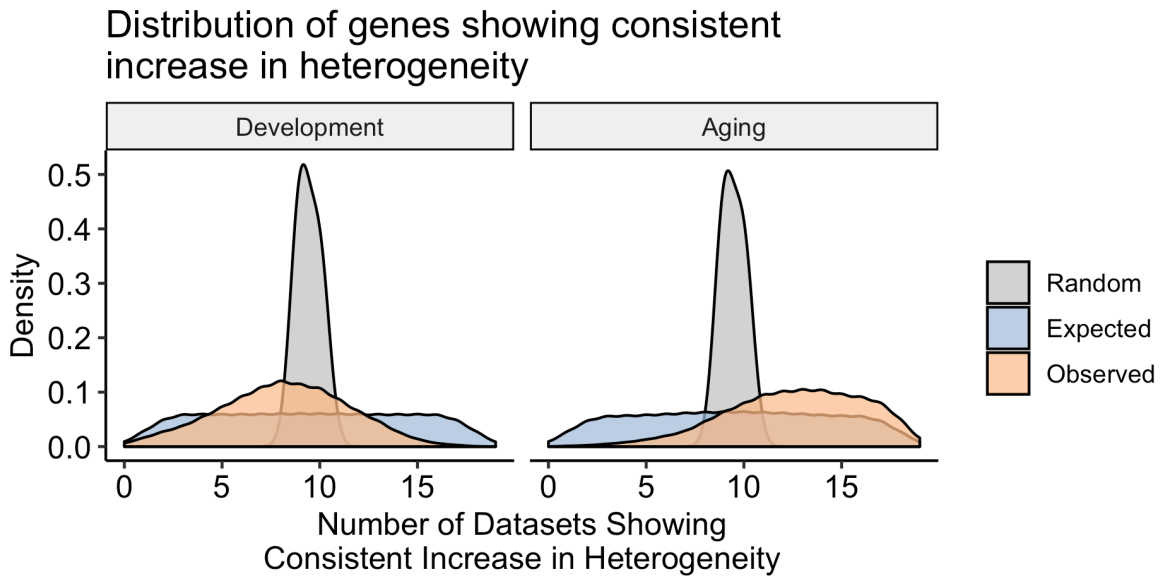


Figure A.10 *Random (expectation calculated with independent permutations), expected (based on permutations taking dataset dependency into account) and observed consistency in the heterogeneity change across datasets in development and ageing. Here we demonstrate the expectation based on independent permutations ('Random'), and demonstrate that it is less stringent than our permutation scheme.*

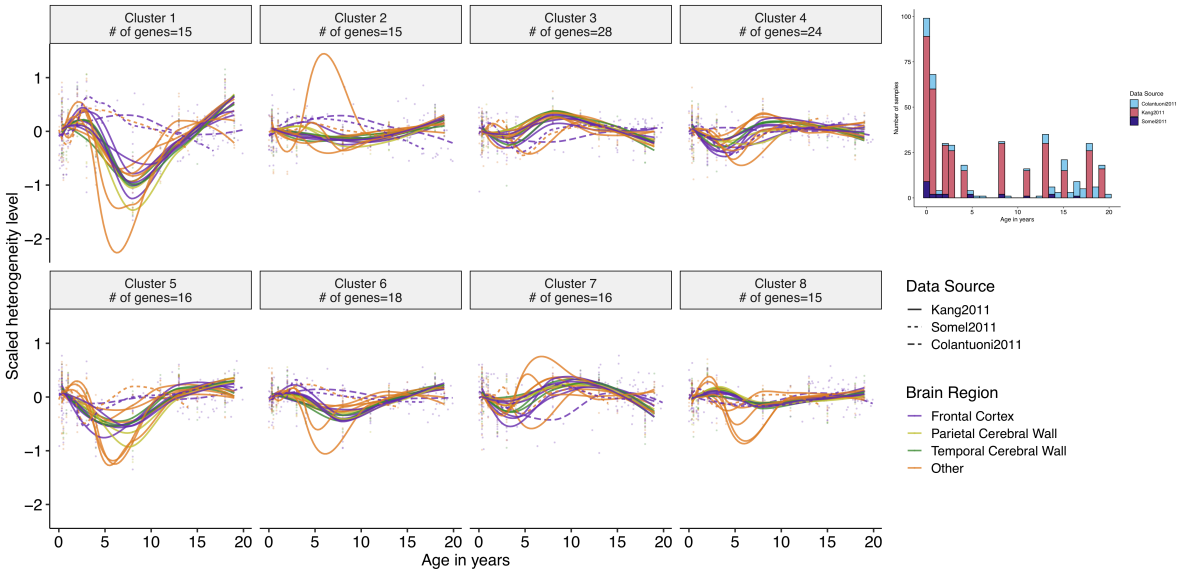


Figure A.11 *Heterogeneity trajectories in development for the genes that were clustered according to their heterogeneities in ageing, using k-means clustering method. The x-axis shows the age and the y-axis shows scaled heterogeneity level (residuals from linear model). Spline curves represent mean age-related heterogeneity changes of genes in each cluster, from each dataset and brain region. The colours and line-types of curves specify different brain regions and data sources, respectively. Different patterns observed between age of 5 to 8 could be biologically relevant but it is important to note that the number of samples in this age range is low. The age distribution of samples in development datasets is given as a bar plot in upper right corner.*

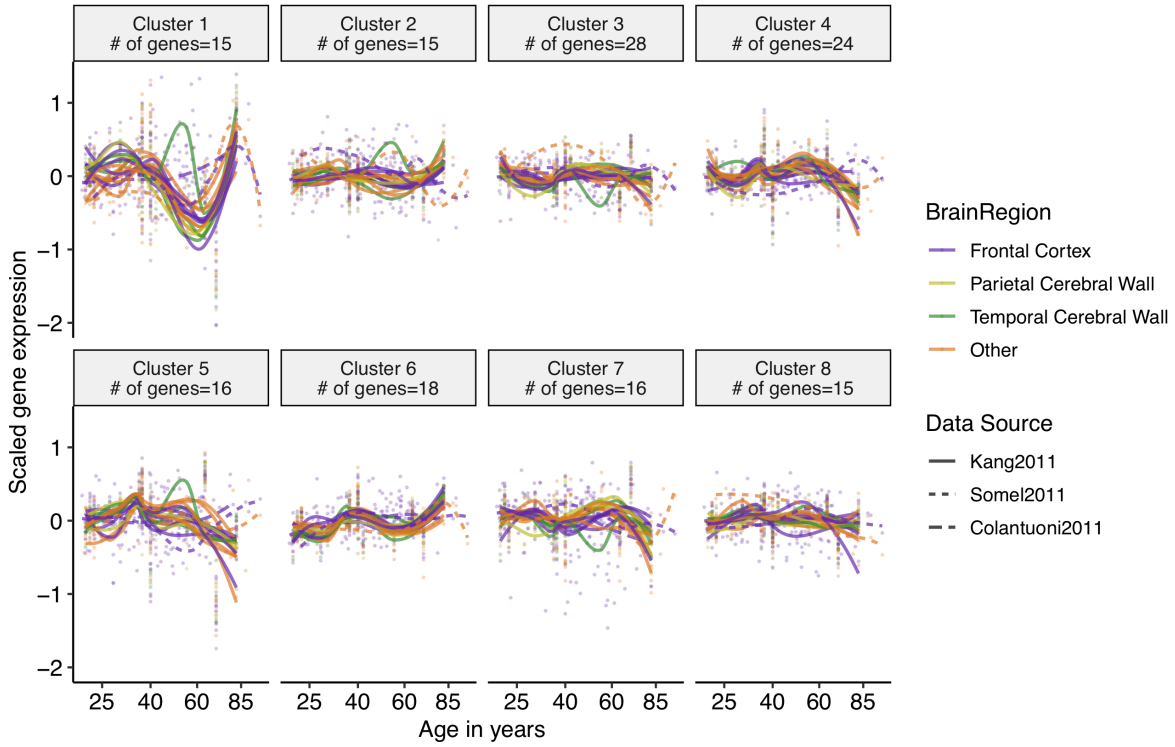


Figure A.12 *Gene expression trajectories of the genes that were clustered according to their heterogeneities, using k-mean clustering method. The x-axis shows the age on the fourth root scale, and the y-axis shows scaled gene expression values. Spline curves represent mean age-related expression changes of genes in each cluster, from each dataset and brain region. The colours and line-types of curves specify different brain regions and data sources, respectively.*

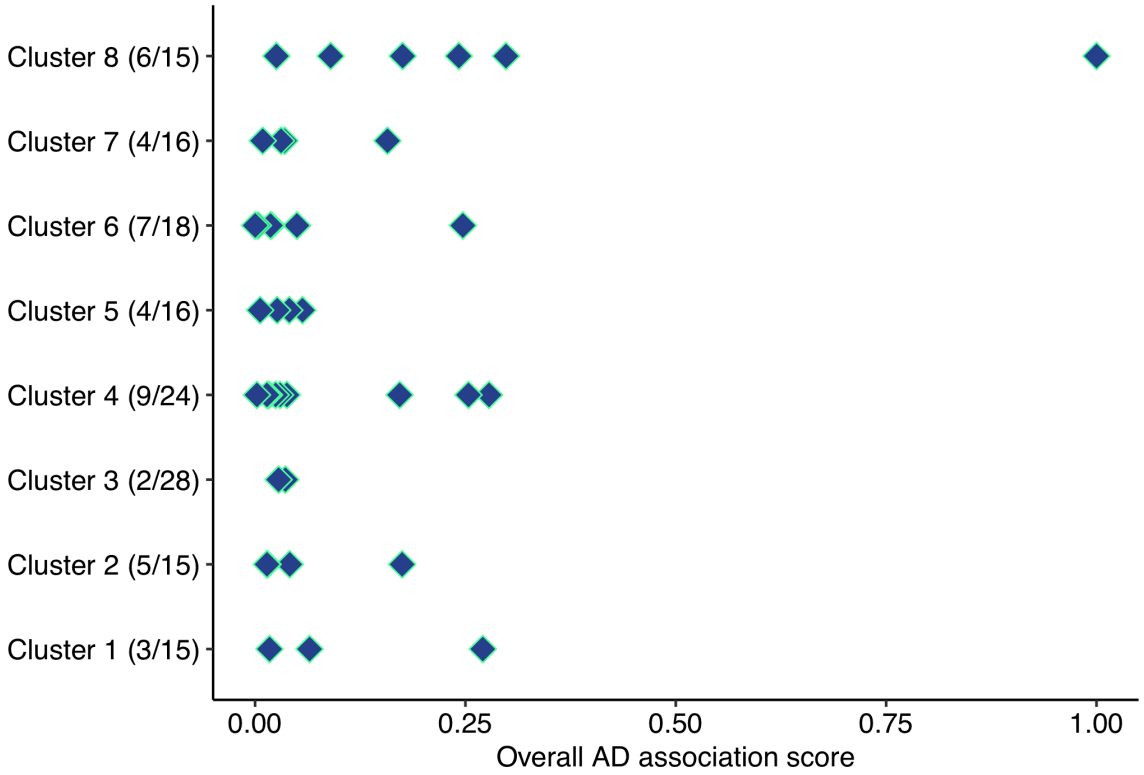


Figure A.13 *Association between the Alzheimer's-related genes and the genes that consistently become more heterogeneous in ageing (n = 147) and belong to different heterogeneity trajectories. The x-axis shows overall score for the association with Alzheimer's Disease, while the y-axis shows different heterogeneity trajectory clusters. Numbers in the parenthesis on the y-axis reflect the proportion of the genes in clusters found to be in association with AD (40/147, overall).*

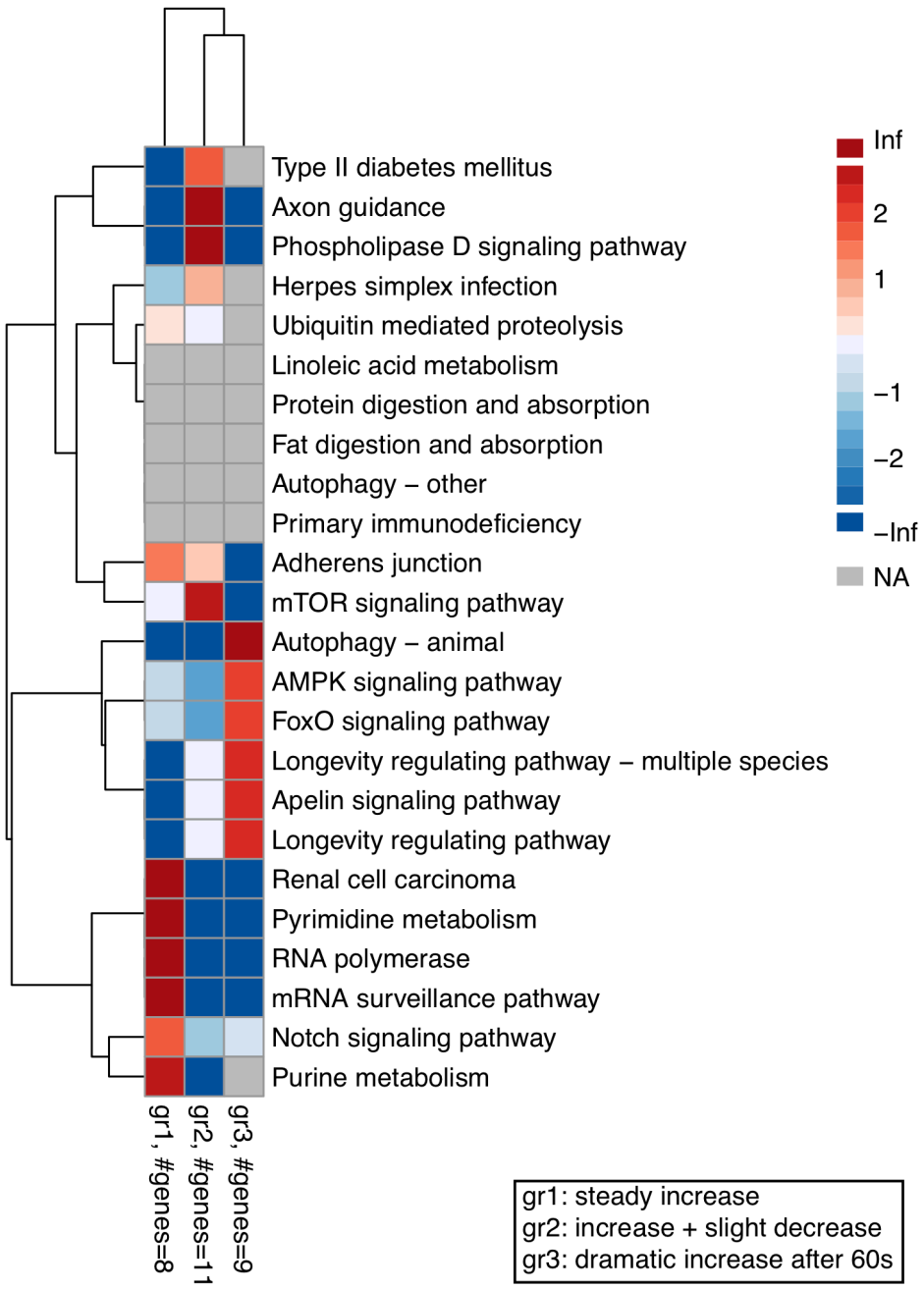


Figure A.14 *Heatmap showing the association between different heterogeneity trajectories and KEGG pathways that are significantly associated with a consistent change in heterogeneity during ageing. Heterogeneity trajectories are based on the definitions in the main text and i) gr1 includes clusters 3 and 7, ii) gr2 includes clusters 4, 5, and 8, and iii) gr3 includes clusters 1, 2, and 6. The colours represent the Odd's ratio, where red shows enrichment and blue shows depletion.*

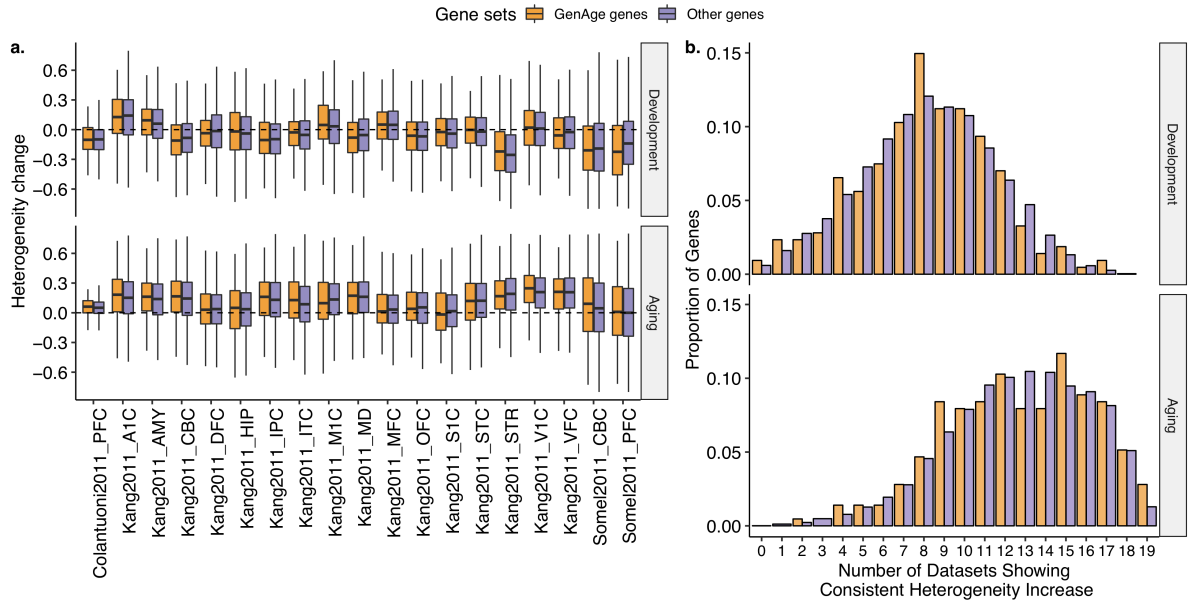


Figure A.15 *Association between GenAge human gene set (n = 214) and age-related heterogeneity. (a) Boxplots show age-related heterogeneity changes (ρ values) (y-axis) of genes from GenAge gene set (orange) and the remaining genes (purple) in different datasets (x-axis) during ageing (lower panel) and development (upper panel). (b) Consistency in age-related heterogeneity increase in genes from GenAge gene set (orange) and the other genes (purple) in ageing (lower panel) and development (upper panel). The x-axis shows the number of datasets among which genes show consistent heterogeneity increase, while the y-axis shows the proportions of the number of genes.*

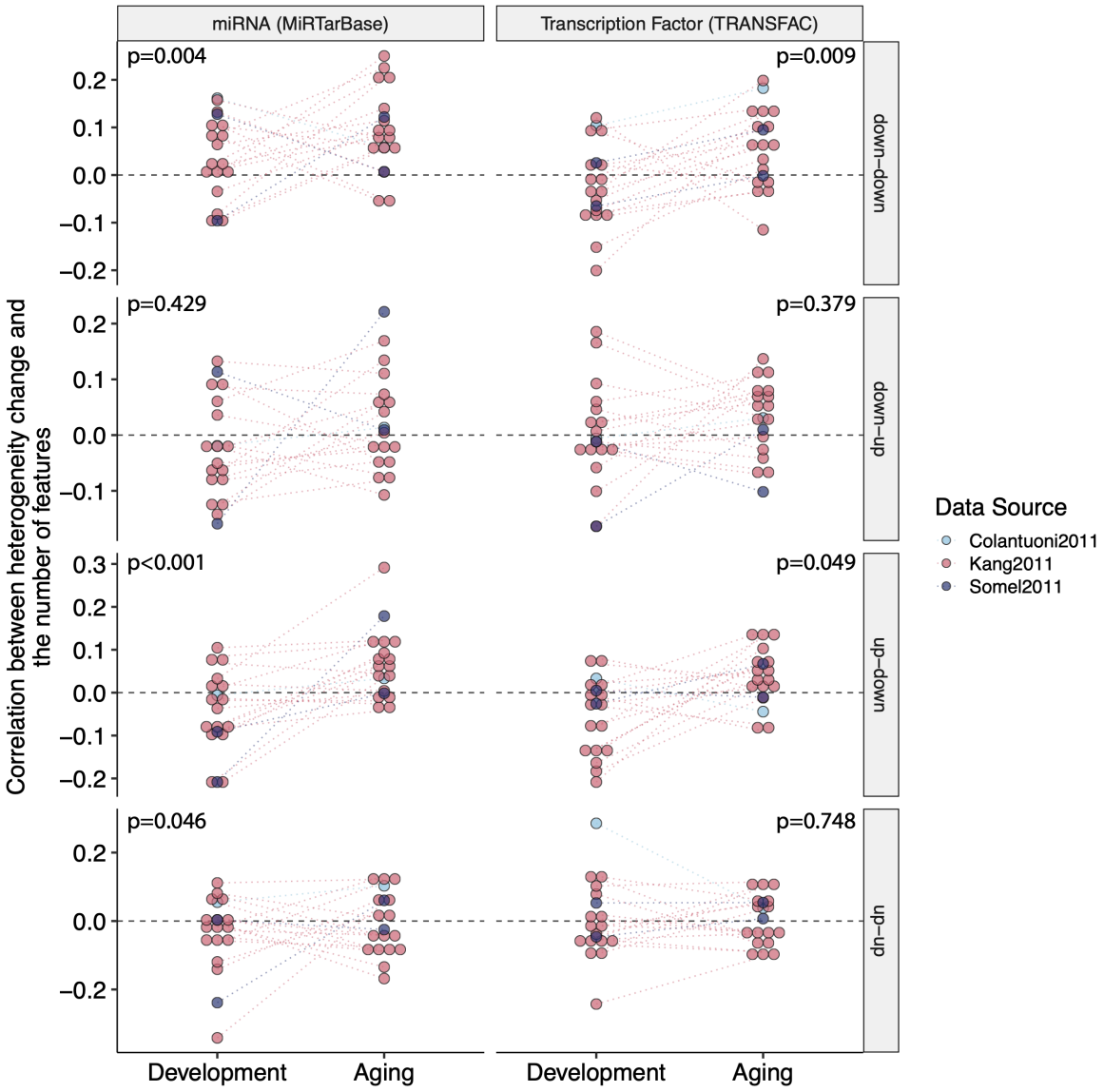
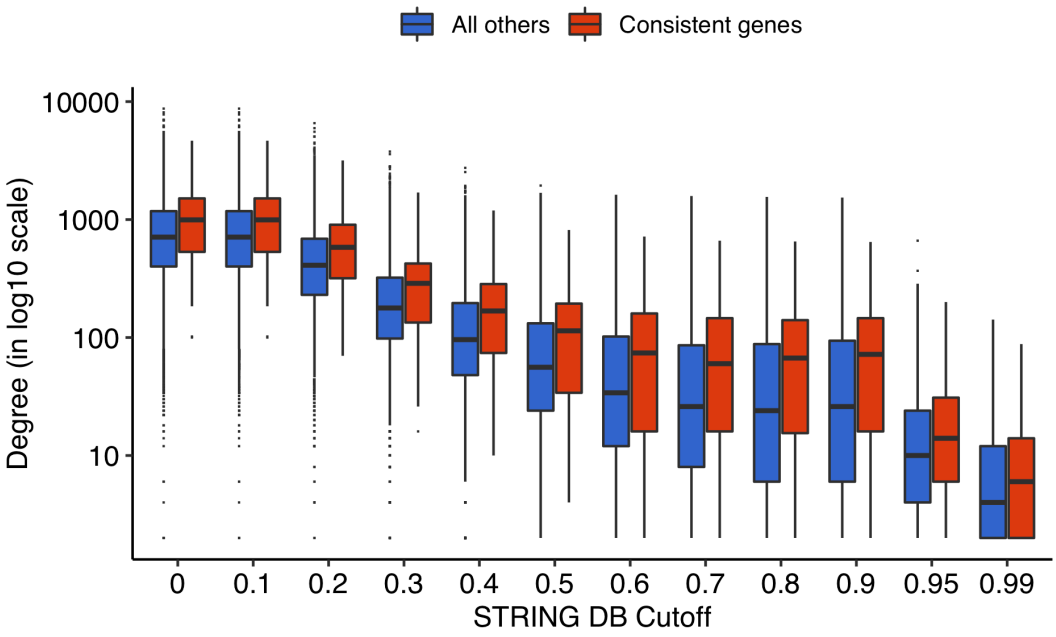


Figure A.16 *Correlation between the change in heterogeneity and number of transcriptional regulators, i.e. miRNA and transcription factors. Each point represents a dataset, and the colour shows the data source. p-values are calculated using a permutation test. The dashed line at y = 0 shows zero correlation. Genes were divided into four sets based on the change in their expression level in development and ageing, e.g. "down-down" includes genes with decreased expression in both development and ageing, whereas "down-up" includes genes with decreasing expression level in development and then increase in ageing.*



Cutoff	Number of Genes	Number of Interactions	p	FDR
0.00	11,016	4,920,316	<0.0001	<0.0001
0.10	11,016	4,920,316	<0.0001	<0.0001
0.20	11,015	2,918,350	0.0001	0.0001
0.30	11,015	1,376,452	<0.0001	<0.0001
0.40	11,008	845,298	<0.0001	<0.0001
0.50	10,955	578,970	<0.0001	<0.0001
0.60	10,665	438,686	<0.0001	<0.0001
0.70	9,881	353,362	<0.0001	<0.0001
0.80	8,588	305,128	<0.0001	<0.0001
0.90	7,386	270,942	0.0001	0.0001
0.95	5,800	58,190	0.0235	0.0256
0.99	3,167	17,660	0.3989	0.3989

Figure A.17 *Degree distributions in protein-protein interaction database (STRING) for consistent genes (red) and all others (blue). The y-axis show degree (number of interactors) in log scale. The x-axis shows different cutoffs for interaction confidence used to filter STRING database. The colour represents two sets of genes; red: consistent genes that show consistent increase in heterogeneity across all ageing datasets, blue: all other genes. Interaction degree is significantly higher in consistent genes across all cutoffs except for 0.99 (permutation test). Details, including the number of genes, interactions and p-values, are given as a table.*

Cell Type	Odds Ratio	FDR	Total Cell–type	Total Heterogeneus	a	b	c	d	p
OLs	5.537	0.179	51	9	2	49	7	953	0.072
Myelin_OLs	3.854	0.179	348	9	6	342	3	660	0.071
OPCs	0.265	0.357	322	9	1	321	8	681	0.286
Neurons	0.000	0.205	272	9	0	272	9	730	0.123
Astrocytes	0.000	1.000	18	9	0	18	9	984	1.000

Figure A.18 *Cell-type specificity analysis for genes that become heterogenous with age across all ageing datasets. Fisher's exact test is used to calculate the association (Odds Ratio) and p-value. FDR corrected p-value (FDR column) is also given. "Total Cell-type" column is the total number of cell-type-specific genes, "Total Heterogeneous" is the total number of genes in the 147 consistently heterogeneous genes and are cell-type specific for at least one cell type, "a" is the overlap between cell-type specific genes and heterogeneous for a given cell-type, "b" is the cell-type specific genes that are not among the 147 genes but in the dataset, "c" is the number of heterogeneous genes that are specific for other cell-types, and "d" is the total number of genes that are specific for at least one cell-type and measured in the age-series datasets.*

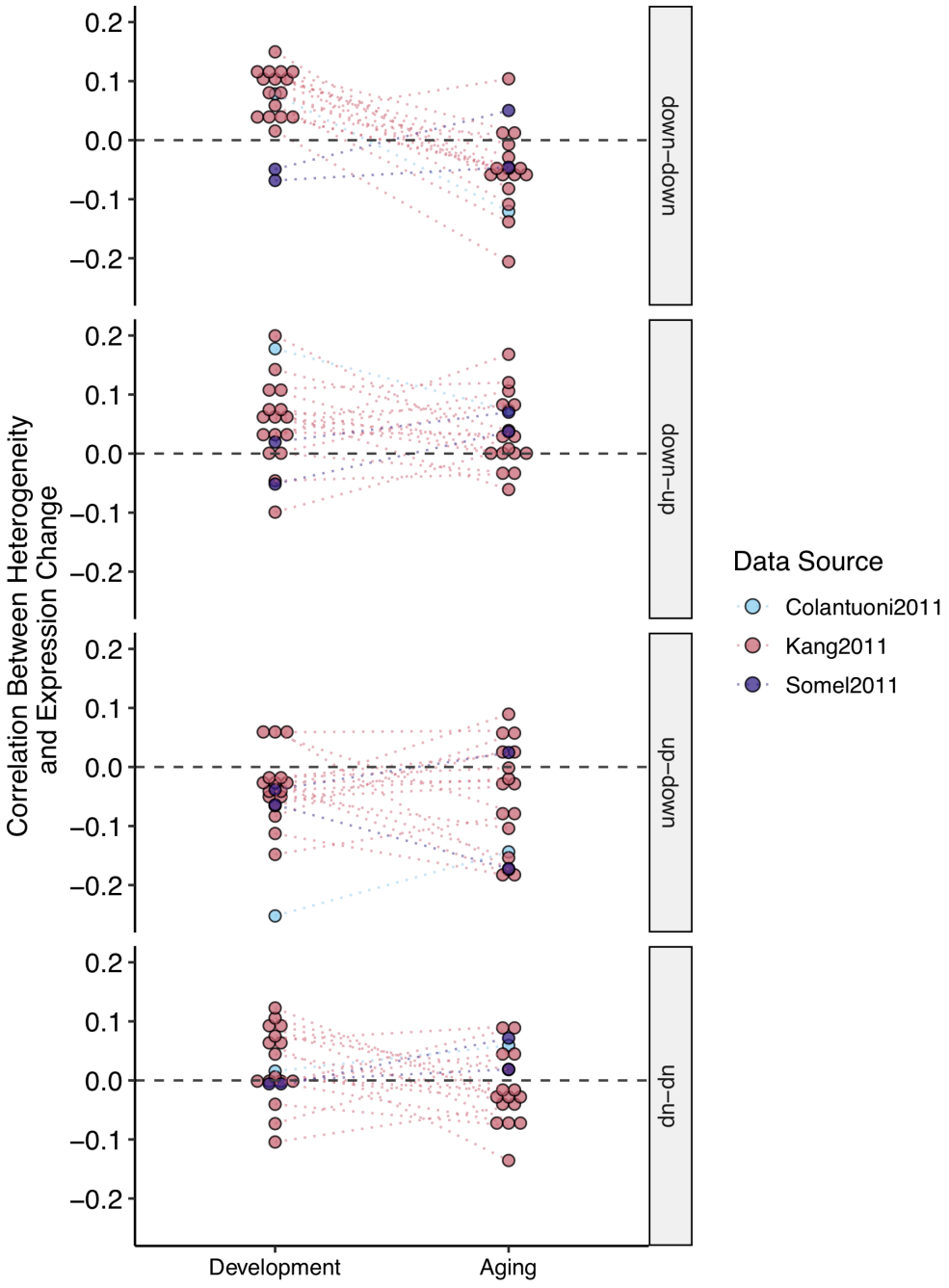


Figure A.19 *a) The change in cell-type proportions with age in different datasets (beta values from a linear model between cell-type proportions and age in fourth root scale), b) The change in cell-type proportions based on the 147 genes that become more heterogeneous with age during ageing, c) The change in cell-type heterogeneity with age (correlation between the absolute value of residuals from the linear model and age), d) The change in cell-type heterogeneity with age, calculated only using the 147 genes.*

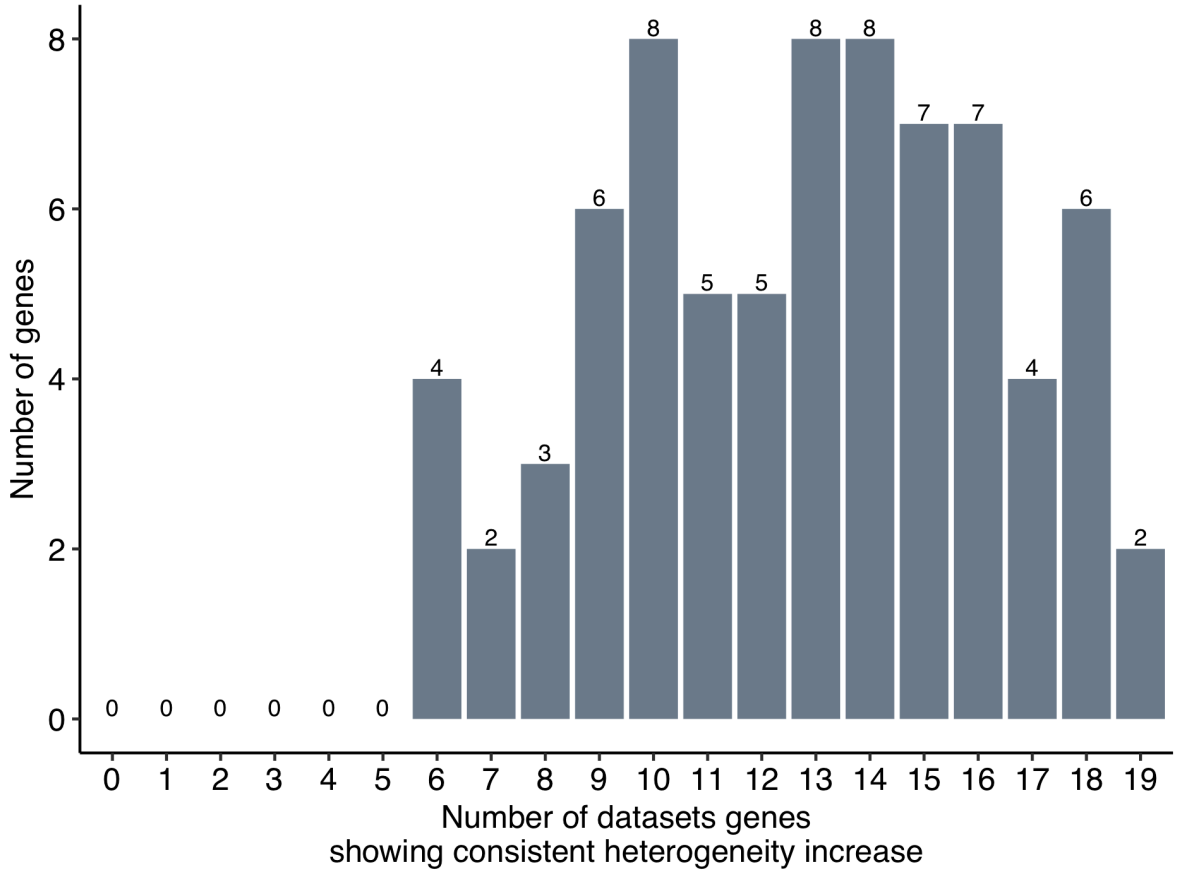


Figure A.20 *Analysis of the consistency in ageing-related heterogeneity of the post-mortem interval (PMI)-associated genes. The x-axis shows the number of datasets in which genes were showing a consistent increase in age-related heterogeneity during ageing, while y-axis reflects the number of genes. There were 105 previously identified PMI associated genes in the human cerebral cortex (see main text), 75 of which were included in our analyses. We asked if these PMI associated genes show more increase in heterogeneity, and found that only 2 of 147 consistent genes (i.e. genes showing an increase in heterogeneity in 19 out of 19 ageing datasets) were associated with PMI.*

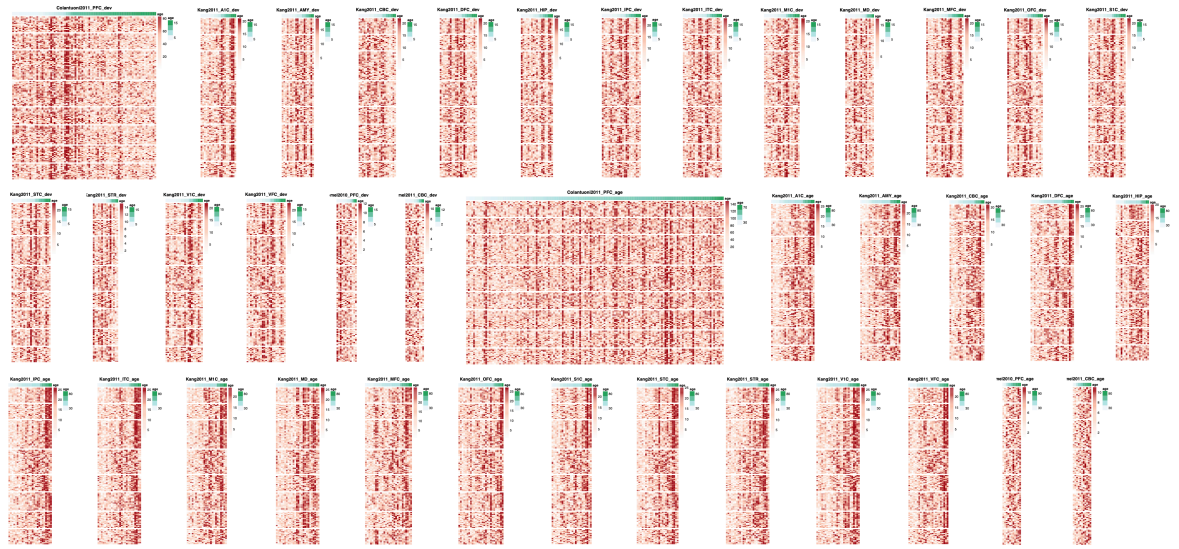


Figure A.21 *Heatmaps for 38 datasets showing the heterogeneity levels (residuals) for each individual (columns) in 147 genes that show consistent increase across all ageing datasets (rows). The colour shows the rank of individuals with respect to their heterogeneity for that particular gene. Darker colours show individuals with the highest heterogeneity for that particular gene. Genes are clustered using the heterogeneity trajectories in Figure 4 and individuals are ordered by age (darker green in column annotation showing older ages).*

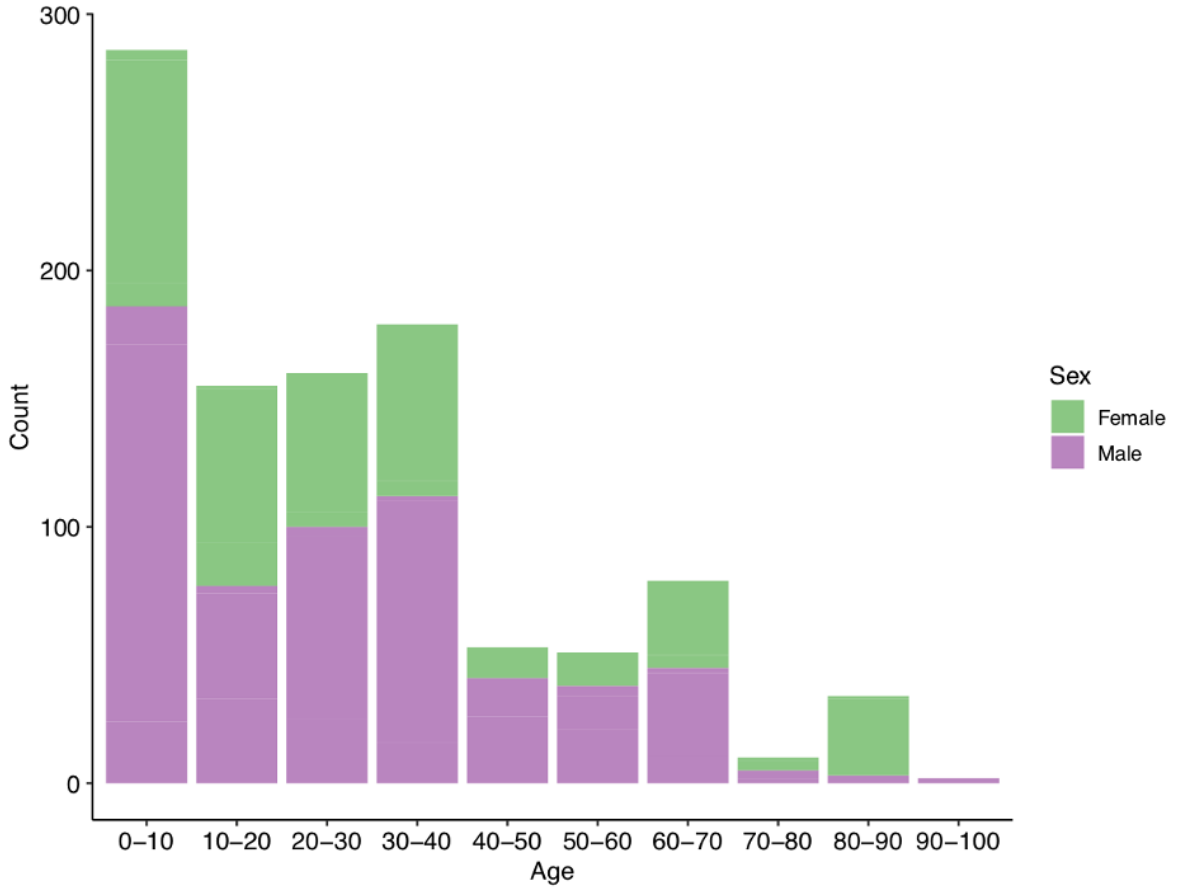


Figure A.22 *The proportions of sexes of samples used in the analysis. The x- and y-axes show age intervals and number of samples, respectively. The colours specify the proportion of male (purple) and female (green) samples.*

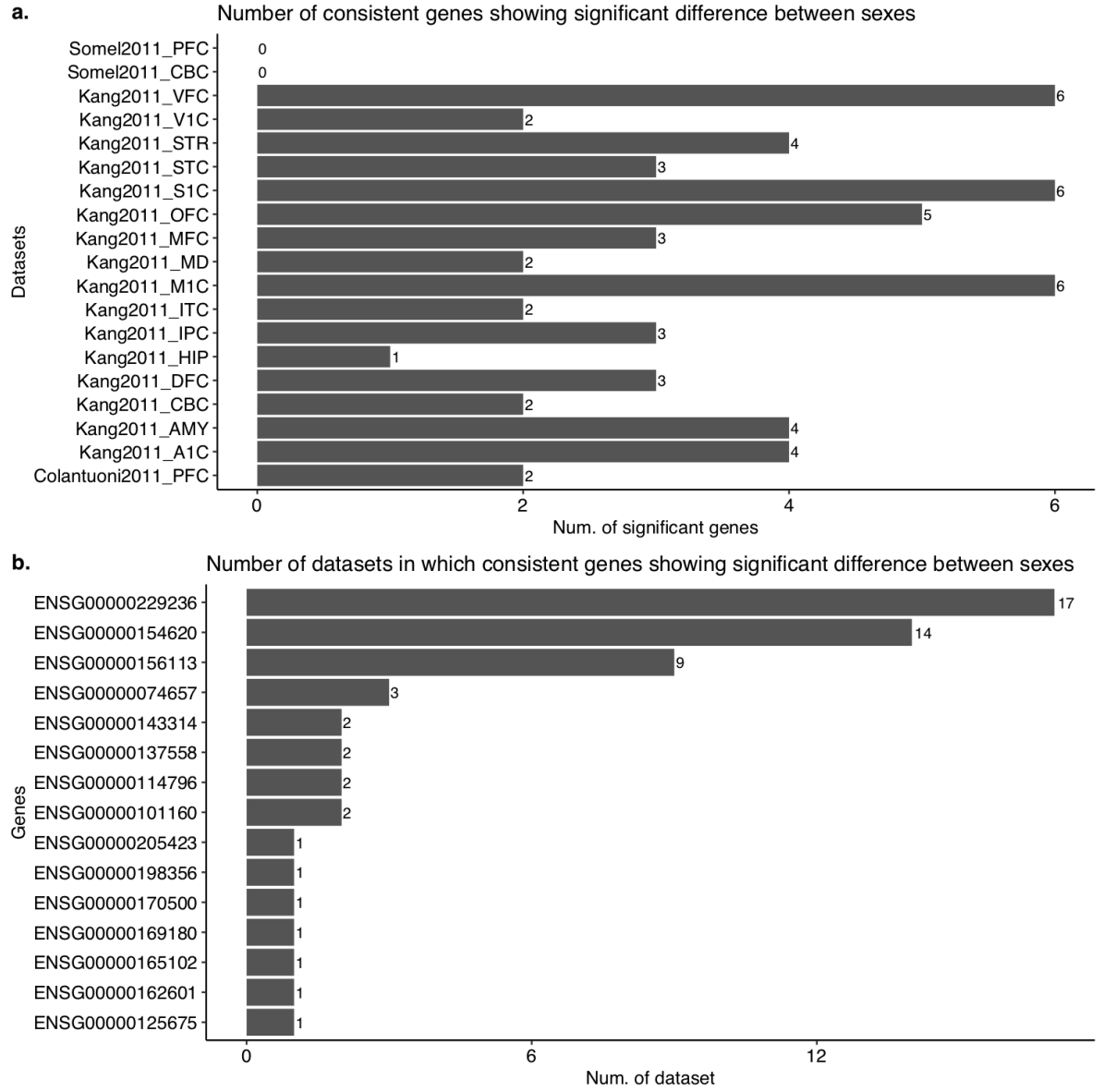


Figure A.23 *a) The change in cell-type proportions with age in different datasets (beta values from a linear model between cell-type proportions and age in fourth root scale), b) The change in cell-type proportions based on the 147 genes that become more heterogeneous with age during ageing, c) The change in cell-type heterogeneity with age (correlation between the absolute value of residuals from the linear model and age), d) The change in cell-type heterogeneity with age, calculated only using the 147 genes.*

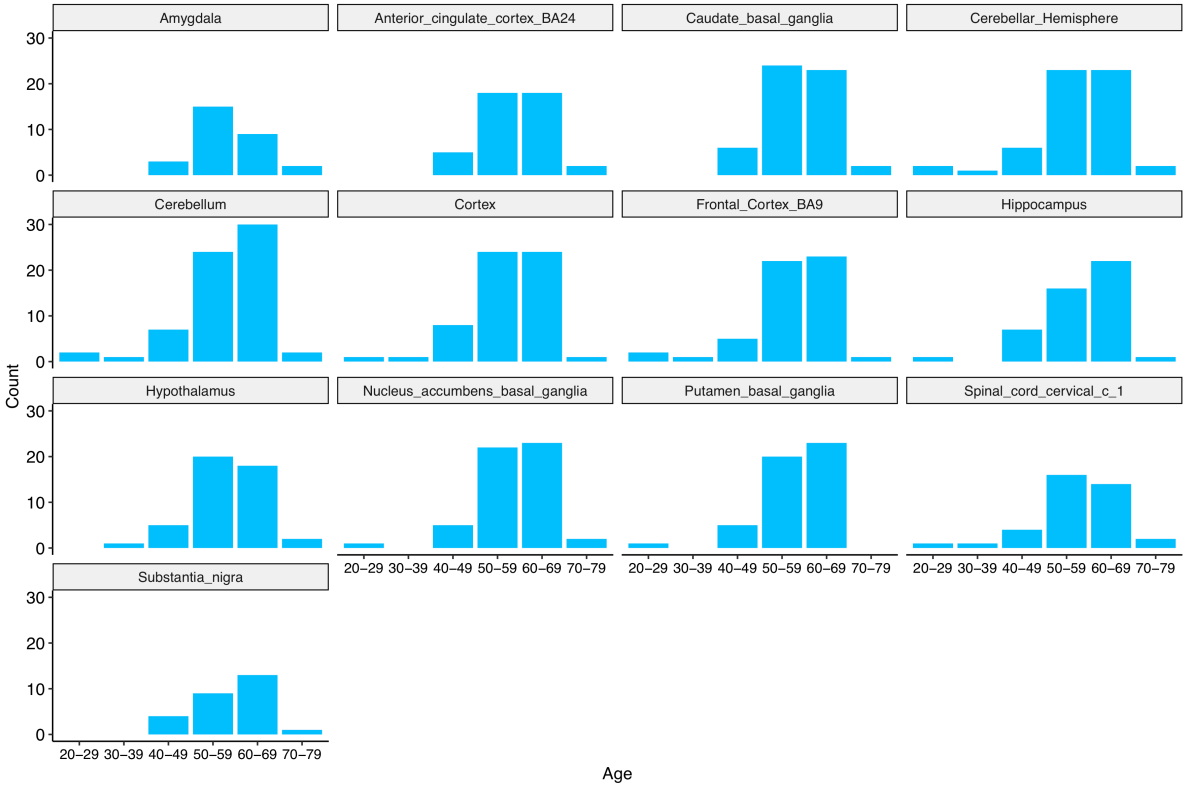


Figure A.24 *Age distribution of the individuals used in GTEx RNA-Seq datasets.*

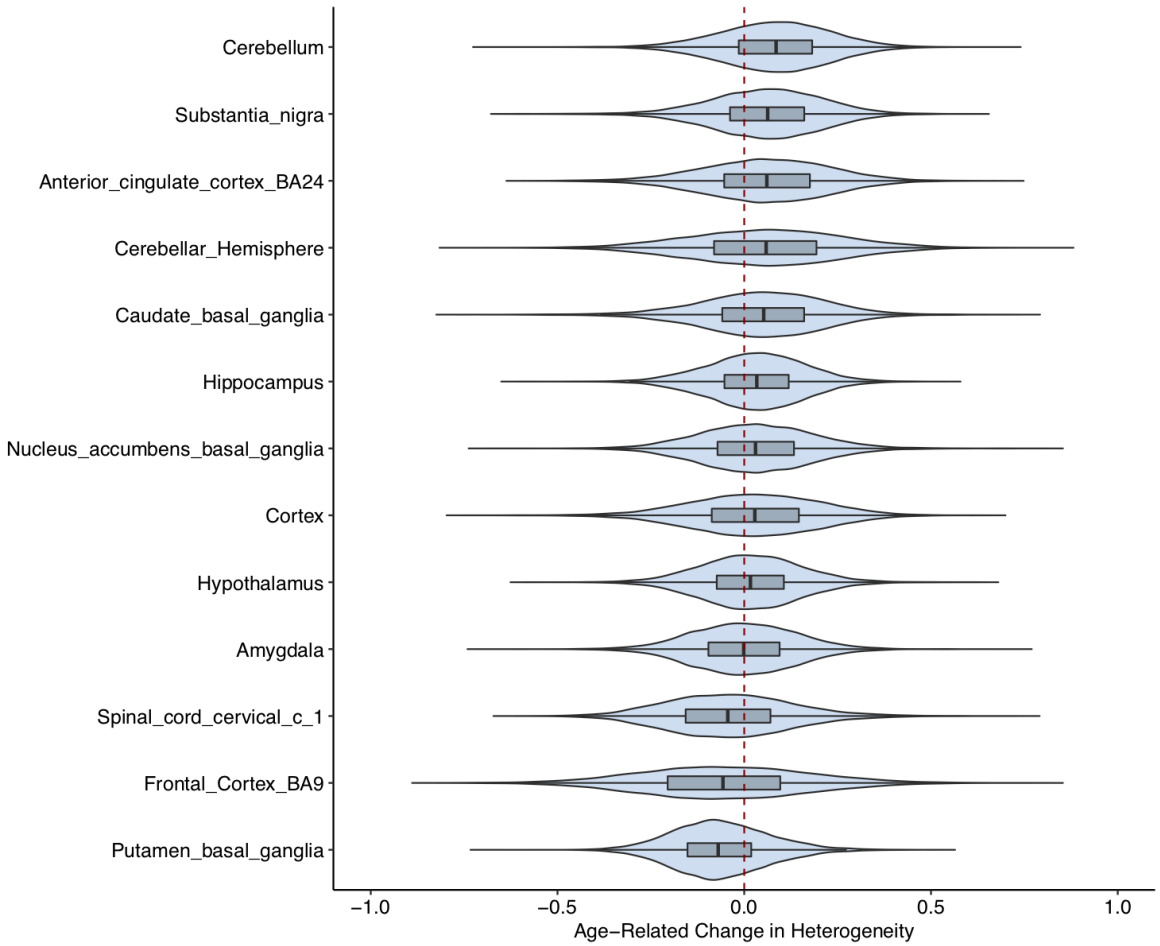


Figure A.25 *Distribution of age-related changes in heterogeneity (ρ values) (x-axis) in different GTEx datasets, corresponding to different brain regions in (y-axis). The red dotted vertical line (at x = 0) reflects no change in heterogeneity.*

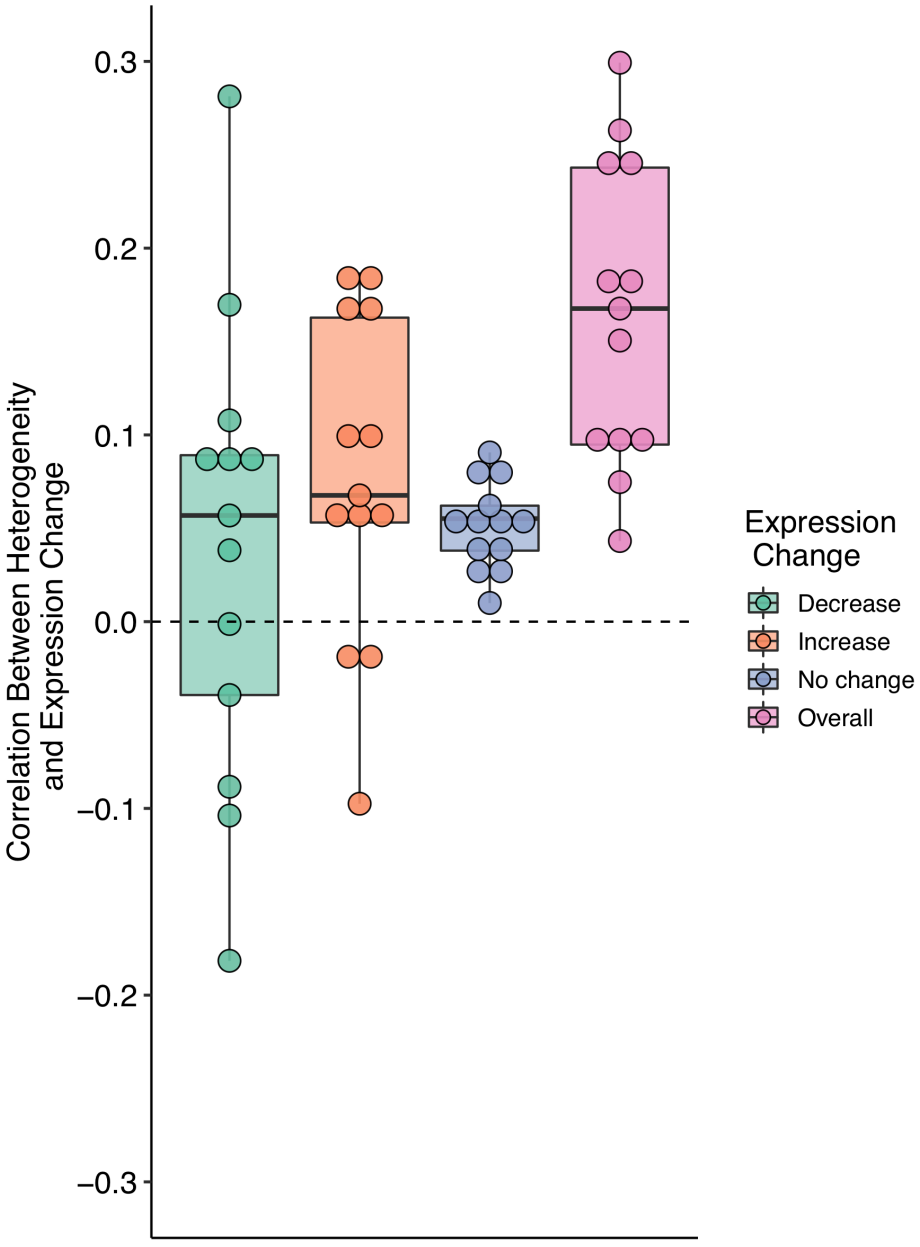


Figure A.26 *The relationship between expression and heterogeneity changes in GTEx datasets. The y-axis shows Spearmen correlation coefficients calculated between age-related expression changes (β values) and heterogeneity changes (ρ values) separately for each dataset. Different colours indicate the direction of expression change. Genes with the beta values within the range of -0.1 and 0.1 were assumed to be no change. The overall includes all the genes irrespective of the direction of their expression change.*

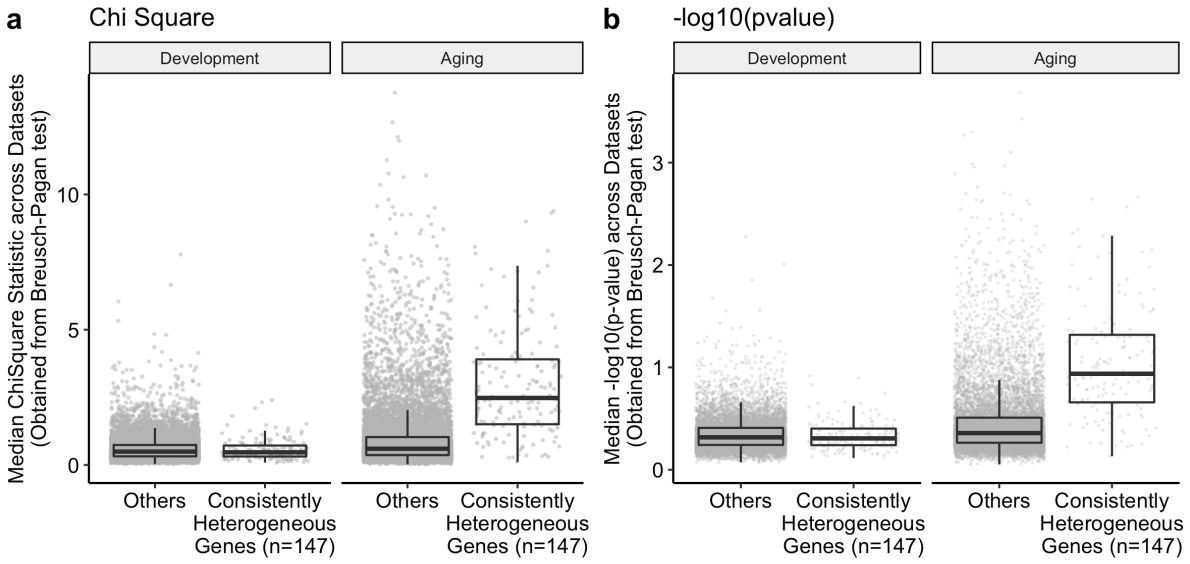


Figure A.27 *Distribution of a) median Chi Square values and b) median -log10 p values for each gene obtained from Breusch-Pagan heteroskedasticity test, stratified for the changes in development and ageing. The 147 genes that become heterogeneous with age consistently in ageing datasets shown separately to compare with other genes.*

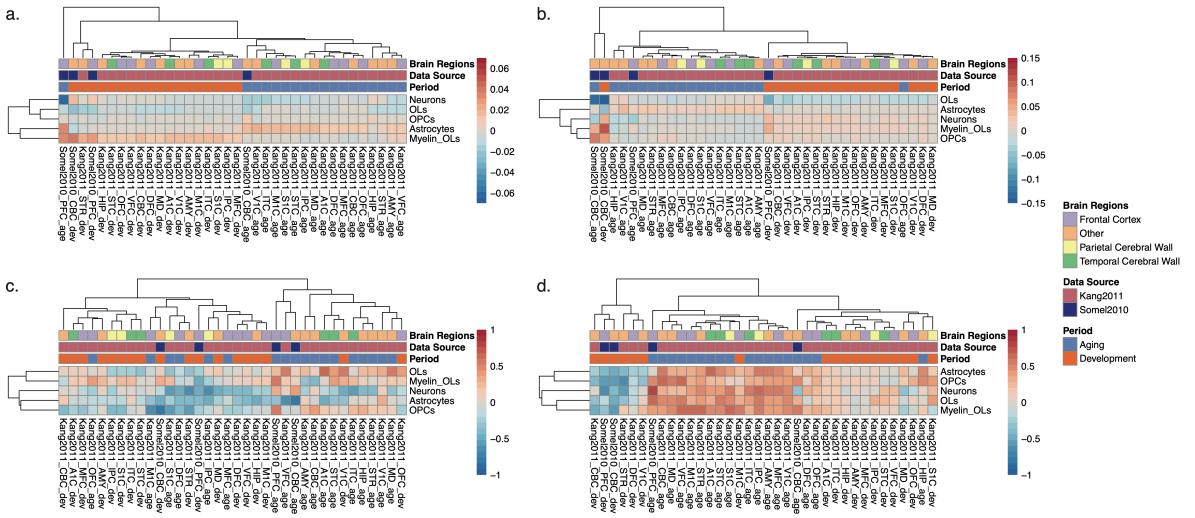


Figure A.28 *a) The change in cell-type proportions with age in different datasets (beta values from a linear model between cell-type proportions and age in fourth root scale), b) The change in cell-type proportions based on the 147 genes that become more heterogeneous with age during ageing, c) The change in cell-type heterogeneity with age (correlation between the absolute value of residuals from the linear model and age), d) The change in cell-type heterogeneity with age, calculated only using the 147 genes.*

A.2 The link between ageing and age-related diseases

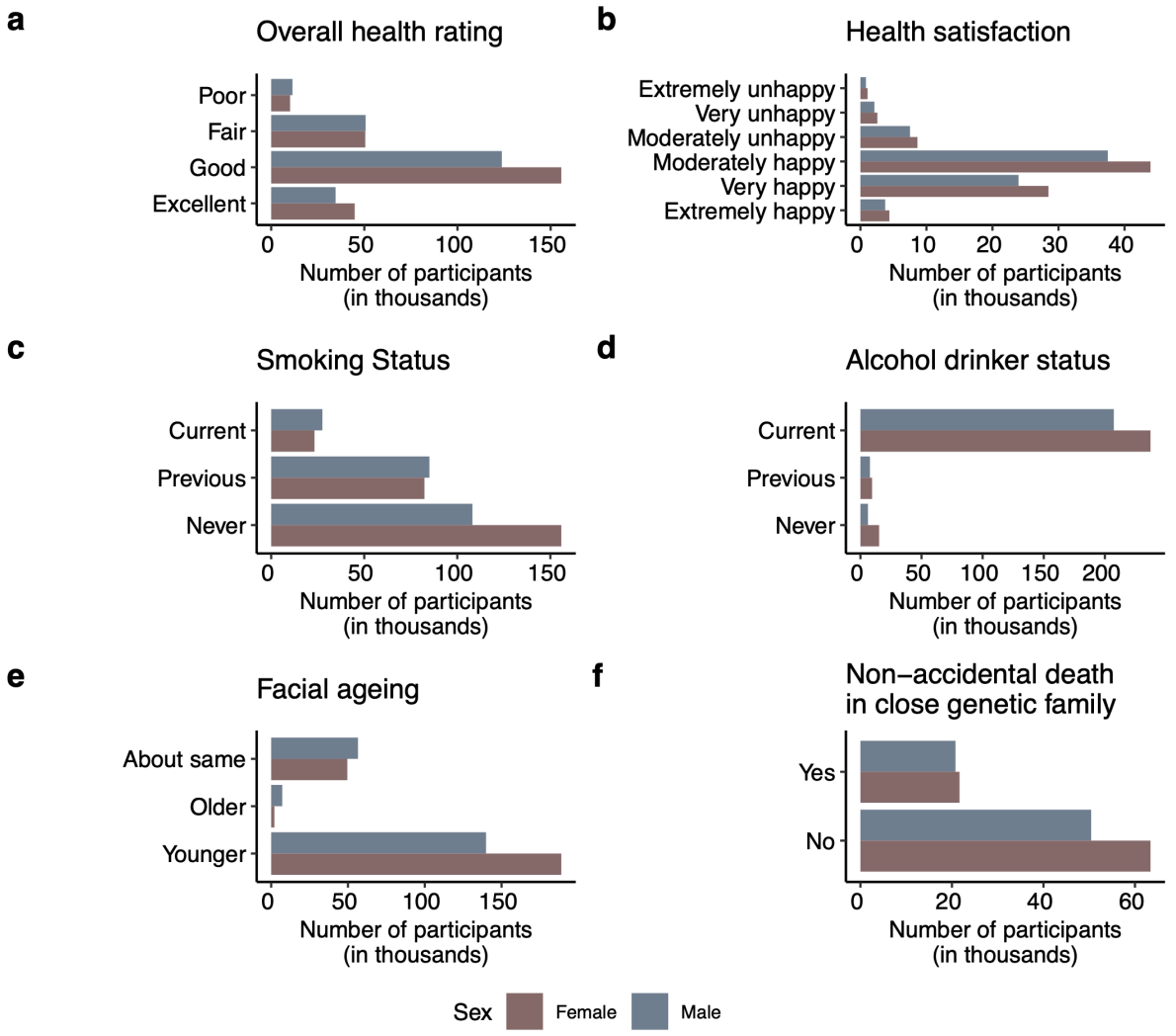


Figure A.29 *The distribution of a) Overall health rating, b) Health satisfaction, c) Smoking status, d) Alcohol drinker status, e) Facial ageing, f) Non-accidental death in close genetic family fields in the UKBB. x-axes show the number of participants, while y-axes are the answers given by the participants.*

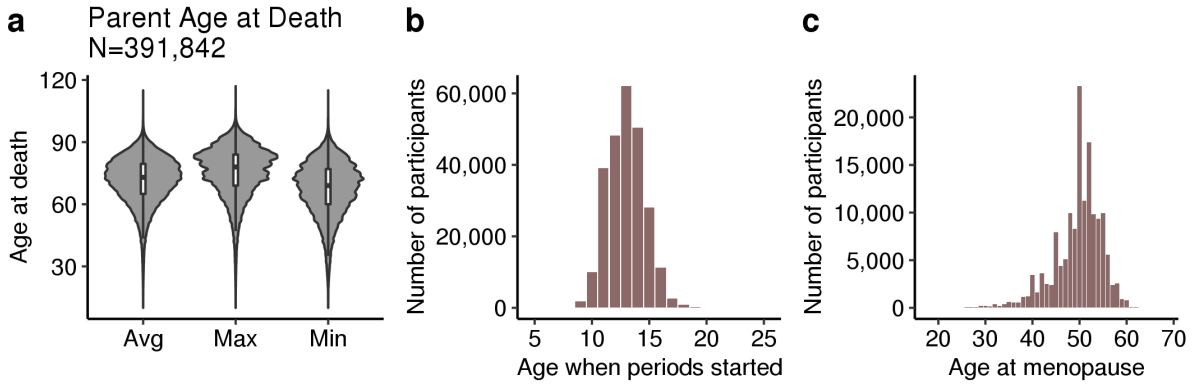


Figure A.30 *Distributions of a) parents' age at death, b) age when periods started (menarche), and c) Age at menopause (last menstrual period).*

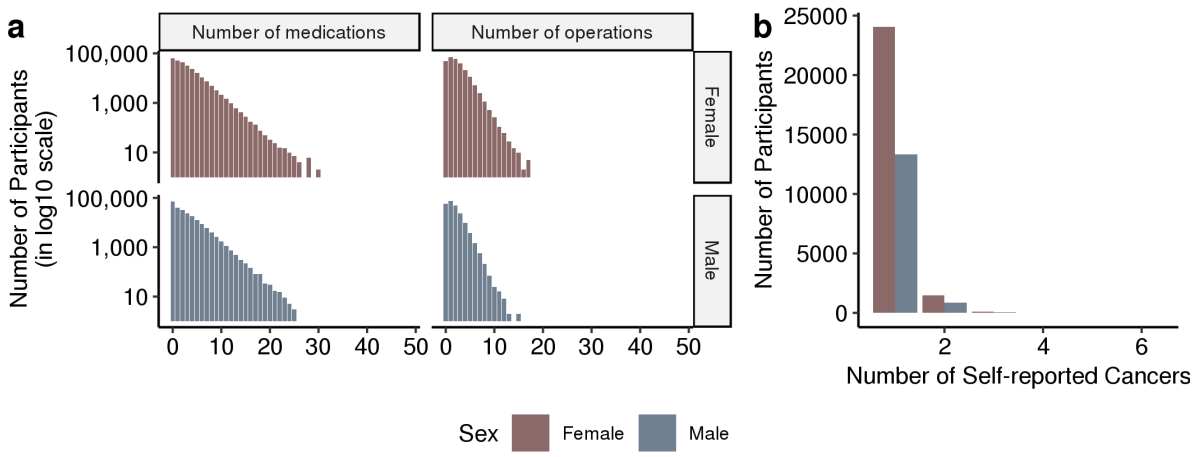


Figure A.31 *Self-reported health data. a) The number of self-reported medications and operations (x-axes) for the participants in the UK Biobank. The y-axis shows the number of participants in log10 scale. b) The number of self-reported cancers (x-axis). Y-axis shows the number of participants.*

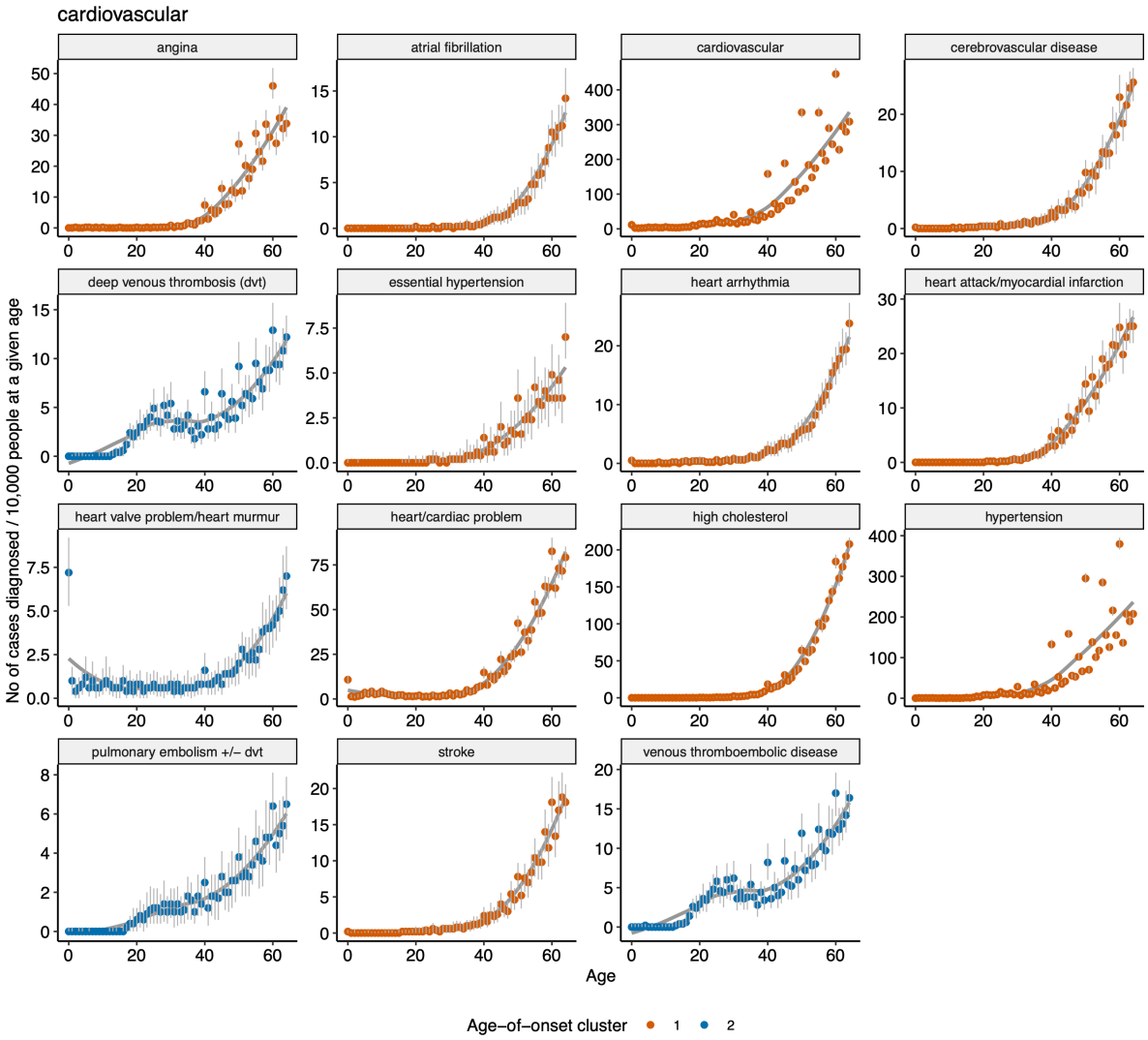


Figure A.32 *Age-of-onset distributions for the cardiovascular diseases. The y-axis shows how many people in 10,000 are diagnosed with that disease at a certain age (x-axis). The plots are also normalized by the number of people that are older than that age so that it is unaffected by the distribution of ages in the UKBB. We run permutations to define confidence intervals for the disease onset rates. We thus randomly select 50,000 participants who passed that particular age for 100 times and calculate the median (points, coloured by the age-of-onset clusters) and 95% range of the all points (grey error lines) A best-fit curve (calculated using loess regression between the medians and age-of-onset) is also displayed.*

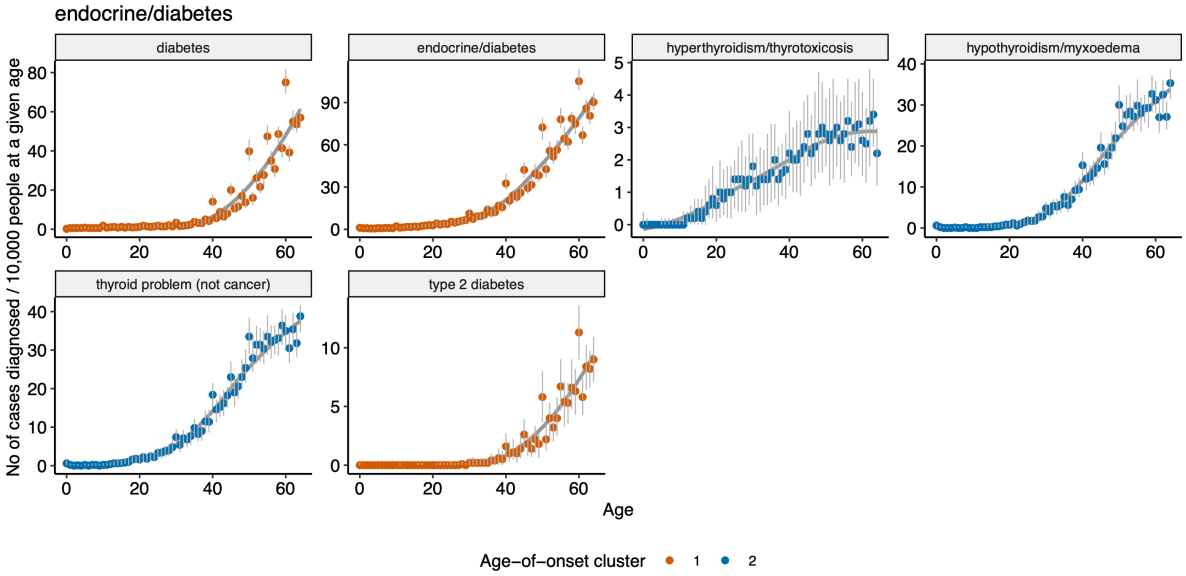


Figure A.33 *Same as the previous figure, but for endocrine / diabetes diseases*

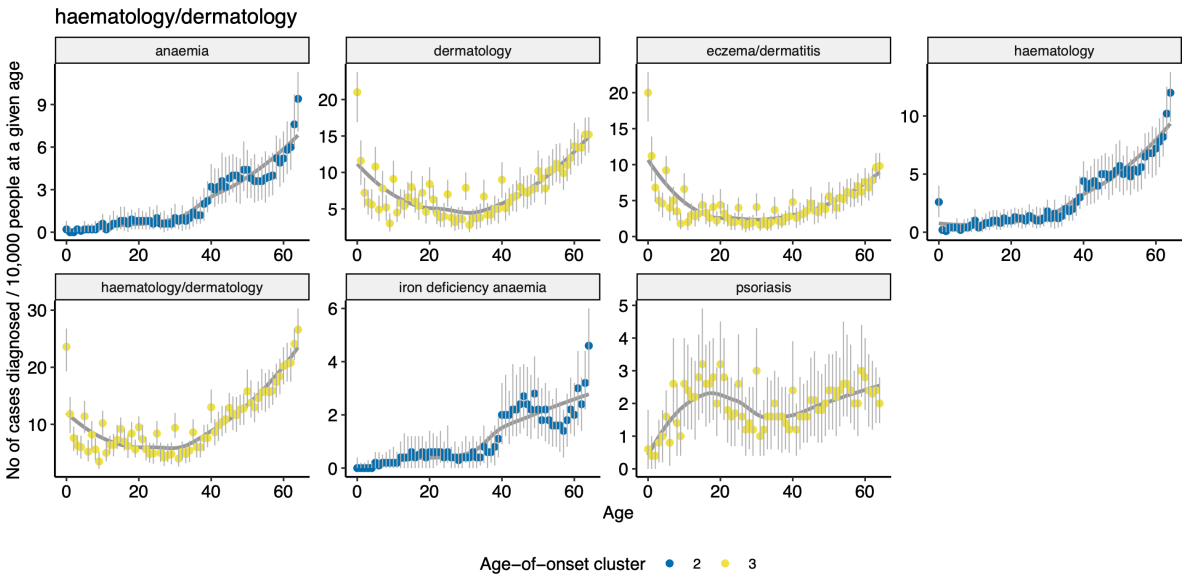


Figure A.34 *Same as the previous figure, but for haematology / dermatology diseases.*

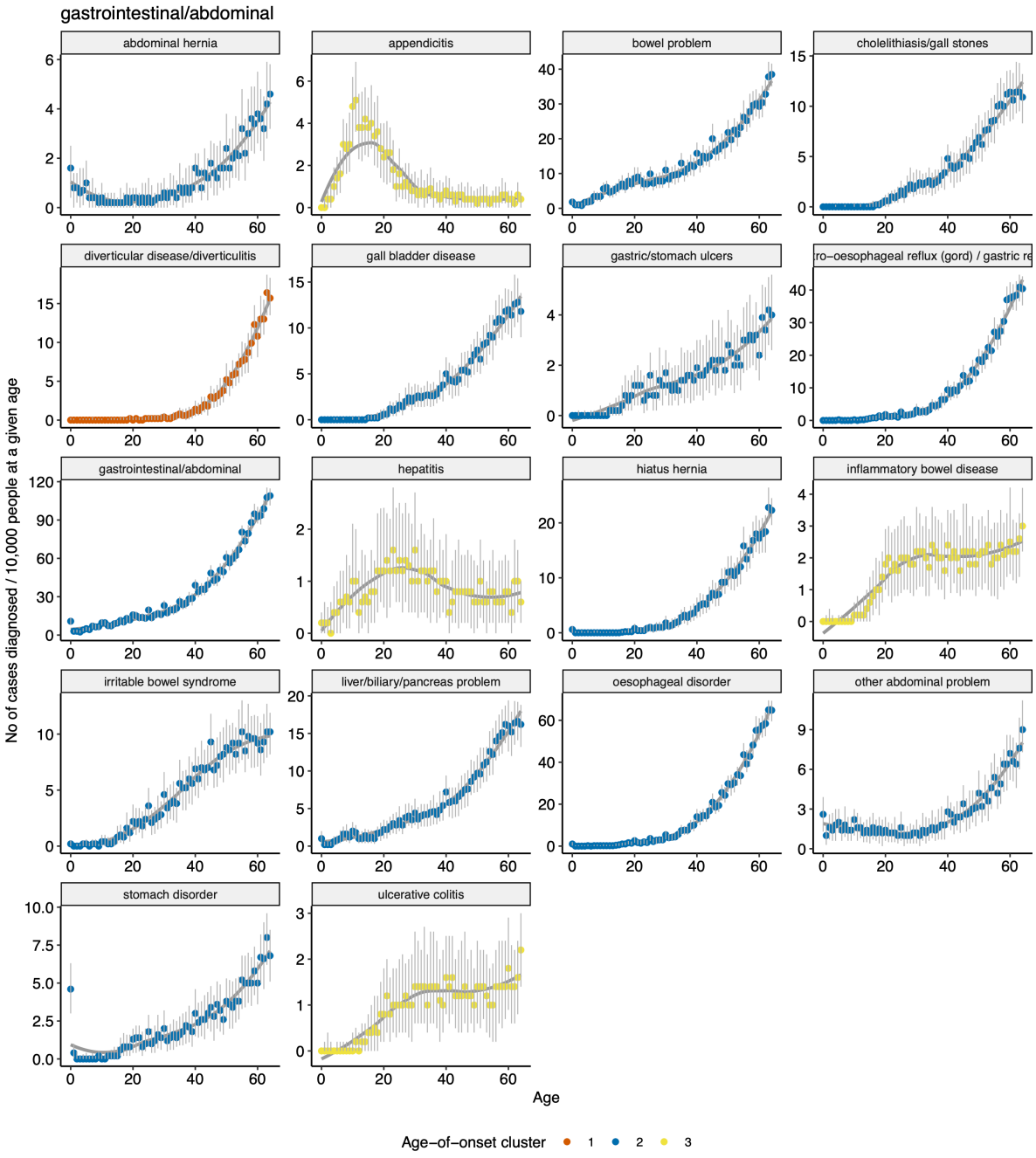


Figure A.35 *Same as the previous figure, but for gastrointestinal / abdominal diseases.*

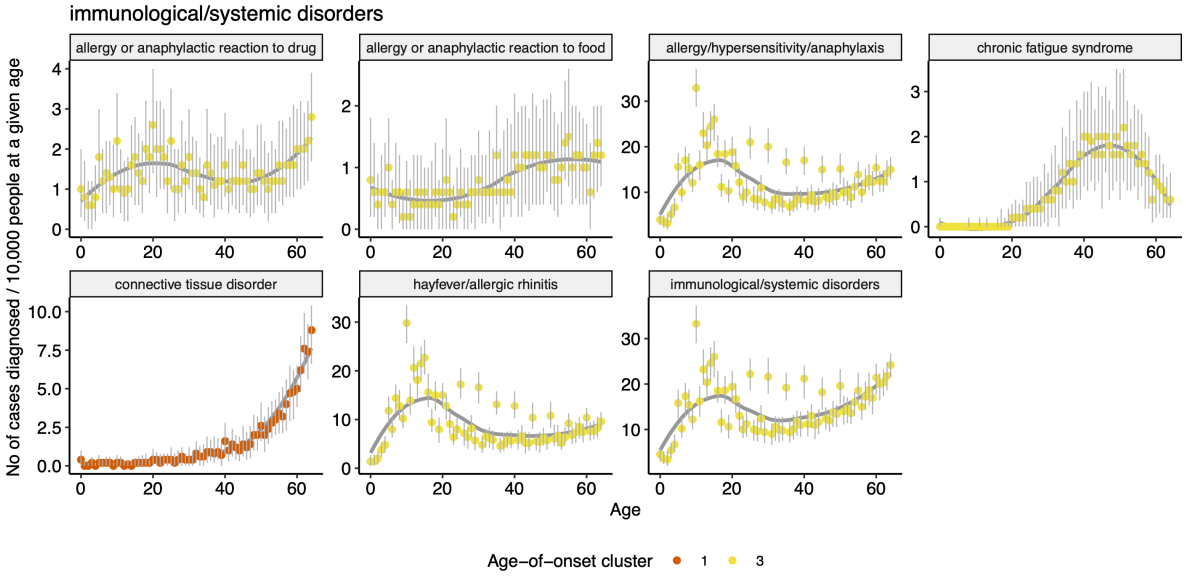


Figure A.36 *Same as the previous figure, but for immunological / systemic disorders.*

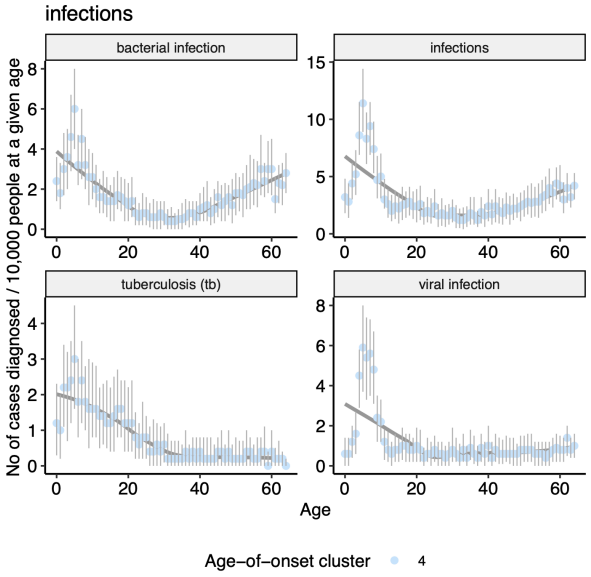


Figure A.37 *Same as the previous figure, but for infections.*

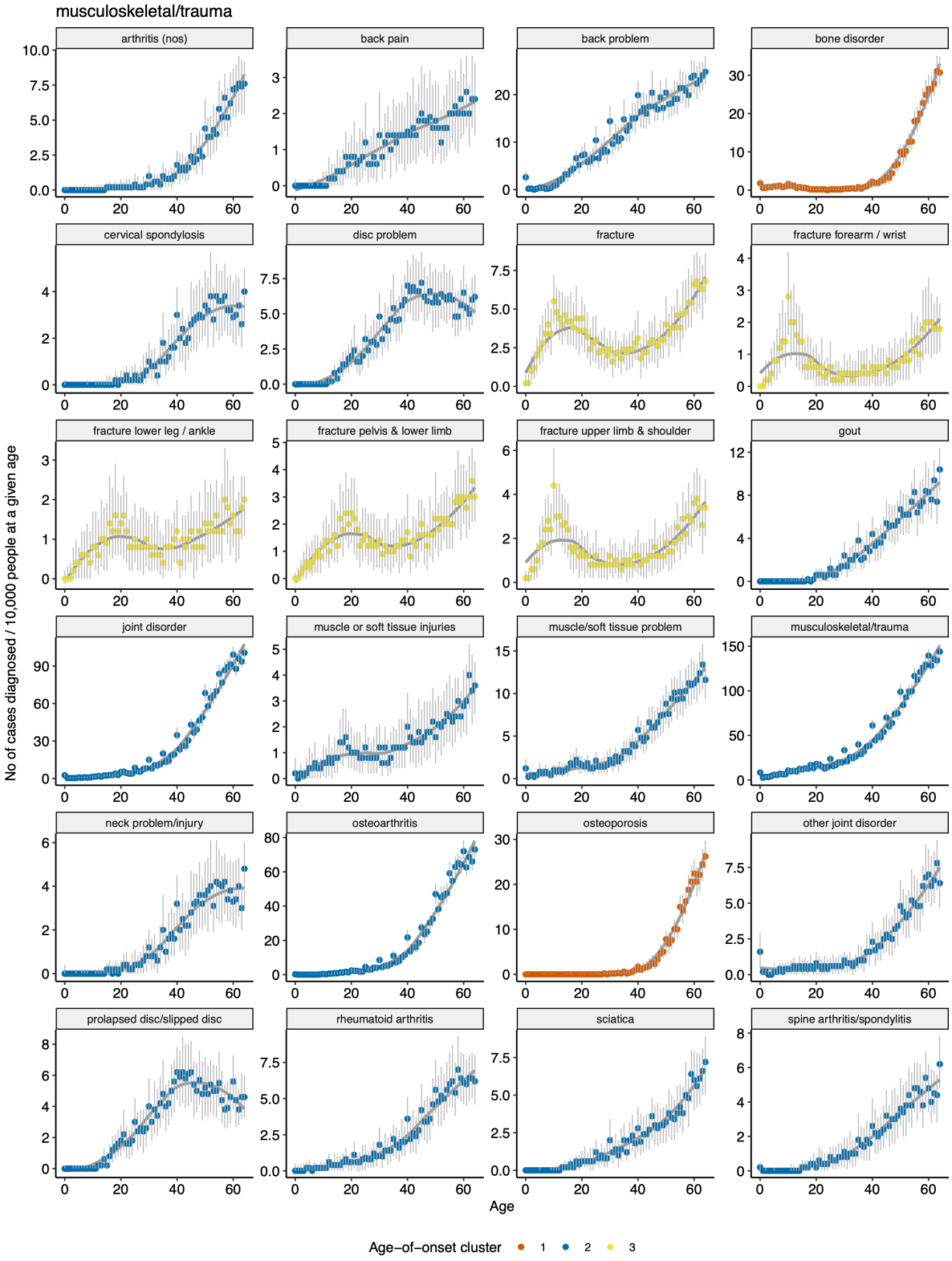


Figure A.38 *Same as the previous figure, but for musculoskeletal / trauma diseases.*

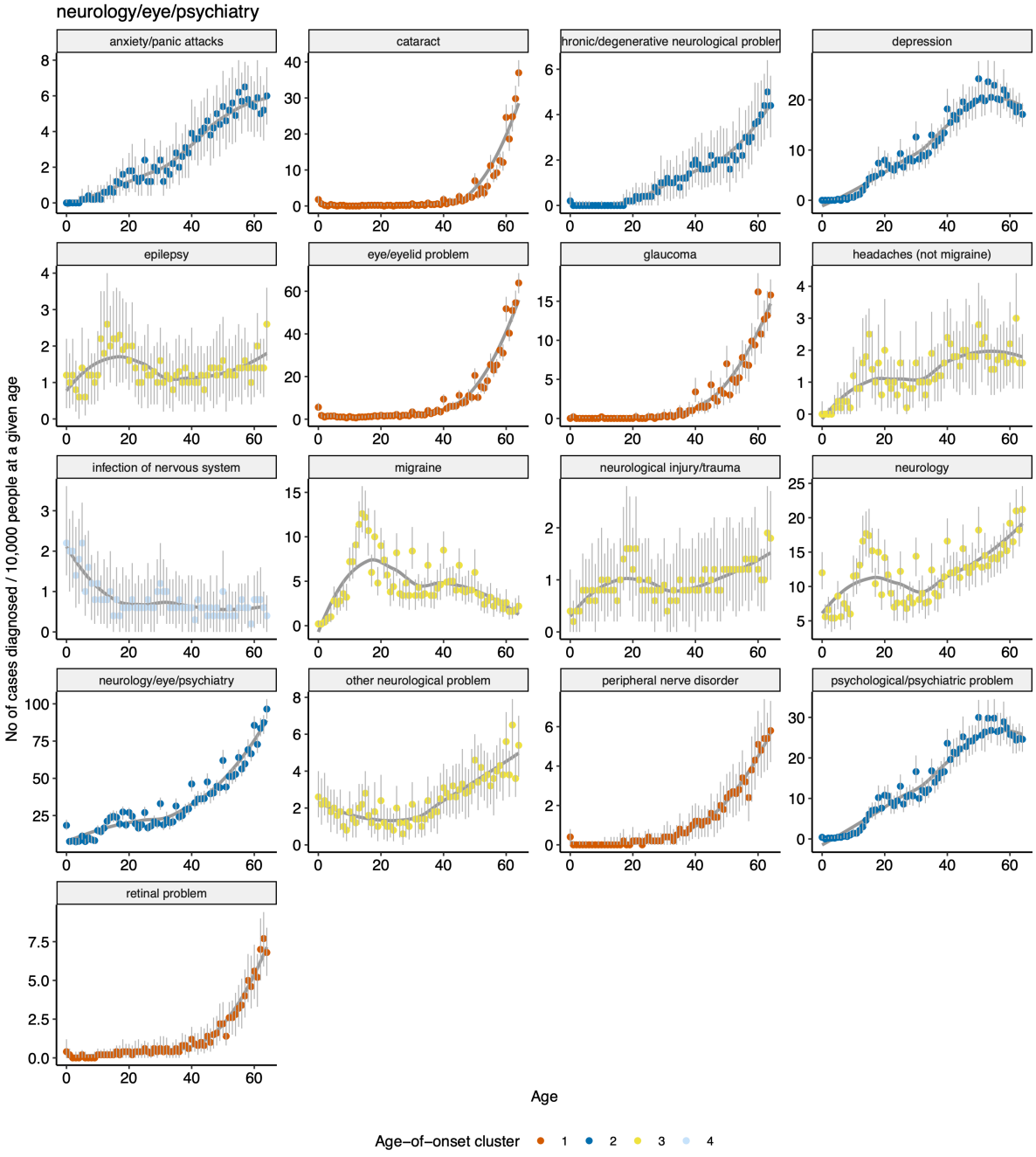


Figure A.39 *Same as the previous figure, but for neurology / eye / psychiatry diseases.*

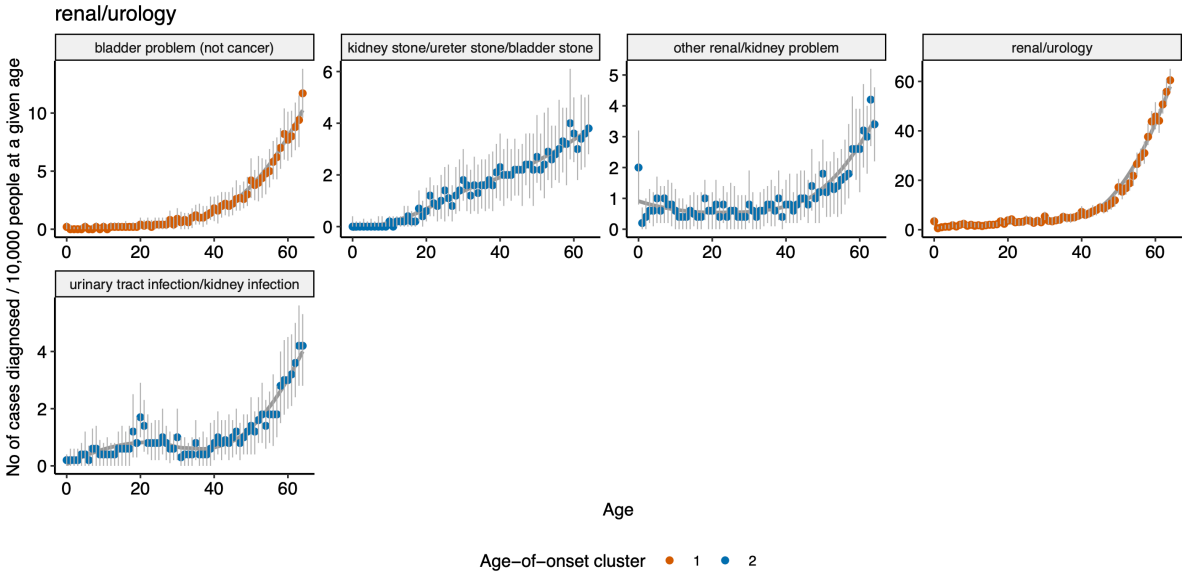


Figure A.40 *Same as the previous figure, but for renal / urology diseases.*

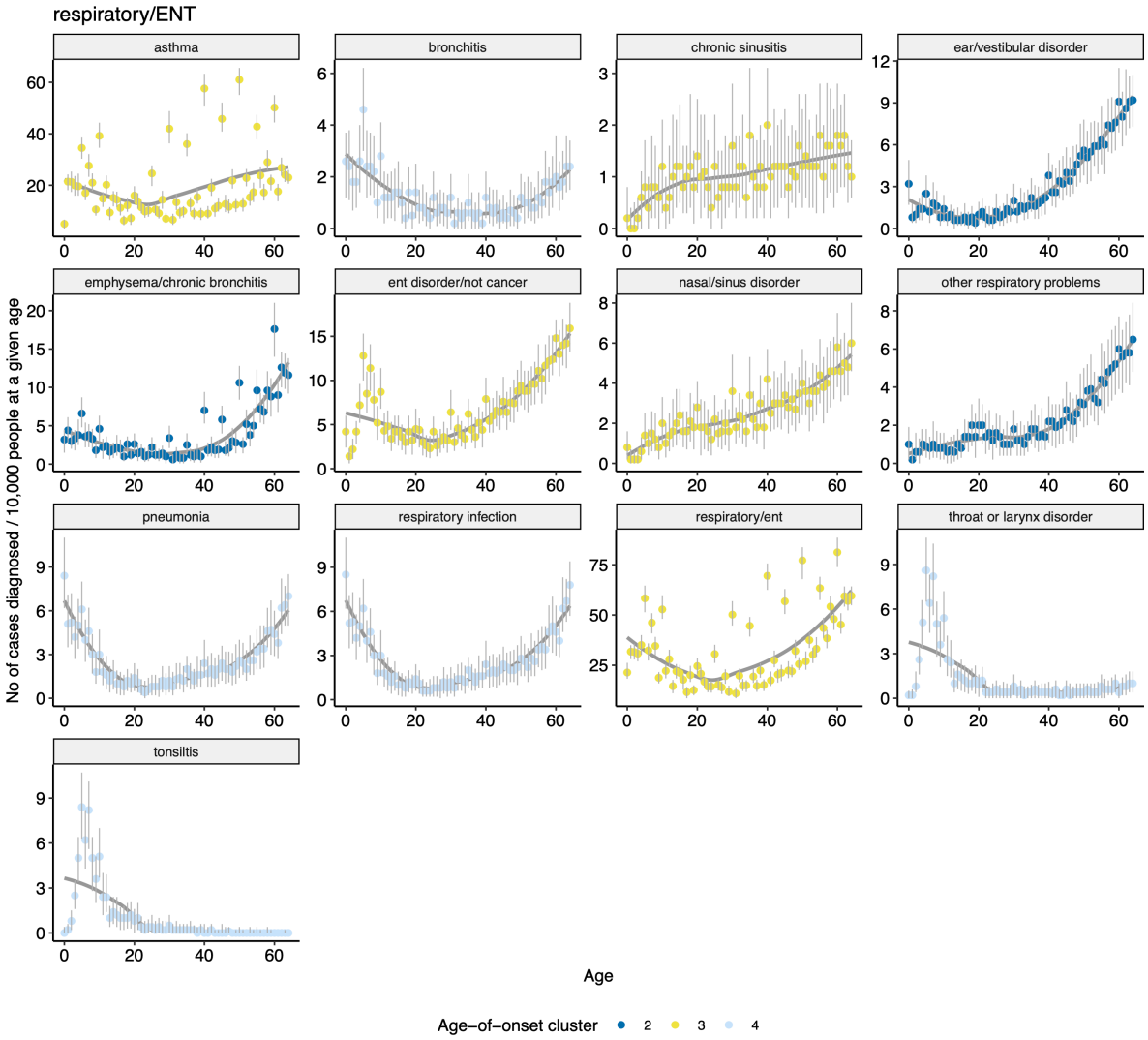


Figure A.41 *Same as the previous figure, but for respiratory / ENT diseases.*

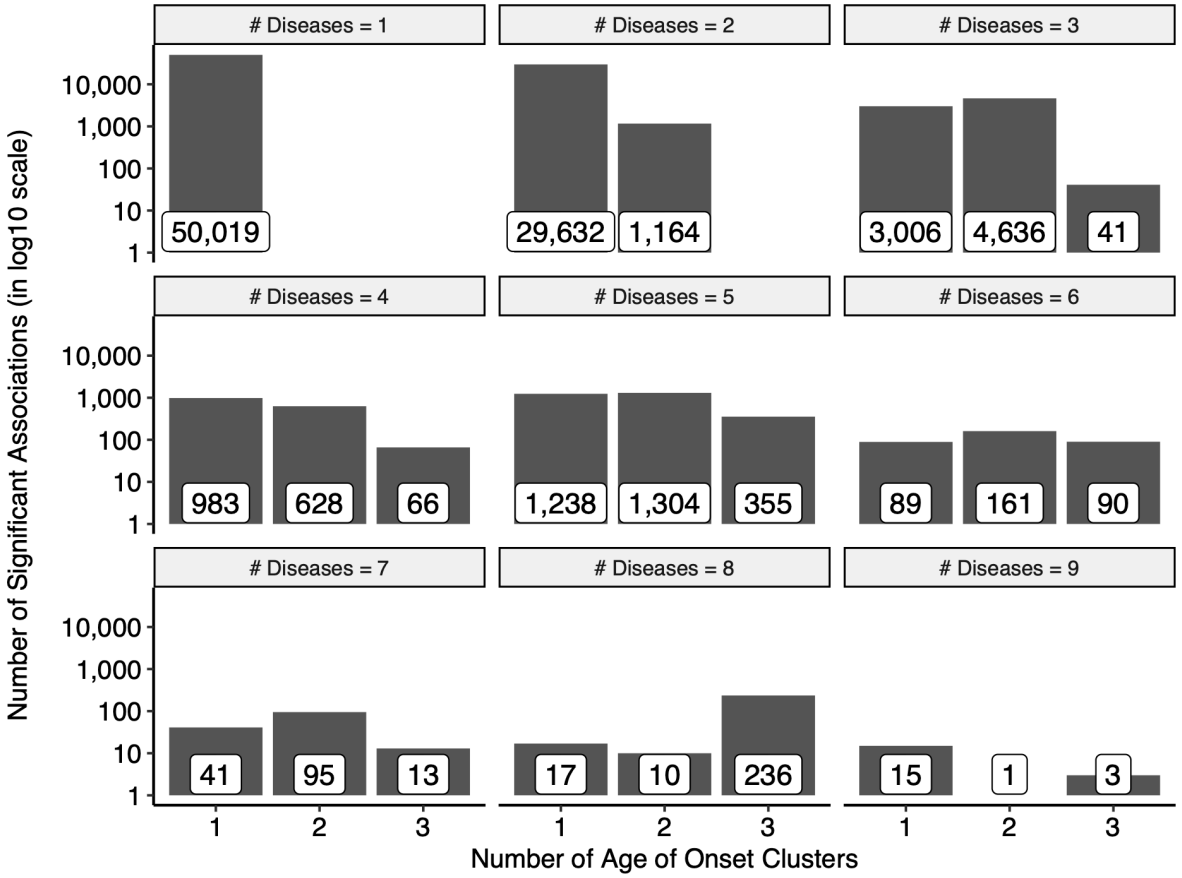


Figure A.42 *Distributions of the number of significant associations (y-axis) according to the number of diseases associated with a given SNP and the number of age-of-onset clusters (x-axis). For example, the upper left plot indicates that 50,019 polymorphisms are significantly associated with one disease in one age-of-onset cluster, while the lower right plot shows that there are 15, 1, and 3 significant SNPs associated with 9 diseases in one, two, or three age-of-onset clusters, respectively.*

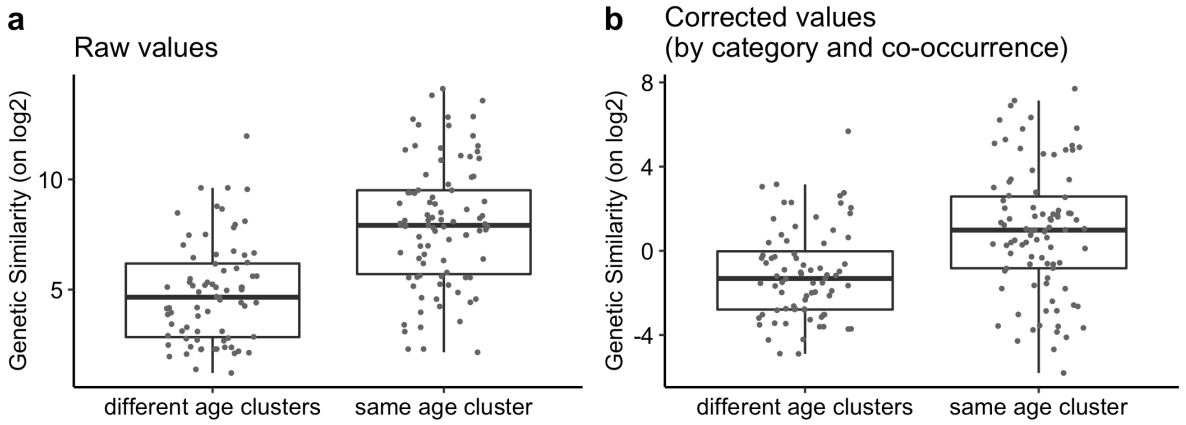


Figure A.43 *a) The difference between genetic similarity within and across age-of-onset clusters. Y-axis shows the genetic similarity on log2 scale (see Supplementary Methods). b) The same as a) but the y-axis is corrected for disease category and co-occurrence using a linear model.*

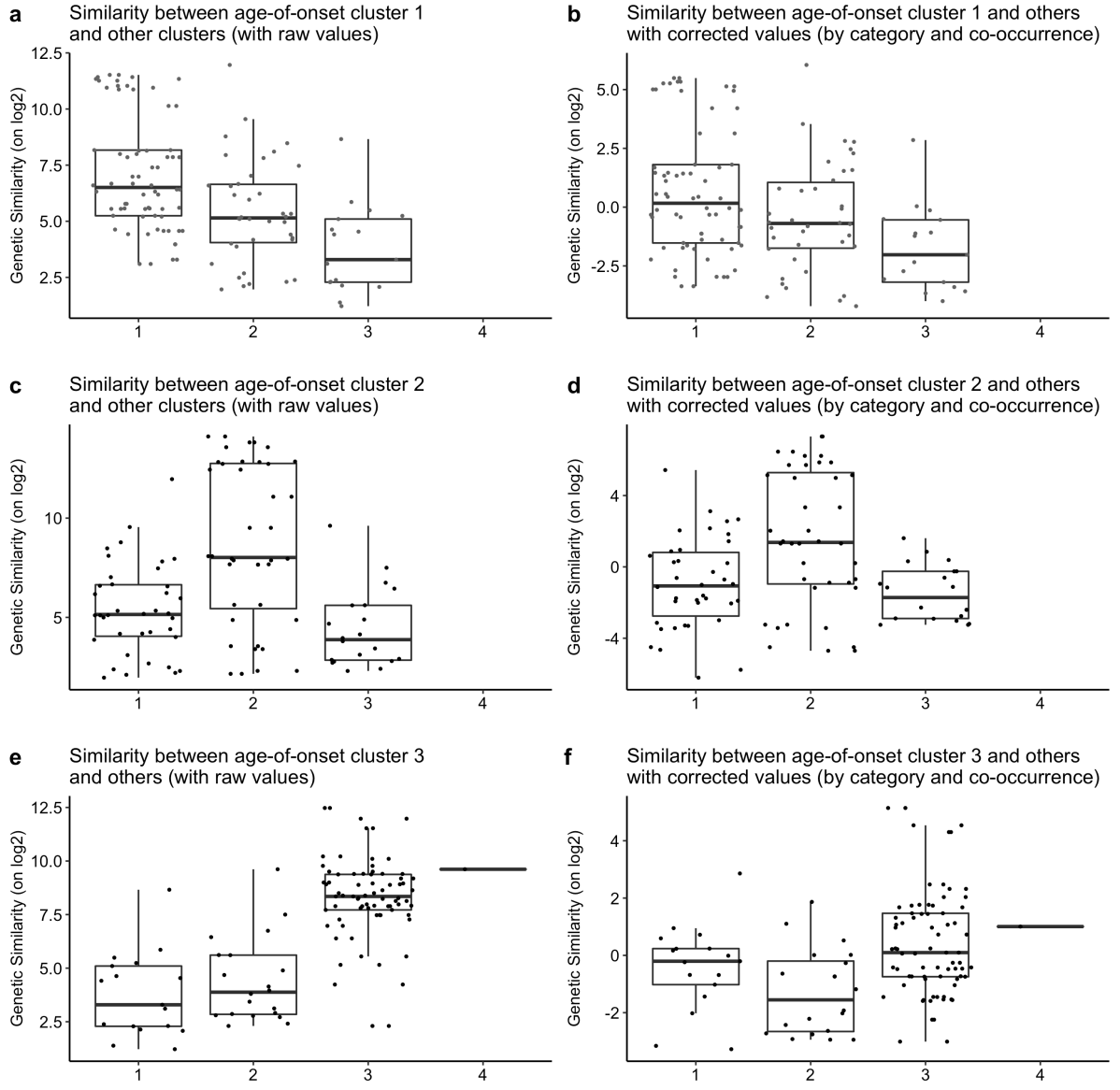


Figure A.44 *Genetic similarities between cluster 1 (a, b), 2 (c, d), 3 (e, f) and other age-of-onset clusters. The y-axis shows the genetic similarity on a log2 scale as the raw values (a, c, e) or as values corrected for disease category and co-occurrence using a linear model (b, d, f).*

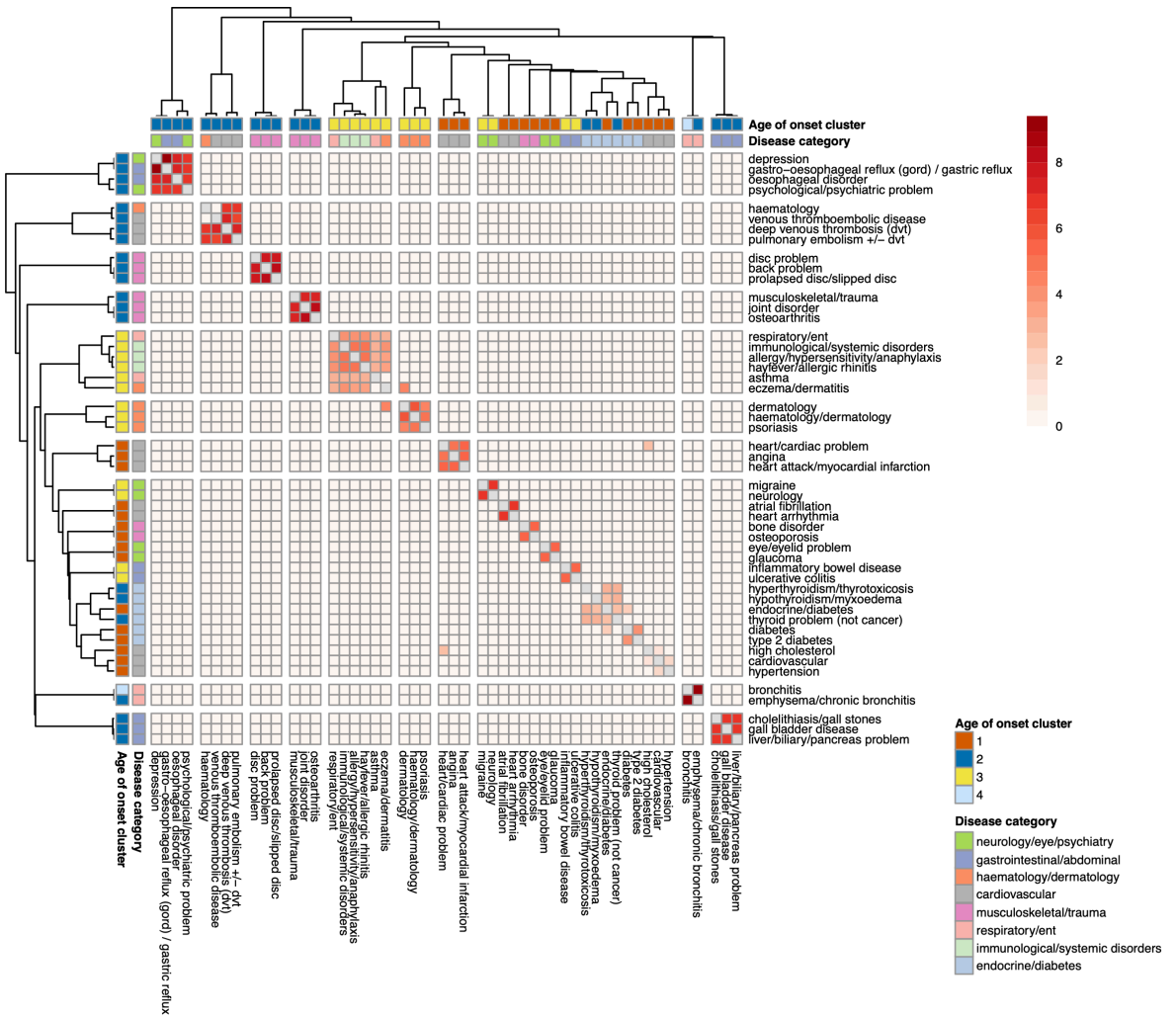


Figure A.45 *Significant genetic similarities (p≤0.01) calculated using independent LD blocks. Diseases (n=50) with at least one significant genetic similarity are displayed. The colour shows the genetic similarity score, darker red means a higher score. Annotation columns show the age-of-onset clusters and disease categories. The diseases are clustered by the hierarchical clustering of genetic similarity scores.*

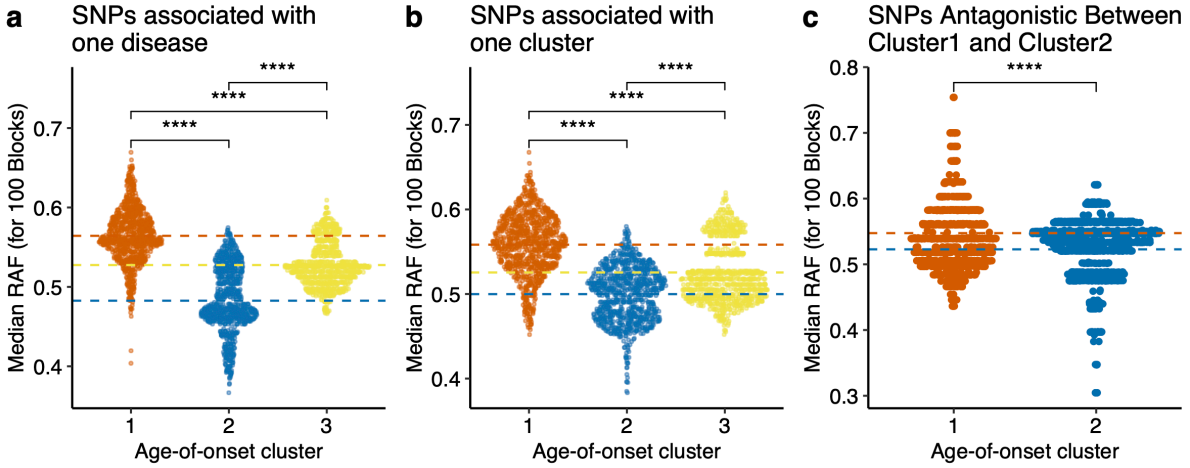


Figure A.46 *The distribution of Median Risk Allele Frequencies (RAF, y-axis) for 100 randomly sampled LD blocks, for 1,000 times, using variants a) associated with one disease, b) associated with one cluster, c) with antagonistic association between cluster 1 and cluster 2. ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001*

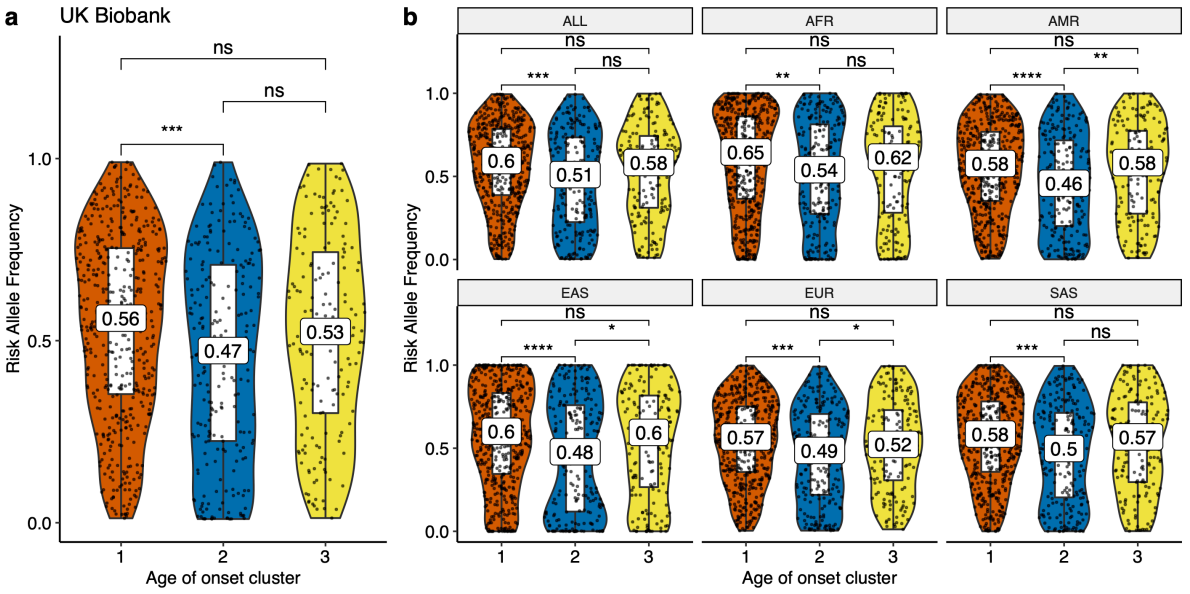


Figure A.47 *a) Risk allele frequency distributions (y-axis) of different age-of-onset clusters (x-axis) in UK Biobank for SNPs associated with one disease. b) The same as panel a but for different 1000 Genomes super-populations (ALL: complete 1000 Genomes cohort, AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian). ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001*

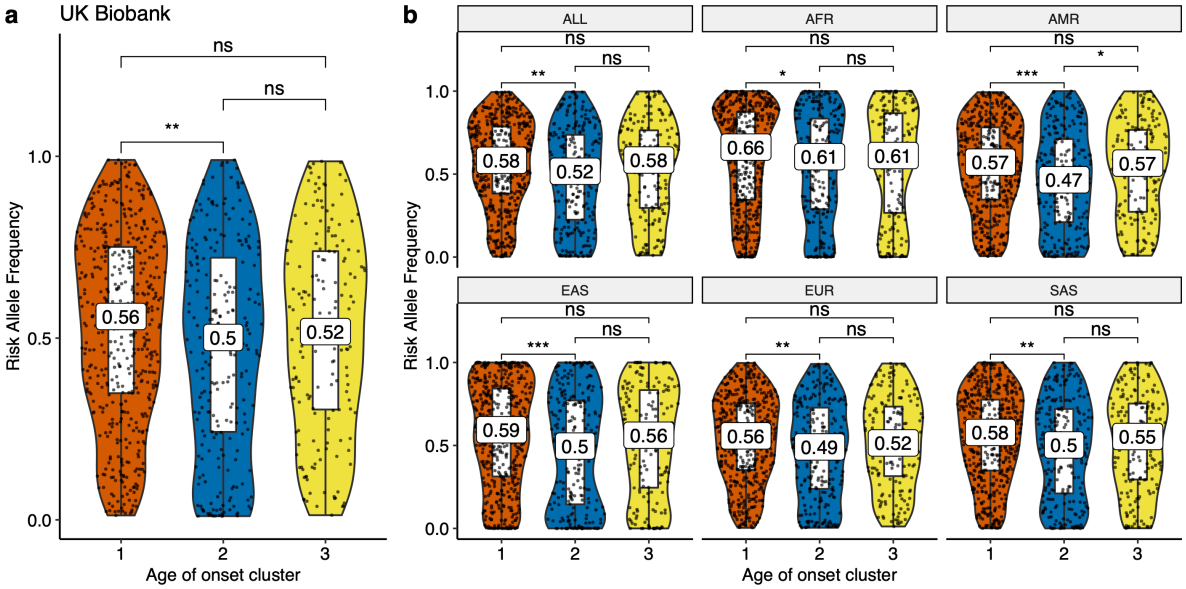


Figure A.48 *a) Risk allele frequency distributions (y-axis) of different age-of-onset clusters (x-axis) in UK Biobank for SNPs associated with one cluster, excluding antagonistic associations. b) The same as a but for different 1000 Genomes super-populations (ALL: complete 1000 Genomes cohort, AFR: African, AMR: Ad Mixed American, EAS: East Asian, EUR: European, SAS: South Asian). ns: p>0.05, *:p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.00001*

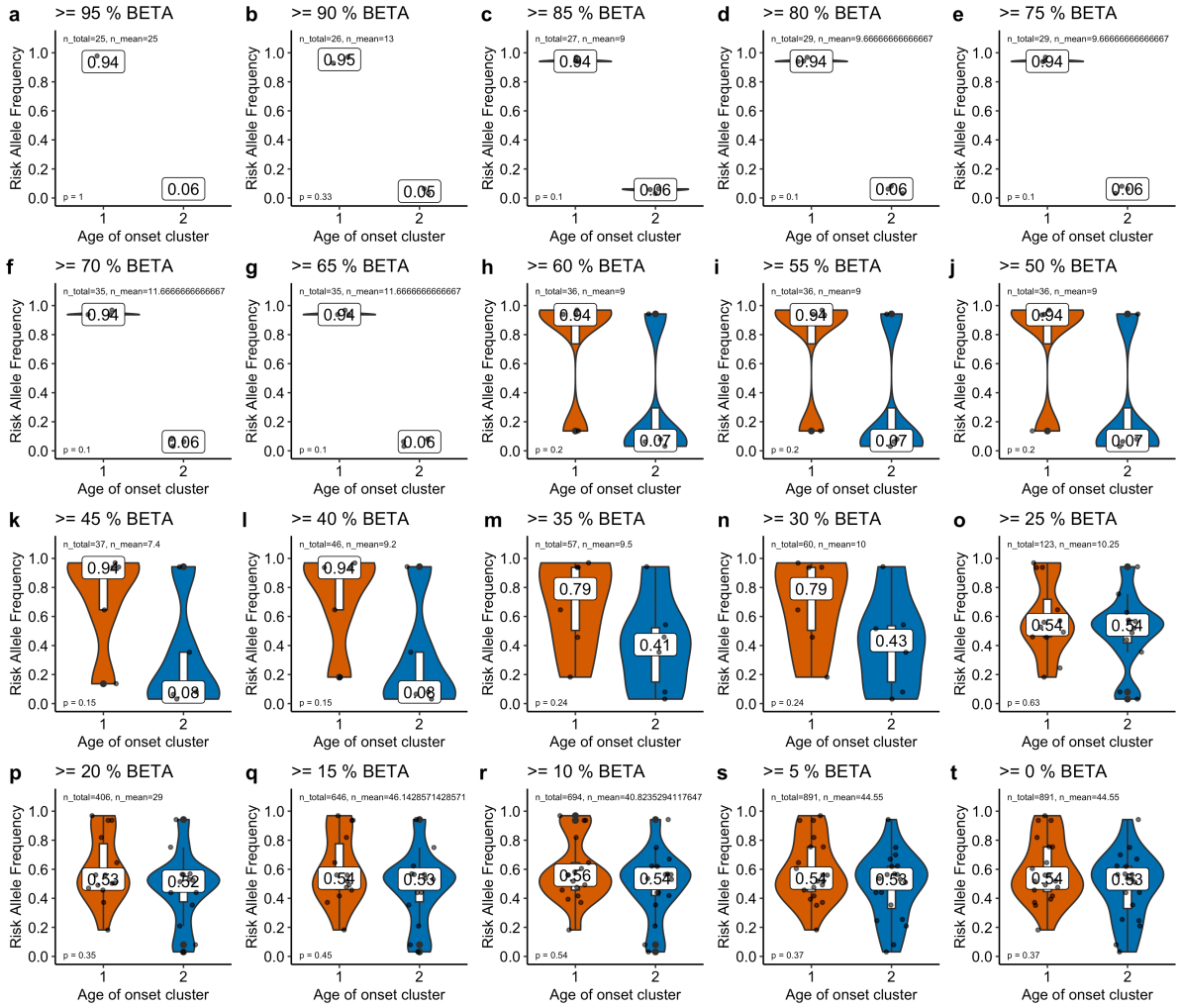


Figure A.49 *Risk allele frequencies in UK Biobank for the loci showing antagonistic associations between cluster 1 and cluster 2 filtered by different effect size cutoffs. The title of each plot shows the cutoff, where e.g. >=95 % BETA means only the SNPs with a BETA (effect size) value higher than 95% of all other antagonistic SNPs are used. >=0 % BETA means no filtering.*

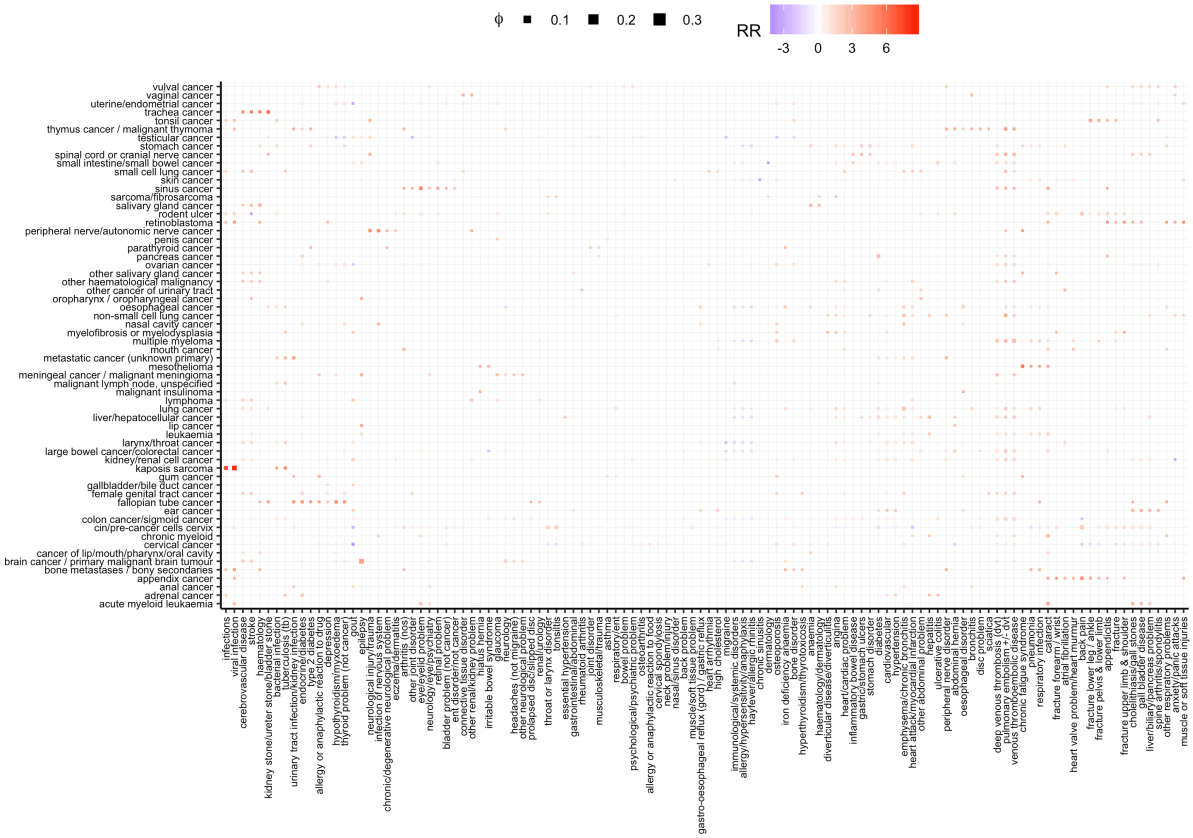


Figure A.50 *Cancer - disease co-occurrence matrix summarizing relative risk scores and correlations. Each row shows a cancer type and column shows a disease. The colour is defined by relative risk scores while the size is determined by ϕ value (the scale is the same as the one used for main figure depicting disease associations between the studied diseases for a better comparison), indicating the robustness. Associations for the 114 diseases and 62 cancers that have at least one relative risk ratio higher than four or lower than minus four ($|log_2 RR| \leq 2$) are plotted.*

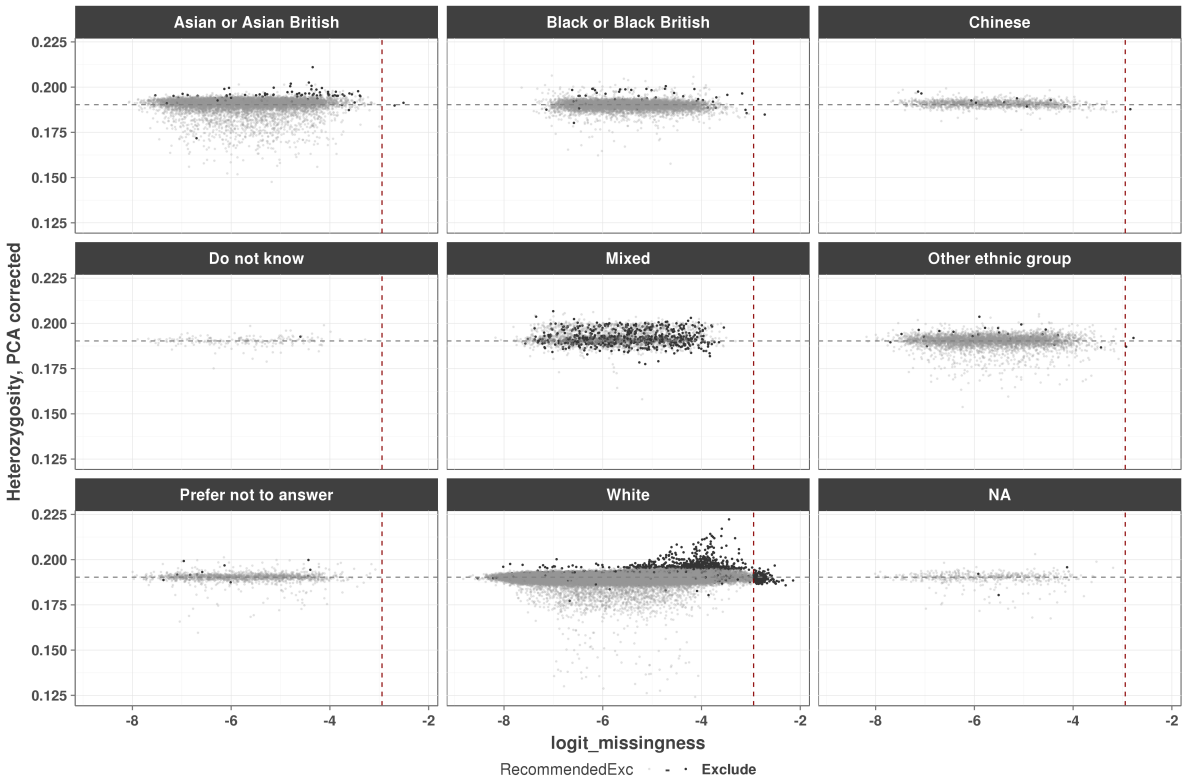


Figure A.51 *Scatter plot between logit(missingness) and PCA corrected heterozygosity measures. Each panel shows a self-declared ethnic background. Vertical red lines show the missing rate of 0.05, and horizontal grey lines show the average heterozygosity in UK Biobank.*

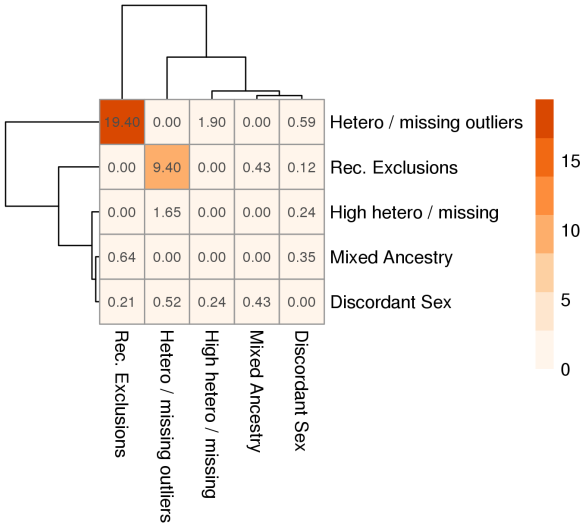


Figure A.52 *A heatmap showing the overlap between exclusions based on different criteria. Values show the percent of the column in the row, e.g. 19.4% of 'Rec. Exclusions' are in 'Hetero / missing outliers' i) 'Hetero / missing outliers': '22027-0.0' (Outliers for heterozygosity or missing rate), ii) 'Rec. Exclusions': field '22010-0.0' (Recommended genomic analysis exclusions), iii) 'High hetero / missing': '22018-0.0', High heterozygosity rate (after correcting for ancestry) or high missing rate, iv) 'Mixed Ancestry': '22018-0.0', Participant self-declared as having a mixed ancestral background, and v) 'Discordant Sex': as described in the sample QC methods.*

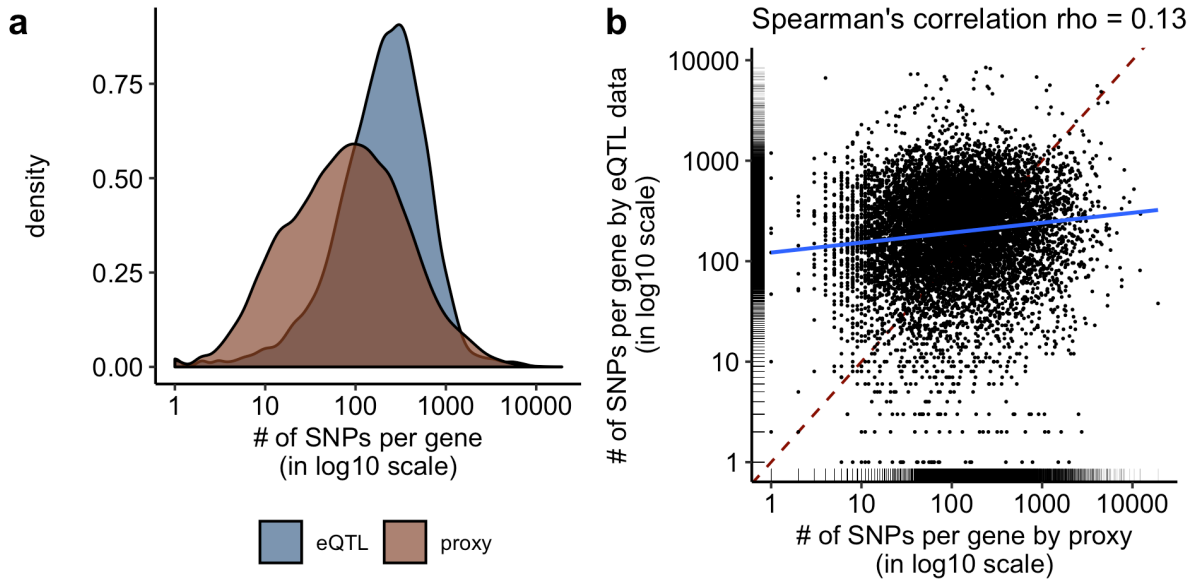


Figure A.53 *(a) Density plots showing the number of SNPs per gene, based on eQTL data (blue) and proximity (brown). (b) Scatter plot between the number of SNPs per gene mapped using genomic proximity (x-axis) or eQTL data (y-axis). Each dot represents a gene and the blue line shows the linear model. Dashed red line shows one-to-one relationship. The rug-plots on the axes show the marginal distribution of genes.*

A.3 Drug repurposing for ageing

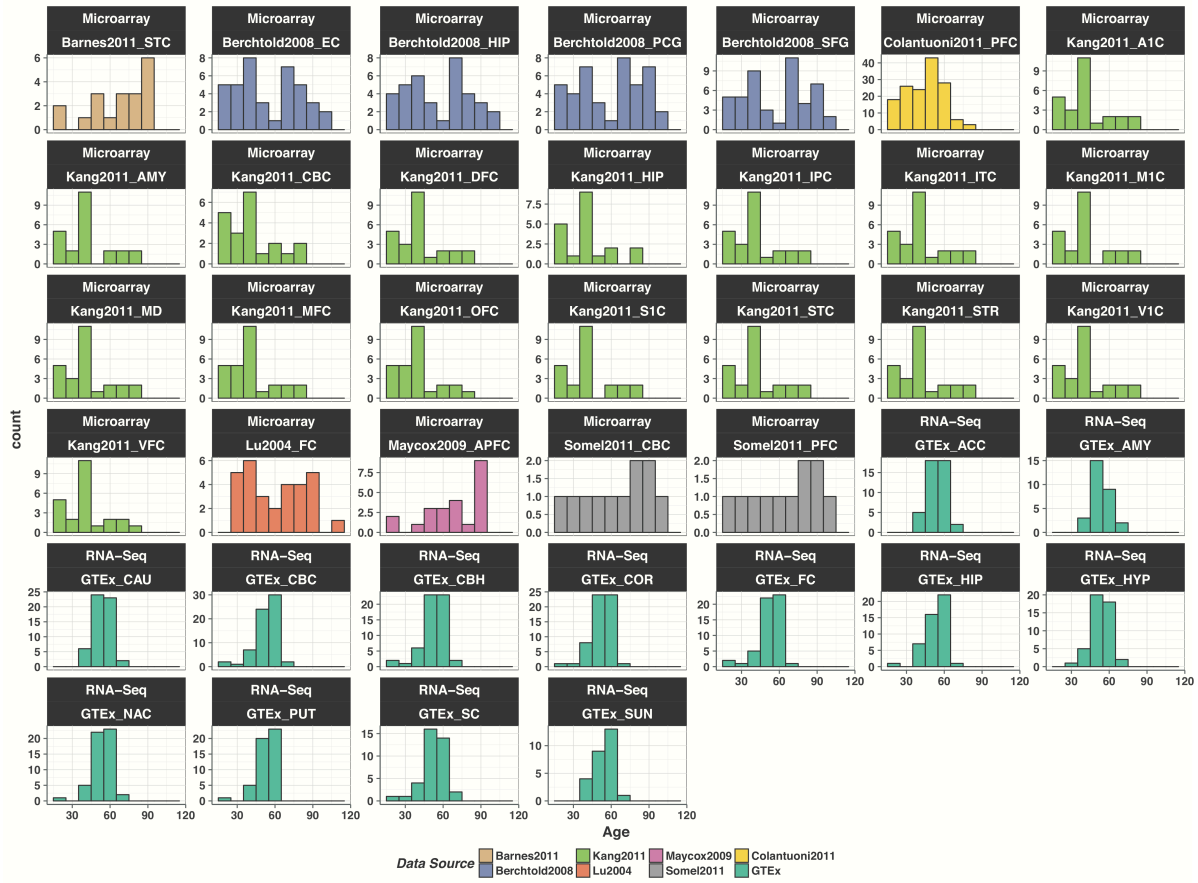


Figure A.54 *a) Age distribution of the sub-datasets. b) PCA plots for the sub-datasets (after outlier removal).*

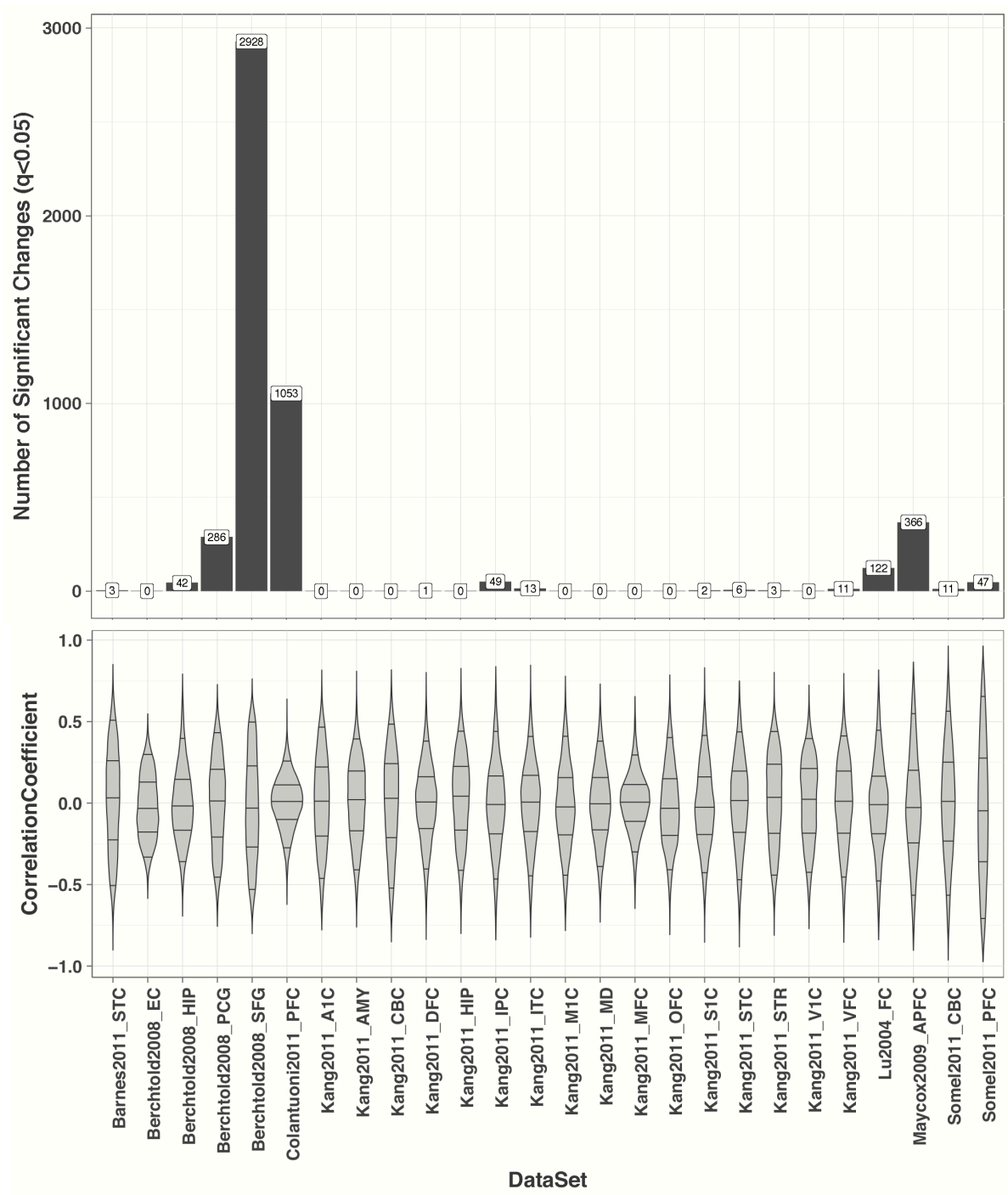


Figure A.55 *The number of significant changes and the distributions of Spearman's correlation coefficient between gene expression and age for each dataset.*

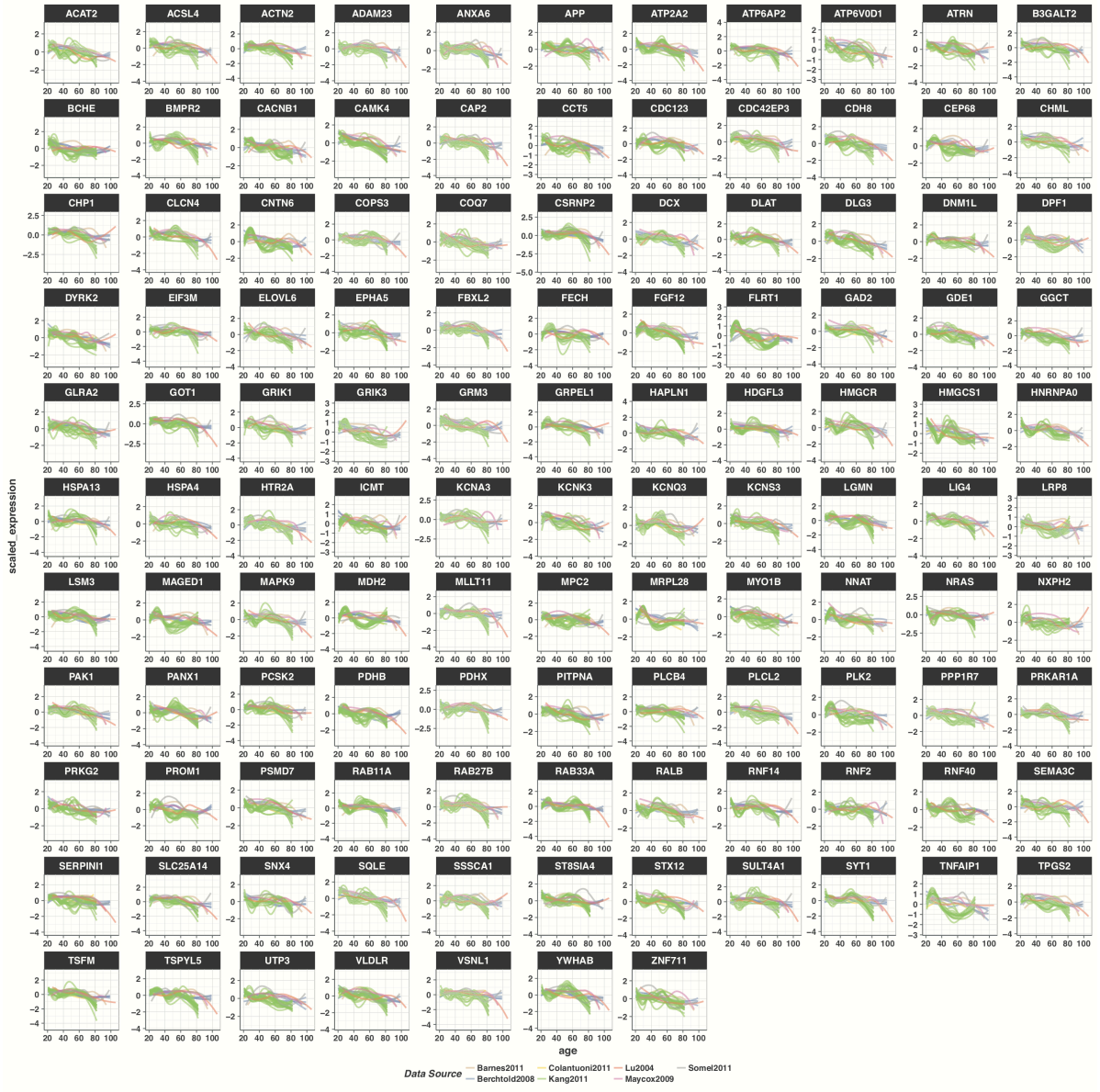


Figure A.56 *Scaled gene expression profile for the down-regulated genes in the microarray ageing profile. Each line shows one dataset, and the colours represent different data sources. The lines are generated using geom_smooth function with method = loess argument in the R ggplot2 package.*

Figure A.57 *Scaled gene expression profile for the up-regulated genes in the microarray ageing profile. Each line shows one dataset, and the colours represent different data sources. The lines are generated using geom_smooth function with method = loess argument in the R ggplot2 package.*

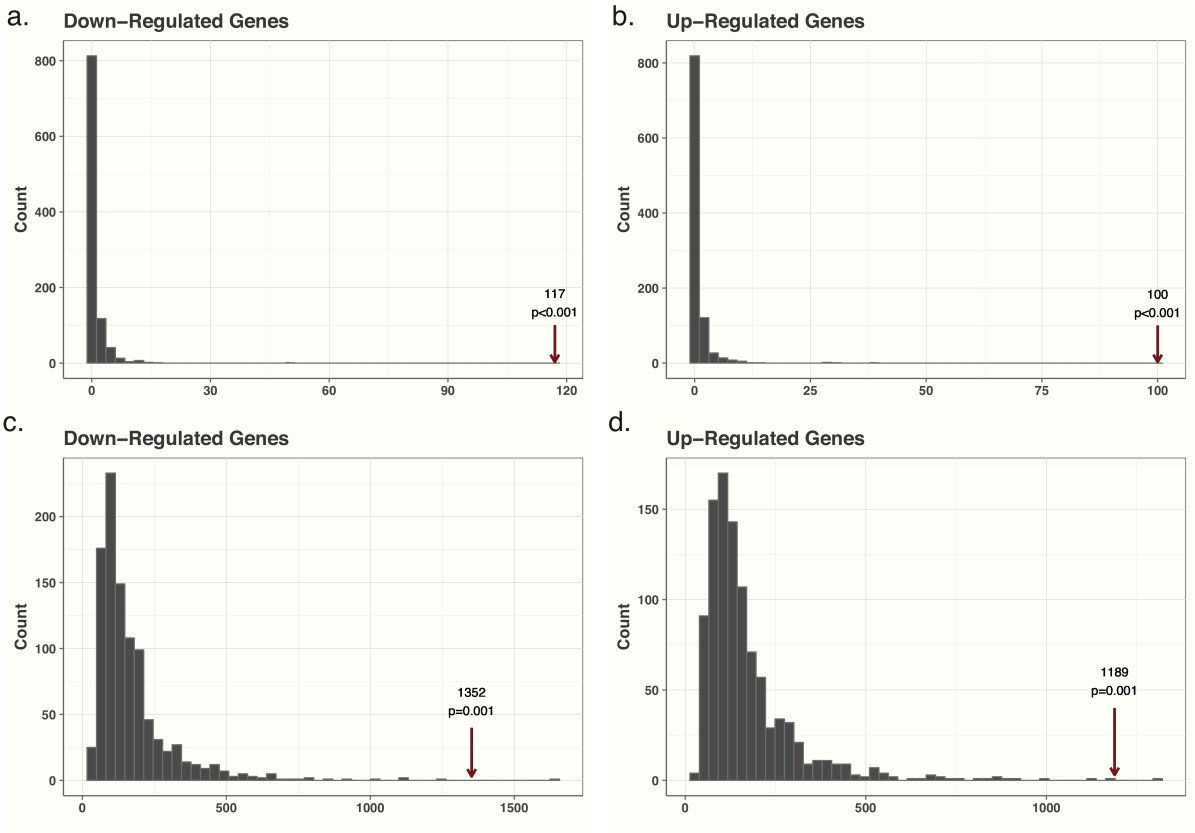


Figure A.58 *Distribution of the number of shared expression changes across datasets in 1000 permutations. a) Shared down-regulation across microarray datasets (expected number=0), b) shared up-regulation across microarray datasets (expected number=0), c) shared down-regulation across GTEx datasets (expected number=127), and d) shared up-regulation across GTEx datasets (expected number=131.5).*

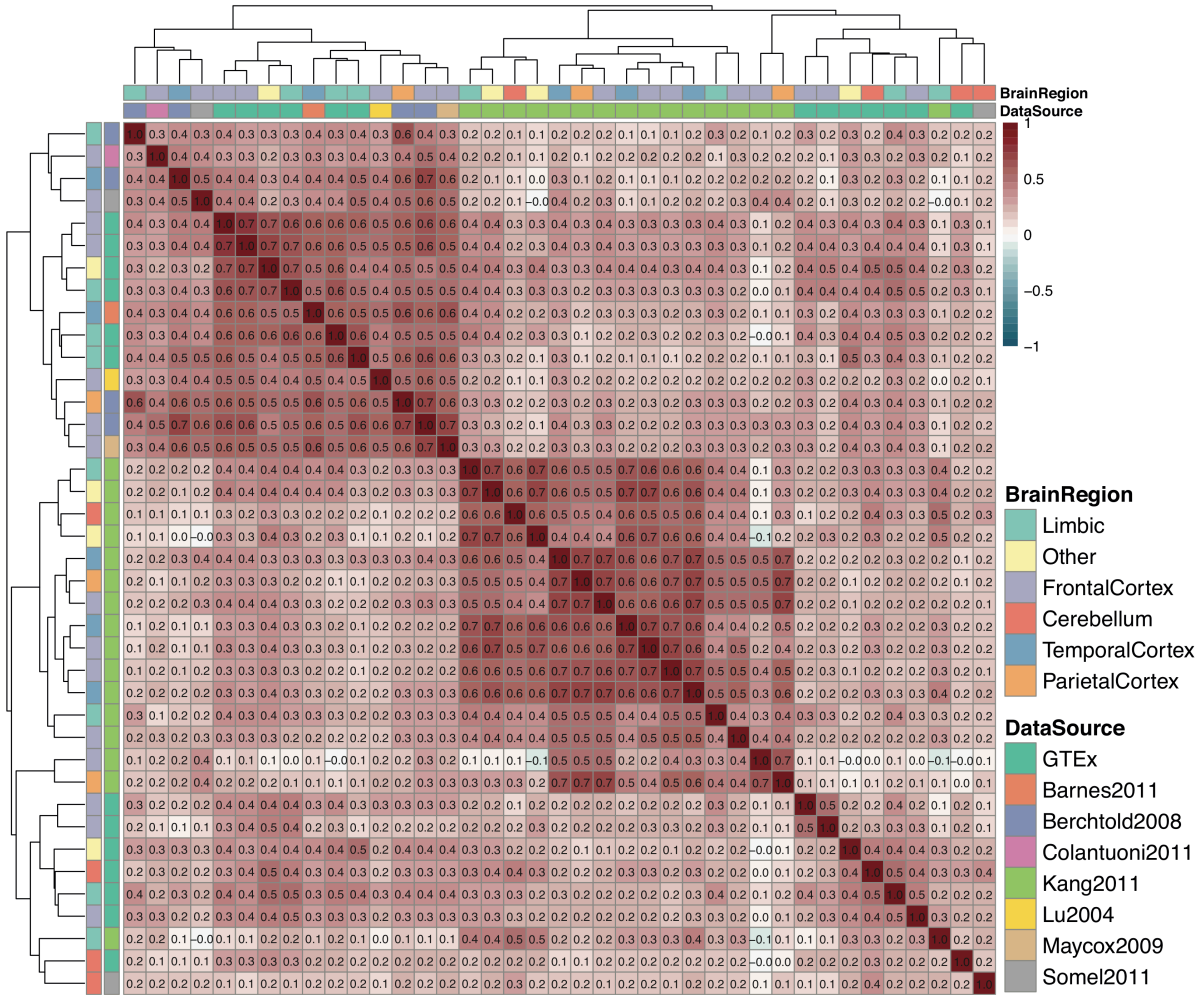


Figure A.59 *Pairwise Spearman's rank correlation coefficients across all datasets, including GTEx. The intensity of the colours on the heatmap shows the magnitude of the correlation coefficient*

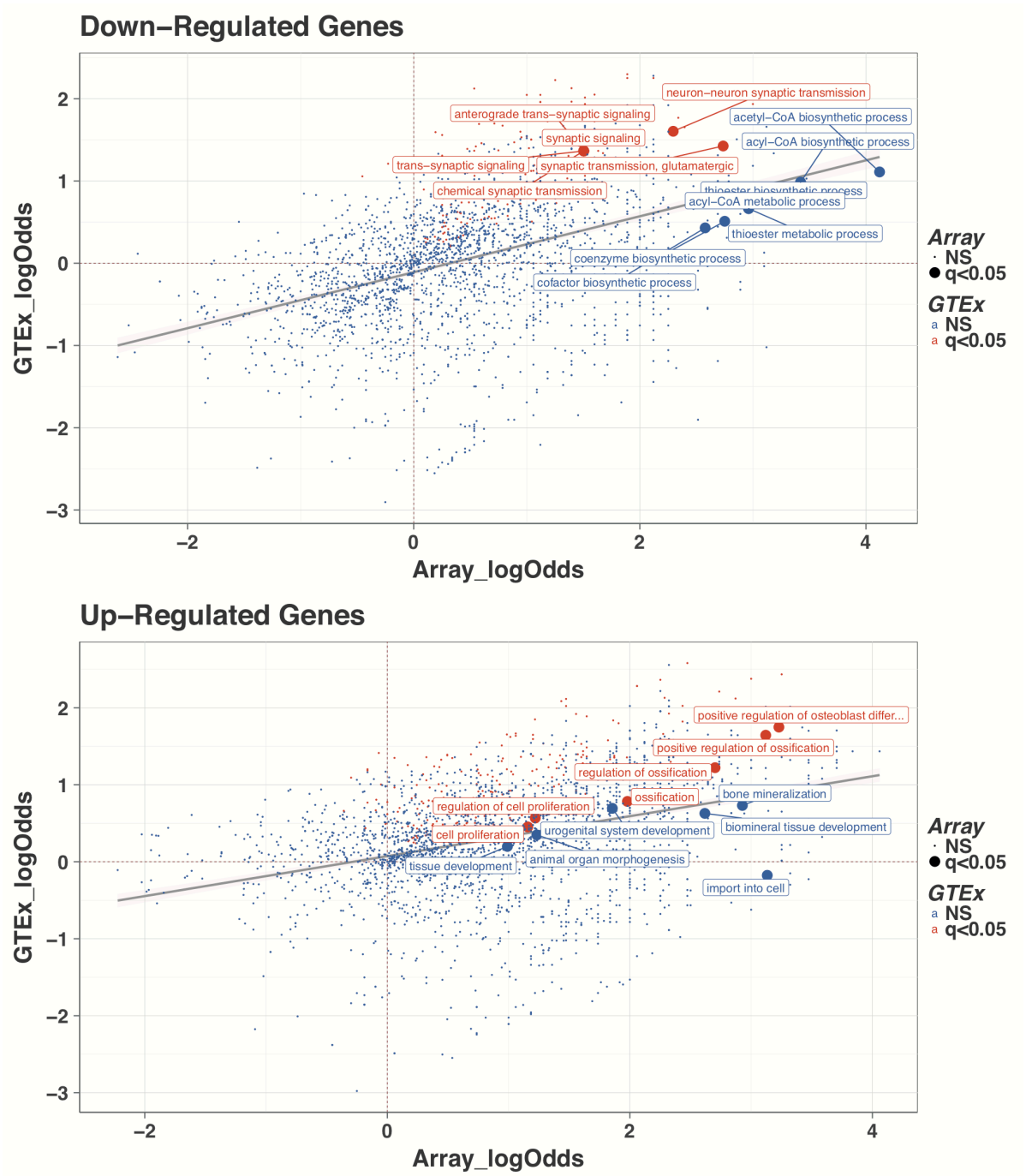


Figure A.60 *Scatter plot for the GO BP category log2 odds ratios calculated for the microarray and GTEx ageing signature, using Fisher's test implemented in the topGO package.*

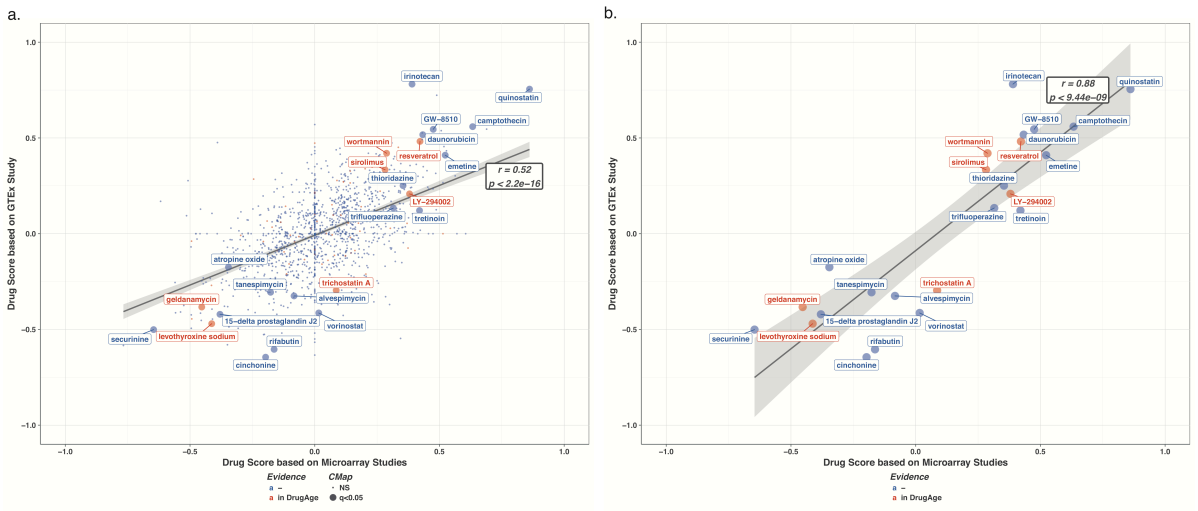
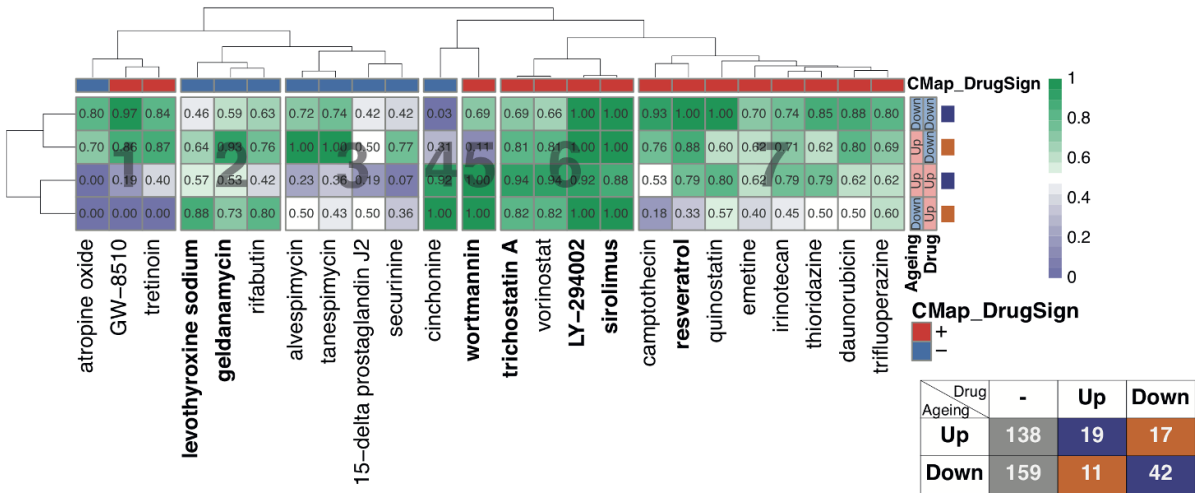


Figure A.61 *Scatter plot of the drug similarity scores for a) all drugs and b) only significant drugs. x-axes show the similarity score based on the microarray ageing signature, whereas y-axes show similarity score calculated using the GTEx ageing signature. The size of the data points represents the statistical significance whereas the colour shows whether a drug is previously tested on model organisms for lifespan extension (based on DrugAge database).*



Drug / Ageing	-	Up	Down
Up	138	19	17
Down	159	11	42

Figure A.62 *Heatmap showing the percent similarity of each drug to the compiled pro-longevity drug profile. The numbers 1-7 show the cluster number based on the hierarchical clustering of the drugs based on the similarities. The annotation rows show the sign of similarity score based on the CMap analysis. The column labels written in bold indicates the drugs in the DrugAge database. Annotation columns show the up- or down-regulation of each category in ageing and pro-longevity drug profile. The small table shows the number of probe-sets in each category we defined to reflect pro-longevity drug profile.*

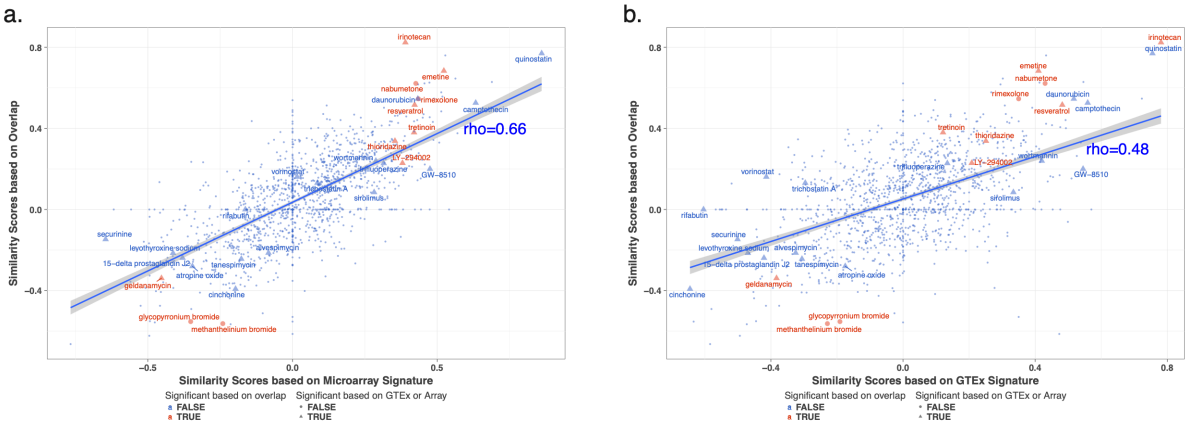
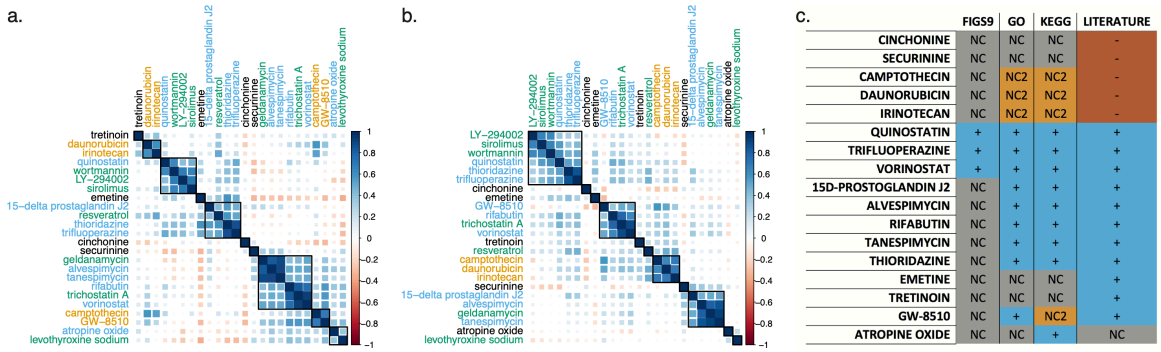


Figure A.63 *Scatter plot of the drug similarity scores based on the overlap of the signatures (50 up and 48 down regulated genes) and a) microarray signature and b) GTEx signature. The size of the data points represents the statistical significance whereas the colour shows whether the drug is significant based on the overlap of the signatures. The shape represents whether the drug is one of the 24 hits reported in the main text.*



	FIGS9	GO	KEGG	LITERATURE
CINCHONINE	NC	NC	NC	-
SECURININE	NC	NC	NC	-
CAMPTOTHECIN	NC	NC2	NC2	-
DAUNORUBICIN	NC	NC2	NC2	-
IRINOTECAN	NC	NC2	NC2	-
QUINOSTATIN	+	+	+	+
TRIFLUOPERAZINE	+	+	+	+
VORINOSTAT	+	+	+	+
15D-PROSTOGLANDIN J2	NC	+	+	+
ALVESPIMYCIN	NC	+	+	+
RIFABUTIN	NC	+	+	+
TANESPIMYCIN	NC	+	+	+
THIORIDAZINE	NC	+	+	+
EMETINE	NC	NC	NC	+
TRETINOIN	NC	NC	NC	+
GW-8510	NC	+	NC2	+
ATROPINE OXIDE	NC	NC	+	NC

Figure A.64 *a-b) Correlation matrices showing drug-drug similarities calculated using pair-wise Spearman's correlation coefficients between normalized enrichment scores for a) KEGG pathways and b) GO Biological Process categories. Drugs are clustered using hierarchical clustering. The rectangles are drawn by cutting the hierarchical tree (the number of clusters is decided based on the visual inspection of the heatmap). Labels written in green are the known prolongevity drugs based on the DrugAge database, in blue are the drugs in the same cluster with at least one pro-longevity drug, in black are the ones which did not cluster with anything and in orange are the ones that did not cluster any pro-longevity drug but are similar to each other. c) Summary matrix showing the outcome of each approach used, KEGG and GO GSEA results from panel (a) and (b) and literature search given at the beginning of this section. The colour code is the same as panel (a) and (b), and dark orange shows likely negative effects, thus pro-ageing drugs.*

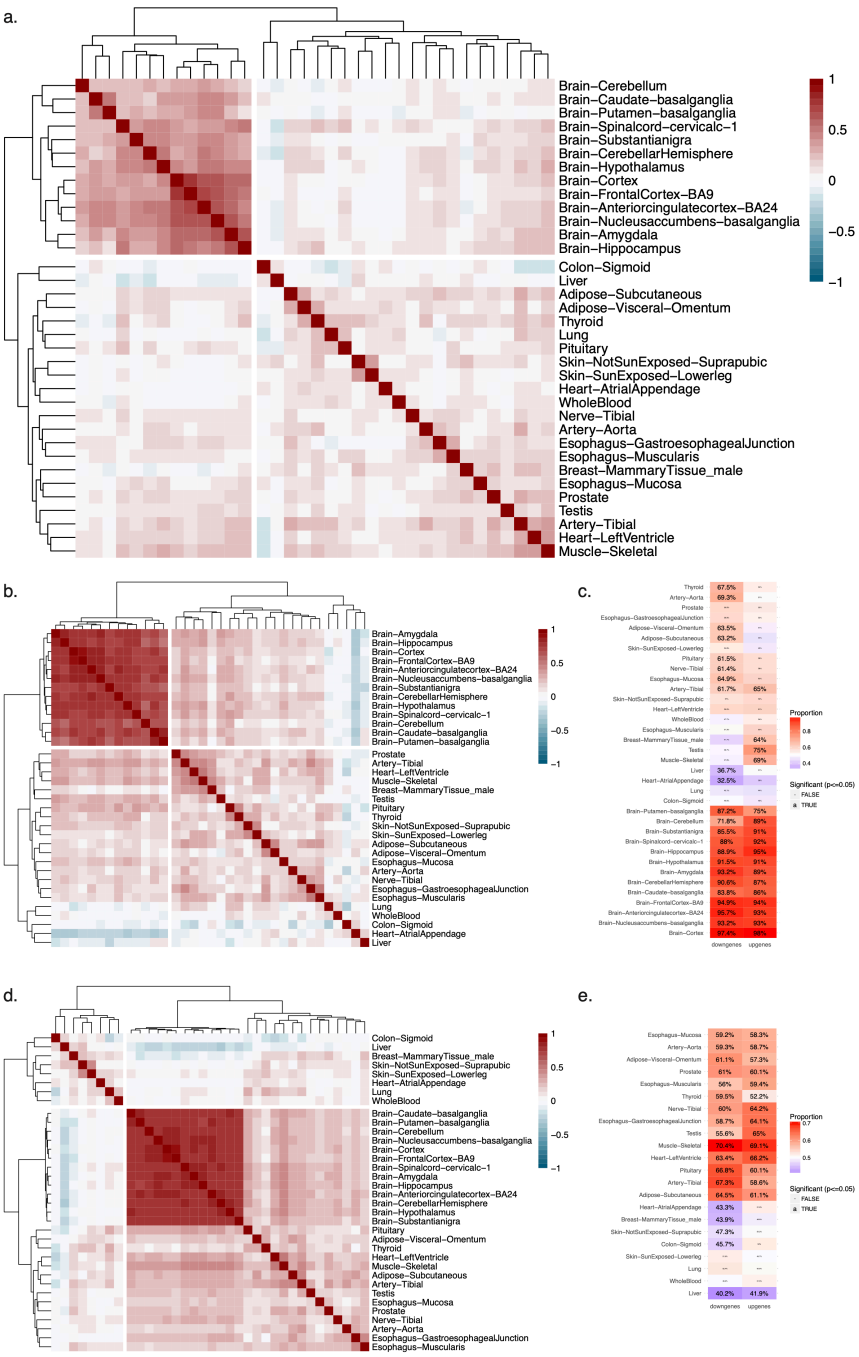


Figure A.65 *a) Heatmap showing the pairwise correlation coefficients among GTEx datasets, corresponding 17 major and 35 minor tissue types. The intensity of the colours on the heatmap shows the magnitude of the correlation coefficient. b and d) The same as (a) but using only the genes in brain ageing signature compiled using microarray (a) and GTEx brain dataset (b). c)Heatmap showing the proportion of the changes in the same direction with ageing signature compiled using microarray signature. The colour shows the similarity (red), and dissimilarity (blue) based on the majority of change. The labels show the proportion of the similar type of change, where the size of the labels shows the statistical significance. e) The same as (c) but for GTEx ageing signature compiled using only the brain data.*

References

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. http://doi.org/10.1038/nature15393

Admasu, T. D., Chaithanya Batchu, K., Barardo, D., Ng, L. F., Lam, V. Y. M., Xiao, L., … Gruber, J. (2018). Drug synergy slows aging and improves healthspan through IGF and SREBP lipid signaling. *Dev. Cell*, *47*(1), 67–79.e5. http://doi.org/10.1016/j.devcel.2018.09.001

Aleman, F. D. D., & Valenzano, D. R. (2019). Microbiome evolution during host aging. *PLoS Pathog.*, *15*(7), e1007727. http://doi.org/10.1371/journal.ppat.1007727

Alexa, A., & Rahnenfuhrer, J. (2019). topGO: Enrichment analysis for gene ontology.

Algire, C., Moiseeva, O., Deschênes-Simard, X., Amrein, L., Petruccelli, L., Birman, E., … Pollak, M. N. (2012). Metformin reduces endogenous reactive oxygen species and associated DNA damage. *Cancer Prev. Res.*, *5*(4), 536–543. http://doi.org/10.1158/1940-6207.CAPR-11-0536

Aliper, A., Belikov, A. V., Garazha, A., Jellen, L., Artemov, A., Suntsova, M., … Zhavoronkov, A. (2016). In search for geroprotectors: In silico screening and in vitro validation of signalome-level mimetics of young healthy state. *Aging*, *8*(9), 2127–2152. http://doi.org/10.18632/aging.101047

Andersen, S. L., Sebastiani, P., Dworkis, D. A., Feldman, L., & Perls, T. T. (2012). Health span approximates life span among many supercentenarians: Compres-

sion of morbidity at the approximate limit of life span. *J. Gerontol. A Biol. Sci. Med. Sci.*, *67*(4), 395–405. http://doi.org/10.1093/gerona/glr223

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nat. Protoc.*, *5*(9), 1564–1573. http://doi.org/10.1038/nprot.2010.116

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., Mayr, C. H., Greiffo, F. R., … Schiller, H. B. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, *10*(1), 963. http://doi.org/10.1038/s41467-019-08831-9

Anisimova, A. S., Meerson, M. B., Gerashchenko, M. V., Kulakovskiy, I. V., Dmitriev, S. E., & Gladyshev, V. N. (2020, January). *Multi-faceted deregulation of gene expression and protein synthesis with age*. *bioRxiv*. http://doi.org/10.1101/2020.01.19.911404

Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., … Dermitzakis, E. T. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. http://doi.org/10.1126/science.1262110

Aris, V. M., Cody, M. J., Cheng, J., Dermody, J. J., Soteropoulus, P., Recce, M., & Tolias, P. P. (2004). Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer. *BMC Bioinformatics*, *5*. http://doi.org/10.1186/1471-2105-5-185

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, *25*(1), 25–29. http://doi.org/10.1038/75556

Austad, S. N. (2004). Is aging programed? *Aging Cell*, *3*(5), 249–251. http://doi.org/10.1111/j.1474-9728.2004.00112.x

Avelar, R. A., Ortega, J. G., Tacutu, R., Tyler, E., Bennett, D., & others. (2019). A multidimensional systems biology analysis of cellular senescence in ageing and disease. *BioRxiv*.

Avramopoulos, D., Szymanski, M., Wang, R., & Bassett, S. (2011). Gene expression reveals overlap between normal aging and alzheimer's disease genes.

Neurobiol. Aging, *32*(12), 2319.e27–34. http://doi.org/10.1016/j.neurobiolaging.2010.04.019

Bahar, R., Hartmann, C. H., Rodriguez, K. A., Denny, A. D., Busuttil, R. A., Dollé, M. E. T., … Vijg, J. (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, *441*(7096), 1011–1014. http://doi.org/10.1038/nature04844

Balaban, R. S., Nemoto, S., & Finkel, T. (2005). Mitochondria, oxidants, and aging. *Cell*, *120*(4), 483–495. http://doi.org/10.1016/j.cell.2005.02.001

Barardo, D., Thornton, D., Thoppil, H., Walsh, M., Sharifi, S., Ferreira, S., … Magalhães, J. P. de. (2017). The DrugAge database of aging-related drugs. *Aging Cell*, *16*(3), 594–597. http://doi.org/10.1111/acel.12585

Barbi, E., Lagona, F., Marsili, M., Vaupel, J. W., & Wachter, K. W. (2018). The plateau of human mortality: Demography of longevity pioneers. *Science*, *360*(6396), 1459–1461. http://doi.org/10.1126/science.aat3119

Barnes, M. R., Huxley-Jones, J., Maycox, P. R., Lennon, M., Thornber, A., Kelly, F., … Belleroche, J. de. (2011). Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. *Journal of Neuroscience Research*. http://doi.org/10.1002/jnr.22647

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., … Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Res.*, *41*(Database issue), D991–5. http://doi.org/10.1093/nar/gks1193

Barroso, G. V., Puzovic, N., & Dutheil, J. Y. (2018). The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level. *Genetics*, *208*(1), 173–189. http://doi.org/10.1534/genetics.117.300467

Barzilai, N., Crandall, J. P., Kritchevsky, S. B., & Espeland, M. A. (2016). Metformin as a Tool to Target Aging. *Cell Metabolism*, *23*(6), 1060–1065. http://doi.org/10.1016/j.cmet.2016.05.011

Batandier, C., Guigas, B., Detaille, D., El-Mir, M.-Y., Fontaine, E., Rigoulet, M., & Leverve, X. M. (2006). The ROS production induced by a reverse-electron flux at respiratory-chain complex 1 is hampered by metformin. *J. Bioenerg. Biomembr.*, *38*(1), 33–42. http://doi.org/10.1007/s10863-006-9003-8

Baur, J. A., & Sinclair, D. A. (2006). Therapeutic potential of resveratrol: The in vivo evidence. *Nat. Rev. Drug Discov.*, *5*(6), 493–506. http://doi.org/10.1038/nrd2060

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. WileyRoyal Statistical Society.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, *29*(4), 1165–1188. http://doi.org/10.1214/aos/1013699998

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., … Overington, J. P. (2014). The ChEMBL bioactivity database: An update. *Nucleic Acids Res.*, *42*(Database issue), D1083–90. http://doi.org/10.1093/nar/gkt1031

Berchtold, N. C., Cribbs, D. H., Coleman, P. D., Rogers, J., Head, E., Kim, R., … Cotman, C. W. (2008). Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proc. Natl. Acad. Sci. U. S. A.*, *105*(40), 15605–15610. http://doi.org/10.1073/pnas.0806883105

Berisa, T., & Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, *32*(2), 283–285. http://doi.org/10.1093/bioinformatics/btv546

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*(1), 235–242. http://doi.org/10.1093/nar/28.1.235

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining* (pp. 359–370). Seattle, WA: AAAI Press.

Bhadra, M., Howell, P., Dutta, S., Heintz, C., & Mair, W. B. (2019). Alternative splicing in aging and longevity. *Hum. Genet.* http://doi.org/10.1007/s00439-019-02094-6

Bien, S. A., & Peters, U. (2019). Moving from one to many: Insights from the growing list of pleiotropic cancer risk genes. *Br. J. Cancer*, *120*(12), 1087–1089. http://doi.org/10.1038/s41416-019-0475-9

Bink, D. I., Lozano-Vidal, N., & Boon, R. A. (2019). Long Non-Coding RNA in vascular disease and aging. *Noncoding RNA*, *5*(1). http://doi.org/10.3390/ncrna5010026

Bitarello, B. D., Filippo, C. de, Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andrés, A. M. (2018). Signatures of Long-Term balancing selection in human genomes. *Genome Biol. Evol.*, *10*(3), 939–955. http://doi.org/10.1093/gbe/evy054

Bjedov, I., Toivonen, J. M., Kerr, F., Slack, C., Jacobson, J., Foley, A., & Partridge, L. (2010). Mechanisms of life span extension by rapamycin in the fruit fly drosophila melanogaster. *Cell Metab.*, *11*(1), 35–46. http://doi.org/10.1016/j.cmet.2009.11.010

Blackburn, E. H., Greider, C. W., & Szostak, J. W. (2006). Telomeres and telomerase: The path from maize, tetrahymena and yeast to human cancer and aging. *Nat. Med.*, *12*(10), 1133–1138. http://doi.org/10.1038/nm1006-1133

Blankenburg, H., Pramstaller, P. P., & Domingues, F. S. (2018). A network-based meta-analysis for characterizing the genetic landscape of human aging. *Biogerontology*, *19*(1), 81–94. http://doi.org/10.1007/s10522-017-9741-5

Bocklandt, S., Lin, W., Sehl, M. E., Sánchez, F. J., Sinsheimer, J. S., Horvath, S., & Vilain, E. (2011). Epigenetic predictor of age. *PLoS One*, *6*(6), e14821. http://doi.org/10.1371/journal.pone.0014821

Bolstad, B. M. (2019). preprocessCore: A collection of pre-processing functions.

Boonekamp, J. J., Simons, M. J. P., Hemerik, L., & Verhulst, S. (2013). Telomere length behaves as biomarker of somatic redundancy rather than biological age. *Aging Cell*, *12*(2), 330–332. http://doi.org/10.1111/acel.12050

Bork, S., Pfister, S., Witt, H., Horn, P., Korn, B., Ho, A. D., & Wagner, W. (2010). DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. *Aging Cell*. http://doi.org/10.1111/j.1474-9726.2009.00535.x

Bové, J., Martínez-Vicente, M., & Vila, M. (2011). Fighting neurodegeneration with rapamycin: Mechanistic insights. *Nat. Rev. Neurosci.*, *12*(8), 437–452. http://doi.org/10.1038/nrn3068

Brinkmeyer-Langford, C. L., Guan, J., Ji, G., & Cai, J. J. (2016). Aging Shapes the Population-Mean and -Dispersion of Gene Expression in Human Brains. *Frontiers in Aging Neuroscience*, *8*, 183. http://doi.org/10.3389/fnagi.2016.00183

Brunet-Rossinni, A. K., & Austad, S. N. (2005). Chapter 9 - senescence in wild populations of mammals and birds. In E. J. Masoro & S. N. Austad (Eds.), *Handbook of the biology of aging (sixth edition)* (pp. 243–266). Burlington: Academic Press. http://doi.org/10.1016/B978-012088387-5/50012-1

Buford, T. W. (2017). (Dis)Trust your gut: The gut microbiome in age-related inflammation, health, and disease. *Microbiome*, *5*(1), 80. http://doi.org/10.1186/s40168-017-0296-0

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., … Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, *47*(11), 1236–1241. http://doi.org/10.1038/ng.3406

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, *47*(D1), D1005–D1012. http://doi.org/10.1093/nar/gky1120

Bushardt, R. L., Massey, E. B., Simpson, T. W., Ariail, J. C., & Simpson, K. N. (2008). Polypharmacy: misleading, but manageable. *Clinical Interventions in Aging*, *3*(2), 383–9.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. http://doi.org/10.1038/s41586-018-0579-z

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., … Barres, B. A. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(1), 264–278. http://doi.org/10.1523/JNEUROSCI.4178-07.2008

Calvert, S., Tacutu, R., Sharifi, S., Teixeira, R., Ghosh, P., & Magalhães, J. P. de. (2016). A network pharmacology approach reveals new candidate caloric restriction mimetics in c. Elegans. *Aging Cell*, *15*(2), 256–266. http://doi.org/10.1111/acel.12432

Camilleri, M. (2019). Leaky gut: Mechanisms, measurement and clinical implications in humans. *Gut*, *68*(8), 1516–1526. http://doi.org/10.1136/gutjnl-2019-318427

Campbell, J. M., Bellman, S. M., Stephenson, M. D., & Lisy., K. (2017). Metformin reduces all-cause mortality and diseases of ageing independent of its effect on diabetes control: A systematic review and meta-analysis. *Ageing Research Reviews*, *40*, 31–44. http://doi.org/10.1016/j.arr.2017.08.003

Campisi, J., & Fagagna, F. d'Adda di. (2007). Cellular senescence: When bad things happen to good cells. *Nat. Rev. Mol. Cell Biol.*, *8*(9), 729–740. http://doi.org/10.1038/nrm2233

Carlson, M. (2019). Org.Hs.eg.db: Genome wide annotation for human.

Carlson, M., & Maintainer, B. P. (2015). TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s).

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, *26*(19), 2363–2367. http://doi.org/10.1093/bioinformatics/btq431

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., … Dunham, I. (2019). Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, *47*(D1), D1056–D1065. http://doi.org/10.1093/nar/gky1133

Castillo-Quan, J. I., Tain, L. S., Kinghorn, K. J., Li, L., Grönke, S., Hinze, Y., … Partridge, L. (2019). A triple drug combination targeting components of the nutrient-sensing network maximizes longevity. *Proc. Natl. Acad. Sci. U. S. A.*, *116*(42), 20817–20819. http://doi.org/10.1073/pnas.1913212116

Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., … Overington, J. P. (2013). UniChem: A unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.*, *5*(1), 3. http://doi.org/10.1186/1758-2946-5-3

Chen, J., Wang, M., Guo, M., Xie, Y., & Cong, Y.-S. (2013). MiR-127 regulates cell proliferation and senescence by targeting BCL6. *PLoS One*, *8*(11), e80266. http://doi.org/10.1371/journal.pone.0080266

Cheung, P., Vallania, F., Warsinske, H. C., Donato, M., Schaffert, S., Chang, S. E., … Kuo, A. J. (2018). Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell*, *173*(6), 1385–1397.e14. http://doi.org/10.1016/j.cell.2018.03.079

Childs, B. G., Gluscevic, M., Baker, D. J., Laberge, R.-M., Marquess, D., Dananberg, J., & Deursen, J. M. van. (2017). Senescent cells: An emerging target for diseases of ageing. *Nat. Rev. Drug Discov.*, *16*(10), 718–735. http://doi.org/10.1038/nrd.2017.116

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., … Huang, H.-D. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research*, *44*(D1), D239–D247. http://doi.org/10.1093/nar/gkv1258

Chouakria, A. D., & Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Adv. Data Anal. Classif.*, *1*(1), 5–21. http://doi.org/10.1007/s11634-006-0004-6

Christensen, K., McGue, M., Petersen, I., Jeune, B., & Vaupel, J. W. (2008). Exceptional longevity does not result in excessive levels of disability. *Proc. Natl. Acad. Sci. U. S. A.*, *105*(36), 13274–13279. http://doi.org/10.1073/pnas.0804931105

Clancy, D. J., Gems, D., Harshman, L. G., Oldham, S., Stocker, H., Hafen, E., … Partridge, L. (2001). Extension of life-span by loss of CHICO, a drosophila insulin receptor substrate protein. *Science*, *292*(5514), 104–106. http://doi.org/10.1126/science.1057991

Colantuoni, C., Lipska, B. K., Ye, T., Hyde, T. M., Tao, R., Leek, J. T., … Kleinman, J. E. (2011). Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature*, *478*(7370), 519–523. http://doi.org/10.1038/nature10524

Colman, R. J., Beasley, T. M., Kemnitz, J. W., Johnson, S. C., Weindruch, R., & Anderson, R. M. (2014). Caloric restriction reduces age-related and all-cause mortality in rhesus monkeys. *Nat. Commun.*, *5*, 3557. http://doi.org/10.1038/ncomms4557

Compound: Acetohexamide. (n.d.-a). https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL1589/.

Compound: ATROPINE OXIDE. (n.d.-b). https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL2146145/.

Conboy, I. M., Conboy, M. J., Wagers, A. J., Girma, E. R., Weissman, I. L., & Rando, T. A. (2005). Rejuvenation of aged progenitor cells by exposure to a

young systemic environment. *Nature*, *433*(7027), 760–764. http://doi.org/10.1038/nature03260

Consortium, T. U. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. http://doi.org/10.1093/nar/gkw1099

Cortes, A., Albers, P. K., Dendrou, C. A., Fugger, L., & McVean, G. (2020). Identifying cross-disease components of genetic risk across hospital data in the UK biobank. *Nat. Genet.*, *52*(1), 126–134. http://doi.org/10.1038/s41588-019-0550-4

Cortes, A., Dendrou, C., Fugger, L., & McVean, G. (2018). Systematic classification of shared components of genetic risk for common human diseases. *bioRxiv*.

Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., … Griffith, M. (2018). DGIdb 3.0: A redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res.*, *46*(D1), D1068–D1073. http://doi.org/10.1093/nar/gkx1143

Crimmins, E. M. (2015). Lifespan and healthspan: Past, present, and promise. *Gerontologist*, *55*(6), 901–911. http://doi.org/10.1093/geront/gnv130

Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., … International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.*, *45*(9), 984–994. http://doi.org/10.1038/ng.2711

Danka Mohammed, C. P., Park, J. S., Nam, H. G., & Kim, K. (2017). MicroRNAs in brain aging. *Mech. Ageing Dev.*, *168*, 3–9. http://doi.org/10.1016/j.mad.2017.01.007

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., … Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.*, *112*(23), 7285–7290. http://doi.org/10.1073/pnas.1507125112

Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., … Aerts, S. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell*, *174*(4), 982–998.e20. http://doi.org/10.1016/j.cell.2018.05.057

Deelen, J., Beekman, M., Capri, M., Franceschi, C., & Slagboom, P. E. (2013). Identifying the genomic determinants of aging and longevity in human popu-

lation studies: Progress and challenges. *Bioessays*, *35*(4), 386–396. http://doi.org/10.1002/bies.201200148

Deelen, J., Evans, D. S., Arking, D. E., Tesi, N., Nygaard, M., Liu, X., … Murabito, J. M. (2019). A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.*, *10*(1), 3669. http://doi.org/10.1038/s41467-019-11558-2

Dong, X., Milholland, B., & Vijg, J. (2016). Evidence for a limit to human lifespan. *Nature*, *538*(7624), 257–259. http://doi.org/10.1038/nature19793

Dorszewska, J. (2013). Cell biology of normal brain aging: synaptic plasticity–cell death. *Aging Clinical and Experimental Research*, *25*(1), 25–34. http://doi.org/10.1007/s40520-013-0004-2

Dönertaş, H. M., Fuentealba, M., Partridge, L., & Thornton, J. M. (2019). Identifying potential Ageing-Modulating drugs in silico. *Trends Endocrinol. Metab.*, *30*(2), 118–131. http://doi.org/10.1016/j.tem.2018.11.005

Dönertaş, H. M., Fuentealba Valenzuela, M., Partridge, L., & Thornton, J. M. (2018). Gene expression-based drug repurposing to target aging. *Aging Cell*, *17*(5), e12819. http://doi.org/10.1111/acel.12819

Dönertaş, H. M., İzgi, H., Kamacıoğlu, A., He, Z., Khaitovich, P., & Somel, M. (2017). Gene expression reversal toward pre-adult levels in the aging human brain and age-related loss of cellular identity. *Sci. Rep.*, *7*(1), 5894. http://doi.org/10.1038/s41598-017-05927-4

Duran-Frigola, M., Mateo, L., & Aloy, P. (2017). Drug repositioning beyond the low-hanging fruits. *Current Opinion in Systems Biology*, *3*, 95–102. http://doi.org/10.1016/j.coisb.2017.04.010

Durieux, J., Wolff, S., & Dillin, A. (2011). The cell-non-autonomous nature of electron transport chain-mediated longevity. *Cell*, *144*(1), 79–91. http://doi.org/10.1016/j.cell.2010.12.016

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, *4*(8), 1184–1191. http://doi.org/10.1038/nprot.2009.97

Ellinghaus, D., Jostins, L., Spain, S. L., Cortes, A., Bethune, J., Han, B., … Franke, A. (2016). Analysis of five chronic inflammatory diseases identifies 27 new as-

sociations and highlights disease-specific patterns at shared loci. *Nat. Genet.*, *48*(5), 510–518. http://doi.org/10.1038/ng.3528

Enderlin, V., Alfos, S., Pallet, V., Garcin, H., Azaïs-Braesco, V., Jaffard, R., & Higueret, P. (1997). Aging decreases the abundance of retinoic acid (RAR) and triiodothyronine (TR) nuclear receptor mRNA in rat brain: Effect of the administration of retinoids. *FEBS Lett.*, *412*(3), 629–632. http://doi.org/10.1016/s0014-5793(97)00845-4

Enge, M., Arda, H. E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., & Quake, S. R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell*, *171*(2), 321–330.e14. http://doi.org/10.1016/j.cell.2017.09.004

Espeland, M. A., Crimmins, E. M., Grossardt, B. R., Crandall, J. P., Gelfond, J. A. L., Harris, T. B., … Barzilai, N. (2016). Clinical Trials Targeting Aging and Age-Related Multimorbidity. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *72*(3), glw220. http://doi.org/10.1093/gerona/glw220

Evans, D. A., Funkenstein, H. H., Albert, M. S., Scherr, P. A., Cook, N. R., Chown, M. J., … Taylor, J. O. (1989). Prevalence of Alzheimer's disease in a community population of older persons. Higher than previously reported. *JAMA*, *262*(18), 2551–6.

Evans, S. E., Goult, B. T., Fairall, L., Jamieson, A. G., Ko Ferrigno, P., Ford, R., … Wagner, S. D. (2014). The ansamycin antibiotic, rifamycin SV, inhibits BCL6 transcriptional repression and forms a complex with the BCL6-BTB/POZ domain. *PLoS One*, *9*(3), e90889. http://doi.org/10.1371/journal.pone.0090889

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., … D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*(D1), D649–D655. http://doi.org/10.1093/nar/gkx1132

Faggioli, F., Wang, T., Vijg, J., & Montagna, C. (2012). Chromosome-specific accumulation of aneuploidy in the aging mouse brain. *Hum. Mol. Genet.*, *21*(24), 5246–5253. http://doi.org/10.1093/hmg/dds375

Fernandes, M., Wan, C., Tacutu, R., Barardo, D., Rajput, A., Wang, J., … Magalhães, J. P. de. (2016). Systematic analysis of the gerontome reveals links

between aging and age-related diseases. *Hum. Mol. Genet.*, *25*(21), 4804–4818. http://doi.org/10.1093/hmg/ddw307

Finkel, T., Serrano, M., & Blasco, M. A. (2007). The common biology of cancer and ageing. *Nature*, *448*(7155), 767–774. http://doi.org/10.1038/nature05985

Flatt, T., & Partridge, L. (2018). Horizons in the evolution of aging. *BMC Biol.*, *16*(1), 93. http://doi.org/10.1186/s12915-018-0562-z

Fontana, L., & Partridge, L. (2015). Promoting health and longevity through diet: From model organisms to humans. *Cell*, *161*(1), 106–118. http://doi.org/10.1016/j.cell.2015.02.020

Fontana, L., Partridge, L., & Longo, V. D. (2010). Extending healthy life span–from yeast to humans. *Science*, *328*(5976), 321–326. http://doi.org/10.1126/science.1172539

Foretz, M., Guigas, B., Bertrand, L., Pollak, M., & Viollet, B. (2014). Metformin: From mechanisms of action to therapies. *Cell Metab.*, *20*(6), 953–966. http://doi.org/10.1016/j.cmet.2014.09.018

Forsberg, L. A., Rasi, C., Razzaghian, H. R., Pakalapati, G., Waite, L., Thilbeault, K. S., … Dumanski, J. P. (2012). Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.*, *90*(2), 217–228. http://doi.org/10.1016/j.ajhg.2011.12.009

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression*.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., … Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, *102*(30), 10604–10609. http://doi.org/10.1073/pnas.0500398102

Franceschi, C., Garagnani, P., Parini, P., Giuliani, C., & Santoro, A. (2018). Inflammaging: A new immune-metabolic viewpoint for age-related diseases. *Nat. Rev. Endocrinol.*, *14*(10), 576–590. http://doi.org/10.1038/s41574-018-0059-4

Freije, J. M. P., & López-Otín, C. (2012). Reprogramming aging and progeria. *Curr. Opin. Cell Biol.*, *24*(6), 757–764. http://doi.org/10.1016/j.ceb.2012.08.009

Friedman, D. B., & Johnson, T. E. (1988). A mutation in the age-1 gene in caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. *Genetics*, *118*(1), 75–86.

Fuentealba, M., Dönertaş, H. M., Williams, R., Labbadia, J., Thornton, J. M., & Partridge, L. (2019). Using the drug-protein interactome to identify anti-ageing compounds for humans. *PLoS Computational Biology*, *15*(1), e1006639. http://doi.org/10.1371/journal.pcbi.1006639

Fuhrmann-Stroissnigg, H., Ling, Y. Y., Zhao, J., McGowan, S. J., Zhu, Y., Brooks, R. W., … Robbins, P. D. (2017). Identification of HSP90 inhibitors as a novel class of senolytics. *Nat. Commun.*, *8*(1), 422. http://doi.org/10.1038/s41467-017-00314-z

Fushan, A. A., Turanov, A. A., Lee, S.-G., Kim, E. B., Lobanov, A. V., Yim, S. H., … Gladyshev, V. N. (2015). Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, *14*(3), 352–365. http://doi.org/10.1111/acel.12283

Gamazon, E. R., Segrè, A. V., Bunt, M. van de, Wen, X., Xi, H. S., Hormozdiari, F., … Ardlie, K. G. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.*, *50*(7), 956–967. http://doi.org/10.1038/s41588-018-0154-4

García-Nieto, P. E., Morrison, A. J., & Fraser, H. B. (2019, June). *The somatic mutation landscape of the human body*. *bioRxiv*. http://doi.org/10.1101/668624

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., … Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.*, *45*(D1), D945–D954. http://doi.org/10.1093/nar/gkw1074

Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). Affy–analysis of affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307–315. http://doi.org/10.1093/bioinformatics/btg405

Glass, D., Viñuela, A., Davies, M. N., Ramasamy, A., Parts, L., Knowles, D., … Spector, T. D. (2013). Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol.*, *14*(7), R75. http://doi.org/10.1186/gb-2013-14-7-r75

Goodell, M. A., & Rando, T. A. (2015). Stem cells and healthy aging. *Science*, *350*(6265), 1199–1204. http://doi.org/10.1126/science.aab3388

Gorbunova, V., Seluanov, A., Mao, Z., & Hine, C. (2007). Changes in DNA repair during aging. *Nucleic Acids Research*, *35*(22), 7466–74. http://doi.org/10.1093/nar/gkm756

Green, D. R., Galluzzi, L., & Kroemer, G. (2011). Mitochondria and the autophagy-inflammation-cell death axis in organismal aging. *Science*, *333*(6046), 1109–1112. http://doi.org/10.1126/science.1201940

Griffith, M., Griffith, O. L., Coffman, A. C., Weible, J. V., Mcmichael, J. F., Spies, N. C., … Wilson, R. K. (2013). DGIdb: Mining the druggable genome. *Nature Methods*, *10*(12), 1209–1210. http://doi.org/10.1038/nmeth.2689

Grillo, F. W., Song, S., Teles-Grilo Ruivo, L. M., Huang, L., Gao, G., Knott, G. W., … De Paola, V. (2013). Increased axonal bouton dynamics in the aging mouse cortex. *Proceedings of the National Academy of Sciences*, *110*(16), E1514–E1523. http://doi.org/10.1073/pnas.1218731110

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. http://doi.org/10.1126/science.1262110

Gu, Q., Dillon, C. F., & Burt, V. L. (2010). Prescription drug use continues to increase: U.S. prescription drug data for 2007-2008. *NCHS Data Brief*, (42), 1–8.

Guthrie, B., Makubate, B., Hernandez-Santiago, V., & Dreischulte, T. (2015). The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995-2010. *BMC Medicine*, *13*, 74. http://doi.org/10.1186/s12916-015-0322-7

Gutiérrez-Sacristán, A., Bravo, À., Giannoula, A., Mayer, M. A., Sanz, F., & Furlong, L. I. (2018). ComoRbidity: An R package for the systematic analysis of disease comorbidities. *Bioinformatics*, *34*(18), 3228–3230. http://doi.org/10.1093/bioinformatics/bty315

Hammond, C. J., Snieder, H., Spector, T. D., & Gilbert, C. E. (2000). Genetic and environmental factors in age-related nuclear cataracts in monozygotic and dizygotic twins. *N. Engl. J. Med.*, *342*(24), 1786–1790. http://doi.org/10.1056/NEJM200006153422404

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., … Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human

aging rates. *Molecular Cell*, *49*(2), 359–367. http://doi.org/10.1016/j.molcel.2012.10.016

Harrison, D. E., Strong, R., Sharp, Z. D., Nelson, J. F., Astle, C. M., Flurkey, K., … Miller, R. A. (2009). Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*, *460*(7253), 392–395. http://doi.org/10.1038/nature08221

Hayflick, L., & Moorhead, P. S. (1961). The serial cultivation of human diploid cell strains. *Exp. Cell Res.*, *25*, 585–621. http://doi.org/10.1016/0014-4827(61)90192-6

Hernando-Herraez, I., Evano, B., Stubbs, T., Commere, P.-H., Jan Bonder, M., Clark, S., … Reik, W. (2019). Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat. Commun.*, *10*(1), 4361. http://doi.org/10.1038/s41467-019-12293-4

Herndon, L. A., Schmeissner, P. J., Dudaronek, J. M., Brown, P. A., Listner, K. M., Sakano, Y., … Driscoll, M. (2002). Stochastic and genetic factors influence tissue-specific decline in ageing C. elegans. *Nature*, *419*(6909), 808–814. http://doi.org/10.1038/nature01135

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., … Flicek, P. (2016). Ensembl comparative genomics resources. *Database*, *2016*, bav096. http://doi.org/10.1093/database/bav096

Hipp, M. S., Kasturi, P., & Hartl, F. U. (2019). The proteostasis network and its decline in ageing. *Nat. Rev. Mol. Cell Biol.*, *20*(7), 421–435. http://doi.org/10.1038/s41580-019-0101-y

Hoeijmakers, J. H. J. (2009). DNA damage, aging, and cancer. *N. Engl. J. Med.*, *361*(15), 1475–1485. http://doi.org/10.1056/NEJMra0804615

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), 1–20. http://doi.org/10.1186/gb-2013-14-10-r115

Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.*, *19*(6), 371–384. http://doi.org/10.1038/s41576-018-0004-3

Inukai, S., Pincus, Z., Lencastre, A. de, & Slack, F. J. (2018). A microRNA feedback loop regulates global microRNA abundance during aging. *RNA*, *24*(2), 159–172. http://doi.org/10.1261/rna.062190.117

Işıldak, U., Somel, M., Thornton, J. M., & Dönertaş, H. M. (2020). Temporal changes in the gene expression heterogeneity during brain development and aging. *Sci. Rep.*, *10*(1), 4080. http://doi.org/10.1038/s41598-020-60998-0

Jagger, C., Gillies, C., Moscone, F., Cambois, E., Van Oyen, H., Nusselder, W., … EHLEIS team. (2008). Inequalities in healthy life years in the 25 countries of the european union in 2005: A cross-national meta-regression analysis. *Lancet*, *372*(9656), 2124–2131. http://doi.org/10.1016/S0140-6736(08)61594-9

Janzen, V., Forkert, R., Fleming, H. E., Saito, Y., Waring, M. T., Dombkowski, D. M., … Scadden, D. T. (2006). Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16INK4a. *Nature*, *443*(7110), 421–426. http://doi.org/10.1038/nature05159

Jaskelioff, M., Muller, F. L., Paik, J.-H., Thomas, E., Jiang, S., Adams, A. C., … Depinho, R. A. (2011). Telomerase reactivation reverses tissue degeneration in aged telomerase-deficient mice. *Nature*, *469*(7328), 102–106. http://doi.org/10.1038/nature09603

Jiang, Y., Ma, S., Shia, B.-C., & Lee, T.-S. (2018). An epidemiological human disease network derived from disease co-occurrence in taiwan. *Sci. Rep.*, *8*(1), 4557. http://doi.org/10.1038/s41598-018-21779-y

Johnson, K., Liu, L., Majdzadeh, N., Chavez, C., Chin, P. C., Morrison, B., … D'Mello, S. R. (2005). Inhibition of neuronal apoptosis by the cyclin-dependent kinase inhibitor GW8510: Identification of 3' substituted indolones as a scaffold for the development of neuroprotective drugs. *J. Neurochem.*, *93*(3), 538–548. http://doi.org/10.1111/j.1471-4159.2004.03004.x

Johnson, S. C., Dong, X., Vijg, J., & Suh, Y. (2015). Genetic evidence for common pathways in human age-related diseases. *Aging Cell*, *14*(5), 809–817. http://doi.org/10.1111/acel.12362

Johnson, S. C., Rabinovitch, P. S., & Kaeberlein, M. (2013). mTOR is a key modulator of ageing and age-related disease. *Nature*, *493*(7432), 338–345. http://doi.org/10.1038/nature11861

Johnson, T. E. (2002). A personal retrospective on the genetics of aging. *Biogerontology*, *3*(1-2), 7–12. http://doi.org/10.1023/a:1015270322517

Kadariya, Y., Tang, B., Wang, L., Al-Saleem, T., Hayakawa, K., Slifker, M. J., & Kruger, W. D. (2013). Germline mutations in mtap cooperate with myc to accelerate tumorigenesis in mice. *PLoS One*, *8*(6), e67635. http://doi.org/10.1371/journal.pone.0067635

Kadariya, Y., Yin, B., Tang, B., Shinton, S. A., Quinlivan, E. P., Hua, X., … Kruger, W. D. (2009). Mice heterozygous for germ-line mutations in methylthioadenosine phosphorylase (MTAP) die prematurely of t-cell lymphoma. *Cancer Res.*, *69*(14), 5961–5969. http://doi.org/10.1158/0008-5472.CAN-09-0145

Kaeberlein, M., Powers, R. W., 3rd, Steffen, K. K., Westman, E. A., Hu, D., Dang, N., … Kennedy, B. K. (2005). Regulation of yeast replicative life span by TOR and sch9 in response to nutrients. *Science*, *310*(5751), 1193–1196. http://doi.org/10.1126/science.1115535

Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D. M., Wendl, M. C., … Ding, L. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.*, *5*, 3156. http://doi.org/10.1038/ncomms4156

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, *47*(D1), D590–D595. http://doi.org/10.1093/nar/gky962

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. http://doi.org/10.1093/nar/gkv1070

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., … Sestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*(7370), 483–489. http://doi.org/10.1038/nature10523

Kapahi, P., Kaeberlein, M., & Hansen, M. (2017). Dietary restriction and lifespan: Lessons from invertebrate models. *Ageing Res. Rev.*, *39*, 3–14. http://doi.org/10.1016/j.arr.2016.12.005

Kapahi, P., Zid, B. M., Harper, T., Koslover, D., Sapin, V., & Benzer, S. (2004). Regulation of lifespan in drosophila by modulation of genes in the TOR signaling

pathway. *Curr. Biol.*, *14*(10), 885–890. http://doi.org/10.1016/j.cub.2004.03.059

Kassambara, A. (2018). ggpubr: 'ggplot2' Based Publication Ready Plots. *R Package Version 0.1.8.*

Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). Partitioning around medoids (program PAM). In *Finding groups in data* (pp. 68–125). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9780470316801.ch2

Kauppila, T. E. S., Kauppila, J. H. K., & Larsson, N.-G. (2017). Mammalian mitochondria and aging: An update. *Cell Metab.*, *25*(1), 57–71. http://doi.org/10.1016/j.cmet.2016.09.017

Kaushik, S., & Cuervo, A. M. (2015). Proteostasis and aging. *Nat. Med.*, *21*(12), 1406–1415. http://doi.org/10.1038/nm.4001

Kedlian, V. R., Donertas, H. M., & Thornton, J. M. (2019). The widespread increase in inter-individual variability of gene expression in the human brain with age. *Aging*. http://doi.org/10.18632/aging.101912

Kennedy, B. K., & Lamming, D. W. (2016). The mechanistic target of rapamycin: The grand ConducTOR of metabolism and aging. *Cell Metab.*, *23*(6), 990–1003. http://doi.org/10.1016/j.cmet.2016.05.009

Kenyon, C., Chang, J., Gensch, E., Rudner, A., & Tabtiang, R. (1993). A c. Elegans mutant that lives twice as long as wild type. *Nature*, *366*(6454), 461–464. http://doi.org/10.1038/366461a0

Kenyon, C. J. (2010). The genetics of ageing. *Nature*, *464*(7288), 504–512. http://doi.org/10.1038/nature08980

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, *17*(2), 93–108. http://doi.org/10.1038/nrg.2015.17

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., … Schriml, L. M. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, *43*(D1), D1071–D1078. http://doi.org/10.1093/nar/gku1011

Kickstein, E., Krauss, S., Thornhill, P., Rutschow, D., Zeller, R., Sharkey, J., … Schweiger, S. (2010). Biguanide metformin acts on tau phosphorylation via mTOR/protein phosphatase 2A (PP2A) signaling. *Proc. Natl. Acad. Sci. U. S. A.*, *107*(50), 21830–21835. http://doi.org/10.1073/pnas.0912793107

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., … Gladyshev, V. N. (2011). Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, *479*(7372), 223–227. http://doi.org/10.1038/nature10533

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., … Bolton, E. E. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.*, *47*(D1), D1102–D1109. http://doi.org/10.1093/nar/gky1033

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., … Bryant, S. H. (2016). PubChem substance and compound databases. *Nucleic Acids Res.*, *44*(D1), D1202–13. http://doi.org/10.1093/nar/gkv951

Kimmel, J. C., Penland, L., Rubinstein, N. D., Hendrickson, D. G., Kelley, D. R., & Rosenthal, A. Z. (2019, June). *A murine aging cell atlas reveals cell identity and tissue-specific trajectories of aging*. *bioRxiv*. http://doi.org/10.1101/657726

Kingston, A., Robinson, L., Booth, H., Knapp, M., & Jagger, C. (2018). OUP accepted manuscript. *Age and Ageing*, *0*, 1–7. http://doi.org/10.1093/ageing/afy003

Kinser, H. E., & Pincus, Z. (2019). MicroRNAs as modulators of longevity and the aging process. *Hum. Genet.* http://doi.org/10.1007/s00439-019-02046-0

Kirkwood, T. B. (1977). Evolution of ageing. *Nature*, *270*(5635), 301–304. http://doi.org/10.1038/270301a0

Kirkwood, T. B. L. (2005). Understanding the odd science of aging. *Cell*, *120*(4), 437–447. http://doi.org/10.1016/j.cell.2005.01.027

Kirkwood, T. B. L., & Melov, S. (2011). On the programmed/non-programmed nature of ageing within the life history. *Curr. Biol.*, *21*(18), R701–7. http://doi.org/10.1016/j.cub.2011.07.020

Klaips, C. L., Jayaraj, G. G., & Hartl, F. U. (2018). Pathways of cellular proteostasis in aging and disease. *J. Cell Biol.*, *217*(1), 51–63. http://doi.org/10.1083/jcb.201709072

Klass, M. R. (1983). A method for the isolation of longevity mutants in the nematode caenorhabditis elegans and initial results. *Mech. Ageing Dev.*, *22*(3-4), 279–286. http://doi.org/10.1016/0047-6374(83)90082-9

Knupp, D., & Miura, P. (2018). CircRNA accumulation: A new hallmark of aging? *Mech. Ageing Dev.*, *173*, 71–79. http://doi.org/10.1016/j.mad.2018.05.001

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12.

Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K., & Ezzati, M. (2017). Future life expectancy in 35 industrialised countries: Projections with a bayesian model ensemble. *Lancet*, *389*(10076), 1323–1335. http://doi.org/10.1016/S0140-6736(16)32381-9

Kowald, A., & Kirkwood, T. B. L. (2016). Can aging be programmed? A critical literature review. *Aging Cell*, *15*(6), 986–998. http://doi.org/10.1111/acel.12510

Kubben, N., & Misteli, T. (2017). Shared molecular and cellular mechanisms of premature ageing and ageing-associated diseases. *Nat. Rev. Mol. Cell Biol.*, *18*(10), 595–609. http://doi.org/10.1038/nrm.2017.68

Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic Acids Res.*, *44*(D1), D1075–9. http://doi.org/10.1093/nar/gkv1075

Kundu, P., Lee, H. U., Garcia-Perez, I., Tay, E. X. Y., Kim, H., Faylon, L. E., … Pettersson, S. (2019). Neurogenesis and prolongevity signaling in young germ-free mice transplanted with the gut microbiota of old mice. *Sci. Transl. Med.*, *11*(518). http://doi.org/10.1126/scitranslmed.aau4760

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, *313*(5795), 1929–1935. http://doi.org/10.1126/science.1132939

Lane, M. A., & Bailey, S. J. (2005). Role of retinoid signalling in the adult brain. *Prog. Neurobiol.*, *75*(4), 275–293. http://doi.org/10.1016/j.pneurobio.2005.03.002

Langmead, C. J., Watson, J., & Reavill, C. (2008). Muscarinic acetylcholine receptors as CNS drug targets. *Pharmacol. Ther.*, *117*(2), 232–243. http://doi.org/10.1016/j.pharmthera.2007.09.009

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., … Wishart, D. S. (2014). DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.*, *42*(Database issue), D1091–7. http://doi.org/10.1093/nar/gkt1068

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, *9*(8), e1003118. http://doi.org/10.1371/journal.pcbi.1003118

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., … Torres, L. C. (2019). *Sva: Surrogate variable analysis*.

Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, *3*(9), 1724–1735. http://doi.org/10.1371/journal.pgen.0030161

Lehallier, B., Gate, D., Schaum, N., Nanasi, T., Lee, S. E., Yousef, H., … Wyss-Coray, T. (2019). Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.*, *25*(12), 1843–1850. http://doi.org/10.1038/s41591-019-0673-2

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., … Horvath, S. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, *10*(4), 573–591. http://doi.org/10.18632/aging.101414

Li, Y., Sun, H., Chen, Z., Xu, H., Bu, G., & Zheng, H. (2016). Implications of GABAergic neurotransmission in alzheimer's disease. *Front. Aging Neurosci.*, *8*, 31. http://doi.org/10.3389/fnagi.2016.00031

Li, Z., Wright, F. A., & Royland, J. (2009). Age-dependent variability in gene expression in male fischer 344 rat retina. *Toxicol. Sci.*, *107*(1), 281–292. http://doi.org/10.1093/toxsci/kfn215

Liu, B., Fan, Z., Edgerton, S. M., Yang, X., Lind, S. E., & Thor, A. D. (2011). Potent anti-proliferative effects of metformin on trastuzumab-resistant breast cancer cells via inhibition of erbB2/IGF-1 receptor interactions. *Cell Cycle*, *10*(17), 2959–2966. http://doi.org/10.4161/cc.10.17.16359

Liu, H., Guo, M., Xue, T., Guan, J., Luo, L., & Zhuang, Z. (2016). Screening lifespan-extending drugs in Caenorhabditis elegans via label propagation on drug-protein networks. *BMC Systems Biology*, *10*(4), 509–519. http://doi.org/10.1186/s12918-016-0362-4

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., … Walsh, C. A. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, *359*(6375), 555–559. http://doi.org/10.1126/science.aao4426

Lodato, M. A., & Walsh, C. A. (2020). Corrigendum: Genome aging: Somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Hum. Mol. Genet.*, *29*(3), 527. http://doi.org/10.1093/hmg/ddz286

Loh, P.-R. (2017). *BOLT-LMM v2. 3.1 user manual*.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., … Price, A. L. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, *47*(3), 284–290. http://doi.org/10.1038/ng.3190

Lombard, D. B., Chua, K. F., Mostoslavsky, R., Franco, S., Gostissa, M., & Alt, F. W. (2005). DNA Repair, Genome Stability, and Aging. *Cell*, *120*(4), 497–512. http://doi.org/10.1016/j.cell.2005.01.028

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., & Kroemer, G. (2013). The hallmarks of aging. *Cell*, *153*(6), 1194–1217. http://doi.org/10.1016/j.cell.2013.05.039

Lu, T., Pan, Y., Kao, S.-Y., Li, C., Kohane, I., Chan, J., & Yankner, B. A. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature*, *429*(6994), 883–891. http://doi.org/10.1038/nature02661

Lucanic, M., Lithgow, G. J., & Alavez, S. (2013). Pharmacological lifespan extension of invertebrates. *Ageing Res. Rev.*, *12*(1), 445–458. http://doi.org/10.1016/j.arr.2012.06.006

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.*, *107*(3), 961–968. http://doi.org/10.1073/pnas.0912629107

Ma, S., Upneja, A., Galecki, A., Tsai, Y.-M., Burant, C. F., Raskind, S., … Gladyshev, V. N. (2016). Cell culture-based profiling across mammals reveals DNA repair and metabolism as determinants of species longevity. *Elife*, *5*. http://doi.org/10.7554/eLife.19130

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). Cluster: Cluster analysis basics and extensions.

Magalhães, J. P. de, Curado, J., & Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, *25*(7), 875–881. http://doi.org/10.1093/bioinformatics/btp073

Maheshri, N., & O'Shea, E. K. (2007). Living with Noisy Genes: How Cells Function Reliably with Inherent Variability in Gene Expression. *Annual Review of Biophysics and Biomolecular Structure*, *36*(1), 413–434. http://doi.org/10.1146/annurev.biophys.36.040306.132705

Mahmoudi, E., & Cairns, M. J. (2019). Circular RNAs are temporospatially regulated throughout development and ageing in the rat. *Sci. Rep.*, *9*(1), 2564. http://doi.org/10.1038/s41598-019-38860-9

Mair, W., & Dillin, A. (2008). Aging and survival: The genetics of life span extension by dietary restriction. *Annu. Rev. Biochem.*, *77*, 727–754. http://doi.org/10.1146/annurev.biochem.77.061206.171059

Mann, A., Miksys, S. L., Gaedigk, A., Kish, S. J., Mash, D. C., & Tyndale, R. F. (2012). The neuroprotective enzyme CYP2D6 increases in the brain with age and is lower in parkinson's disease patients. *Neurobiol. Aging*, *33*(9), 2160–2171. http://doi.org/10.1016/j.neurobiolaging.2011.08.014

Mannick, J. B., Del Giudice, G., Lattanzi, M., Valiante, N. M., Praestgaard, J., Huang, B., … Klickstein, L. B. (2014). mTOR inhibition improves immune function in the elderly. *Sci. Transl. Med.*, *6*(268), 268ra179. http://doi.org/10.1126/scitranslmed.3009892

Marengoni, A., Angleman, S., Melis, R., Mangialasche, F., Karp, A., Garmen, A., … Fratiglioni, L. (2011). Aging with multimorbidity: A systematic review of the literature. *Ageing Research Reviews*, *10*(4), 430–439. http://doi.org/10.1016/j.arr.2011.03.003

Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, *349*(6255), 1483–1489. http://doi.org/10.1126/science.aab4082

Martinez-Jimenez, C. P., Eling, N., Chen, H.-C., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., … Odom, D. T. (2017). Aging increases cell-to-cell transcriptional

variability upon immune stimulation. *Science*, *355*(6332), 1433–1436. http://doi.org/10.1126/science.aah4115

Martins, R., Lithgow, G. J., & Link, W. (2016). Long live FOXO: unraveling the role of FOXO proteins in aging and longevity. *Aging Cell*, *15*(2), 196–207. http://doi.org/10.1111/acel.12427

Marttila, S., Chatsirisupachai, K., Palmer, D., & Magalhães, J. P. de. (2020). Ageing-associated changes in the expression of lncRNAs in human tissues reflect a transcriptional modulation in ageing pathways. *Mech. Ageing Dev.*, *185*, 111177. http://doi.org/10.1016/j.mad.2019.111177

Matheu, A., Maraver, A., Collado, M., Garcia-Cao, I., Cañamero, M., Borras, C., … Serrano, M. (2009). Anti-aging activity of the Ink4/Arf locus. *Aging Cell*, *8*(2), 152–161. http://doi.org/10.1111/j.1474-9726.2009.00458.x

Mattison, J. A., Roth, G. S., Beasley, T. M., Tilmont, E. M., Handy, A. M., Herbert, R. L., … Cabo, R. de. (2012). Impact of caloric restriction on health and survival in rhesus monkeys from the NIA study. *Nature*, *489*(7415), 318–321. http://doi.org/10.1038/nature11432

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., … Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, *31*(1), 374–8. http://doi.org/10.1093/nar/gkg108

Max Roser, E. O.-O., & Ritchie, H. (2020). Life expectancy. *Our World in Data*.

Maycox, P. R., Kelly, F., Taylor, A., Bates, S., Reid, J., Logendra, R., … Belleroche, J. de. (2009). Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol. Psychiatry*, *14*(12), 1083–1094. http://doi.org/10.1038/mp.2009.18

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C.-S., Jan, Y. N., … Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat. Genet.*, *36*(2), 197–204. http://doi.org/10.1038/ng1291

McClellan, A. J., Xia, Y., Deutschbauer, A. M., Davis, R. W., Gerstein, M., & Frydman, J. (2007). Diverse cellular functions of the hsp90 molecular chaperone uncovered using systems approaches. *Cell*, *131*(1), 121–135. http://doi.org/10.1016/j.cell.2007.07.036

McDonald, P., Maizi, B. M., & Arking, R. (2013). Chemical regulation of mid- and late-life longevities in drosophila. *Exp. Gerontol.*, *48*(2), 240–249. http://doi.org/10.1016/j.exger.2012.09.006

Medawar, P. B. (1953). Unsolved problem of biology. *Med. J. Aust.*, *1*(24), 854–855.

Medvedev, Z. A. (1990). An attempt at a rational classification of theories of ageing. *Biol. Rev. Camb. Philos. Soc.*, *65*(3), 375–398.

Menni, C., Kastenmüller, G., Petersen, A. K., Bell, J. T., Psatha, M., Tsai, P.-C., … Valdes, A. M. (2013). Metabolomic markers reveal novel pathways of ageing and early development in human populations. *Int. J. Epidemiol.*, *42*(4), 1111–1119. http://doi.org/10.1093/ije/dyt094

Mering, C. von, Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., … Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, *33*(Database issue), D433–7. http://doi.org/10.1093/nar/gki005

Metformin. (n.d.). https://www.drugbank.ca/drugs/DB00331.

MHC region of the human genome - genome reference consortium. (n.d.). https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37.

Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., … Lein, E. S. (2014). Transcriptional landscape of the prenatal human brain. *Nature*, *508*(7495), 199–206. http://doi.org/10.1038/nature13185

Mofidifar, S., Sohraby, F., Bagheri, M., & Aryapour, H. (2018). Repurposing existing drugs for new AMPK activators as a strategy to extend lifespan: a computer-aided drug discovery study. *Biogerontology*. http://doi.org/10.1007/s10522-018-9744-x

Montero, P., & Vilar, J. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software, Articles*, *62*(1), 1–43. http://doi.org/10.18637/jss.v062.i01

Morrison, J. H., & Baxter, M. G. (2012). The ageing cortical synapse: hallmarks and implications for cognitive decline. *Nature Reviews. Neuroscience*, *13*(4), 240–250. http://doi.org/10.1038/nrn3200

Moskalev, A., Chernyagina, E., Magalhães, J. P. de, Barardo, D., Thoppil, H., Shaposhnikov, M., … Zhavoronkov, A. (2015). Geroprotectors.org: A new, structured and curated database of current therapeutic interventions in aging and age-related disease. *Aging*, *7*(9), 616–628. http://doi.org/10.18632/aging.100799

Mukherjee, S., Date, A., Patravale, V., Korting, H. C., Roeder, A., & Weindl, G. (2006). Retinoids in the treatment of skin aging: An overview of clinical efficacy and safety. *Clin. Interv. Aging*, *1*(4), 327–348. http://doi.org/10.2147/ciia.2006.1.4.327

Mungan, Z., & Pınarbaşı Şimşek, B. (2017). Which drugs are risk factors for the development of gastroesophageal reflux disease? *Turk. J. Gastroenterol.*, *28*(Suppl 1), S38–S43. http://doi.org/10.5152/tjg.2017.11

Murabito, J. M., Yuan, R., & Lunetta, K. L. (2012). The search for longevity and healthy aging genes: Insights from epidemiological studies and samples of long-lived individuals. *J. Gerontol. A Biol. Sci. Med. Sci.*, *67*(5), 470–479. http://doi.org/10.1093/gerona/gls089

Nadon, N. L., Strong, R., Miller, R. A., & Harrison, D. E. (2016). In Focus NIA Interventions Testing Program: Investigating Putative Aging Intervention Agents in a Genetically Heterogeneous Mouse Model. http://doi.org/10.1016/j.ebiom.2016.11.038

Naumova, O. Y., Palejev, D., Vlasova, N. V., Lee, M., Rychkov, S. Y., Babich, O. N., … Grigorenko, E. L. (2012). Age-related changes of gene expression in the neocortex: Preliminary data on RNA-Seq of the transcriptome in three functionally distinct cortical areas. *Dev. Psychopathol.*, *24*(4), 1427–1442. http://doi.org/10.1017/S0954579412000818

Nelson, G., Wordsworth, J., Wang, C., Jurk, D., Lawless, C., Martin-Ruiz, C., & Zglinicki, T. von. (2012). A senescent cell bystander effect: Senescence-induced senescence. *Aging Cell*, *11*(2), 345–349. http://doi.org/10.1111/j.1474-9726.2012.00795.x

Niccoli, T., & Partridge, L. (2012). Ageing as a risk factor for disease. *Curr. Biol.*, *22*(17), R741–52. http://doi.org/10.1016/j.cub.2012.07.024

Nishimura, M., Ocorr, K., Bodmer, R., & Cartry, J. (2011). Drosophila as a model to study cardiac aging. *Exp. Gerontol.*, *46*(5), 326–330. http://doi.org/10.1016/j.exger.2010.11.035

Novelle, M. G., Wahl, D., Diéguez, C., Bernier, M., & Cabo, R. de. (2015). Resveratrol supplementation: Where are we now and where should we go? *Ageing Res. Rev.*, *21*, 1–15. http://doi.org/10.1016/j.arr.2015.01.002

Nussey, D. H., Froy, H., Lemaitre, J.-F., Gaillard, J.-M., & Austad, S. N. (2013). Senescence in natural populations of animals: Widespread evidence and its implications for bio-gerontology. *Ageing Res. Rev.*, *12*(1), 214–225. http://doi.org/10.1016/j.arr.2012.07.004

Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., & Morgan, M. (2014). VariantAnnotation: A bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, *30*(14), 2076–2078. http://doi.org/10.1093/bioinformatics/btu168

O'Connor, L. J., & Price, A. L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat. Genet.*, *50*(12), 1728–1734. http://doi.org/10.1038/s41588-018-0255-0

Oeppen, J., & Vaupel, J. W. (2002). Demography. Broken limits to life expectancy. *Science*, *296*(5570), 1029–1031. http://doi.org/10.1126/science.1069675

Olovnikov, A. M. (1996). Telomeres, telomerase, and aging: Origin of the theory. *Exp. Gerontol.*, *31*(4), 443–448. http://doi.org/10.1016/0531-5565(96)00005-8

Olshansky, S. J. (2016). Ageing: Measuring our narrow strip of life. *Nature*, *538*(7624), 175–176. http://doi.org/10.1038/nature19475

Olshansky, S. J., Carnes, B. A., & Cassel, C. (1990). In search of methuselah: Estimating the upper limits to human longevity. *Science*, *250*(4981), 634–640. http://doi.org/10.1126/science.2237414

Olshansky, S. J., Carnes, B. A., & Désesquelles, A. (2001). Demography. Prospects for human longevity. *Science*, *291*(5508), 1491–1492. http://doi.org/10.1126/science.291.5508.1491

Ori, A., Toyama, B. H., Harris, M. S., Bock, T., Iskar, M., Bork, P., … Beck, M. (2015). Integrated transcriptome and proteome analyses reveal Organ-Specific

proteome deterioration in old rats. *Cell Systems*. http://doi.org/10.1016/j.cels.2015.08.012

O'Toole, P. W., & Jeffery, I. B. (2015). Gut microbiota and aging. *Science*, *350*(6265), 1214–1215. http://doi.org/10.1126/science.aac8469

Pal, S., & Tyler, J. K. (2016). Epigenetics and aging. *Sci Adv*, *2*(7), e1600584. http://doi.org/10.1126/sciadv.1600584

Pan, H., & Finkel, T. (2017). Key proteins and pathways that regulate lifespan. *J. Biol. Chem.*, *292*(16), 6452–6460. http://doi.org/10.1074/jbc.R116.771915

Panagiotou, O. A., Ioannidis, J. P. A., & Genome-Wide Significance Project. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.*, *41*(1), 273–286. http://doi.org/10.1093/ije/dyr178

Parameswaran Nair, N., Chalmers, L., Peterson, G. M., Bereznicki, B. J., Castelino, R. L., & Bereznicki, L. R. (2016). Hospitalization in older patients due to adverse drug reactions -the need for a prediction tool. *Clinical Interventions in Aging*, *11*, 497–505. http://doi.org/10.2147/CIA.S99097

Park, J., Lee, D.-S., Christakis, N. A., & Barabási, A.-L. (2009). The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.*, *5*, 262. http://doi.org/10.1038/msb.2009.16

Parkes, M., Cortes, A., Heel, D. A. van, & Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat. Rev. Genet.*, *14*(9), 661–673. http://doi.org/10.1038/nrg3502

Partridge, L., Deelen, J., & Slagboom, P. E. (2018). Facing up to the global challenges of ageing. *Nature*, *561*(7721), 45–56. http://doi.org/10.1038/s41586-018-0457-8

Pearson, K. J., Baur, J. A., Lewis, K. N., Peshkin, L., Price, N. L., Labinskyy, N., … Cabo, R. de. (2008). Resveratrol delays age-related deterioration and mimics transcriptional aspects of dietary restriction without extending life span. *Cell Metab.*, *8*(2), 157–168. http://doi.org/10.1016/j.cmet.2008.06.011

Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. (2007). Estimation of the multiple testing burden for genomewide association studies of common variants. *Nature Precedings*. http://doi.org/10.1038/npre.2007.359.1

Peters, M. J., Joehanes, R., Pilling, L. C., Schurmann, C., Conneely, K. N., Powell, J., … Johnson, A. D. (2015). The transcriptional landscape of age in human peripheral blood. *Nat. Commun.*, *6*, 8570. http://doi.org/10.1038/ncomms9570

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., & Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.*, *48*(7), 709–717. http://doi.org/10.1038/ng.3570

Piegholdt, S., Rimbach, G., & Wagner, A. E. (2016). The phytoestrogen prunetin affects body composition and improves fitness and lifespan in male drosophila melanogaster. *FASEB J.*, *30*(2), 948–958. http://doi.org/10.1096/fj.15-282061

Pilling, L. C., Kuo, C.-L., Sicinski, K., Tamosauskaite, J., Kuchel, G. A., Harries, L. W., … Melzer, D. (2017). Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging*, *9*(12), 2504–2520. http://doi.org/10.18632/aging.101334

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., & Furlong, L. I. (2019). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* http://doi.org/10.1093/nar/gkz1021

Poduri, A., Evrony, G. D., Cai, X., & Walsh, C. A. (2013). Somatic mutation, genomic variation, and neurological disease. *Science*, *341*(6141), 1237758. http://doi.org/10.1126/science.1237758

Polleux, F., Ince-Dunn, G., & Ghosh, A. (2007). Transcriptional regulation of vertebrate axon guidance and synapse formation. *Nature Reviews Neuroscience*, *8*(5), 331–340. http://doi.org/10.1038/nrn2118

Poulain, M., Herm, A., & Pes, G. (2013). The blue zones: Areas of exceptional longevity around the world. *Vienna Yearb. Popul. Res.*, *11*, 87–108.

Rakyan, V. K., Down, T. A., Maslau, S., Andrew, T., Yang, T.-P., Beyan, H., … Spector, T. D. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.*, *20*(4), 434–439. http://doi.org/10.1101/gr.103101.109

Rando, T. A., & Chang, H. Y. (2012). Aging, rejuvenation, and epigenetic reprogramming: Resetting the aging clock. *Cell*, *148*(1-2), 46–57. http://doi.org/10.1016/j.cell.2012.01.003

Resveratrol. (n.d.). https://www.drugbank.ca/drugs/DB02709.

Risques, R. A., & Kennedy, S. R. (2018). Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet.*, *14*(1), e1007108. http://doi.org/10.1371/journal.pgen.1007108

Rissman, R. A., De Blas, A. L., & Armstrong, D. M. (2007). GABA(A) receptors in aging and alzheimer's disease. *J. Neurochem.*, *103*(4), 1285–1292. http://doi.org/10.1111/j.1471-4159.2007.04832.x

Robida-Stubbs, S., Glover-Cutter, K., Lamming, D. W., Mizunuma, M., Narasimhan, S. D., Neumann-Haefelin, E., … Blackwell, T. K. (2012). TOR signaling and rapamycin influence longevity by regulating SKN-1/Nrf and DAF-16/FoxO. *Cell Metab.*, *15*(5), 713–724. http://doi.org/10.1016/j.cmet.2012.04.007

Rodier, F., & Campisi, J. (2011). Four faces of cellular senescence. *J. Cell Biol.*, *192*(4), 547–556. http://doi.org/10.1083/jcb.201009094

Rodin, R. E., & Walsh, C. A. (2018). Somatic mutation in pediatric neurological diseases. *Pediatr. Neurol.*, *87*, 20–22. http://doi.org/10.1016/j.pediatrneurol.2018.08.008

Rodríguez, J. A., Marigorta, U. M., Hughes, D. A., Spataro, N., Bosch, E., & Navarro, A. (2017). Antagonistic pleiotropy and mutation accumulation influence human senescence and disease. *Nat Ecol Evol*, *1*(3), 55. http://doi.org/10.1038/s41559-016-0055

Rossi, D. J., Bryder, D., Seita, J., Nussenzweig, A., Hoeijmakers, J., & Weissman, I. L. (2007). Deficiencies in DNA damage repair limit the function of haematopoietic stem cells with age. *Nature*, *447*(7145), 725–729. http://doi.org/10.1038/nature05862

Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., & Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, *2016*, baw100. http://doi.org/10.1093/database/baw100

Rubinsztein, D. C., Mariño, G., & Kroemer, G. (2011). Autophagy and Aging. *Cell*, *146*(5), 682–695. http://doi.org/10.1016/j.cell.2011.07.030

Russell, S. J., & Kahn, C. R. (2007). Endocrine regulation of ageing. *Nat. Rev. Mol. Cell Biol.*, *8*(9), 681–691. http://doi.org/10.1038/nrm2234

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiol. Aging*, *30*(4), 507–514. http://doi.org/10.1016/j.neurobiolaging.2008.09.023

Sanchez-Valle, J., Tejero, H., Fernandez, J. M., Juan, D., Capella, S., Al-Shahrour, F., … Valencia, A. (2018, October). *Unveiling the molecular basis of disease co-occurrence: Towards personalized comorbidity profiles*. *bioRxiv*. http://doi.org/10.1101/431312

Schmid, M., Malicki, D., Nobori, T., Rosenbach, M. D., Campbell, K., Carson, D. A., & Carrera, C. J. (1998). Homozygous deletions of methylthioadenosine phosphorylase (MTAP) are more frequent than p16INK4A (CDKN2) homozygous deletions in primary non-small cell lung cancers (NSCLC). *Oncogene*, *17*(20), 2669–2675. http://doi.org/10.1038/sj.onc.1202205

Schratt, G. (2009). microRNAs at the synapse. *Nature Reviews Neuroscience*, *10*(12), 842–849. http://doi.org/10.1038/nrn2763

Seidel, J., & Valenzano, D. R. (2018). The role of the gut microbiome during host ageing. *F1000Res.*, *7*. http://doi.org/10.12688/f1000research.15121.1

Seim, I., Ma, S., Zhou, X., Gerashchenko, M. V., Lee, S.-G., Suydam, R., … Gladyshev, V. N. (2014). The transcriptome of the bowhead whale balaena mysticetus reveals adaptations of the longest-lived mammal. *Aging*, *6*(10), 879–899. http://doi.org/10.18632/aging.100699

Shamanna, R. A., Lu, H., Croteau, D. L., Arora, A., Agarwal, D., Ball, G., … Bohr, V. A. (2016). Camptothecin targets WRN protein: Mechanism and relevance in clinical breast cancer. *Oncotarget*, *7*(12), 13269–13284. http://doi.org/10.18632/oncotarget.7906

Shay, J. W. (2018). Telomeres and aging. *Curr. Opin. Cell Biol.*, *52*, 1–7. http://doi.org/10.1016/j.ceb.2017.12.001

Sirolimus. (n.d.). https://www.drugbank.ca/drugs/DB00877.

Slieker, R. C., Iterson, M. van, Luijk, R., Beekman, M., Zhernakova, D. V., Moed, M. H., … Heijmans, B. T. (2016). Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.*, *17*(1), 191. http://doi.org/10.1186/s13059-016-1053-6

Smith, J. M. (1976). Group selection. *Q. Rev. Biol.*, *51*(2), 277–283.

Smith, P., Willemsen, D., Popkes, M., Metge, F., Gandiwa, E., Reichard, M., & Valenzano, D. R. (2017). Regulation of life span by the gut microbiota in the short-lived african turquoise killifish. *Elife*, *6*. http://doi.org/10.7554/eLife.27014

Snell, T. W., Johnston, R. K., Matthews, A. B., Zhou, H., Gao, M., & Skolnick, J. (2018). Repurposed FDA-approved drugs targeting genes influencing aging can extend lifespan and healthspan in rotifers. *Biogerontology*. http://doi.org/10.1007/s10522-018-9745-9

Snell, T. W., Johnston, R. K., Srinivasan, B., Zhou, H., Gao, M., & Skolnick, J. (2016). Repurposing FDA-approved drugs for anti-aging therapies. *Biogerontology*, *17*(5-6), 907–920. http://doi.org/10.1007/s10522-016-9660-x

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nat. Rev. Genet.*, *14*(7), 483–495. http://doi.org/10.1038/nrg3461

Somel, M., Guo, S., Fu, N., Yan, Z., Hu, H. Y., Xu, Y., … Khaitovich, P. (2010). MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.*, *20*(9), 1207–1218. http://doi.org/10.1101/gr.106849.110

Somel, M., Khaitovich, P., Bahn, S., Pääbo, S., & Lachmann, M. (2006). Gene expression becomes heterogeneous with age. *Curr. Biol.*, *16*(10), R359–60. http://doi.org/10.1016/j.cub.2006.04.024

Somel, M., Liu, X., Tang, L., Yan, Z., Hu, H., Guo, S., … Khaitovich, P. (2011). MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biol.*, *9*(12), e1001214. http://doi.org/10.1371/journal.pbio.1001214

Sousa-Franco, A., Rebelo, K., Rocha, S. T. da, & Bernardes de Jesus, B. (2019). LncRNAs regulating stemness in aging. *Aging Cell*, *18*(1), e12870.

Sowell, E. R., Thompson, P. M., & Toga, A. W. (2004). Mapping Changes in the Human Cortex throughout the Span of Life. *The Neuroscientist*, *10*(4), 372–392. http://doi.org/10.1177/1073858404263960

Stefani, G., & Slack, F. J. (2008). Small non-coding RNAs in animal development. *Nature Reviews Molecular Cell Biology*, *9*(3), 219–230. http://doi.org/10.1038/nrm2347

Stegeman, R., & Weake, V. M. (2017). Transcriptional signatures of aging. *J. Mol. Biol.*, *429*(16), 2427–2437. http://doi.org/10.1016/j.jmb.2017.06.019

Stephenson, J. D., Laskowski, R. A., Nightingale, A., Hurles, M. E., & Thornton, J. M. (2019). VarMap: A web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/btz482

Strong, R., Miller, R. A., Astle, C. M., Baur, J. A., Cabo, R. de, Fernandez, E., … Harrison, D. E. (2013). Evaluation of resveratrol, green tea extract, curcumin, oxaloacetic acid, and medium-chain triglyceride oil on life span of genetically heterogeneous mice. *J. Gerontol. A Biol. Sci. Med. Sci.*, *68*(1), 6–16. http://doi.org/10.1093/gerona/gls070

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, *102*(43), 15545–15550. http://doi.org/10.1073/pnas.0506580102

Szklarczyk, D., Santos, A., Von Mering, C., Jensen, L. J., Bork, P., & Kuhn, M. (2016). STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, *44*(D1), D380–D384. http://doi.org/10.1093/nar/gkv1277

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., … Magalhães, J. P. de. (2018). Human ageing genomic resources: New and updated databases. *Nucleic Acids Res.*, *46*(D1), D1083–D1090. http://doi.org/10.1093/nar/gkx1042

Takauji, Y., Wada, T., Takeda, A., Kudo, I., Miki, K., Fujii, M., & Ayusawa, D. (2016). Restriction of protein synthesis abolishes senescence features at cellular and organismal levels. *Sci. Rep.*, *6*, 18722. http://doi.org/10.1038/srep18722

Talens, R. P., Christensen, K., Putter, H., Willemsen, G., Christiansen, L., Kremer, D., … Heijmans, B. T. (2012). Epigenetic variation during the adult lifespan: Cross-sectional and longitudinal data on monozygotic twin pairs. *Aging Cell*, *11*(4), 694–703. http://doi.org/10.1111/j.1474-9726.2012.00835.x

Tebbenkamp, A. T. N., Willsey, A. J., State, M. W., & Šestan, N. (2014). The developmental transcriptome of the human brain. *Current Opinion in Neurology*, *27*(2), 149–156. http://doi.org/10.1097/WCO.0000000000000069

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., … Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.*, *20*(4), 440–446. http://doi.org/10.1101/gr.103606.109

Teschendorff, A. E., West, J., & Beck, S. (2013). Age-associated epigenetic drift: Implications, and a case of epigenetic thrift? *Hum. Mol. Genet.*, *22*(R1), R7–R15. http://doi.org/10.1093/hmg/ddt375

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. http://doi.org/10.1093/nar/gky1055

The Gene Ontology Consortium, & The Gene Ontology Consortium. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gky1055

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. http://doi.org/10.1111/1467-9868.00293

Tsurumi, A., & Li, W. X. (2012). Global heterochromatin loss: A unifying theory of aging? *Epigenetics*, *7*(7), 680–688. http://doi.org/10.4161/epi.20540

Valdes, A. M., Glass, D., & Spector, T. D. (2013). Omics technologies and the study of human ageing. *Nat. Rev. Genet.*, *14*(9), 601–607. http://doi.org/10.1038/nrg3553

Van Cauwenberghe, C., Van Broeckhoven, C., & Sleegers, K. (2016). The genetic landscape of alzheimer disease: Clinical implications and perspectives. *Genet. Med.*, *18*(5), 421–430. http://doi.org/10.1038/gim.2015.117

Vaughan, K. L., Kaiser, T., Peaden, R., Anson, R. M., Cabo, R. de, & Mattison, J. A. (2017). Caloric restriction study design limitations in rodent and nonhuman primate studies. *J. Gerontol. A Biol. Sci. Med. Sci.*, *73*(1), 48–53. http://doi.org/10.1093/gerona/glx088

Vaupel, J. W., Carey, J. R., & Christensen, K. (2003). Aging. It's never too late. *Science*, *301*(5640), 1679–1681. http://doi.org/10.1126/science.1090529

Vellai, T., Takacs-Vellai, K., Zhang, Y., Kovacs, A. L., Orosz, L., & Müller, F. (2003). Genetics: Influence of TOR kinase on lifespan in c. Elegans. *Nature*, *426*(6967), 620. http://doi.org/10.1038/426620a

Victoria, B., Nunez Lopez, Y. O., & Masternak, M. M. (2017). MicroRNAs and the metabolic hallmarks of aging. *Mol. Cell. Endocrinol.*, *455*, 131–147. http://doi.org/10.1016/j.mce.2016.12.021

Vijg, J. (2004). Impact of genome instability on transcription regulation of aging and senescence. *Mechanisms of Ageing and Development*, *125*(10-11), 747–753. http://doi.org/10.1016/j.mad.2004.07.004

Villeda, S. A., Luo, J., Mosher, K. I., Zou, B., Britschgi, M., Bieri, G., … Wyss-Coray, T. (2011). The ageing systemic milieu negatively regulates neurogenesis and cognitive function. *Nature*, *477*(7362), 90–94. http://doi.org/10.1038/nature10357

Viñuela, A., Brown, A. A., Buil, A., Tsai, P.-C., Davies, M. N., Bell, J. T., … Small, K. S. (2018). Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Human Molecular Genetics*, *27*(4), 732–741. http://doi.org/10.1093/hmg/ddx424

Violan, C., Foguet-Boreu, Q., Flores-Mateo, G., Salisbury, C., Blom, J., Freitag, M., … Valderas, J. M. (2014). Prevalence, determinants and patterns of multimorbidity in primary care: A systematic review of observational studies. *PLoS ONE*, *9*(7), e102149. http://doi.org/10.1371/journal.pone.0102149

Vorinostat. (n.d.).

Waite, L. J. (2004). *Aging, health, and public policy: Demographic and economic perspectives* (Vol. 30). Population Council New York, NY, USA.

Walker, R., Gurven, M., Hill, K., Migliano, A., Chagnon, N., De Souza, R., … Yamauchi, T. (2006). Growth rates and life histories in twenty-two small-scale societies. *Am. J. Hum. Biol.*, *18*(3), 295–311. http://doi.org/10.1002/ajhb.20510

Wang, J., Zhang, S., Wang, Y., Chen, L., & Zhang, X.-S. (2009). Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. *PLoS Comput. Biol.*, *5*(9), e1000521. http://doi.org/10.1371/journal.pcbi.1000521

Wang, Z., Monteiro, C. D., Jagodnik, K. M., Fernandez, N. F., Gundersen, G. W., Rouillard, A. D., … Ma'ayan, A. (2016). Extraction and analysis of signatures

from the Gene Expression Omnibus by the crowd. *Nature Communications*, *7*, 12846. http://doi.org/10.1038/ncomms12846

Warren, L. A., Rossi, D. J., Schiebinger, G. R., Weissman, I. L., Kim, S. K., & Quake, S. R. (2007). Transcriptional instability is not a universal attribute of aging. *Aging Cell*. http://doi.org/10.1111/j.1474-9726.2007.00337.x

Wickham, H. (2017). ggplot2 – Elegant Graphics for Data Analysis. *Journal of Statistical Software*, *77*(April), 3–5. http://doi.org/10.18637/jss.v077.b02

Wiley, C. D., Flynn, J. M., Morrissey, C., Lebofsky, R., Shuga, J., Dong, X., … Campisi, J. (2017). Analysis of individual cells identifies cell-to-cell variability following induction of cellular senescence. *Aging Cell*. http://doi.org/10.1111/acel.12632

Williams, G. C. (1957). Pleiotropy, natural selection, and the evolution of senescence. *Evolution*, *11*(4), 398–411. http://doi.org/10.2307/2406060

Wolfson, M., Budovsky, A., Tacutu, R., & Fraifeld, V. (2009). The signaling hubs at the crossroad of longevity and age-related disease networks. *Int. J. Biochem. Cell Biol.*, *41*(3), 516–520. http://doi.org/10.1016/j.biocel.2008.08.026

World report on Ageing And Health. (2015).

Wright, K. M., Rand, K. A., Kermany, A., Noto, K., Curtis, D., Garrigan, D., … Ruby, J. G. (2019). A prospective analysis of genetic variants associated with human lifespan. *G3*, *9*(9), 2863–2878. http://doi.org/10.1534/g3.119.400448

Xiao, R., Zhang, B., Dong, Y., Gong, J., Xu, T., Liu, J., & Xu, X. Z. S. (2013). A genetic program promotes c. Elegans longevity at cold temperatures via a thermosensitive TRP channel. *Cell*, *152*(4), 806–817. http://doi.org/10.1016/j.cell.2013.01.020

Ximerakis, M., Lipnick, S. L., Innes, B. T., Simmons, S. K., Adiconis, X., Dionne, D., … Rubin, L. L. (2019). Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.*, *22*(10), 1696–1708. http://doi.org/10.1038/s41593-019-0491-3

Ximerakis, M., Lipnick, S. L., Simmons, S. K., Adiconis, X., Innes, B. T., Dionne, D., … Rubin, L. L. (2018). Single-cell transcriptomics of the aged mouse brain reveals convergent, divergent and unique aging signatures. *bioRxiv*, 440032. http://doi.org/10.1101/440032

Xu, J., Chang, W.-H., Fong, L. W. R., Weiss, R. H., Yu, S.-L., & Chen, C.-H. (2019). Targeting the insulin-like growth factor-1 receptor in MTAP-deficient renal cell carcinoma. *Signal Transduct Target Ther*, *4*, 2. http://doi.org/10.1038/s41392-019-0035-z

Xue, H., Xian, B., Dong, D., Xia, K., Zhu, S., Zhang, Z., … Han, J.-D. J. (2007). A modular network model of aging. *Mol. Syst. Biol.*, *3*, 147. http://doi.org/10.1038/msb4100189

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., … Tu, Z. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience*, *42*(1), 353–372. http://doi.org/10.1007/s11357-019-00106-x

Ye, X., Linton, J. M., Schork, N. J., Buck, L. B., & Petrascheck, M. (2014). A pharmacological network for lifespan extension in caenorhabditis elegans. *Aging Cell*, *13*(2), 206–215. http://doi.org/10.1111/acel.12163

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol.*, *11*(2), R14. http://doi.org/10.1186/gb-2010-11-2-r14

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS*, *16*(5), 284–287. http://doi.org/10.1089/omi.2011.0118

Zhang, C., & Cuervo, A. M. (2008). Restoration of chaperone-mediated autophagy in aging liver improves cellular maintenance and hepatic function. *Nat. Med.*, *14*(9), 959–965. http://doi.org/10.1038/nm.1851

Zhang, G., Li, J., Purkayastha, S., Tang, Y., Zhang, H., Yin, Y., … Cai, D. (2013). Hypothalamic programming of systemic ageing involving IKK-β, NF-κB and GnRH. *Nature*, *497*(7448), 211–216. http://doi.org/10.1038/nature12143

Zhang, L., & Vijg, J. (2018). Somatic mutagenesis in mammals and its implications for human disease and aging. *Annu. Rev. Genet.*, *52*, 397–419. http://doi.org/10.1146/annurev-genet-120417-031501

Zhou, H., Gao, M., & Skolnick, J. (2015). Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Scientific Reports*, *5*(1), 11090. http://doi.org/10.1038/srep11090

Zhu, Y., Wang, L., Yin, Y., & Yang, E. (2017). Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Scientific Reports*, *7*(1), 5435. http://doi.org/10.1038/s41598-017-05882-0

Ziehm, M., Kaur, S., Ivanov, D. K., Ballester, P. J., Marcus, D., Partridge, L., & Thornton, J. M. (2017). Drug repurposing for aging research using model organisms. *Aging Cell*, *16*(5), 1006–1015. http://doi.org/10.1111/acel.12626

Zierer, J., Menni, C., Kastenmüller, G., & Spector, T. D. (2015). Integration of 'omics' data in aging research: From biomarkers to systems biology. *Aging Cell*, *14*(6), 933–944. http://doi.org/10.1111/acel.12386