META-ANALYSIS OF GENE EXPRESSION HETEROGENEITY IN BRAIN DEVELOPMENT AND AGING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ULAŞ IŞILDAK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOLOGY

SEPTEMBER 2022

Approval of the thesis:

META-ANALYSIS OF GENE EXPRESSION HETEROGENEITY IN BRAIN DEVELOPMENT AND AGING

submitted by **ULAŞ IŞILDAK** in partial fulfillment of the requirements for the degree of **Master of Science in Biology Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşe Gül Gözen
Head of Department, **Biology** _____

Prof. Dr. Mehmet Somel
Supervisor, **Biology Dept., METU** _____

Examining Committee Members:

Assist. Prof. Dr. İdil Yet
Bioinformatics, Hacettepe University _____

Prof. Dr. Mehmet Somel
Biology, METU _____

Assist. Prof. Dr. Nihal Terzi Çizmecioğlu
Biology, METU _____

Date: 01.09.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Ulaş Işıldak

Signature :

ABSTRACT

META-ANALYSIS OF GENE EXPRESSION HETEROGENEITY IN BRAIN DEVELOPMENT AND AGING

Işıldak, Ulaş
M.S., Department of Biology
Supervisor: Prof. Dr. Mehmet Somel

September 2022, 93 pages

Aging is a complex process associated with the accumulation of stochastic genetic and epigenetic alterations, leading to functional decline and increased risk for disease and death. Although some previous studies demonstrated a tendency towards increased inter-individual heterogeneity during aging, whether it is a function of time that starts at the beginning of life is unknown. Its functional consequences and regulations have also not been systematically studied. In this study, I addressed these questions by the meta-analysis of 19 microarray age-series datasets, comprising 17 brain regions of 298 individuals. Investigating the age-related gene expression heterogeneity changes, I found that there is a significant shift towards increased heterogeneity consistency during aging (20 to 98 years of age) compared to the post-natal development period (0 to 20 years of age). Moreover, the genes that become more heterogeneous consistently across all aging datasets were found to be associated with biological processes and pathways that are related to neuronal function (i.e., axon guidance, postsynaptic specialization) and longevity (i.e., autophagy, mTOR signaling). Gene set enrichment analysis for transcriptional regulators (i.e., miRNAs and transcription factors) further revealed a positive correlation between the number of

regulators and consistent changes in heterogeneity, indicating the possible role of transcriptional regulators in the underlying mechanism.

Overall, the results presented here demonstrate that increased inter-individual expression heterogeneity is a general characteristic of the aging human brain, which is associated with multiple lifespans and disease-related pathways and processes, suggesting that increased heterogeneity may contribute to the emergence of aging-associated phenotypes.

Keywords: aging, development, gene expression, heterogeneity, brain

ÖZ

BEYİN GELİŞİMİ VE YAŞLANMASINDA GEN ANLATIMI HETEROJENLİĞİNİN META-ANALİZİ

Işıldak, Ulaş
Yüksek Lisans, Biyoloji Bölümü
Tez Yöneticisi: Prof. Dr. Mehmet Somel

Eylül 2022 , 93 sayfa

Yaşlanma, stokastik genetik ve epigenetik değişikliklerin birikimi ile ilişkili karmaşık bir süreçtir. Yaşa bağlı gen anlatımı değişikliklerini inceleyen önceki bazı çalışmalar, bireyler arası heterojenliğin artmasına yönelik bir eğilimin olduğunu gösterse de, bu artışın tam olarak ne zaman başladığı ve fonksiyonel sonuçlarının ne olduğu sistematik olarak çalışılmamıştır. Bu çalışmada, 298 bireyin 17 beyin bölgesini içeren 19 mikrodizi yaş serisi veri setinin meta-analizi ile bu soruları ele aldım. Yaşa bağlı gen anlatımı heterojenlik değişikliklerine inceleyerek, yaşlanma sırasında (20 ila 98 yaş aralığı) doğum sonrası gelişim dönemine (0-20 yaş aralığı) kıyasla heterojenlik tutarlılığının artışı yönünde istatistiksel olarak anlamlı bir kayma olduğu bulunmuştur. Bu sonuç artan heterojenliğin sadece zamanın genel etkisinden kaynaklanmadığını, daha ziyade yaşlanma döneminin spesifik bir etkisi olduğuna işaret ediyor. Ayrıca, bu tutarlı artış gösteren genlerin nöronal fonksiyon (akson rehberliği, postsinaptik uzmanlaşma) ve yaşlanmayla (otofaji, mTOR sinyali) ilgili biyolojik süreçler ve yolaklar ile ilişkili olduğu bulunmuştur. Transkripsiyonel regülatörler (miRNA'lar ve transkripsiyon faktörleri) için gen seti zenginleştirme analizi ayrıca şunu ortaya koydu: artan

heterojenlik sadece belirli regülatörlerle ilişkili değil, aynı zamanda regülatörlerin sayısı ile heterojenlik artışı arasında da pozitif bir ilişki var.

Genel olarak, burada sunulan sonuçlar, artan bireyler arası gen anlatımı heterojenliğinin yaşlanan insan beyninin genel bir özelliği olduğunu göstermektedir. Ayrıca, bu artışın yaşlanma ve hastalıklarla ilgili yolaklar ve biyolojik işlevlerle ilişkili olması, artan heterojenliğin yaşlanma ile ilişkili fenotiplerin ortaya çıkmasına katkıda bulunabileceğini düşündürmektedir.

Anahtar Kelimeler: yaşlanma, gelişim, gen anlatımı, heterojenlik, beyin

To Melfi

ACKNOWLEDGMENTS

This work would not have been possible without the help and support of many people. First and foremost, I would like to thank my supervisor Mehmet Somel for his guidance and support throughout my undergraduate and graduate studies. I am grateful for him giving me a chance to do research in his group. The lessons he taught me will help me a great deal in my career henceforth. I would also like to thank my unofficial co-advisor Handan Melike Dönertaş. This work would not have been possible without her exceptional guidance, support, and scientific contribution. I also want to express my gratitude to Janet Thornton for her critical comments and discussions.

I would like to thank all members of METU CompEvo Lab for their fruitful discussions and friendship. I should especially acknowledge Handan Melike Dönertaş, Hamit İzgi, Gözde Turan, and Poorya Parvizi for their continued willingness to discuss all the questions I had. I also want to thank METU CompEvo Lab and Kıvılcım Başak Vural for computational support.

I want to thank my family for their endless love and generosity. They always supported me and provided me with everything to successfully complete my Master's degree. I can never thank them enough for all of their sacrifices and support.

Lastly, İrem Güner deserves a special recognition for her constant love and companionship. I am grateful to her for sharing my excitement and standing my complaining when things didn't work out so well. This work would not be possible without her endless support.

The research I conducted as part of this thesis was previously published in (Isildak et al., 2020).

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

A1C	Primary Auditory Cortex
AMY	Amygdala
CBC	Cerebellum
DFC	Dorsolateral Prefrontal Cortex
HIP	Hippocampus
IPC	Posterior Inferior Parletal Cortex
ITC	Inferior Temporal Cortex
M1C	Primary Motor Cortex
MD	Mediodorsal Nucleusofthe Thalamus
MFC	Medial Prefrontal Cortex
OFC	Orbital Prefrontal Cortex
PFC	Prefrontal Cortex
S1C	Primary Somatosensory Cortex
STC	Superior Temporal Cortex
STR	Striatum
V1C	Primary Visual Cortex
VFC	Ventrolateral Prefrontal Cortex
RMA	Robust Multi-Array Analysis
PCA	Principal Component Analysis
NES	Normalized Enrichment Scores
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
PMI	Post-Mortem Interval

CHAPTER 1

INTRODUCTION

1.1 Aging

Aging can be defined as a time-dependent deterioration of multiple biological functions and processes. Decreasing the capacity of an organism to maintain homeostasis, aging is associated with an increased vulnerability to many diseases including cancer, cardiovascular and neurodegenerative disorders (Niccoli & Partridge, 2012).

Aging is accompanied by the accumulation of damage at different levels of an organism, from simple molecules to organs. It is well known that replication, transcription and translation are error-prone processes. Furthermore, the protective responses against these damages are also known to be vulnerable and tend to produce other forms of damages (Gladyshev, 2016). These damages are exacerbated during the aging period and further contribute to aging and age-associated diseases. While both stochastic and deterministic factors contribute to this damage, the deterministic (i.e., genetic) component is considered to be more important for distantly related species. The effect of the genetic component, for example, can be observed in the comparison of the lifespans of a human and a mouse living in the same environment. The stochastic component, on the other hand, is suggested to be an important factor in explaining aging-related differences within homogenous populations (Gladyshev, 2016). The stochastic factors account for the damage to nucleic acids (i.e., DNA), proteins, lipids and metabolites, as well as other age-related deleterious changes.

1.1.1 Start of aging period

Although in everyday life, 65 years-of-age is often used to denote the beginning of old age, the reason for choosing this age is actually historical, rather than being biological (Covey, 1992). In the biological context, however, the age-related changes were considered to manifest themselves after the organism reaches its maximum reproductive capacity (Vijg, 2009). In human populations, the age of 20 years approximately corresponds to the age of reproductive maturity (Walker et al., 2006).

Moreover, earlier studies also showed that the aging-associated structural changes in the human brain begin to exhibit themselves when individuals are in their 20s. These changes include declines in regional brain volumes (Sowell et al., 2003), myelin integrity (Sullivan & Pfefferbaum, 2006), cortical thickness (Magnotta, 1999; Salat, 2004), serotonin receptor binding (Sheline et al., 2002), and concentration of several brain metabolitees (Kadota et al., 2001; Salthouse, 2009). Recent studies analyzing transcriptional patterns in the human brain further revealed that the age of 20 years is a turning point in age-related gene expression trajectories, suggesting that it roughly corresponds to the point at which the aging-associated changes started (Colantuoni et al., 2011; Dönertaş et al., 2017; Somel et al., 2010).

1.1.2 Hallmarks of aging

In a 2013 paper, the authors reviewed 9 hallmarks of aging that were suggested to be the contributors to the emergence of aging-associated phenotypes (López-Otín et al., 2013). The authors used three criteria while considering candidate hallmarks. First, a hallmark must exhibit itself during normal aging. Second, experimental worsening of a hallmark must accelerate aging. Third, experimental amelioration of a hallmark must slow down the aging process.

Moreover, the hallmarks were further divided into 3 categories:

1. **Primary hallmarks**, including genomic instability, telomere attrition, epigenetic alterations and loss of proteostasis, were considered to be primary causes of damage.

2. **Antagonistic hallmarks**, including deregulated nutrient sensing, mitochondrial dysfunction and cellular senescence, were promoted (or accelerated) by the primary hallmarks, and they contribute to further accumulation of damages.

3. **Integrative hallmarks**, including stem cell exhaustion and altered intercellular communication, were suggested to arise due to the accumulation of damages caused by the primary and antagonistic hallmarks, and they further affect tissue homeostasis and function.

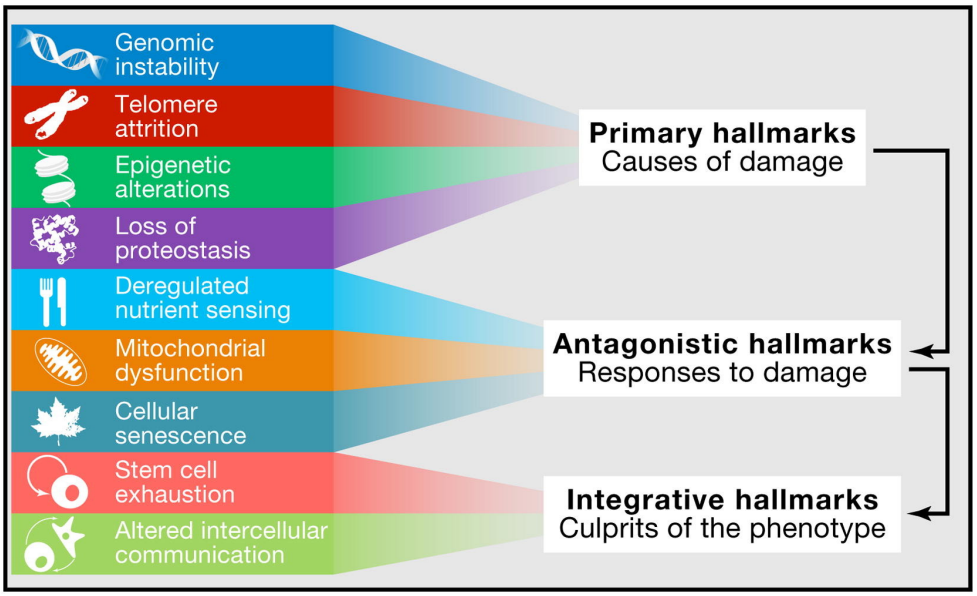

Figure 1.1: The hallmarks of aging (López-Otín et al., 2013).

Each hallmark of aging is explained in detail in the following subsections.

1.1.2.1 Genomic instability

Genomic instability was suggested to be one of the major stochastic mechanisms of aging. Many studies previously demonstrated the accumulation of somatic mutations during aging in humans and other model organisms (Lodato et al., 2018; Lombard et al., 2005; Lu et al., 2004; Moskalev et al., 2013; Vijg, 2004). The cumulative effect of somatic mutations hinders the normal functioning of essential genes and transcriptional pathways, contributing to aging and aging-associated diseases. Moreover, deficiencies in the DNA repair mechanism were found to be associated with acceler-

ated aging, whereas increased expression of DNA repair genes resulted in extended lifespan (Stead & Bjedov, 2021; Symphorien & Woodruff, 2003).

1.1.2.2 Telomere attrition

Telomeres are repetitive and protective DNA sequences found at the end of the chromosomes. Since most mammalian cells do not express telomerase, an enzyme responsible for maintaining telomere length, telomeres tend to shorten as the cell divide (i.e., as the organism age). Therefore, telomere length limits the proliferation capacity of somatic cells that do not express telomerase enzyme (Blasco, 2007). Moreover, telomere dysfunction was shown to be associated with accelerated aging (Armanios et al., 2009), whereas experimental induction of telomerase was suggested to lead to delayed aging (Tomás-Loba et al., 2008).

1.1.2.3 Epigenetic alterations

A number of epigenetic alterations including alterations in DNA methylation patterns, chromatin remodeling and post-transcriptional modifications of histones, were suggested to constitute aging-associated epigenetic marks, significantly affecting the normal functioning of cells. The most notable effect of age-related epigenetic alterations is on transcriptional outcomes, given the key role of epigenetic factors in transcriptional regulation. It was suggested that epigenetic alterations may cause abnormal production and maturation of some mRNAs, and even further lead to increased transcriptional variation (Ashapkin et al., 2017).

1.1.2.4 Loss of proteostasis

Proteostasis, mechanisms that involve preserving the stability and functionality of the proteome, is suggested to be altered during aging (Koga et al., 2011). Age-associated impairment of proteostasis leads to continuous expression of misfolded and aggregated proteins, whose accumulation during aging further contributes to the development of age-associated diseases, including Alzheimer's disease and Parkin-

son's disease (Powers et al., 2009).

1.1.2.5 Deregulated nutrient sensing

Nutrient-sensing pathways play important role in sensing the presence or absence of extra- and intra-cellular nutrients, and they further regulate their intake. The Insulin/Insulin-like Growth Factor Signaling (IIS) pathway is one of the nutrient-sensing pathways that was found to be regulating aging. The downstream intracellular effectors of the IIS pathway include AKT, mTOR and FoxO, all of which were suggested to be associated with aging (Barzilai et al., 2012; Fontana et al., 2010; Kenyon, 2010). Decreased activity of the IIS and mTOR signaling pathways, for example, was found to extend the lifespan in many model organisms (Fontana et al., 2010).

1.1.2.6 Mitochondrial dysfunction

During normal aging, the mitochondrial machinery becomes rusty, leading to increased electron leakage and reduced ATP generation (Green et al., 2011). In addition to accumulation of damages in the nuclear DNA (see **Section 1.1.2.1**), mitochondrial DNA (mtDNA) is also considered to be vulnerable to somatic mutations due to limited repair mechanisms and oxidative microenvironment, leading to impaired functionality of mitochondria. Moreover, an age-related increase in reactive oxygen species (ROS) was suggested to cause global cellular damage after a certain threshold, further contributing to the emergence of aging-associated phenotypes (Hekimi et al., 2011).

1.1.2.7 Cellular senescence

Cellular senescence can be defined as permanent arrest of the cell cycle. It can be triggered by a number of factors including, telomere shortening, DNA damage and a number of mitogenic alterations. The number of senescent cells was shown to increase with age. Although cellular senescence is originally a protective mechanism preventing the proliferation of damaged cells, their accumulation during aging results

in deleterious effects on tissue homeostasis, further contributing to aging (Biran et al., 2017; Yousefzadeh et al., 2020).

1.1.2.8 Stem cell exhaustion

It is long known that aging is accompanied by a decline in stem cell numbers and renewal capacity, contributing to the declined homeostatic and regenerative capacity of aged tissues (Oh et al., 2014). The factors contributing to stem cell exhaustion include DNA damage (**Section 1.1.2.1**), epigenetic alterations (**Section 1.1.2.3**), aggregation of damaged proteins (**Section 1.1.2.4**), accumulation of toxic metabolites (i.e., ROS) and mithochondrial dysfunction (**Section 1.1.2.6**). Moreover, experimental rejuvenation of stem cells was found to reset the aging clock, meaning that it has potential to reverse aging-associated phenotypes (Rando & Chang, 2012).

1.1.2.9 Altered intercellular communication

In addition to cell-autonomous hallmarks, the last hallmark of aging according to López-Otín et al. (2013) is related to altered communication of cells in terms of endocrine, neuroendocrine and neuronal signaling (Russell & Kahn, 2007). A harmonious intercellular communication was suggested to be a key factor for stress response, cell survival and maintaining homeostasis (Tan et al., 2021). The aging period is associated with increased inflammatory reactions, decreased immunosurveillance and a changed extracellular environment, all of which contribute to the deregulation of neurohormonal signaling. Specifically, senescent cells were shown to produce an altered secretome, which is rich in proinflammatory cytokines, which in turn contributes to the emergence of aging-associated phenotypes (Childs et al., 2016; Kuilman et al., 2010).

1.2 Age-related gene expression changes

As high-throughput technologies become more affordable and widely accessible, the past two decades have seen a dramatic increase in studies that focus on gene expres-

sion changes in the brain during the aging period. One of the earlier studies conducted by Lu and colleagues found that the expression levels of the genes that play important role in synaptic plasticity and mitochondrial tended to decrease in aging human brain (Lu et al., 2004). They further demonstrated that this decrease is also accompanied by increased promotor damage, suggesting that DNA damage may reduce the expression of genes involved in neuronal functioning, possibly contributing to the emergence of aging-associated pathologies.

A 2008 study, employing a microarray-based approach, found that the majority of genes in the aging human brain tend to decrease in expression (Berchtold et al., 2008). However, they also identified a set of up-regulated genes during aging, which was found to be related to immune activation and inflammation. These results were also confirmed by an independent study, where they found that, although the majority of genes were down-regulated during aging, the genes involved in immune and inflammatory responses were found to increase in expression (Lu et al., 2004). Combined, dysregulation of immune system genes might be a characteristic of the aging human brain (Frenk & Houseley, 2018), and it may be associated with aging-related phenotypes in the human brain.

Another common gene expression signature of aging is the downregulation of genes encoding mitochondrial ribosomal proteins and components of the electron transport chain. In their 2013 study, Kumar *et al.* analyzed microarray data generated from the frontal lobe of the cerebral cortex and cerebellum, and found that genes encoding mitochondrial components tend to be downregulated during aging (Kumar et al., 2013). Similar trends were also observed in other model organisms including rodents, flies and worms, and across different tissues from skin to muscle (Frenk & Houseley, 2018), suggesting that the downregulation of genes encoding mitochondrial proteins may be a characteristic of aging, and may contribute to the age-related mitochondrial dysfunction (**Section 1.1.2.6**).

Other studies investigated the relationship between the age-related gene expression change patterns during development and aging. Somel *et al.* analyzed gene and miRNA expression trajectories and found that the majority of expression changes observed in the aging period represent reversals or extensions of developmental pat-

terns (Somel et al., 2010). Colantuoni *et al.* also found a similar effect in human brain during development (Colantuoni et al., 2011). In a more recent study, the authors performed a meta-analysis on gene expression reversals and identified a set of genes showing the up-down pattern (i.e., increase during development, decrease during aging). They further demonstrated that these genes are involved in neuronal and synaptic functions, suggesting that decreased expression levels during aging may be associated with the stochastic nature of aging (Dönertaş et al., 2017).

There are also more recent studies that investigated age-related gene expression change patterns using RNA-Sequencing based approaches. Yang *et al.*, for example, analyzed RNA-Sequencing data across different human tissues and identified a set of age-associated genes (Yang et al., 2015). They further showed that aging-associated genes showing down-regulation were associated with mitochondrial function, whereas up-regulated age-associated genes were found to be associated with cell death and inflammation reponse (Yang et al., 2015). Another RNA-Seq based study also found that the genes that are down-regulated during aging were associated with neuronal development and the transmission of nerve impulses (Naumova et al., 2012). Thus, these studies demonstrate that independent RNA-Seq based approaches also confirm the previous results obtained from microarray-based studies.

1.2.1 Age-related heterogeneity changes

It was previously suggested that the aging period is also associated with dysregulation of gene expression and mRNA processing (Frenk & Houseley, 2018), suggesting a possible increase in age-related gene expression heterogeneity (also called noise or variation) between individuals. A number of studies were published reporting increased variation between individuals and cells during aging.

In one of the earlier studies investigating age-related heterogeneity change, Somel *et al.* demonstrated an increase in age-related gene expression variation during aging, which was suggested to be a result of the accumulation of stochastic errors during aging (Somel et al., 2006). However, they failed to show the functional consequences of increased heterogeneity as they could not identify a set of genes that become more heterogeneous with age. Consistently, another 2006 study also reported an increase

in cell-to-cell variation in mouse heart during the aging period (Bahar et al., 2006). In a more recent study conducted by Kedlian & Donertas *et al.*, the authors investigated microarray data from the human prefrontal cortex and revealed that increased heterogeneity is a weak but consistent pattern, which is also associated with a wide range of pathways (Kedlian et al., 2019).

Some more recent studies leveraged the availability of more advanced technology, single-cell RNA sequencing, which provides a resolution that is unattainable by previous sequencing methodologies. In a 2017 study, Martinez-Jimenez *et al.* analyzed single-cell sequencing data of T cells of mice and found that the cell-to-cell heterogeneity increases during aging among immune cells (Martinez-Jimenez et al., 2017). Another study analyzed gene expression data from human pancreas generated by single-cell sequencing technologies, which also allows the detection of age-related stochastic errors. They found that aging is associated with a gradual accumulation of stochastic errors, which leads to increased cell-to-cell heterogeneity during aging (Enge et al., 2017). Similar results were also observed in mice lung by Angelidis *et al.* where they found increased transcriptional heterogeneity during aging (Angelidis et al., 2019).

Increased heterogeneity during aging was not only observed at the transcriptome level. A 2002 study that analyzed cellular changes during aging in *C. elegans* using the electron-microscopic data also found increased variance in age-related cellular decline (Herndon et al., 2002). The authors further discussed that the observed effect might be a result of weakened gene regulation in post-reproductive life, suggesting the significant role of stochastic factors. In a more recent study, the authors analyzed chromatin modifications in human immune cells at the single cell level, and found an increased heterogeneity between both cells and individuals (Cheung et al., 2018). They further demonstrated that age-related chromatin alterations are largely driven by non-heritable factors. Another study analyzed age-related heterogeneity changes in monozygotic twins using both transcriptomic and epigenetic datasets. They found that the heterogeneity in the epigenetic modification patterns between monozygotic twins increases with age. Moreover, by analyzing gene expression patterns, they demonstrated that 50 years old twins display significantly different gene expression profiles, whereas 3 years old twins have almost identical expression profiles (Fraga

et al., 2005).

Given all the above, it should also be of note that the generalizability of this effect still remains unclear. For example, analysis of single-cell profiles of aging mouse brain suggested that aging may not be broadly associated with increased transcriptional heterogeneity (Ximerakis et al., 2019). Moreover, Vinuela *et al. et al.* analyzed age-dependent heterogeneity changes in a twin cohort and observed a decrease in gene expression heterogeneity for the majority of genes (Viñuela et al., 2018).

1.3 Research Objectives

Traditionally, molecular aging studies have studied senescence as a monomorphic process. In recent years, however, a number of studies have gone beyond this approach, and have been reporting changes in gene expression heterogeneity with age. Still, the generality of increase in heterogeneity with age remains contentious. Moreover, whether age-related heterogeneity change is a function of time that starts early in development or is limited to the aging period has not been systematically explored. Also, the functional role of increased heterogeneity and its potential contribution to the emergence of aging-associated phenotypes in the human brain still remains unclear.

In this study, I aimed to address these standing questions by analyzing 19 microarray datasets from 3 independent studies covering diverse human brain regions. The previous research mainly focused on significant changes in individual datasets, which is sensitive to sample size and highly affected by confounding factors. Thus, in this study, I adapted a meta-analysis approach to analyze consistent changes in gene expression heterogeneity across multiple datasets. Using this approach, I managed to reduce the effects of confounding factors and technical noise, and identify weak but consistent patterns across datasets.

CHAPTER 2

MATERIALS AND METHODS

2.1 Datasets

In this study, I analyzed 19 microarray age-series datasets to investigate age-related gene expression heterogeneity change in the human brain during development and aging. The datasets were retrieved from 3 independent sources, containing microarray data for the human brain (Colantuoni et al., 2011; Kang et al., 2011; Somel et al., 2010; Somel et al., 2011). Overall, the datasets include 1,010 samples from 298 individuals spanning 17 different brain regions, which are not mutually exclusive. All datasets have samples covering whole lifespan with ages ranging from 0 to 98 years (**Figure 2.1**). A summary of datasets used in this study is shown in **Table 2.1**.

It should also be noted that the Kang2011 datasets contain samples from the left and right hemispheres of the same individual. These samples were analyzed as biological replicates, meaning that samples were not separated into two different datasets, for three reasons. First, it was previously suggested that the left and right hemispheres of the brain may show asymmetric age-related changes (Dolcos et al., 2002; Sun et al., 2005). Second, the other datasets do not contain hemisphere information. Last, previous studies analyzing this dataset, including the original study, also treated them as biological replicates (Dönertaş et al., 2017; Kang et al., 2011).

Additionally, the Somel2011_PFC dataset includes two pairs of technical replicates, between which the correlation was high. Therefore, the mean of expression values

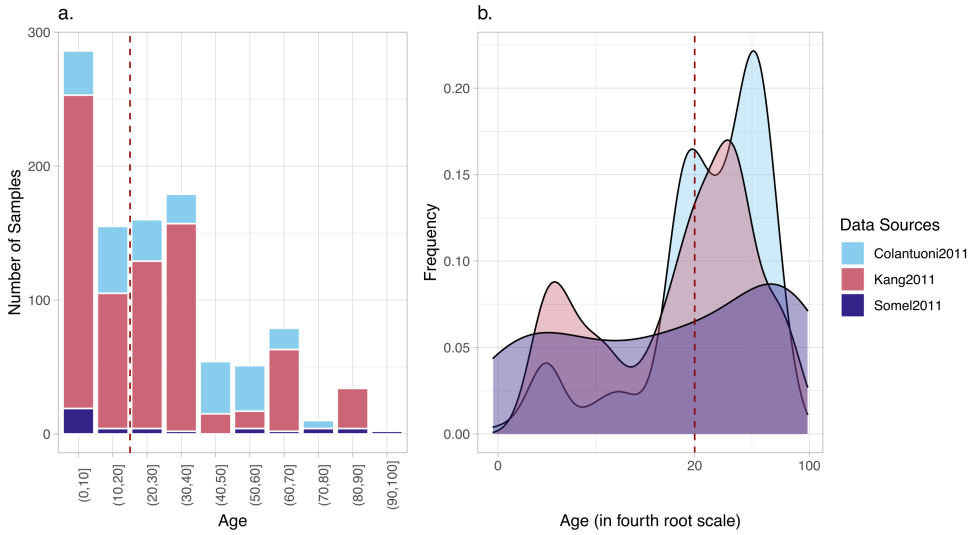

Figure 2.1: The distribution of ages (in years) of the samples. (a) Number of samples included in age intervals. (b) Distribution of ages. The color coding reflect different data sources.

was used in the downstream analysis.

The datasets were downloaded from the NCBI Gene Expression Omnibus (GEO) database using the accession codes given in the **Table 2.1**. All analyses was performed in R programming environment (R Core Team, 2020).

2.1.1 Dataset selection

The age-series datasets analyzed in this study are all microarray-based. Although there was one other RNA Sequencing-based dataset that covers the whole lifespan (Mazin et al., 2013), I chose not to include it in this analysis for two reasons. First, the samples were already included in the Somel2011 dataset. Second, it is an underpowered dataset with data from only 35 individuals that cannot reliably lead to a conclusion.

There were also RNA-Sequencing datasets containing samples from only development or aging periods. Since combining independent development and aging datasets may confound the biological effects that I aimed to examine, this study was limited to the meta-analysis of 19 microarray-based datasets.

Table 2.1: The list of microarray human brain gene expression datasets. The sample sizes were calculated after the removal of outliers.

GEO Acc.	**Source**	**Brain Region**	**Sample Size**
GSE30272	Colantuoni2011	PFC	231
GSE25219	Kang2011	A1C	47
GSE25219	Kang2011	AMY	43
GSE25219	Kang2011	CBC	47
GSE25219	Kang2011	DFC	48
GSE25219	Kang2011	HIP	39
GSE25219	Kang2011	IPC	49
GSE25219	Kang2011	ITC	49
GSE25219	Kang2011	M1C	45
GSE25219	Kang2011	MD	43
GSE25219	Kang2011	MFC	50
GSE25219	Kang2011	OFC	48
GSE25219	Kang2011	S1C	46
GSE25219	Kang2011	STC	48
GSE25219	Kang2011	STR	41
GSE25219	Kang2011	V1C	48
GSE25219	Kang2011	VFC	47
GSE22569	Somel2011	PFC	23
GSE18069	Somel2011	CBC	22

2.1.2 Seperating development and aging datasets

The aim of this study is to investigate age-related gene expression changes during development and aging. Thus, all the datasets were separated into two datasets: development (0 to 20 years of age) and aging (20 to 98 years of age). The age of 20 years was used to separate development and aging for the following reasons:

1. The age of 20 was shown to correspond approximately to the age of reproduction in human societies (Walker et al., 2006).
2. Previous studies investigating age-related gene expression trajectories demonstrated that 20 years of age corresponds to a turning point of gene expression patterns (Colantuoni et al., 2011; Dönertaş et al., 2017; Somel et al., 2010).
3. Earlier research connected the structural changes occurring in the human brain after the age of 20 to age-related phenotypes (Sowell et al., 2004).

As a result, I obtained: (i) one development and one aging dataset for Colantuoni2011; (ii) 16 development and 16 aging datasets for Kang2011; and (iii) two development and two aging datasets for Somel2011. Overall, both development and aging datasets resulted in a comparable number of samples ($n_{development} = 441$; $n_{aging} = 569$).

Moreover, it is important to note that I excluded samples from the prenatal development period, since gene expression trajectories were shown to be discontinuous between prenatal and postnatal development period (Colantuoni et al., 2011; Kang et al., 2011), and since the scope of this study is limited to investigating changes in gene expression heterogeneity during aging compared to pre-adulthood.

2.2 Dataset Preprocessing

Microarray technology is a widely used tool to quantify the expression level of gene transcripts from a given sample. A microarray chip contains known sequences of oligonucleotides -known as probes- that are located on a solid surface. Typically, each transcript is represented by a set of 11-20 pairs of probes, called the probe-set of that transcript, in Affymetrix microarray platforms. The cDNAs derived from the

mRNA transcripts of the sample are hybridized to target probes labeled by detectable fluorochrome molecules, where the amount of hybridization is reflected by the light intensity levels. The quantification of expression is then performed by measuring the light intensity levels of each probe, which are stored in CEL files.

The Kang2011 and Somel2011 datasets were generated by Affymetrix HuEx-1_0-st and HuGene-1_0-st microarray platforms, respectively. Colantuoni2011 dataset, on the other hand, was generated using the HEEBO-7 set (Human 49K oligo array), which is an Illumina-based array. Since there is no public R library available to process Illumina-based data from Colantuoni2011, I used the expression data preprocessed by the authors of the original study (Colantuoni et al., 2011). For the datasets from Kang2011 and Somel2011 sources, I downloaded CEL files from GEO database (Barrett et al., 2013). The preprocessing of Kang2011 and Somel2011 datasets can be summarized in four steps: (1) RMA convolution, (2) probe-set summarization, (3) log2 transformation, and (4) quantile normalization. Distribution of expression values after each preprocessing step is shown in **Figure 2.1**. For the Colantuoni2011 dataset, quantile normalization was performed on the preprocessed data.

2.2.1 RMA correction

The very first step of microarray analysis is the removal of noise and biases from the raw data obtained from light intensities. There can be a number of factors contributing to background errors, such as optical noise, unspecific hybridization and incomplete washing (Bengtsson & Hössjer, 2006). Nevertheless, low-level preprocessing and normalization, having a significant effect on the downstream analysis, were suggested to be one of the most important steps in any microarray data analysis (Bengtsson & Hössjer, 2006).

In this study, background normalization was performed by the Robust Multiarray Average (RMA) convolution method, which is one of the most commonly used methods to perform background normalization on microarray data. The RMA method involves the removal of technical artifacts so that the measurements from neighboring probes do not interfere with each other (Irizarry et al., 2003).

Apart from background normalization, the RMA algorithm also performs probe-to-probe-set summarization. Since each transcript is represented by a set of 11-20 probes, it is necessary to summarize probe-level data into probe-sets, by grouping probes corresponding to the same transcript. I used the R "oligo" library to perform RMA correction (Carvalho & Irizarry, 2010). As previously stated, RMA correction was performed only on Kang2011 and Somel2011 datasets.

2.2.2 Probe-set summarization

I next summarized the probe-set expression values into gene expression values. This step is required to combine and analyze data from different platforms since we need to have expression values that are defined universally (i.e., by gene IDs). Thus, converting probe-sets into gene IDs allows us to compare expression levels of the genes among different platforms.

However, probe-set to gene ID conversion is not one-to-one in many platforms. There can be multiple probe-sets that correspond to the same gene, while it is also possible to have a probe-set that maps to multiple genes. In this study, probe-sets that correspond to more than one gene were removed, since keeping these samples may lead to a pseudo-replication problem, where the expression level of multiple genes would not be independent. For the genes having multiple probe-set data, the expression values were calculated by taking the average of the expression values of the probe-sets corresponding to the gene.

For the dataset generated by HuGene-1_0-st (Somel2011 datasets), Ensembl v.84 annotations (Yates et al., 2016) were retrieved through "biomaRt" library in R (Durinck et al., 2009). For the Kang2011 dataset generated by the HuEx-1_0-st platform, the GPL file deposited in the GEO database was used since the IDs of probe-sets were not complete in Ensembl. Lastly, for the Colantuoni2011 dataset, the gene IDs were retrieved from the GPL file deposited in the GEO database.

2.2.3 Log2 transformation

The Log_2 transformation is the most widely used transformation in microarray data analysis to remove the correlation between mean and variance, and to make the variance more comparable. Moreover, in the data produced by microarray platforms, there are typically many genes having lower expression values, whereas there are fewer genes having high expression levels, leading to a right-skewed distribution (**Figure 2.2**). Making distributions more similar, Log_2 transformation allows us to perform parametric statistical tests, as most of them assume equal variance. Additionally, the Log_2 transformation also allows us to visualize the data more easily.

Therefore, the Log_2 transformation was applied to the expression data from Kang2011 and Somel2011 sources.

2.2.4 Quantile normalization

The microarray platforms are susceptible to technical variation from different sources, hampering the meta-analysis of multiple datasets from different sources. Quantile normalization is one of the commonly used methods used to minimize technical variation (Zhao et al., 2020).

Quantile normalization assumes the same distribution for all samples. Therefore, any significant variation in the distribution shape is regarded as unwanted and non-biological noise and is eliminated. However, it is also important to note that quantile normalization should be used with caution as it may remove signals that can be of biological interest and introduce false signals as well (Hicks & Irizarry, 2014). Despite this danger, the quantile normalization was performed on all datasets for three reasons. First, all the datasets analyzed in this study contain samples only from the human brain. Second, sample collection was performed under similar conditions (i.e., from healthy individuals), combined with the first reason, they indicate that overall expression distributions should be similar. Third, this study mainly focuses on consistent patterns among different datasets, rather than focusing on significant changes in individual genes from a single dataset. We expect quantile normalization to eliminate only random confounding factors that are not shared among datasets. Then,

the shared patterns among different datasets were considered as a potential biological signal.

The "preprocessCore" R package was used to perform quantile normalization (Bolstad, 2021).

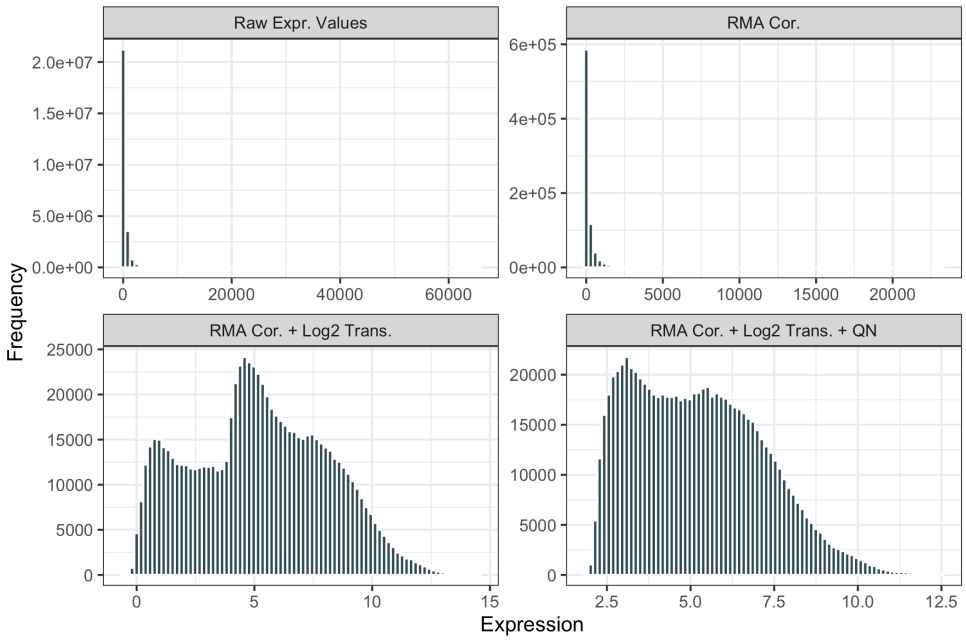

Figure 2.2: Summary of preprocessing steps for Somel2011_CBC dataset at different preprocessing steps. The top left histogram shows the distribution of raw probe expression levels. The top right histogram shows probe-set expression values after RMA correction. The bottom left plot shows the gene expression values after RMA correction and Log2 transformation. The bottom right plot shows the gene expression values after RMA correction, Log2 transformation and quantile normalization.

2.2.5 Scaling

Next, the expression levels for each gene for each dataset were scaled to $mean = 0$ and $standard\ deviation = 1$. Since linear regression analysis was performed in the downstream analysis, it is important to scale the genes before model fitting in order to obtain comparable residuals. Scaling was performed by using the following formula.

$$Scaled\ Expr_i = \frac{Expr_i - mean(Expr)}{Standard\ Deviation(Expr)} \tag{2.1}$$

where $Expr_i$ and $Scaled\ Expr_i$ represent expression and scaled expression value of the sample i, respectively, whereas $Expr$ is the expression values of the gene for all samples in a dataset.

2.2.6 Batch-effect correction

There was one additional normalization step applied to the Somel2011_PFC dataset to correct the batch effect, which refers to non-biological differences in the distribution of expression values of different groups of samples that are processed separately. The correction was performed in three steps as follows:

1. For each probe-set, mean expression values were calculated.
2. Each batch was scaled separately to $mean = 0$ and $standard\ deviation = 1$ using the equation 2.1.
3. The mean values calculated at step 1 were added to each value.

2.3 Outlier removal

In addition to technical variance introduced by microarray analysis, there can be some samples showing greater divergence from the rest of the samples due to a number of factors, such as different disease backgrounds or different diets. Including such samples (i.e., outliers) in the analysis would introduce extra noise and would obstruct the identification of the true relationship between age and expression levels.

To identify outliers, Principal Component Analysis (PCA) was used. PCA is a dimension reduction method that allows us to visualize high-dimensional data and detect outliers (see **Section 2.7** for a more detailed explanation of PCA). The first two principal components, which are the most important components explaining the largest variance, were used to visualize data and identify outliers.

Consistent with the previous studies (Dönertaş et al., 2017; Dönertaş et al., 2018), the following 7 samples were removed from the analysis:

1. 3 years old *GSM705108* from Kang2011 dataset (A1C brain region);
2. 37 years old *GSM704438* from Kang2011 dataset (CBC brain region);
3. 42 years old *GSM705202* from Kang2011 dataset (CBC brain region);
4. 0 year old *GSM704567* from Kang2011 dataset (HIP brain region);
5. 40 year old *GSM704627* from Kang2011 dataset (HIP brain region);
6. 70 year old *GSM704226* from Kang2011 dataset (HIP brain region);
7. 70 year old *GSM704227* from Kang2011 dataset (HIP brain region).

After removing outliers, we obtained 1,010 samples from 298 individuals (**Table 2.1**). Lastly, the common genes among all datasets were selected for the downstream analysis ($n = 11,137$), and the genes for which we do not have a measurement in at least one dataset were removed.

2.4 Investigating age-related expression change

Having preprocessed datasets for both development and aging periods, I next sought to characterize age-related gene expression changes. Linear regression was used to quantify the relationship between expression levels and age. The following linear model was fitted to each gene for each time period separately.

$$Expr_i = \beta_{i0} + \beta_{i1} * Age^{1/4} + \epsilon_i \tag{2.2}$$

where, $Expr_i$ is the scaled expression value of the ith gene, β_{i0} is the intercept, β_{i1} is the slope, Age is the age in days, and ϵ_i is the residual. The β_{i1} values were considered as the measure of age-related expression change. Throughout this study, β_{i1} values will be referred to as simply 'beta' values. Since it is the slope of the linear model, a negative beta value indicates a decrease in gene expression with age, while

positive values indicate an age-related increase in the expression of the corresponding gene. The left panels of **Figure 2.3** show an example of quantification of age-related expression changes for two genes.

It is also worth noting that we transformed the individual ages into the fourth root scale to obtain an approximately uniform distribution of ages across the lifespan. To test the effect of the age scale, I performed the same analysis using different age scales and confirmed that they also resulted in quantitatively similar results.

The linear regression was performed by using the "lm" function in base R for each gene in each dataset (development and aging datasets, separately).

2.4.1 Multiple testing correction

I obtained p-values for each gene from the linear regression model, showing the significance of observed age-related expression change. In this study, we used a significance level of $\alpha = 0.05$, meaning that there is a 5% chance that there will be a false-positive result for a single statistical test. Since the regression analysis was performed on each gene independently, the false positive rate increased dramatically as a result of multiple comparison problem. Therefore, the p-values obtained from linear regression were adjusted for multiple testing to eliminate the accumulation of false-positive results.

While there are several other methods available, I used Benjamini & Hochberg method throughout this study to adjust p-values (Benjamini & Hochberg, 1995). It is the standard method that is commonly used to control the false discovery rate. Compared to other methods, B&H is one of the less conservative (and thus more powerful) methods.

The correction was performed by using the "p.adjust" function in base R.

2.5 Age-related heterogeneity change

Given that the age-related expression change display a linear trend, the residual obtained from the **Equation 2.2** (i.e., the ϵ values) reflects the deviation of that sample (see dashed vertical lines on the left panels of **Figure 2.3**). Thus, the absolute value of residuals was used as a measure of the heterogeneity of the corresponding sample. To characterize the age-related change in gene expression heterogeneity, Spearman's correlation test was performed between absolute values of residuals and the fourth root of age for each gene and each dataset separately (see the right panels in **Figure 2.3**).

Then, the Spearman's correlation coefficients (ρ) were considered as a measure of heterogeneity change, where positive values indicate an increase in heterogeneity with age and negative values reflect a decrease in age-related heterogeneity.

The correlation test was performed by using the "cor.test" function in base R and the p-values were adjusted for multiple testing using the B&H method, as explained in **Section 2.4.1**.

2.6 Correlation among datasets

To evaluate the correlation among datasets in age-related gene expression (and heterogeneity) change, I performed pairwise Spearman's correlation test among beta values (rho values) for each pair of datasets, using the 'cor.test' function in R. The heatmap visualizations (see **Figure 3.1a** and **Figure 3.2a**) were created by using "pheatmap" library in R (Kolde, 2019).

A random permutation test was used to test the significance of correlations among datasets (see **Section 2.8.1** for details).

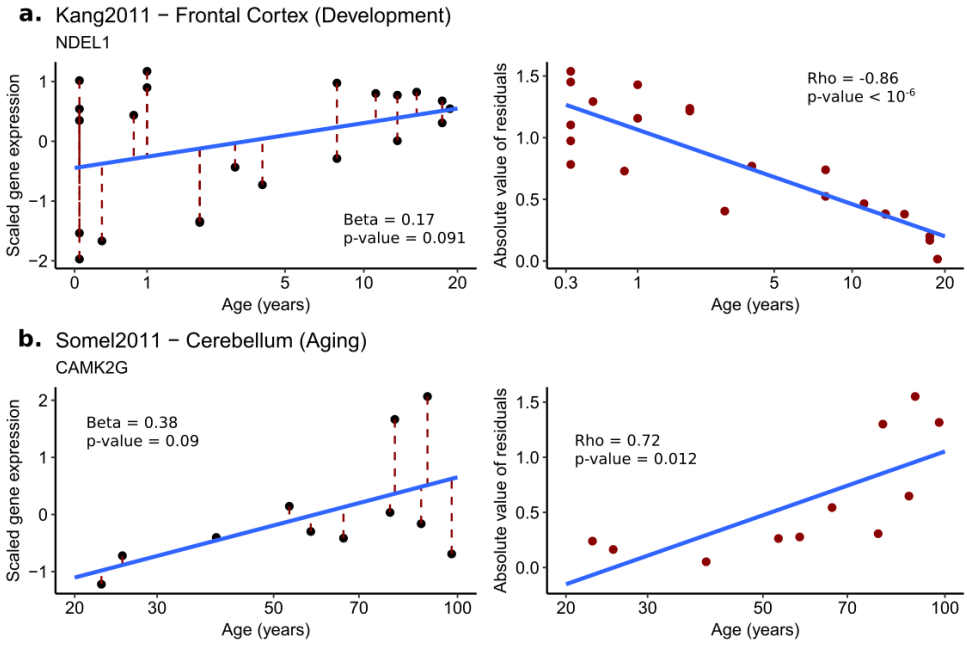

Figure 2.3: Summary of the method used to calculate age-related expression change (left panels) and age-related expression heterogeneity change (right panels) during development (a) and aging (b). Each point represents a sample. The *Beta* values were obtained from linear regression analysis, while the *Rho* values were obtained from Spearman's correlation test. The *p-value* shows the FDR corrected p-values. The figure is adapted from (Isildak et al., 2020).

2.7 Principal component analysis

Since biological data, including gene expression data, is typically high-dimensional, it is difficult to visualize and discover patterns in the data. Principal component analysis (PCA) is a dimension reduction method that is widely used in biological data analysis for summarization and visualization of the data. Basically, the PCA algorithm creates linear combinations of variables, known as principal components (PCs). The principal components were sorted based on their capacity to explain the variance, where the first principal component explains the most variance. Usually, the first two principal component axes are used for visualization purposes. Therefore, PCA allows us to visualize and summarize high-dimensional data in 2-dimensional space.

Apart from the outlier removal step (see **Section 2.3**), the PCA algorithm is also

used to visualize age-related expression (**Figure 3.1b**) and heterogeneity changes (**Figure 3.2b**). I performed PCA using the "prcomp" function in base R.

2.8 Permutation test

The random permutation test is a statistical method used for hypothesis testing. The main advantage of the permutation test over other parametric statistical tests is that the permutation test does not have distributional assumptions. Additionally, conventional parametric tests require each sample to be independent, while the permutation test is more flexible, allowing us to design a permutation procedure that accounts for non-independence. Since there are multiple samples from the same individuals in some sub-datasets, meaning that not all observations are independent, the permutation test was used to assess significance throughout this study.

I used the permutation schema that was developed by earlier studies, taking into account that Kang2011 and Somel2011 datasets contain samples from different brain regions of the same individual (Dönertaş et al., 2017; Dönertaş et al., 2018). Specifically:

1. The ages were randomly permuted among individuals for each data source, by using the "sample" function in R. It is important to note that ages were not randomized among samples, but among individuals to maintain the dependency between samples.

2. The randomized ages of individuals were mapped to corresponding samples. All the samples obtained from the same individual were assigned to the same age.

3. The age-related gene expression and heterogeneity changes were calculated by using permuted ages of the samples as follows:

    (a) To calculate age-related expression change, linear regression was performed between randomized ages and expression levels, using the **Equation 2.2** as explained in **Section 2.4**

(b) To calculate age-related heterogeneity change, Spearman's correlation test was performed between randomized ages and absolute value of residuals obtained from **Equation 2.2** computed with non-randomized ages.

4. Steps 1-3 were repeated 1,000 times to obtain the null distribution for the test statistic.

The basic idea of this procedure is to simulate the null hypothesis, where there is no relationship between age and expression (heterogeneity). After having 1,000 expression (heterogeneity) change values for each gene in each dataset, we can compare observed values with this null distribution to assess the significance of observed tendency.

Using the expression and heterogeneity change estimates, I tested the significance of (i) dataset correlations, (ii) overall increase in heterogeneity, and (iii) expected consistency in the heterogeneity change. All permutation tests were performed one-tailed.

2.8.1 Dataset correlations

Using the permuted ages, I first examined the coordination in expression and heterogeneity changes among datasets by computing correlation coefficients for expression and heterogeneity change within the development and aging datasets. One approach to test correlations would be to calculate pairwise correlations for each pair of datasets for each permutation, and calculate the expected median correlation among development and aging datasets. However, since 16 of 19 datasets analyzed in both time periods were retrieved from the same data source (i.e., Kang2011) and contain samples from different brain regions of the same individual, this approach would lead to excess false positives due to a high number of dependent pairwise comparisons. Therefore, we adopted an alternative approach, where I first calculated a median correlation coefficient value by performing Spearman's correlation test among all pairwise combinations of three subsets of datasets: one dataset from Kang2011, one dataset from Somel2011, and the Colantuoni2011 dataset. I performed this 1,000 using the permuted ages and generated a null distribution of medians for development and aging

datasets, separately. Then, the observed median values were compared against the null distribution of medians to assess the significance of the observed median correlation among development and aging datasets.

Next, the difference in correlations between development and aging datasets was tested. For each permutation, I calculated a median difference between the correlations of development and aging datasets. Similarly, observed median differences were compared against the distribution of median differences obtained from permutations.

2.8.2 Significant expression and heterogeneity changes

To test if there is a significant difference in the number of genes showing the significant age-related change in expression (heterogeneity) between development and aging datasets, I calculated the difference in the number of significantly changing genes between development and aging datasets for each permutation to construct the null distribution. The p-values were obtained by comparing observed differences against the null distribution.

2.8.3 Overall increase in heterogeneity

To test if the overall increase in heterogeneity during aging is significantly higher than development, I calculated the median difference between median heterogeneity change of development and aging for each dataset. The median value was computed 1,000 times for each permutation to construct the null distribution. Then, I compared the observed median difference against the null distribution to obtain an empirical p-value.

2.8.4 Testing consistency in heterogeneity changes

Lastly, I used heterogeneity change values computed using the permuted ages to estimate the randomly expected consistency in heterogeneity increase, and to test the significance of observed consistency. For each permutation, I calculated the number

of genes showing a consistent increase in the X number of datasets, where X is an integer having values from 0 to 19. Then, the observed number of genes showing a consistent increase in heterogeneity in the X number of datasets were compared against the expected number of genes for each X value ranging from 0 to 19.

2.9 Clustering

Clustering of genes based on shared heterogeneity patterns was performed using the k-means clustering algorithm on 147 genes showing a consistent increase in heterogeneity in all 19 aging datasets. The heterogeneity levels (i.e., absolute residuals obtained from **Equation 2.2**) were first scaled to have the same mean and standard deviation. Then, using the scaled heterogeneity levels, spline curves were fitted for each gene with a degree of freedom of 3, by using the "smooth.spline" R function. Within each dataset, the smallest sample size was used to interpolate the data so that the age points are evenly spaced. Lastly, I used interpolated values to perform k-means clustering (k = 8), using the "kmeans" function in R. An alternative approach would be to directly use scaled heterogeneity values to perform clustering. However, this approach would fail to represent all data points equally due to the varying sample sizes of different datasets.

2.10 Functional enrichment analysis

Gene set enrichment analysis (GSEA) was performed using Gene Ontology (GO) Biological Process categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways to determine the functional role of increased heterogeneity. Specifically, the common genes were first ordered based on the number of datasets they show a consistent increase. Then, by using the "gseGO" and "gseKEGG" functions from the "clusterProfiler" R package (Wu et al., 2021), I investigated whether the genes showing increased consistency in heterogeneity change are associated with specific GO categories or KEGG pathways. The gene sets with a size between 5 and 500 were employed to run GSEA analysis and the resulting Normalized Enrichment Scores (NES) were used to assess the strength of the association. The p-values were adjusted

for multiple testing using the B&H method, consistent with the previous sections (**Section 2.4.1**).

Since I want to test if genes with consistent heterogeneity increase are associated with certain functionalities, the genes were ranked according to the number of datasets in which they show consistent heterogeneity increase, meaning that a number between 0 and 19 was assigned to each gene. One problem with this approach is that not all genes can be ranked unambiguously due to many ties observed in the dataset. To address this issue, I ran GSEA 1,000 times and calculated the number of times that each pathway was detected as "significant" (see 'significanceIn1000' column in **Table C.1**).

GO Biological Process enrichment results were visualized by using the "treemap" function from "treemap" package (Tennekes, 2021), whereas significant KEGG pathways were visualized by using the "KEGGgraph" package in R (Zhang & Wiemann, 2009).

2.11 Transcriptional regulation enrichment analysis

Next, the gene enrichment analysis was also performed for transcriptional regulators, namely miRNAs and transcription factors. The Harmonizome database was used to retrieve transcription-regulator association information (Rouillard et al., 2016). Specifically, the miRNA-target interaction information was obtained from miRTarBase (Chou et al., 2016), containing information for 12,086 genes and 596 miRNAs. The information about transcription factors and their binding sites were curated by the TRANSFAC database for 13,216 genes and 201 transcription factors (Matys et al., 2006). To run gene set enrichment analysis on these custom gene sets, the "fgsea" package in R was used (Korotkevich et al., 2019). Similar to functional enrichment analysis, the genes were ranked based on the number of datasets in which they show a consistent increase. The regulators with 10 to 500 targets were used to perform enrichment analysis.

Additionally, I performed a correlation test between the heterogeneity change and the number of transcriptional regulators. Specifically, Spearman's correlation test

was performed between the number of datasets with a consistent increase in heterogeneity and the number of regulators (i.e., miRNAs and transcription factors), for development and aging periods separately. To test the significance of the difference in correlations between development and aging, the permutation test was used. In detail, I randomly permuted the number of regulators 1,000 times and calculated Spearman's correlation coefficients between the number of datasets with a consistent increase and permuted number of regulators. Then, for each permutation, I computed the proportion of datasets where the correlation in the aging dataset is higher than the development. The empirical p-value was obtained by comparing the observed proportion against the distribution of expected proportions.

CHAPTER 3

RESULTS

In this study, I investigated the age-related gene expression heterogeneity change by analyzing 19 microarray datasets containing 1,010 samples from diverse brain regions of 298 individuals, and covering the whole lifespan. To compare heterogeneity change occurring before and after adulthood, each dataset was first separated into two datasets: one development dataset (including samples with ages ranging from 0 to 20 years old) and one aging dataset (including samples with ages ranging from 20 to 98 years old). Only the common genes (n = 11,137) were used in the downstream analysis.

3.1 Age-related gene expression change

While this study aims to investigate age-related changes in gene expression heterogeneity, I first sought to characterize the age-related change in gene expression. Linear regression was performed to characterize the relationship between gene expression and age for each gene and dataset separately (**Section 2.4**). The beta values (i.e., β_{i1} from **Equation 2.2**) obtained from linear regression were considered as a measure of age-related expression change.

To investigate the coordination in age-related expression change between datasets, I calculated pairwise Spearman's correlation test for each pair of datasets among the beta value. The heatmap in the **Figure 3.1a** shows the Spearman's correlation coefficients (ρ) for each pair of datasets. The significance of correlation among datasets

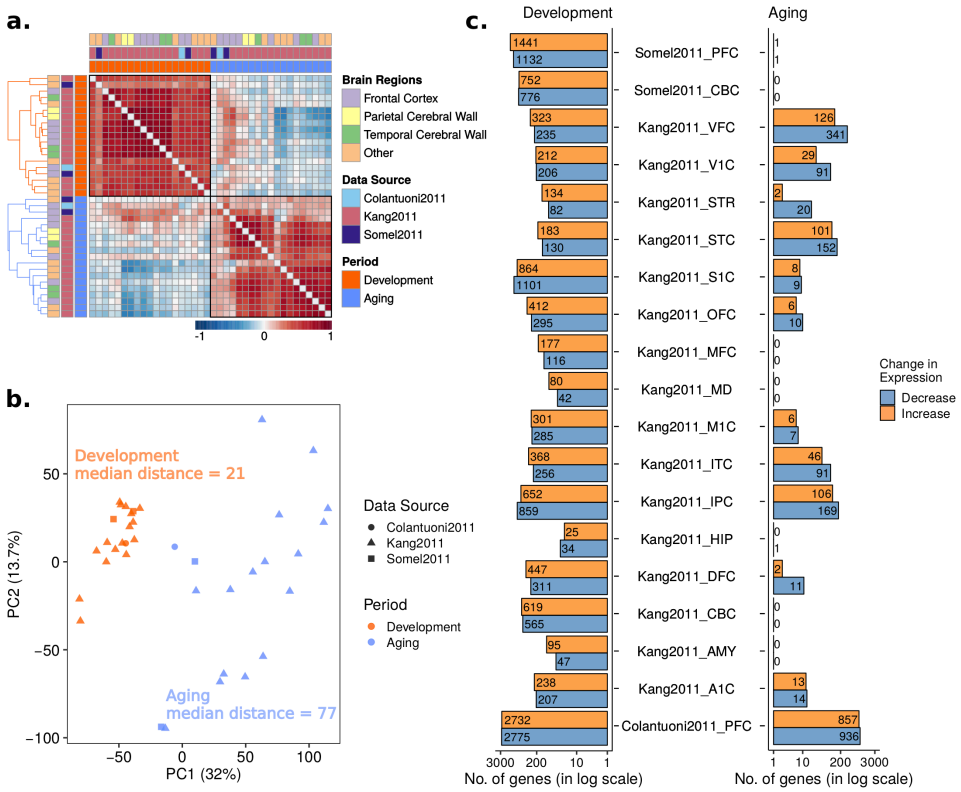

Figure 3.1: Age-related gene expression change. (a) The heatmap shows Spearman's correlation coefficient values for each pair of datasets. The row and column annotations reflect the brain region, data source and time period of the corresponding dataset. The datasets of each period were clustered separately. (b) Principal component analysis of age-related expression changes. Each point shows a dataset, where color-coding reflects time period and point-shape shows the data source. The values in parenthesis (on the axis labels) reflect the explained variance by the corresponding dataset. The values shown on the plot show the median euclidian distance among development and aging datasets, calculated using PC1 and PC2. (c) Barplots show the number of genes showing significant age-related expression change (FDR p-value < 0.05) during development (left side) and aging (right side). Color-coding reflects the direction of expression change, which is determined by the sign of the beta value. Adapted from (Isildak et al., 2020).

was tested by performing a permutation test (see **Section 2.8**). Overall, I found that both development (permutation test p-value < 0.001, median ρ = 0.56) and aging (permutation test p-value = 0.003, median ρ = 0.43) datasets display a modest correlation with the datasets from the same time period, whereas the difference between the correlation of development and aging datasets was not significant (permutation test p-value = 0.1). Still, the weaker correlation among aging datasets compared to development may indicate the stochasticity of the aging period.

I next performed principal component analysis (PCA), which is a dimension reduction algorithm that is widely used in transcriptome data analysis. The beta values obtained for each gene for each dataset (development and aging datasets, separately) were used to calculate principal components. The first two principal components (PC1 and PC2), which also explain the most variance, were used to visualize the data. **Figure 3.1b** shows the development and aging datasets projected on the PC1 and PC2 axis, where the first principal component separates development and aging datasets, suggesting that expression changes may display different patterns in development and aging. By calculating the median euclidian distance, I found that the median distance among development datasets is 21, whereas the median distance among aging datasets is 77. The fact that development datasets are clustered more closely compared to aging datasets may reflect an increase in heterogeneity during the aging period.

Then, the number of genes showing significant age-related gene expression change was calculated for development and aging datasets (**Figure 3.1c**). The p-values obtained from linear regression analysis for each gene in each dataset were adjusted for multiple testing using the B&H method (see **Section 2.4.1**). Analyzing significant changes, I first found that there are significantly more genes in the development datasets showing significant change compared to aging datasets (permutation test p-value = 0.003), which may again reflect the higher heterogeneity during aging compared to development. Second, the direction of change in development datasets was mainly in the positive direction (14 of 19 datasets showed more increase), whereas genes showing significant change during aging tended to decrease in expression (13 of 19 datasets showed more decrease).

Combined, my analysis of age-related expression changes demonstrated that the age-

related gene expression changes may be less coordinated during aging, leading to a higher inter-individual variability during aging, which may be the first clue to increased age-related heterogeneity during aging.

3.2 Age-related expression heterogeneity change

I next characterized age-related change in gene expression heterogeneity by performing Spearman's correlation test between absolute value of residuals and fourth root of age (see **Section 2.5**). Similar to the previous section, I first examined the correlations in gene expression heterogeneity change among development and aging datasets (**Figure 3.2a**). By conducting Spearman's correlation test between all possible pairs of datasets among rho values, I found that aging datasets show higher correlation (median $\rho = 0.21$, permutation test p-value = 0.24) in heterogeneity change compared to development datasets (median $\rho = 0.11$, permutation test p-value = 0.25), although both were not significant. Whilst the difference in correlation of heterogeneity change between development and aging datasets was also not significant (permutation test p-value = 0.2), this observation may still suggest that aging datasets show more similar changes in heterogeneity compared to development datasets.

Principal component analysis of heterogeneity changes (i.e., rho values) was performed using a similar procedure as the previous section, except rho values were used instead of beta values. Similar to the PCA of expression changes, PCA of heterogeneity changes showed that development and aging datasets can be differentiated from each other by heterogeneity change (**Figure 3.2b**). However, unlike expression change, development datasets tended to cluster more sparsely (median euclidian distance = 44), compared to aging datasets (median euclidian distance = 41).

I then investigated the genes showing significant changes in expression heterogeneity. The p-values obtained from Spearman's correlation test for each gene for each dataset were corrected for multiple testing using the B&H method (see **Section 2.5**). **Figure 3.2c** shows the number of genes showing significant heterogeneity change during development and aging. First, I found that there is a less number of genes showing significant change during development compared to aging (permutation test

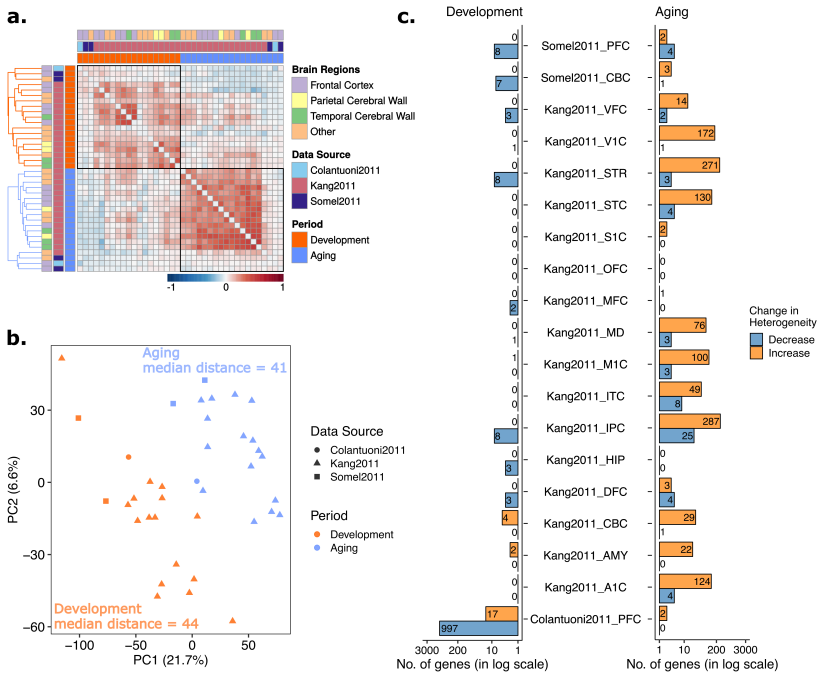

Figure 3.2: Age-related heterogeneity change. (a) The heatmap shows Spearman's correlation coefficient values for each pair of datasets. The row and column annotations reflect the brain region, data source and time period of the corresponding dataset. The datasets of each period were clustered separately. (b) Principal component analysis of age-related heterogeneity changes. Each point shows a dataset, where color-coding reflects time period and point-shape shows the data source. The values in parenthesis (on the axis labels) reflect the explained variance by the corresponding dataset. The values shown on the plot show the median euclidian distance among development and aging datasets, calculated using PC1 and PC2. (c) Barplots show the number of genes showing significant age-related heterogeneity change (FDR p-value < 0.05) during development (left side) and aging (right side). Color-coding reflects the direction of heterogeneity change, which is determined by the sign of the rho value. Adapted from (Isildak et al., 2020).

p-value < 0.05). Moreover, genes showing significant heterogeneity change during development tended to decrease in heterogeneity, whereas they showed a tendency to increase in aging datasets. It is also important to note the high number of genes showing a significant change in Colantuoni2011, which is a result of the large sample size leading to increased statistical power.

Overall, this analysis demonstrated that there may be a higher consistency in heterogeneity change during aging compared to development, suggesting that increased heterogeneity may be a characteristic of the aging period.

3.3 Consistent increase in heterogeneity during aging

Having observed both that heterogeneity changes can differentiate between time periods, and that there is a tendency towards more similar changes in heterogeneity during aging, I next focused on identifying those genes showing the consistent change in heterogeneity. As previously stated, in this study, I chose to focus on genes showing consistent trends among different datasets, rather than focusing on significant changes within individual datasets. The most important advantage of this approach is that, using the advantage of having multiple datasets, it enables me to detect genes that would otherwise not pass the significance threshold due to low sample size.

I first analyzed the distribution of heterogeneity changes (i.e., ρ values) for each dataset and time period (**Figure 3.3a**). The 18 of 19 aging datasets showed more increase in age-related heterogeneity (median $\rho > 0$), while the remaining one dataset showed no change in heterogeneity (median $\rho = 0$). During development, on the other hand, I found that 14 of 19 datasets displayed a decrease in heterogeneity (median $\rho < 0$). Moreover, a comparison of median heterogeneity changes between development and aging datasets revealed that 18 of 19 datasets showed more increase during aging compared to development. The permutation test further demonstrated that the overall increase in heterogeneity during aging compared to development is significant (permutation test p-value < 0.001).

Although I observed increased heterogeneity during aging compared to development, one possible explanation may be related to the positive correlation between mean and

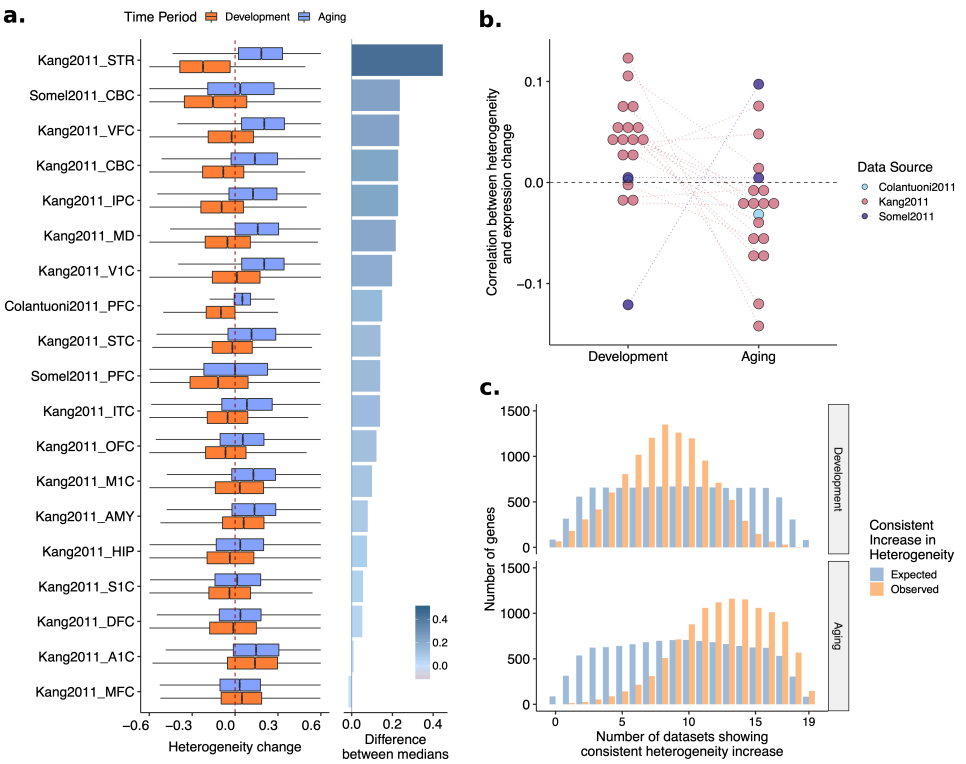

Figure 3.3: (a) Boxplots on the left side shows the distribution of rho values, whereas the barplots on the right shows the difference in medians between development and aging datasets. (b) Distribution of Spearman's correlation coefficients between expression and heterogeneity change of each dataset. The color-coding reflects the data source, whereas (c) Barplots showing expected and observed numbers of genes showing a consistent increase in heterogeneity among datasets during development (upper panel) and aging (lower panel). Adapted from (Isildak et al., 2020).

variation, where increased heterogeneity may be a result of increased expression levels during aging, rather than being a biological effect. To test this, I analyzed the correlations between expression and heterogeneity changes for each dataset and time period (**Figure 3.3b**). Spearman's correlation test revealed that there is no significant dependence between expression and heterogeneity change. In fact, development datasets showed higher positive correlations compared to aging datasets. Nevertheless, this analysis showed that the observed overall increase in heterogeneity during aging is not a result of mean-variance dependence.

Given that we observed an overall increase in heterogeneity during aging compared to development (**Figure 3.3a**), and this increase is not a technical artifact (**Figure 3.3b**), I next focus on detecting those genes showing a consistent change in heterogeneity. For each gene, I calculated the number of datasets in which they show a consistent increase in heterogeneity (in development and aging periods separately), irrespective of whether the increase is significant (**Figure 3.3c**). The expected consistency in heterogeneity change was calculated using the random permutations (see **Section 2.8.4** for details). By comparing observed consistency in heterogeneity increase to expected consistency, I first observed a shift towards increased heterogeneity consistency during the aging period, whereas no such shift was observed in development datasets. Moreover, 147 genes were identified that consistently increase in heterogeneity among all 19 aging datasets (permutation test p-value < 0.001), whereas there was only 1 gene showing a consistent decrease in all datasets during the aging period. The full list of the consistent 147 is given in the **Appendix A**. However, it is also important to note that, according to permutations, the number of genes that would show a consistent increase in heterogeneity under the null hypothesis (i.e., by chance) is 84, indicating a 40% true positive rate. Although I couldn't confidently identify a gene set showing increased heterogeneity consistently among all datasets, this analysis still demonstrated a clear shift toward increased heterogeneity during aging compared to development.

3.4 Clustering

I next examined if 147 consistent genes (i.e., showing consistent heterogeneity increase in all 19 datasets) display certain trajectories of heterogeneity change. To test this, I grouped these consistent genes according to their heterogeneity change pattern using the k-means clustering algorithm into 8 clusters (**Figure 3.4**). Overall, the clustering analysis revealed three diverse patterns of heterogeneity change during aging:

1. A steady increase in heterogeneity throughout the aging period (genes in clusters 3 and 7).

2. A steady increase until the age of 60 years, followed by a slight fall in heterogeneity (genes in clusters 4, 5 and 8).

3. A sharp increase in heterogeneity around the age of 60 years (genes in clusters 1, 2 and 6).

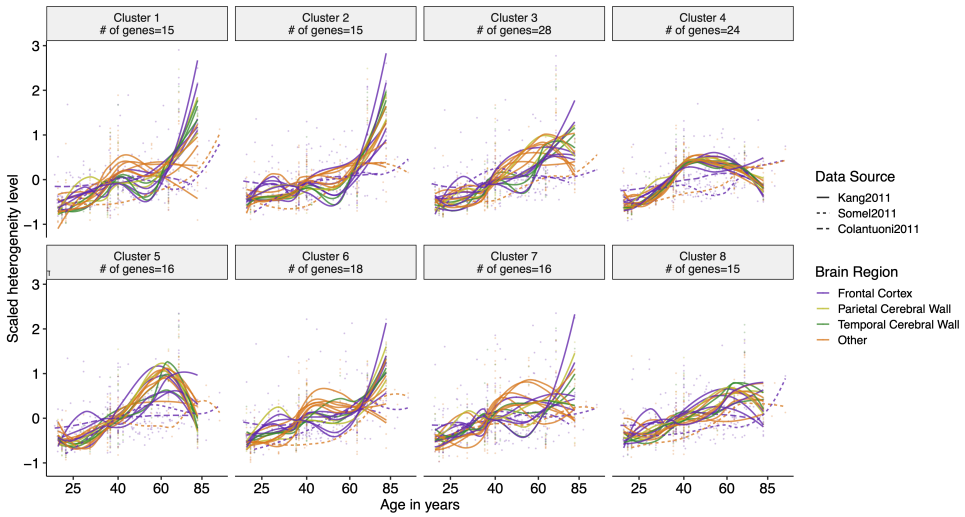

Figure 3.4: Trajectories of different clusters of 147 genes showing consistent increase in all 19 datasets during aging. The x-axis shows age in years, while the y-axis shows scaled heterogeneity changes (i.e., rho values). The spline lines show the mean heterogeneity change level for genes. Adapted from (Isildak et al., 2020).

3.5 Functional enrichmenet analysis

To investigate the functional role of increased heterogeneity during aging, I next performed gene set enrichment analysis using Gene Ontology (GO) Biological Process categories (Consortium, 2019) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al., 2019). The gene set enrichment analysis was performed on the genes that were ranked by the number of datasets they show a consistent increase in heterogeneity.

3.5.1 GO Biological Process enrichment analysis

Gene Ontology Biological Process enrichment analysis was performed on development and aging datasets, separately. While there was no GO term significantly enriched for the consistent heterogeneity changes during development, I identified 111 significantly enriched GO Biological Process terms for the consistent changes in aging. The full list of enriched GO terms and normalized enrichment scores can be found in **Table B.1**, while **Figure 3.5** shows the REVIGO summarization.

One of the most important enriched biological process GO term was autophagy (GO term ID: GO:0006914), which was previously suggested to play important role in aging and aging-related diseases (Rubinsztein et al., 2011). Moreover, another important enriched group of genes belonged to the axon regeneration (GO term ID: GO:0048679), which also demonstrated to decrease during aging (Belin et al., 2014). Additionally, the regulation of T-helper cell differentiation and cellular response to virus terms were found to be significantly enriched. Given the weakened immune system in aging, this result also suggests that increased heterogeneity may have important consequences in aging. Other significantly enriched groups included terms related to metabolic processes, mRNA processing and localization. Overall, GO Biological Process enrichment analysis of genes showing a consistent increase in heterogeneity indicates that increased heterogeneity may be associated with the aging-related phenotypes.

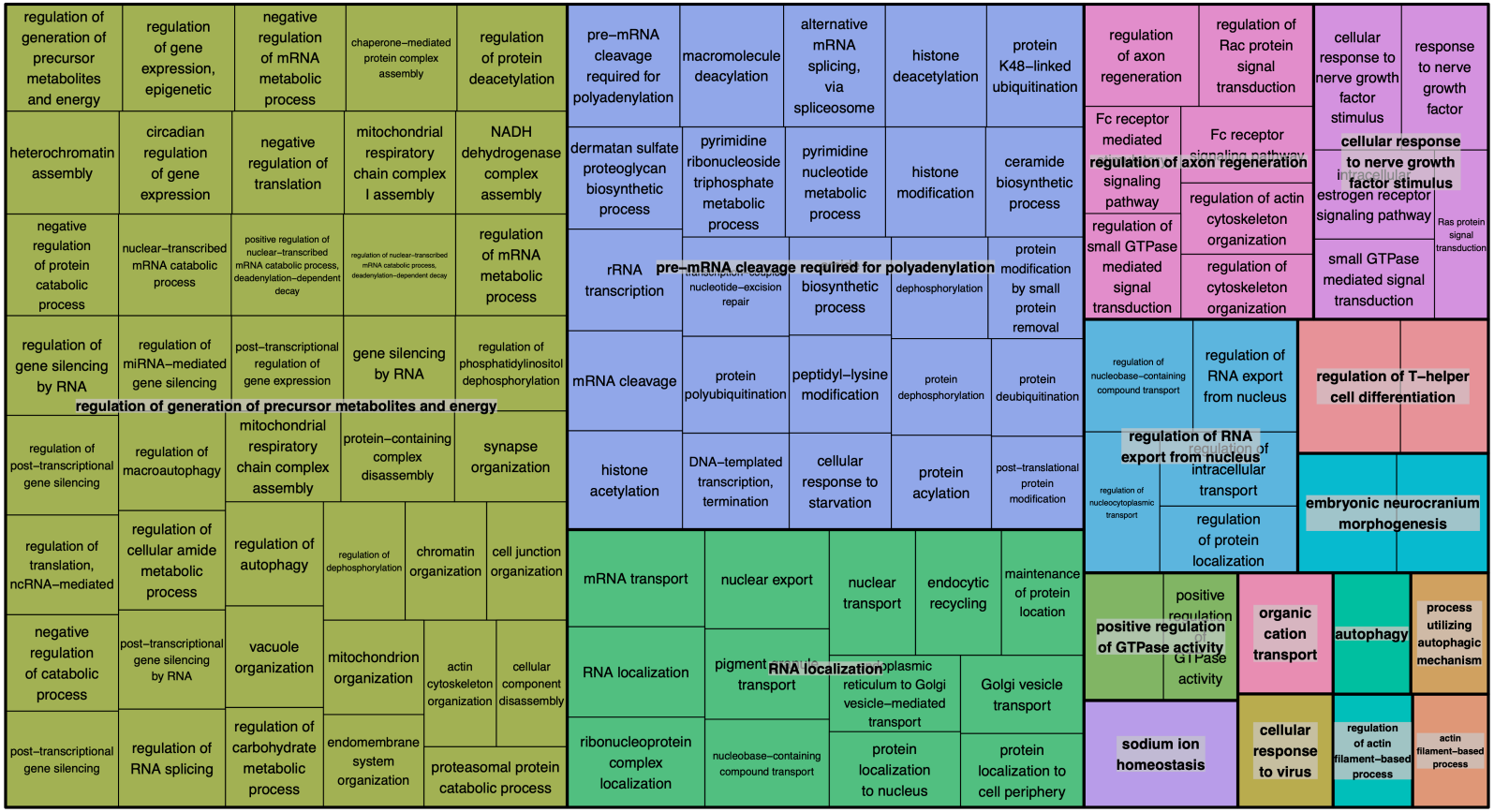

Figure 3.5: GO Biological Process enrichment analysis results for the genes showing consistent increase in heterogeneity, summarized by REVIGO (Supek et al., 2011).

3.5.2 KEGG pathway enrichment analysis

Next, I performed a KEGG pathway enrichment analysis. Similarly, there was no KEGG pathway found to be significantly enriched in the development period. In aging, on the other hand, there were 21 KEGG pathways significantly enriched for the genes that become consistently heterogeneous. The full list of significantly enriched KEGG pathways can be found in **Table C.1**, while **Figure 3.6a** shows the distribution of consistency of those genes that belong to significantly enriched pathways. Among the significantly enriched KEGG pathways, the most notable ones were longevity regulating pathway, autophagy, mTOR signaling and FoxO signaling, all of which were shown to be related to aging and aging-related diseases (Johnson et al., 2013; Martins et al., 2016; Rubinsztein et al., 2011). Additionally, the difference in the distribution of the number of datasets with an increase in heterogeneity between development and aging can be clearly identified, where the aging period has a higher number of datasets with heterogeneity increase compared to development (**Figure 3.6**). Importantly, 4/21 pathways were found to have negative enrichment scores: protein digestion and absorption pathway, primary immunodeficiency pathway, linoleic acid metabolism, and fat digestion and absorption pathway. However, due to the skewed distribution of observed consistency (see the lower panel of **Figure 3.3c**), negative scores do not automatically indicate a decrease in heterogeneity consistency. **Figure 3.6b** further demonstrates the heterogeneity consistency of genes in the longevity regulating pathway (KEGG Pathway ID: hsa04211) for development (upper panel) and aging (lower panel).

Overall, the functional enrichment analysis of genes that become consistently heterogeneous suggested that increased heterogeneity may have important functional consequences during the aging period, possibly contributing to the aging-related phenotypes.

3.6 Transcriptional regulation enrichment analysis

I next asked if there are specific transcriptional regulators associated with the genes that become consistently heterogeneous with age. I performed gene set enrichment

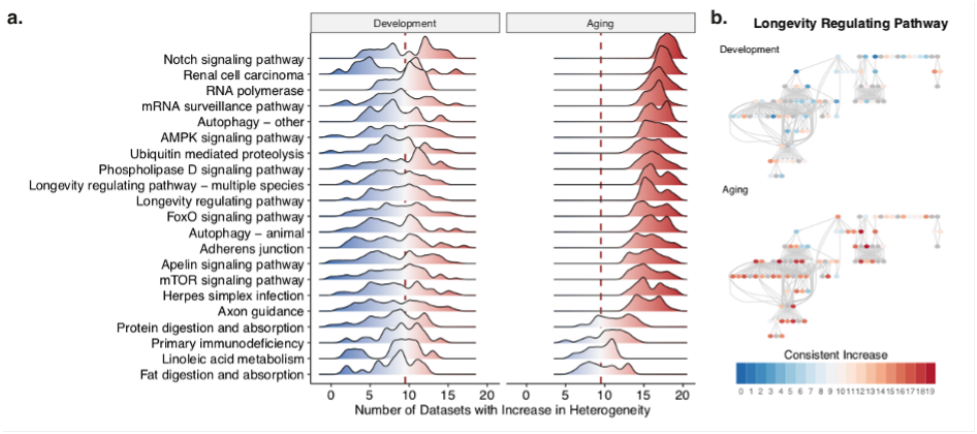

Figure 3.6: KEGG pathways enrichment analysis results for the genes showing consistent increase in heterogeneity. (a) Significantly enriched KEGG pathways (on the y-axis) and the distribution of the number of datasets in which genes show a consistent increase in heterogeneity (x-axis). (b) Demonstration of Longevity Regulating Pathway as an example during development (upper panel) and aging (lower panel). Nodes and edges represent the genes and their relationship, respectively, while color-coding reflects the consistency in heterogeneity increase. Adapted from (Isildak et al., 2020).

analysis for both miRNAs and transcription factors.

Performing gene set enrichment analysis for transcription factors, I found 30 significantly enriched transcription factors in aging (see **Table D.1** for the full list), while no transcription factor was found to be significantly enriched in development. The significantly enriched transcription factors included transcription factors belonging to the Early Growth Response factor (EGR) and Forkhead box class O (FoxO) families, both of which are known to be associated with longevity and tissue homeostasis.

Then, the same gene set enrichment analysis was performed for miRNAs. Overall, I found only 2 miRNAs significantly enriched for the heterogeneity changes in development (see **Table E.1**), while there were 99 miRNAs that are significantly associated with heterogeneity changes during aging. While **Table E.2** shows the full list of miRNAs, the most notable ones are those belong to miR-34 family, which is a highly conserved family that was shown to be important modulator of brain ag-

ing (Kennerdell et al., 2018). Thus, the GSEA analysis for transcriptional regulators also indicated the potential relevance of increased heterogeneity in brain aging.

3.6.1 Association between the number of regulators and increased heterogeneity

As earlier studies suggested the association between the number of regulators and gene expression noise (Barroso et al., 2018; Sharon et al., 2014), I next sought to characterize the relationship between the number of regulators and gene expression heterogeneity for development and aging, separately (**Figure 3.7**). Spearman's correlation coefficient was calculated for miRNAs and transcription factors, separately. In the aging dataset, the correlation was mainly in the positive direction, where the 18/19 and 15/19 datasets showed a positive correlation for miRNAs and transcription factors, respectively. Moreover, the permutation test revealed that the difference between development and aging datasets is also significant for both miRNAs (permutation test p-value = 0.007) and transcription factors (permutation test p-value = 0.045).

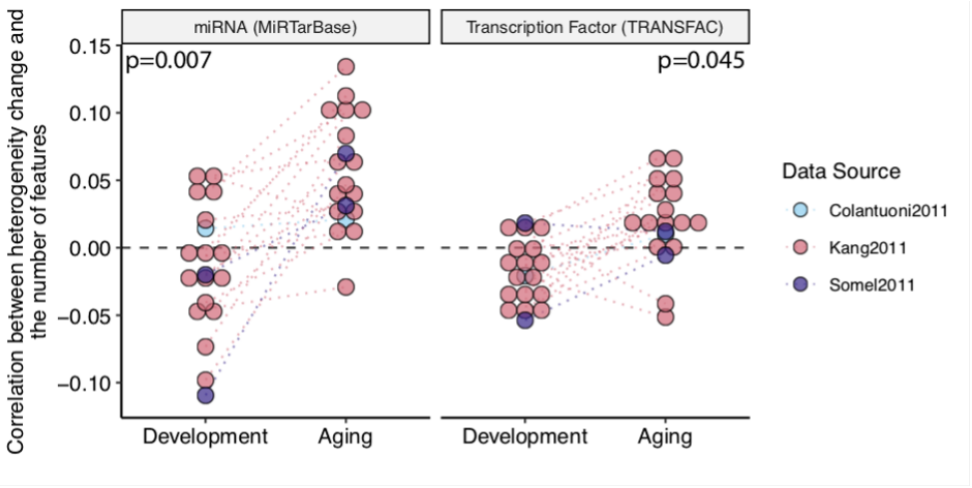

Figure 3.7: The distribution of correlation coefficients between number of transcriptional regulators and heterogeneity changes. The p-values were computed by permutation test. Adapted from (Isildak et al., 2020).

CHAPTER 4

DISCUSSION

In this study, I aimed to investigate the changes in gene expression heterogeneity during development and aging periods. The dataset that I analyzed included 19 time-series microarray datasets containing gene expression measurements for the human brain from 3 independent sources. Overall, 1,010 samples from 17 different brain regions of 298 individuals (**Table 2.1**, **Figure 2.1**) were included in this study.

The datasets, containing samples covering the whole lifespan (ages from 0 to 98 years old), were first divided into development and aging datasets, using the age of 20 years as a separation point (see **Section 2.1.2**), which was previously shown to be the global turning point of gene expression trajectories (Colantuoni et al., 2011; Dönertaş et al., 2017; Somel et al., 2010). Overall, I obtained 19 development datasets including samples whose ages range from 0 to 20 years old (n = 441). It is also important to note that pre-natal samples were excluded from the downstream analysis since the gene expression trajectories were suggested to be discontinous between pre and post-natal development, and the scope of this study was to compare heterogeneity changes during postnatal development and aging. The aging datasets (n = 19 datasets), on the other hand, included samples whose ages range from 20 to 98 (n = 569). Only the common genes (i.e., the genes for which I have measurement across all datasets) were included in the downstream analysis (n = 11,137).

Using the advantage of having multiple datasets, this study focused on consistent changes that are shared across the datasets, rather than focusing on significant changes in a single dataset, which is highly dependent on the sample sizes. Thus, this approach was able to capture weak but shared signals that would otherwise fail to pass the significance treshold in individual datasets.

4.1 Correlations among datasets in expression and heterogeneity changes

After performing preprocessing on microarray datasets (see **Section 2.2**), I first sought to characterize the age-related changes in gene expression, by performing a linear regression analysis between scaled expression values and fourth root of ages in days as shown in the left panels of **Figure 2.3**. The β_{i1} values obtained from **Equation 2.2** were considered as the measure of age-related expression change. The regression analysis was performed for each gene and for each time period, separately (see **Section 2.4** for details).

Then I investigated the coordination in expression change between all possible pairs of datasets by calculating the Spearman correlation coefficient and found that the correlation among development datasets is significantly higher than the correlation among aging datasets. Furthermore, more genes showed significant changes during development compared to the aging period, and genes showing significant changes during aging mostly tended to decrease in expression (**Figure 3.1**). One possible explanation of these results might be related to the stochastic nature of aging. As previously suggested, the accumulation of random detrimental effects (i.e., mutations) during aging may cause reduced gene expressions, and in turn, causing an increased level of heterogeneity in aging (Lu et al., 2004). Consistent with earlier findings, my initial analysis of gene expression changes also suggests that the changes in development are well-regulated, and further supports the view of aging as a stochastic process.

Next, the change in gene expression heterogeneity was characterized by performing Spearman's correlation test between absolute value of residuals obtained from **Equation 2.2** and the fourth root of age for each gene and time period separately (see **Section 2.5**). Analyzing the correlations among development and aging datasets in heterogeneity change, I found that aging datasets display a higher correlation compared to development datasets, reflecting a more consistent heterogeneity change in aging. Further analysis of genes showing significant heterogeneity changes revealed that there are more genes showing significant changes in heterogeneity during aging, compared to development (**Figure 3.2b**). Moreover, the significant changes in heterogeneity are mostly in the positive direction during aging, suggesting an increase in

heterogeneity.

4.2 Increased heterogeneity consistency in aging

Having observed more consistent heterogeneity change and more significant heterogeneity increase in aging, I next investigated heterogeneity changes in individual datasets and found an overall increase in heterogeneity during aging (i.e., 18 of 19 aging datasets display higher median heterogeneity change compared to development, see **Figure 3.3a**). An analysis of consistent heterogeneity change further revealed that there is a significant shift towards increased heterogeneity consistency during aging compared to random expectation, while no such shift was observed for development datasets (**Figure 3.3c**).

There are a number of factors that can explain increased heterogeneity during aging compared to development. First, many studies previously demonstrated that the stochastic accumulation of somatic mutations may cause genomic instability, which may, in turn, lead to increased heterogeneity in the aging period (Lodato et al., 2018; Lombard et al., 2005; Lu et al., 2004; Vijg, 2004). Second, Cheung *et al.*, analyzing a twin cohort, demonstrated the stochastic nature of age-related changes in chromatin, leading to increase variation between both individuals and cells in aging (Cheung et al., 2018). The third factor might be related to the transcriptional regulation, which was suggested to be an inherently stochastic process due to the randomness of biochemical reactions (Barroso et al., 2018; Maheshri & O'Shea, 2007). Previous studies found that the variability in gene expression is positively correlated with the number of transcription factors that control is regulation (Barroso et al., 2018; Sharon et al., 2014). While the first two factors could not be tested in this study since the datasets did not contain a somatic mutation or epigenetic regulation information, the investigation of the third factor revealed a mainly positive correlation between the number of regulators and heterogeneity change during aging (**Figure 3.7**). Therefore, the results obtained in this study also supported the view that transcriptional regulation may be a factor in the underlying mechanism of increased heterogeneity in aging.

4.3 Increased heterogeneity may have important functional consequences

Then, I investigated the functional consequences of increased heterogeneity in aging by conducting gene set enrichment analysis for GO Biological Processes and KEGG pathways. The significantly enriched KEGG pathways included many pathways that were known to be important in aging, including longevity regulating pathway, authophagy, and mTOR signaling pathways (**Figure 3.3**). Moreover, the significantly enriched GO terms also included terms that are related to aging and aging-related diseases, suggesting the functional significance of increased heterogeneity in aging and aging-related diseases. Additionally, GO terms related to neuronal and synaptic functions were also enriched for genes showing increased heterogeneity consistency during aging, indicating the potential role of increased heterogeneity in age-related cognitive decline, which was suggested to be mainly a result of synaptic dysfunction (Morrison & Baxter, 2012).

I also performed gene set enrichment analysis for transcriptional regulators. Transcription factors found to be significantly associated with increased heterogeneity during aging included FoxO and EGR family of transcription factors, which were shown to be regulating genes important for synaptic homeostasis, stress resistance, cell cycle arrest and apoptosis. Moreover, I identified 99 miRNAs that were significantly enriched for the genes showing consistent increase in heterogeneity, which may again have functional consequences, given the previous evidences suggesting the important role of miRNAs in regulating longevity (Liu et al., 2012; Shen et al., 2012; Smith-Vikos & Slack, 2012). Combined, gene set enrichment analysis revealed a potentially important role of increased heterogeneity in human brain aging.

4.4 Increased heterogeneity is a biological signal

I next confirmed that the observed increase in heterogeneity consistency is a biological signal, rather than being technical artifact or a result of low statistical power, given the similar sample sizes of development and aging periods (**Figure 2.1**).

One technical factor that can explain increased heterogeneity during aging might be

related to the dependence between mean and variance, where the accompanying increase in expression levels may cause an increased variance, which in turn was detected as increased heterogeneity. To address this issue, I performed a correlation test between heterogeneity changes and expression changes for each dataset and each period, and found that the correlation coefficients calculated for aging datasets are mostly negative (**Figure 3.3b**), suggesting that the observed increase in heterogeneity was not caused by mean-variance dependence.

Another technical factor that can cause increased inter-individual variance may be related to post-mortem interval (PMI), which measures the time between death and sample collection. It was previously suggested that PMI-related mRNA degradation is gene-specific, leading to a bias in downstream analysis (Zhu et al., 2017). To confirm that increased heterogeneity was not a result of PMI-related mRNA degradation, I used previously identified 107 PMI-associated genes (Zhu et al., 2017), 75 of which were included in this analysis. Specifically, I tested if the 75 PMI-associated genes show more increase in heterogeneity during aging, and found that only 2 of 147 consistent genes were PMI-associated, suggesting that PMI by itself was not enough to explain the observed increase in heterogeneity (**Figure F.1**).

One other factor that can affect the main results presented here might be related to age scales. As previously stated, the fourth root of age scale was used in this study to obtain a relatively uniform distribution of ages across the lifespan. However, whether the observed increase in heterogeneity depends on the use of specific age scales remained unanswered. To assess the effect of using different age scales on the downstream analysis, I repeated the analysis using 3 additional age scales: (1) age in days, (2) age in log2 scale, and (3) age in years (**Figure F.2**). Overall, I found that using different age scales also yields quantitatively similar results. In fact, the use of the log2 age scale resulted in a higher number of genes showing a consistent increase in heterogeneity across all 19 datasets (**Figure F.2b**, lower left panel). Nevertheless, this analysis indicated that the observed increase in heterogeneity is not a result of the use of a specific age scale.

It was previously suggested a sex-specific difference in human brain aging, where males showed more changes in gene expression (Berchtold et al., 2008). In this anal-

ysis, however, both males and females were combined to calculate expression and heterogeneity changes, raising a question about the possible confound of sex with age. To address this question, I retrieved the real values of residuals obtained from **Equation 2.2** (not absolute values) for 147 genes showing a consistent increase in heterogeneity. Then, for each gene and each aging dataset, two-sample Mann–Whitney U test was performed on residuals to test if there is a significant difference between males and females. The obtained p-values were corrected for multiple testing by B&H method (**Section 2.4.1**). Overall, I found that there are only 15 out of 147 consistent genes that show a significant difference between sexes in at least one dataset (**Figure F.3**), suggesting that the increased heterogeneity cannot be explained solely by sex-specific differences in brain aging.

In this study, a permutation scheme that takes into account the dependency of Kang2011 and Somel2011 datasets was employed to calculate expectation of heterogeneity consistency (**Section 2.8**). To test the effect of chosen permutation scheme, I also performed random permutations to calculate expected consistency in heterogeneity increase, and found that the scheme used in this study was more strict than the random permutations (**Figure F.4**).

One important assumption of this study is that the relationship between scaled expression levels and the fourth root of age is linear, as linear regression was used to characterize the age-related expression change. To ensure that this assumption did not have a significant effect on the downstream results, I re-calculated heterogeneity changes using the residuals obtained from loess regression and found a high correlation between heterogeneity changes calculated using linear regression and loess regression (**Figure F.5**). Yet, the heterogeneity changes calculated from loess regression did not include in the downstream analysis since both the model parameters and sample sizes have a significant effect on the estimates of loess regression.

The last factor that can cause the observed increase in heterogeneity might be related to the outliers in the datasets. For example, one older individual having too low or high expression value (i.e., having a higher absolute value of residual) can drive the heterogeneity estimates up. To investigate the effect of outliers, I plotted the absolute value of residuals for 147 consistent genes (**Figure F.5**). A visual inspection revealed

that there was no significant outlier sample that can explain the observed increase in heterogeneity.

Overall, these extra analyses demonstrated that the increased heterogeneity reported in this study cannot be explained by low statistical power and technical factors, but rather it is indeed a biological signal.

4.5 Limitations & future perspectives

1. Microarray platforms, unlike RNA-Sequencing, are unable to measure the absolute abundance of mRNAs. Rather, the light intensities obtained from the microarray platform reflect the relative expression levels. In their 2018 paper, Davie *et al.* found that the total abundance of mRNAs tends to decrease with age (Davie et al., 2018). Moreover, it was previously shown that the genes with lower gene expression values are susceptible to having higher variance (Aris et al., 2004). In this respect, the datasets analyzed in this study were unable to detect the contribution of mRNA decay to the observed increase in heterogeneity, and further research is needed to understand the contribution of total mRNA decay to the observed increase in gene expression heterogeneity.

2. Another limitation is also related to microarray datasets containing bulk mRNA expression data. Since the expression datasets analyzed in this study contain the average measurements for many cells, the results presented here only reflect increased heterogeneity among individuals, not cells. Future research using single-cell RNA-Sequencing data is required to investigate heterogeneity changes between cells (Ximerakis et al., 2019).

3. Although 19 different datasets were analyzed in this study, it is important to note that they originated from only 3 independent sources, where Somel2011 and Kang2011 datasets contain measurements from different brain regions of the same individual.

4. Although a significant overall shift was observed towards increased heterogeneity consistency during aging (**Figure 3.3c, lower panel**), a gene set that becomes significantly more heterogeneous across all datasets was not confidently

identified due to 40% true positive rate.

5. One other limitation is related to unequal sample sizes across datasets. Specifically, the Colantuoni2011 dataset had a markedly higher sample size compared to all other datasets, leading to higher statistical power, and subsequent identification of more significant genes in both expression (**Figure 3.1c**) and heterogeneity changes (**Figure 3.2c**).

6. While two possible explanations of increased heterogeneity are related to the accumulation of somatic mutations and epigenetic regulations, I could not test their effect due to a lack of data. More comprehensive studies incorporating different types of data are needed to reveal the source of increased heterogeneity.

CHAPTER 5

CONCLUSION

Aging is a complex process characterized by a gradual functional decline. Moreover, aging is considered to pose a major risk factor for many diseases, including cancer, and cardiovascular and neurodegenerative disorders. Transcriptome studies focusing on age-related changes in the human brain have been offering novel insights into the understanding of underlying mechanisms of aging-associated changes.

In this study, I conducted a meta-analysis of 19 microarray datasets containing 1,010 samples from 17 brain regions and covering the whole lifespan. Specifically, I investigated the changes in inter-individual gene expression heterogeneity during aging in comparison to the development period. The main findings of this study were summarized in the following bullet points below.

- There are more genes showing significant changes in gene expression during development compared to aging.
- In development, the majority of genes showing significant expression change decrease in expression.
- Development datasets show higher coordination in gene expression changes, compared to aging datasets.
- Gene expression heterogeneity consistently increases with age during aging (20 to 98 years of age) but not in postnatal development (0 to 20 years of age).
- The heterogeneity increase observed in aging comes with physiological consequences, such that the genes showing consistent effects are associated with biological processes important for life- and health-span regulation (e.g. autophagy,

mTOR signaling), as well as for cognitive functions (e.g. axon guidance, post-synaptic specialization).

- Not only specific regulators (miRNAs and transcription factors) but also the number of regulators is positively associated with consistent changes in heterogeneity.

Overall, the results presented in this thesis showed that gene expression heterogeneity between individuals increases with age in human brain. Further, this increase is limited to aging period, suggesting that increased heterogeneity is not only a function of time that starts at developmental stages. Moreover, genes showing consistent increase in heterogeneity are associated with pathways important for neuronal function and aging, highlighting the possible significance of increased heterogeneity in aging. Future work incorporating larger datasets can provide a deeper understanding of the underlying mechanisms of increased heterogeneity.

REFERENCES

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., Mayr, C. H., Greiffo, F. R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T.-M., Nagendran, M., Desai, T., Eickelberg, O., Mann, M., Theis, F. J., & Schiller, H. B. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, *10*, 963. https://doi.org/10.1038/s41467-019-08831-9

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*. https://doi.org/10.1186/1471-2105-5-185

Armanios, M., Alder, J. K., Parry, E. M., Karim, B., Strong, M. A., & Greider, C. W. (2009). Short telomeres are sufficient to cause the degenerative defects associated with aging. *The American Journal of Human Genetics*, *85*, 823–832. https://doi.org/10.1016/j.ajhg.2009.10.028

Ashapkin, V. V., Kutueva, L. I., & Vanyushin, B. F. (2017). Aging as an epigenetic phenomenon. *Current Genomics*, *18*, 385. https://doi.org/10.2174/1389202918666170412112130

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

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*, D991–D995. https://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*, 173–189. https://doi.org/10.1534/genetics.117.300467

Barzilai, N., Huffman, D. M., Muzumdar, R. H., & Bartke, A. (2012). The critical role of metabolic pathways in aging. *Diabetes*, *61*, 1315–1322. https://doi.org/10.2337/db11-1300

Belin, S., Norsworthy, M., & He, Z. (2014). Independent control of aging and axon regeneration. *Cell metabolism*, *19*, 354. https://doi.org/10.1016/J.CMET.2014.02.014

Bengtsson, H., & Hössjer, O. (2006). Methodological study of affine transformations of gene expression data with proposed robust non-parametric multi-dimensional normalization method. *BMC bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-100

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

Berchtold, N. C., Cribbs, D. H., Coleman, P. D., Rogers, J., Head, E., Kim, R., Beach, T., Miller, C., Troncoso, J., Trojanowski, J. Q., Zielke, H. R., & Cotman, C. W. (2008). Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 15605–15610. https://doi.org/10.1073/PNAS.0806883105/SUPPL_FILE/ST9.XLS

Biran, A., Zada, L., Karam, P. A., Vadai, E., Roitman, L., Ovadya, Y., Porat, Z., & Krizhanovsky, V. (2017). Quantitative identification of senescent cells in aging and disease. *Aging Cell*, *16*, 661–671. https://doi.org/10.1111/acel.12592

Blasco, M. A. (2007). Telomere length, stem cells and aging. *Nature Chemical Biology*, *3*, 640–649. https://doi.org/10.1038/nchembio.2007.38

Bolstad, B. (2021). *Preprocesscore: a collection of pre-processing functions* [R package version 1.56.0].

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics (Oxford, England)*, *26*, 2363–2367. https://doi.org/10.1093/BIOINFORMATICS/BTQ431

Cheung, P., Vallania, F., Warsinske, H. C., Donato, M., Schaffert, S., Chang, S. E., Dvorak, M., Dekker, C. L., Davis, M. M., Utz, P. J., Khatri, P., & Kuo, A. J. (2018). Single-cell chromatin modification profiling reveals increased epigenetic variations with aging. *Cell*, *173*, 1385–1397.e14. https://doi.org/10.1016/j.cell.2018.03.079

Childs, B. G., Baker, D. J., Wijshake, T., Conover, C. A., Campisi, J., & van Deursen, J. M. (2016). Senescent intimal foam cells are deleterious at all stages of atherosclerosis. *Science*, *354*, 472–477. https://doi.org/10.1126/science.aaf6659

Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., Tsai, T.-R., Ho, S.-Y., Jian, T.-Y., Wu, H.-Y., Chen, P.-R., Lin, N.-C., Huang, H.-T., Yang, T.-L., Pai, C.-Y., ... Huang, H.-D. (2016). Mirtarbase 2016: updates to the experimentally validated mirna-target interactions database. *Nucleic Acids Research*, *44*, D239–D247. https://doi.org/10.1093/nar/gkv1258

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

Consortium, T. G. O. (2019). The gene ontology resource: 20 years and still going strong. *Nucleic Acids Research*, *47*, D330–D338. https://doi.org/10.1093/nar/gky1055

Covey, H. C. (1992). The definitions of the beginning of old age in history. *International journal of aging and human development*, *34*, 325–37. https://doi.org/10.2190/GBXB-BE1F-1BU1-7FKK

Davie, K., Janssens, J., Koldere, D., Waegeneer, M. D., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., González-Blas, C. B., Poovathingal, S., Hulselmans, G., Spanier, K. I., Moerman, T., Vanspauwen, B., Geurs, S., Voet, T., Lammertyn, J., Thienpont, B., ... Aerts, S. (2018). A single-cell transcriptome atlas of the aging drosophila brain. *Cell*, *174*, 982–998.e20. https://doi.org/10.1016/j.cell.2018.05.057

Dolcos, F., Rice, H. J., & Cabeza, R. (2002). Hemispheric asymmetry and aging: right hemisphere decline or asymmetry reduction. *Neuroscience and biobehavioral reviews*, *26*, 819–825. https://doi.org/10.1016/S0149-7634(02)00068-4

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. *Scientific Reports*, *7*, 5894. https://doi.org/10.1038/s41598-017-05927-4

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

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols 2009 4:8*, *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97

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*, 321–330.e14. https://doi.org/10.1016/j.cell.2017.09.004

Fontana, L., Partridge, L., & Longo, V. D. (2010). Extending healthy life span–from yeast to humans. *Science (New York, N.Y.)*, *328*, 321–6. https://doi.org/10.1126/science.1172539

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suñer, D., Cigudosa, J. C., Urioste, M., Benitez, J., Boix-Chornet, M., Sanchez-Aguilera, A., Ling, C., Carlsson, E., Poulsen, P., Vaag, A., Stephan, Z., Spector, T. D., Wu, Y.-Z., ... Esteller, M. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, *102*, 10604–10609. https://doi.org/10.1073/pnas.0500398102

Frenk, S., & Houseley, J. (2018). Gene expression hallmarks of cellular ageing. *Biogerontology*, *19*, 547–566. https://doi.org/10.1007/s10522-018-9750-z

Gladyshev, V. N. (2016). Aging: progressive decline in fitness due to the rising deleteriome adjusted by genetic, environmental, and stochastic processes. *Aging Cell*, *15*, 594–602. https://doi.org/10.1111/acel.12480

Green, D. R., Galluzzi, L., & Kroemer, G. (2011). Mitochondria and the autophagy-inflammation-cell death axis in organismal aging. *Science (New York, N.Y.)*, *333*, 1109–12. https://doi.org/10.1126/science.1201940

Hekimi, S., Lapointe, J., & Wen, Y. (2011). Taking a "good" look at free radicals in the aging process. *Trends in cell biology*, *21*, 569–576. https://doi.org/10.1016/J.TCB.2011.06.008

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

Hicks, S. C., & Irizarry, R. A. (2014). When to use quantile normalization? *bioRxiv*, 012203. https://doi.org/10.1101/012203

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic acids research*, *31*, e15. https://doi.org/10.1093/NAR/GNG015

Isildak, U., Somel, M., Thornton, J. M., & Dönertaş, H. M. (2020). Temporal changes in the gene expression heterogeneity during brain development and aging. *Scientific Reports 2020 10:1*, *10*, 1–15. https://doi.org/10.1038/s41598-020-60998-0

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

Kadota, T., Horinouchi, T., & Kuroda, C. (2001). Development and aging of the cerebrum: assessment with proton mr spectroscopy. *AJNR. American journal of neuroradiology*, *22*, 128–35.

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

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A. M. M., Pletikos, M., Meyer, K. A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M. B., Krsnik, Ž., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S. N., Vortmeyer, A., … Šestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature*, *478*, 483–489. https://doi.org/10.1038/nature10523

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*, *11*, 2253–2280. https://doi.org/10.18632/aging.101912

Kennerdell, J. R., Liu, N., & Bonini, N. M. (2018). Mir-34 inhibits polycomb repressive complex 2 to modulate chaperone expression and promote healthy brain aging. *Nature Communications*, *9*, 4188. https://doi.org/10.1038/s41467-018-06592-5

Kenyon, C. J. (2010). The genetics of ageing. *Nature*, *464*, 504–12. https://doi.org/10.1038/nature08980

Koga, H., Kaushik, S., & Cuervo, A. M. (2011). Protein homeostasis and aging: the importance of exquisite quality control. *Ageing Research Reviews*, *10*, 205–215. https://doi.org/10.1016/j.arr.2010.02.001

Kolde, R. (2019). *Preprocesscore: a collection of pre-processing functions* [R package version 1.0.12].

Korotkevich, G., Sukhov, V., & Sergushichev, A. (2019). Fast gene set enrichment analysis. *bioRxiv*. https://doi.org/10.1101/060012

Kuilman, T., Michaloglou, C., Mooi, W. J., & Peeper, D. S. (2010). The essence of senescence. *Genes and development*, *24*, 2463–79. https://doi.org/10.1101/gad.1971610

Kumar, A., Gibbs, J. R., Beilina, A., Dillman, A., Kumaran, R., Trabzuni, D., Ryten, M., Walker, R., Smith, C., Traynor, B. J., Hardy, J., Singleton, A. B., & Cookson, M. R. (2013). Age-associated changes in gene expression in human brain and isolated neurons. *Neurobiology of Aging*, *34*, 1199–1209. https://doi.org/10.1016/j.neurobiolaging.2012.10.021

Liu, N., Landreh, M., Cao, K., Abe, M., Hendriks, G.-J., Kennerdell, J. R., Zhu, Y., Wang, L.-S., & Bonini, N. M. (2012). The microrna mir-34 modulates ageing and neurodegeneration in drosophila. *Nature*, *482*, 519–523. https://doi.org/10.1038/nature10810

Lodato, M. A., Rodin, R. E., Bohrson, C. L., Coulter, M. E., Barton, A. R., Kwon, M., Sherman, M. A., Vitzthum, C. M., Luquette, L. J., Yandava, C. N., Yang, P., Chittenden, T. W., Hatem, N. E., Ryu, S. C., Woodworth, M. B., Park, P. J., & Walsh, C. A. (2018). Aging and neurodegeneration are associated

with increased mutations in single human neurons. *Science*, *359*, 555–559. https://doi.org/10.1126/science.aao4426

Lombard, D. B., Chua, K. F., Mostoslavsky, R., Franco, S., Gostissa, M., & Alt, F. W. (2005). Dna repair, genome stability, and aging. *Cell*, *120*, 497–512. https://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*, 1194–1217. https://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*, 883–891. https://doi.org/10.1038/nature02661

Magnotta, V. A. (1999). Quantitative in vivo measurement of gyrification in the human brain: changes associated with aging. *Cerebral Cortex*, *9*, 151–160. https://doi.org/10.1093/cercor/9.2.151

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*, 413–434. https://doi.org/10.1146/annurev.biophys.36.040306.132705

Martinez-Jimenez, C. P., Eling, N., Chen, H.-C., Vallejos, C. A., Kolodziejczyk, A. A., Connor, F., Stojic, L., Rayner, T. F., Stubbington, M. J. T., Teichmann, S. A., de la Roche, M., Marioni, J. C., & Odom, D. T. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science*, *355*, 1433–1436. https://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*, 196–207. https://doi.org/10.1111/acel.12427

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., & Wingender, E. (2006). Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, *34*, D108–10. https://doi.org/10.1093/nar/gkj143

Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Chen, Y. P. P., Li, N., Hu, Y., Fu, N., Ning, Z., Zeng, R., Yang, H., Chen, W.,

Gelfand, M., & Khaitovich, P. (2013). Widespread splicing changes in human brain development and aging. *Molecular Systems Biology*, *9*, 633. https://doi.org/10.1038/MSB.2012.67

Morrison, J. H., & Baxter, M. G. (2012). The ageing cortical synapse: hallmarks and implications for cognitive decline. *Nature Reviews Neuroscience*, *13*, 240–250. https://doi.org/10.1038/nrn3200

Moskalev, A. A., Shaposhnikov, M. V., Plyusnina, E. N., Zhavoronkov, A., Budovsky, A., Yanai, H., & Fraifeld, V. E. (2013). The role of dna damage and repair in aging through the prism of koch-like criteria. *Ageing Research Reviews*, *12*, 661–684. https://doi.org/10.1016/j.arr.2012.02.001

Naumova, O. Y., Palejev, D., Vlasova, N. V., Lee, M., Rychkov, S. Y., Babich, O. N., Vaccarino, F. M., & 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. *Development and Psychopathology*, *24*, 1427–1442. https://doi.org/10.1017/S0954579412000818

Niccoli, T., & Partridge, L. (2012). Ageing as a risk factor for disease. *Current Biology*, *22*, R741–R752. https://doi.org/10.1016/j.cub.2012.07.024

Oh, J., Lee, Y. D., & Wagers, A. J. (2014). Stem cell aging: mechanisms, regulators and therapeutic opportunities. *Nature medicine*, *20*, 870–80. https://doi.org/10.1038/nm.3651

Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W., & Balch, W. E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. *Annual Review of Biochemistry*, *78*, 959–991. https://doi.org/10.1146/annurev.biochem.052308.114844

R Core Team. (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rando, T. A., & Chang, H. Y. (2012). Aging, rejuvenation, and epigenetic reprogramming: resetting the aging clock. *Cell*, *148*, 46–57. https://doi.org/10.1016/j.cell.2012.01.003

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. https://doi.org/10.1093/database/baw100

Rubinsztein, D. C., Mariño, G., & Kroemer, G. (2011). Autophagy and aging. *Cell*, *146*, 682–695. https://doi.org/10.1016/j.cell.2011.07.030

Russell, S. J., & Kahn, C. R. (2007). Endocrine regulation of ageing. *Nature Reviews Molecular Cell Biology*, *8*, 681–691. https://doi.org/10.1038/nrm2234

Salat, D. H. (2004). Thinning of the cerebral cortex in aging. *Cerebral Cortex*, *14*, 721–730. https://doi.org/10.1093/cercor/bhh032

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*, 507–514. https://doi.org/10.1016/j.neurobiolaging.2008.09.023

Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z., & Segal, E. (2014). Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Research*, *24*, 1698–1706. https://doi.org/10.1101/gr.168773.113

Sheline, Y. I., Mintun, M. A., Moerlein, S. M., & Snyder, A. Z. (2002). Greater loss of 5-ht _{2a} receptors in midlife than in late life. *American Journal of Psychiatry*, *159*, 430–435. https://doi.org/10.1176/appi.ajp.159.3.430

Shen, Y., Wollam, J., Magner, D., Karalay, O., & Antebi, A. (2012). A steroid receptor-microrna switch regulates life span in response to signals from the gonad. *Science*, *338*, 1472–1476. https://doi.org/10.1126/science.1228967

Smith-Vikos, T., & Slack, F. J. (2012). Micrornas and their roles in aging. *Journal of cell science*, *125*, 7–17. https://doi.org/10.1242/jcs.099200

Somel, M., Guo, S., Fu, N., Yan, Z., Hu, H. Y., Xu, Y., Yuan, Y., Ning, Z., Hu, Y., Menzel, C., Hu, H., Lachmann, M., Zeng, R., Chen, W., & Khaitovich, P. (2010). Microrna, mrna, and protein expression link development and aging in human and macaque brain. *Genome Research*, *20*, 1207–1218. https://doi.org/10.1101/gr.106849.110

Somel, M., Khaitovich, P., Bahn, S., Pääbo, S., & Lachmann, M. (2006). Gene expression becomes heterogeneous with age. *Current Biology*, *16*, R359–R360. https://doi.org/10.1016/j.cub.2006.04.024

Somel, M., Liu, X., Tang, L., Yan, Z., Hu, H., Guo, S., Jiang, X., Zhang, X., Xu, G., Xie, G., Li, N., Hu, Y., Chen, W., Pääbo, S., & Khaitovich, P. (2011). Microrna-driven developmental remodeling in the brain distinguishes humans

from other primates (D. Penny, Ed.). *PLoS Biology*, *9*, e1001214. https://doi.org/10.1371/journal.pbio.1001214

Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., & Toga, A. W. (2003). Mapping cortical change across the human life span. *Nature Neuroscience*, *6*, 309–315. https://doi.org/10.1038/nn1008

Sowell, E. R., Thompson, P. M., & Toga, A. W. (2004). Mapping changes in the human cortex throughout the span of life. *The Neuroscientist*, *10*, 372–392. https://doi.org/10.1177/1073858404263960

Stead, E. R., & Bjedov, I. (2021). Balancing dna repair to prevent ageing and cancer. *Experimental Cell Research*, *405*, 112679. https://doi.org/10.1016/j.yexcr.2021.112679

Sullivan, E. V., & Pfefferbaum, A. (2006). Diffusion tensor imaging and aging. *Neuroscience and Biobehavioral Reviews*, *30*, 749–761. https://doi.org/10.1016/j.neubiorev.2006.06.002

Sun, T., Patoine, C., Abu-Khalil, A., Visvader, J., Sum, E., Cherry, T. J., Orkin, S. H., Geschwind, D. H., & Walsh, C. A. (2005). Early asymmetry of gene transcription in embryonic human left and right cerebral cortex. *Science (New York, N.Y.)*, *308*, 1794–1798. https://doi.org/10.1126/SCIENCE.1110324

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms (C. Gibas, Ed.). *PLoS ONE*, *6*, e21800. https://doi.org/10.1371/journal.pone.0021800

Symphorien, S., & Woodruff, R. C. (2003). Effect of dna repair on aging of transgenic drosophila melanogaster: i. mei-41 locus. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *58*, B782–B787. https://doi.org/10.1093/gerona/58.9.B782

Tan, Q., Liang, N., Zhang, X., & Li, J. (2021). Dynamic aging: channeled through microenvironment. *Frontiers in Physiology*, *12*, 1089. https://doi.org/10.3389/FPHYS.2021.702276/BIBTEX

Tennekes, M. (2021). *Treemap: treemap visualization* [R package version 2.4-3].

Tomás-Loba, A., Flores, I., Fernández-Marcos, P. J., Cayuela, M. L., Maraver, A., Tejera, A., Borrás, C., Matheu, A., Klatt, P., Flores, J. M., Viña, J., Serrano, M., & Blasco, M. A. (2008). Telomerase reverse transcriptase delays aging in

cancer-resistant mice. *Cell*, *135*, 609–622. https://doi.org/10.1016/j.cell.2008.09.034

Vijg, J. (2004). Impact of genome instability on transcription regulation of aging and senescence. *Mechanisms of Ageing and Development*, *125*, 747–753. https://doi.org/10.1016/j.mad.2004.07.004

Vijg, J. (2009). Snp'ing for longevity. *Aging*, *1*, 442–443. https://doi.org/10.18632/aging.100048

Viñuela, A., Brown, A. A., Buil, A., Tsai, P.-C., Davies, M. N., Bell, J. T., Dermitzakis, E. T., Spector, T. D., & Small, K. S. (2018). Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Human Molecular Genetics*, *27*, 732–741. https://doi.org/10.1093/hmg/ddx424

Walker, R., Gurven, M., Hill, K., Migliano, A., Chagnon, N., Souza, R. D., Djurovic, G., Hames, R., Hurtado, A. M., Kaplan, H., Kramer, K., Oliver, W. J., Valeggia, C., & Yamauchi, T. (2006). Growth rates and life histories in twenty-two small-scale societies. *American Journal of Human Biology*, *18*, 295–311. https://doi.org/10.1002/ajhb.20510

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., xiaochong Fu, Liu, S., Bo, X., & Yu, G. (2021). Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

Ximerakis, M., Lipnick, S. L., Innes, B. T., Simmons, S. K., Adiconis, X., Dionne, D., Mayweather, B. A., Nguyen, L., Niziolek, Z., Ozek, C., Butty, V. L., Isserlin, R., Buchanan, S. M., Levine, S. S., Regev, A., Bader, G. D., Levin, J. Z., & Rubin, L. L. (2019). Single-cell transcriptomic profiling of the aging mouse brain. *Nature Neuroscience*, *22*, 1696–1708. https://doi.org/10.1038/s41593-019-0491-3

Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C. V., Schadt, E. E., Zhu, J., & Tu, Z. (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific Reports*, *5*, 15145. https://doi.org/10.1038/srep15145

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan,

S., Lavidas, I., … Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, *44*, D710–D716. https://doi.org/10.1093/NAR/GKV1157

Yousefzadeh, M. J., Zhao, J., Bukata, C., Wade, E. A., McGowan, S. J., Angelini, L. A., Bank, M. P., Gurkar, A. U., McGuckian, C. A., Calubag, M. F., Kato, J. I., Burd, C. E., Robbins, P. D., & Niedernhofer, L. J. (2020). Tissue specificity of senescent cell accumulation during physiologic and accelerated aging of mice. *Aging Cell*, *19*, e13094. https://doi.org/10.1111/acel.13094

Zhang, J. D., & Wiemann, S. (2009). Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, *25*(11), 1470–1471.

Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports 2020 10:1*, *10*, 1–11. https://doi.org/10.1038/s41598-020-72664-6

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*, 5435. https://doi.org/10.1038/s41598-017-05882-0

Appendix A

LIST OF 147 GENES SHOWING CONSISTENT AGE-RELATED HETEROGENEITY INCREASE AMONG ALL 19 AGING DATASETS

Table A.1: List of 147 genes showing consistent age-related heterogeneity increase in all 19 datasets during aging period and their cluster numbers.

Ensembl Gene ID	HGNC Symbol	Cluster
ENSG00000005339	CREBBP	3
ENSG00000003056	M6PR	6
ENSG00000005339	CREBBP	3
ENSG00000009844	VTA1	7
ENSG00000011243	AKAP8L	1
ENSG00000015153	YAF2	3
ENSG00000021574	SPAST	5
ENSG00000026652	AGPAT4	8
ENSG00000031003	FAM13B	6
ENSG00000039523	RIPOR1	1
ENSG00000048740	CELF2	8
ENSG00000065150	IPO5	7
ENSG00000065989	PDE4A	7
ENSG00000067182	TNFRSF1A	4
ENSG00000073910	FRY	5
ENSG00000074657	ZNF532	8
ENSG00000075303	SLC25A40	1
ENSG00000076201	PTPN23	6
ENSG00000078804	TP53INP2	6

Table A.1 (continued)

ENSG00000082701	GSK3B	4
ENSG00000083544	TDRD3	3
ENSG00000086102	NFX1	5
ENSG00000086598	TMED2	3
ENSG00000086758	HUWE1	3
ENSG00000088538	DOCK3	5
ENSG00000088812	ATRN	4
ENSG00000088888	MAVS	5
ENSG00000092010	PSME1	2
ENSG00000092847	AGO1	3
ENSG00000095397	WHRN	7
ENSG00000100345	MYH9	1
ENSG00000101160	CTSZ	3
ENSG00000101384	JAG1	8
ENSG00000102317	RBM3	4
ENSG00000102858	MGRN1	1
ENSG00000104442	ARMC1	3
ENSG00000104904	OAZ1	8
ENSG00000105397	TYK2	6
ENSG00000105497	ZNF175	4
ENSG00000105784	RUNDC3B	7
ENSG00000107263	RAPGEF1	3
ENSG00000107669	ATE1	3
ENSG00000107736	CDH23	2
ENSG00000107862	GBF1	5
ENSG00000110092	CCND1	1
ENSG00000110274	CEP164	5
ENSG00000111348	ARHGDIB	4
ENSG00000111676	ATN1	6
ENSG00000111707	SUDS3	4
ENSG00000111880	RNGTT	3
ENSG00000112996	MRPS30	4
ENSG00000113273	ARSB	4
ENSG00000113638	TTC33	3
ENSG00000114054	PCCB	8

Table A.1 (continued)

ENSG00000114796	KLHL24	4
ENSG00000115306	SPTBN1	2
ENSG00000117713	ARID1A	1
ENSG00000119878	CRIPT	1
ENSG00000120306	CYSTM1	8
ENSG00000123178	SPRYD7	6
ENSG00000123243	ITIH5	2
ENSG00000125107	CNOT1	5
ENSG00000125675	GRIA3	4
ENSG00000127337	YEATS4	3
ENSG00000128191	DGCR8	5
ENSG00000128268	MGAT3	2
ENSG00000129219	PLD2	4
ENSG00000129250	KIF1C	6
ENSG00000130158	DOCK6	6
ENSG00000132692	BCAN	6
ENSG00000134283	PPHLN1	7
ENSG00000135424	ITGA7	2
ENSG00000137558	PI15	4
ENSG00000138078	PREPL	4
ENSG00000138162	TACC2	3
ENSG00000138629	UBL7	3
ENSG00000138663	COPS4	3
ENSG00000139180	NDUFA9	2
ENSG00000140391	TSPAN3	8
ENSG00000140464	PML	4
ENSG00000140829	DHX38	1
ENSG00000141644	MBD1	2
ENSG00000142892	PIGK	3
ENSG00000143314	MRPL24	8
ENSG00000144834	TAGLN3	2
ENSG00000144867	SRPRB	3
ENSG00000146830	GIGYF1	3
ENSG00000147649	MTDH	4
ENSG00000147669	POLR2K	3

Table A.1 (continued)

ENSG00000149294	NCAM1	5
ENSG00000149636	DSN1	7
ENSG00000151067	CACNA1C	8
ENSG00000151240	DIP2C	8
ENSG00000154620	TMSB4Y	4
ENSG00000155034	FBXL18	7
ENSG00000156113	KCNMA1	4
ENSG00000157103	SLC6A1	2
ENSG00000158805	ZNF276	6
ENSG00000158869	FCER1G	4
ENSG00000160007	ARHGAP35	3
ENSG00000160191	PDE9A	2
ENSG00000160460	SPTBN4	1
ENSG00000161202	DVL3	3
ENSG00000162601	MYSM1	5
ENSG00000164070	HSPA4L	6
ENSG00000164332	UBLCP1	6
ENSG00000165102	HGSNAT	5
ENSG00000165516	KLHDC2	3
ENSG00000166582	CENPV	8
ENSG00000166825	ANPEP	7
ENSG00000168056	LTBP3	1
ENSG00000169180	XPO6	5
ENSG00000170500	LONRF2	4
ENSG00000171055	FEZ2	6
ENSG00000171105	INSR	8
ENSG00000171509	RXFP1	7
ENSG00000172037	LAMB2	3
ENSG00000172260	NEGR1	8
ENSG00000172262	ZNF131	7
ENSG00000172534	HCFC1	1
ENSG00000174669	SLC29A2	5
ENSG00000174780	SRP72	7
ENSG00000174842	GLMN	7
ENSG00000177728	TMEM94	6

Table A.1 (continued)

ENSG00000178913	TAF7	5
ENSG00000179021	C3orf38	3
ENSG00000179152	TCAIM	3
ENSG00000181929	PRKAG1	2
ENSG00000182197	EXT1	4
ENSG00000182400	TRAPPC6B	6
ENSG00000182492	BGN	3
ENSG00000182541	LIMK2	4
ENSG00000182872	RBM10	2
ENSG00000184677	ZBTB40	4
ENSG00000185236	RAB11B	7
ENSG00000185608	MRPL40	2
ENSG00000185650	ZFP36L1	6
ENSG00000185721	DRG1	7
ENSG00000185950	IRS2	1
ENSG00000186350	RXRA	1
ENSG00000196352	CD55)	8
ENSG00000196498	NCOR2	1
ENSG00000198356	ASNA1)	5
ENSG00000198898	CAPZA2	3
ENSG00000205423	CNEP1R1	7
ENSG00000213246	SUPT4H1	2
ENSG00000221869	CEBPD	6
ENSG00000229236	TTTY10	4

Appendix B

LIST OF SIGNIFICANTLY ENRICHED GO BIOLOGICAL PROCESS CATEGORIES

Table B.1: The list of GO Biological Process categories significantly enriched (adjusted p-value < 0.1)

ID	Description	setSize	NES
GO:0098789	pre-mRNA cleavage required for polyadenylation	8	2.471
GO:0051028	mRNA transport	99	2.388
GO:0050657	nucleic acid transport	129	2.330
GO:0050658	RNA transport	129	2.330
GO:0006611	protein export from nucleus	123	2.311
GO:1990090	cellular response to nerve growth factor stimulus	38	2.287
GO:0035601	protein deacylation	83	2.276
GO:0098732	macromolecule deacylation	83	2.276
GO:0006476	protein deacetylation	79	2.275
GO:0098787	mRNA cleavage involved in mRNA processing	10	2.264
GO:1990089	response to nerve growth factor	40	2.264
GO:0000381	regulation of alternative mRNA splicing, via spliceosome	28	2.257
GO:0051236	establishment of RNA localization	132	2.253
GO:0006406	mRNA export from nucleus	74	2.243
GO:0071427	mRNA-containing ribonucleoprotein..	74	2.243

Table B.1 (continued)

GO:0006405	RNA export from nucleus	94	2.237
GO:0043467	regulation of generation of precursor metabolites and energy	75	2.234
GO:0000380	alternative mRNA splicing, via spliceosome	39	2.232
GO:0050684	regulation of mRNA processing	77	2.230
GO:0006403	RNA localization	156	2.222
GO:0016575	histone deacetylation	69	2.222
GO:0071166	ribonucleoprotein complex localization	85	2.219
GO:0071426	ribonucleoprotein complex export from nucleus	85	2.219
GO:0051168	nuclear export	134	2.206
GO:0032239	regulation of nucleobase-containing compound transport	14	2.193
GO:0000288	nuclear-transcribed mRNA catabolic process..	51	2.192
GO:0040029	regulation of gene expression, epigenetic	191	2.144
GO:0070936	protein K48-linked ubiquitination	34	2.136
GO:1903312	negative regulation of mRNA metabolic process	55	2.124
GO:0051131	chaperone-mediated protein complex assembly	12	2.122
GO:0000289	nuclear-transcribed mRNA poly(A) tail shortening	25	2.118
GO:0048024	regulation of mRNA splicing, via spliceosome	49	2.116
GO:0048702	embryonic neurocranium morphogenesis	6	2.112
GO:0090311	regulation of protein deacetylation	31	2.106
GO:0046831	regulation of RNA export from nucleus	12	2.105
GO:0050651	dermatan sulfate proteoglycan biosynthetic process	13	2.105
GO:0050655	dermatan sulfate proteoglycan metabolic process	13	2.105
GO:0048679	regulation of axon regeneration	16	2.098

Table B.1 (continued)

GO:1903313	positive regulation of mRNA metabolic process	44	2.093
GO:0035020	regulation of Rac protein signal transduction	14	2.092
GO:0009303	rRNA transcription	21	2.092
GO:0046051	UTP metabolic process	7	2.090
GO:0006342	chromatin silencing	55	2.090
GO:0032922	circadian regulation of gene expression	38	2.077
GO:0017148	negative regulation of translation	145	2.064
GO:0033119	negative regulation of RNA splicing	19	2.061
GO:0006379	mRNA cleavage	20	2.058
GO:0060968	regulation of gene silencing	76	2.052
GO:0010257	NADH dehydrogenase complex assembly	39	2.051
GO:0032981	mitochondrial respiratory chain complex I assembly	39	2.051
GO:0042177	negative regulation of protein catabolic process	76	2.051
GO:0034249	negative regulation of cellular amide metabolic process	158	2.051
GO:0000956	nuclear-transcribed mRNA catabolic process	107	2.047
GO:1900151	regulation of nuclear-transcribed mRNA..	12	2.038
GO:1900153	positive regulation of nuclear-transcribed mRNA..	12	2.038
GO:0016573	histone acetylation	111	2.038
GO:1903311	regulation of mRNA metabolic process	186	2.027
GO:0016458	gene silencing	156	2.026
GO:0070932	histone H3 deacetylation	14	2.024
GO:0009208	pyrimidine ribonucleoside triphosphate metabolic process	12	2.022
GO:0006228	UTP biosynthetic process	6	2.021
GO:0006475	internal protein amino acid acetylation	115	2.014
GO:0060147	regulation of posttranscriptional gene silencing	58	2.011

Table B.1 (continued)

GO:0060966	regulation of gene silencing by RNA	58	2.011
GO:0051904	pigment granule transport	18	2.011
GO:0035278	miRNA mediated inhibition of translation	9	2.011
GO:0040033	negative regulation of translation, ncRNA-mediated	9	2.011
GO:0045974	regulation of translation, ncRNA-mediated	9	2.011
GO:0018393	internal peptidyl-lysine acetylation	112	2.009
GO:0051084	'de novo' posttranslational protein folding	21	2.002
GO:0018394	peptidyl-lysine acetylation	118	1.994
GO:0045815	positive regulation of gene expression, epigenetic	35	1.992
GO:0050686	negative regulation of mRNA processing	23	1.990
GO:1903363	negative regulation of cellular protein catabolic process	44	1.989
GO:0035195	gene silencing by miRNA	80	1.985
GO:0043967	histone H4 acetylation	49	1.978
GO:0006220	pyrimidine nucleotide metabolic process	35	1.970
GO:0016570	histone modification	325	1.970
GO:0006913	nucleocytoplasmic transport	234	1.964
GO:0043966	histone H3 acetylation	47	1.964
GO:0048675	axon extension	94	1.961
GO:0015931	nucleobase-containing compound transport	157	1.959
GO:0009895	negative regulation of catabolic process	180	1.948
GO:0006417	regulation of translation	280	1.943
GO:0016441	posttranscriptional gene silencing	86	1.941
GO:0051169	nuclear transport	236	1.936
GO:0046513	ceramide biosynthetic process	38	1.934
GO:0031124	mRNA 3'-end processing	64	1.934
GO:0016569	covalent chromatin modification	334	1.933
GO:0098586	cellular response to virus	29	1.930
GO:0032456	endocytic recycling	28	1.930

Table B.1 (continued)

GO:0051196	regulation of coenzyme metabolic process	59	1.926
GO:0030520	intracellular estrogen receptor signaling pathway	39	1.925
GO:0060964	regulation of gene silencing by miRNA	55	1.923
GO:0045814	negative regulation of gene expression, epigenetic	66	1.917
GO:0031330	negative regulation of cellular catabolic process	147	1.912
GO:0046822	regulation of nucleocytoplasmic transport	79	1.909
GO:0010608	posttranscriptional regulation of gene expression	360	1.908
GO:0031047	gene silencing by RNA	102	1.906
GO:0060304	regulation of phosphatidylinositol dephosphorylation	6	1.905
GO:0016241	regulation of macroautophagy	122	1.901
GO:0034248	regulation of cellular amide metabolic process	317	1.899
GO:0035194	posttranscriptional gene silencing by RNA	85	1.896
GO:0006283	transcription-coupled nucleotide-excision repair	42	1.892
GO:0043484	regulation of RNA splicing	78	1.877
GO:0033108	mitochondrial respiratory chain complex assembly	52	1.873
GO:0006473	protein acetylation	135	1.860
GO:0002431	Fc receptor mediated stimulatory signaling pathway	57	1.857
GO:0045185	maintenance of protein location	80	1.854
GO:0006412	translation	426	1.848
GO:0000209	protein polyubiquitination	205	1.848
GO:0043624	cellular protein complex disassembly	148	1.845
GO:0006353	DNA-templated transcription, termination	72	1.841

Table B.1 (continued)

GO:0051056	regulation of small GTPase mediated signal transduction	225	1.832
GO:0032386	regulation of intracellular transport	288	1.828
GO:0050808	synapse organization	215	1.828
GO:0010506	regulation of autophagy	235	1.823
GO:0000375	RNA splicing, via transesterification reactions	234	1.821
GO:0038093	Fc receptor signaling pathway	118	1.819
GO:0043087	regulation of GTPase activity	309	1.811
GO:0000377	RNA splicing..	231	1.799
GO:0000398	mRNA splicing, via spliceosome	231	1.799
GO:0006888	ER to Golgi vesicle-mediated transport	133	1.793
GO:0043604	amide biosynthetic process	497	1.791
GO:0018205	peptidyl-lysine modification	267	1.786
GO:0006397	mRNA processing	334	1.779
GO:0007033	vacuole organization	103	1.778
GO:0046578	regulation of Ras protein signal transduction	165	1.776
GO:0034504	protein localization to nucleus	173	1.775
GO:0016236	macroautophagy	203	1.775
GO:0043043	peptide biosynthetic process	442	1.770
GO:0090501	RNA phosphodiester bond hydrolysis	94	1.769
GO:0006606	protein import into nucleus	95	1.767
GO:0009267	cellular response to starvation	104	1.764
GO:0033157	regulation of intracellular protein transport	184	1.763
GO:0016311	dephosphorylation	306	1.757
GO:0006109	regulation of carbohydrate metabolic process	116	1.750
GO:0009142	nucleoside triphosphate biosynthetic process	115	1.745
GO:0070646	protein modification by small protein removal	198	1.742
GO:0007264	small GTPase mediated signal transd..	378	1.732

Table B.1 (continued)

GO:0035303	regulation of dephosphorylation	137	1.728
GO:0048193	Golgi vesicle transport	239	1.727
GO:0006470	protein dephosphorylation	201	1.727
GO:0042594	response to starvation	135	1.723
GO:0006325	chromatin organization	499	1.722
GO:0032388	positive regulation of intracellular transport	174	1.721
GO:0034330	cell junction organization	201	1.705
GO:0043547	positive regulation of GTPase activity	249	1.698
GO:0006402	mRNA catabolic process	194	1.698
GO:0007005	mitochondrion organization	371	1.697
GO:0032984	protein-containing complex disassembly	210	1.694
GO:0032956	regulation of actin cytoskeleton organization	222	1.691
GO:0043543	protein acylation	164	1.689
GO:0006914	autophagy	346	1.684
GO:0061919	process utilizing autophagic mechanism	346	1.684
GO:0008380	RNA splicing	296	1.680
GO:0031331	positive regulation of cellular catabolic process	239	1.678
GO:0010256	endomembrane system organization	289	1.668
GO:0045216	cell-cell junction organization	175	1.663
GO:1903827	regulation of cellular protein localization	359	1.645
GO:0030036	actin cytoskeleton organization	423	1.639
GO:0007265	Ras protein signal transduction	305	1.633
GO:1990778	protein localization to cell periphery	186	1.629
GO:0006401	RNA catabolic process	218	1.616
GO:0016579	protein deubiquitination	186	1.614
GO:0042176	regulation of protein catabolic process	240	1.613
GO:0032970	regulation of actin filament-based process	258	1.608
GO:0022411	cellular component disassembly	400	1.604
GO:0072657	protein localization to membrane	336	1.582
GO:0010498	proteasomal protein catabolic process	310	1.551
GO:0009896	positive regulation of catabolic process	278	1.550
GO:0043687	post-translational protein modification	283	1.545

Table B.1 (continued)

GO:0051493	regulation of cytoskeleton organization	344	1.545
GO:0030029	actin filament-based process	487	1.533
GO:0043632	modification-dependent macromolecule catabolic process	385	1.522
GO:0006511	ubiquitin-dependent protein catabolic process	371	1.522
GO:0019941	modification-dependent protein catabolic process	378	1.504
GO:0048667	cell morphogenesis involved in neuron differentiation	428	1.500
GO:0061564	axon development	381	1.477
GO:0051603	proteolysis involved in cellular protein catabolic process	441	1.473
GO:0015695	organic cation transport	29	-2.061
GO:0045622	regulation of T-helper cell differentiation	23	-2.107
GO:0046638	positive regulation of alpha-beta T cell differentiation	32	-2.170
GO:0071638	negative regulation of monocyte chemotactic protein-1 production	5	-2.461
GO:0055078	sodium ion homeostasis	38	2.939

Appendix C

LIST OF SIGNIFICANTLY ENRICHED KEGG PATHWAYS

Table C.1: The list of KEGG pathways significantly enriched (adjusted p-value < 0.1)

ID	Description	setSize	NES	significanceIn1000
hsa03020	RNA polymerase	17	2.260	1000
hsa04136	Autophagy - other	19	2.255	1000
hsa04140	Autophagy - animal	87	2.095	1000
hsa03015	mRNA surveillance pathway	64	2.014	1000
hsa04152	AMPK signaling pathway	92	2.009	1000
hsa04120	Ubiquitin mediated proteolysis	99	1.973	1000
hsa04520	Adherens junction	54	1.896	1000
hsa04330	Notch signaling pathway	36	1.889	1000
hsa04213	Longevity regulating pathway - multiple species	42	1.875	1000
hsa04211	Longevity regulating pathway	65	1.833	1000
hsa04150	mTOR signaling pathway	103	1.804	1000
hsa05211	Renal cell carcinoma	48	1.792	972
hsa04068	FoxO signaling pathway	93	1.778	1000
hsa04371	Apelin signaling pathway	100	1.748	1000
hsa04360	Axon guidance	144	1.744	1000
hsa05168	Herpes simplex infection	105	1.710	1000
hsa04072	Phospholipase D signaling pathway	102	1.665	962
hsa04974	Protein digestion and absorption	55	-1.891	624
hsa04975	Fat digestion and absorption	25	-1.972	875
hsa00591	Linoleic acid metabolism	16	-1.974	837
hsa05340	Primary immunodeficiency	23	-2.060	981

Appendix D

LIST OF TRANSCRIPTION FACTORS SIGNIFICANTLY ASSOCIATED WITH GENES SHOWING A CONSISTENT CHANGE IN HETEROGENEITY

Table D.1: The list of transcription factors that are significantly associated with consistent heterogeneity changes during aging (adjusted p-value < 0.1)

TF	NES	medianConsistentIncraese
EGR1	1.3843181445017199	15
EGR4	1.452929028663498	15
MTF1	1.453022442390886	15
AHR	1.274268306161655	14
ATF2	1.2014669947238317	14
E2F1	1.3065910229773126	14
EGR3	1.2537402078100792	14
ESR1	1.210294254579839	14
ETS1	1.2380923577641698	14
ETV7	1.2810222653234309	14
FOXI1	1.386944270455295	14
GABPB1	1.272353825093829	14
GCM1	1.274710009968939	14
GTF3A	1.240026791735294	14
HIF1A	1.27926757984335	14
MIF	1.270685102125093	14
NFIL3	1.272677244827733	14
PAX3	1.2829216919553512	14

Table D.1 (continued)

PDX1	1.3193194856375423	14
RFX1	1.243956817058612	14
SP3	1.3372318882184395	14
SPZ1	1.2855761272705426	14
TFDP1	1.2070059004606133	14
ZBTB14	1.27749693066955	14
ZIC2	1.2668847069583584	14
FOXJ1	1.2289997622477544	13
FOXO1	1.2165775192487729	13
FOXO3	1.2673790810612913	13
RB1	1.2985174722108035	13
TEF	1.2621044253697316	13

Appendix E

LIST OF MIRNAS SIGNIFICANTLY ASSOCIATED WITH GENES SHOWING CONSISTENT CHANGE IN HETEROGENEITY

Table E.1: The list of miRNAs that are significantly associated with consistent heterogeneity changes during development (adjusted p-value < 0.1)

miRNA	**NES**	**medianConsistentIncrease**
hsa-miR-34b-3p	1.863	12
hsa-miR-200c-3p	1.712	6

Table E.2: The list of miRNAs that are significantly associated with consistent heterogeneity changes during aging (adjusted p-value < 0.1)

miRNA	**NES**	**medianConsistentIncrease**
hsa-miR-1227-3p	1.457	16
hsa-miR-140-5p	1.735	16
hsa-miR-199a-3p	1.594	16
hsa-miR-425-3p	1.531	16
hsa-miR-532-3p	1.480	16
hsa-miR-301a-3p	1.567	15.5
hsa-let-7c-5p	1.546	15
hsa-let-7d-5p	1.556	15
hsa-let-7f-5p	1.552	15

Table E.2 (continued)

hsa-miR-106b-3p	1.367	15
hsa-miR-1296-5p	1.706	15
hsa-miR-150-5p	1.474	15
hsa-miR-17-5p	1.583	15
hsa-miR-183-5p	1.380	15
hsa-miR-18a-3p	1.479	15
hsa-miR-28-5p	1.420	15
hsa-miR-324-3p	1.547	15
hsa-miR-346	1.398	15
hsa-miR-361-5p	1.684	15
hsa-miR-374b-5p	1.716	15
hsa-miR-378a-5p	1.598	15
hsa-miR-505-3p	1.393	15
hsa-miR-671-5p	1.363	15
hsa-miR-744-5p	1.422	15
hsa-miR-766-3p	1.500	15
hsa-miR-92b-3p	1.537	15
hsa-miR-96-5p	1.565	15
hsa-miR-99a-5p	1.559	15
hsa-miR-181b-5p	1.404	14.5
hsa-miR-195-5p	1.404	14.5
hsa-miR-25-3p	1.500	14.5
hsa-miR-378a-3p	1.442	14.5
hsa-miR-504-5p	1.435	14.5
hsa-miR-877-5p	1.545	14.5
hsa-let-7a-5p	1.390	14
hsa-let-7e-5p	1.431	14
hsa-miR-100-5p	1.245	14
hsa-miR-101-3p	1.296	14
hsa-miR-106b-5p	1.311	14
hsa-miR-10a-5p	1.297	14
hsa-miR-10b-5p	1.292	14
hsa-miR-1226-3p	1.268	14
hsa-miR-1229-3p	1.256	14
hsa-miR-125b-5p	1.407	14

Table E.2 (continued)

hsa-miR-1260b	1.401	14
hsa-miR-128-3p	1.249	14
hsa-miR-130b-3p	1.339	14
hsa-miR-148a-3p	1.382	14
hsa-miR-149-5p	1.473	14
hsa-miR-15a-5p	1.329	14
hsa-miR-15b-5p	1.366	14
hsa-miR-181a-5p	1.327	14
hsa-miR-185-5p	1.552	14
hsa-miR-186-5p	1.310	14
hsa-miR-197-3p	1.400	14
hsa-miR-20a-5p	1.386	14
hsa-miR-218-5p	1.483	14
hsa-miR-221-3p	1.201	14
hsa-miR-222-3p	1.391	14
hsa-miR-23a-3p	1.315	14
hsa-miR-23b-3p	1.288	14
hsa-miR-24-3p	1.259	14
hsa-miR-26a-5p	1.363	14
hsa-miR-296-3p	1.233	14
hsa-miR-30a-5p	1.326	14
hsa-miR-30b-5p	1.380	14
hsa-miR-30c-5p	1.390	14
hsa-miR-31-5p	1.398	14
hsa-miR-32-5p	1.338	14
hsa-miR-320a	1.295	14
hsa-miR-324-5p	1.377	14
hsa-miR-331-3p	1.371	14
hsa-miR-33a-5p	1.281	14
hsa-miR-342-3p	1.318	14
hsa-miR-34a-5p	1.303	14
hsa-miR-375	1.205	14
hsa-miR-421	1.300	14
hsa-miR-423-3p	1.267	14
hsa-miR-423-5p	1.268	14

Table E.2 (continued)

hsa-miR-455-3p	1.383	14
hsa-miR-503-5p	1.344	14
hsa-miR-590-3p	1.403	14
hsa-miR-652-3p	1.320	14
hsa-miR-760	1.285	14
hsa-miR-769-3p	1.293	14
hsa-miR-769-5p	1.363	14
hsa-miR-877-3p	1.351	14
hsa-miR-93-5p	1.402	14
hsa-miR-98-5p	1.195	14
hsa-miR-122-5p	1.248	13.5
hsa-miR-9-5p	1.207	13.5
hsa-miR-19b-3p	1.170	13
hsa-miR-21-5p	1.242	13
hsa-miR-215-5p	1.242	13
hsa-miR-7-5p	1.228	13
hsa-miR-34b-3p	1.432	12
hsa-miR-548b-3p	1.503	12
hsa-miR-625-5p	1.503	10.5
hsa-miR-194-5p	1.552	10

Appendix F

SUPPLEMENTARY FIGURES

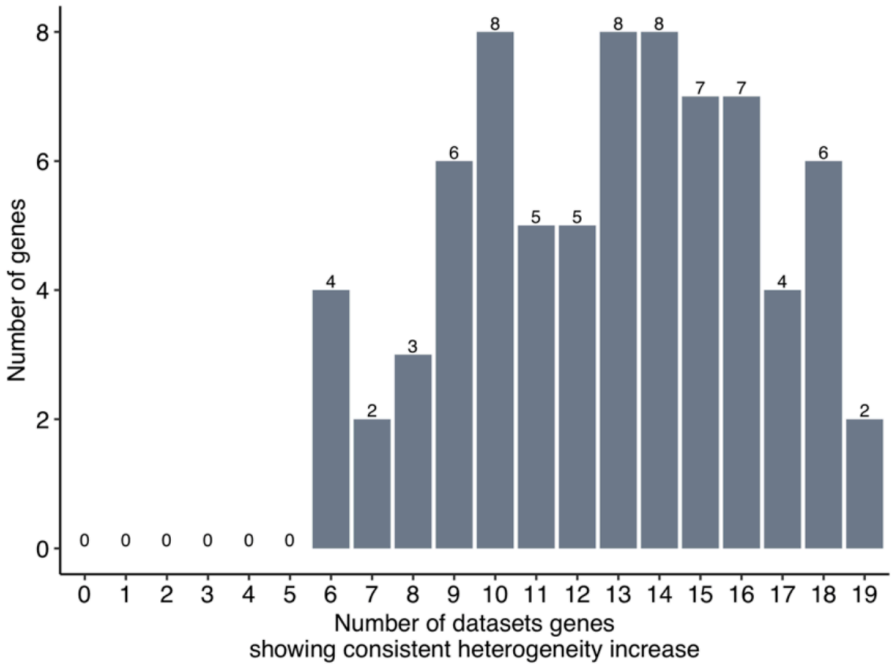

Figure F.1: The association between PMI and increased heterogeneity. The barplot shows the number of datasets 75 PMI-associated genes showing consistent heterogeneity increase. Adapted from (Isildak et al., 2020).

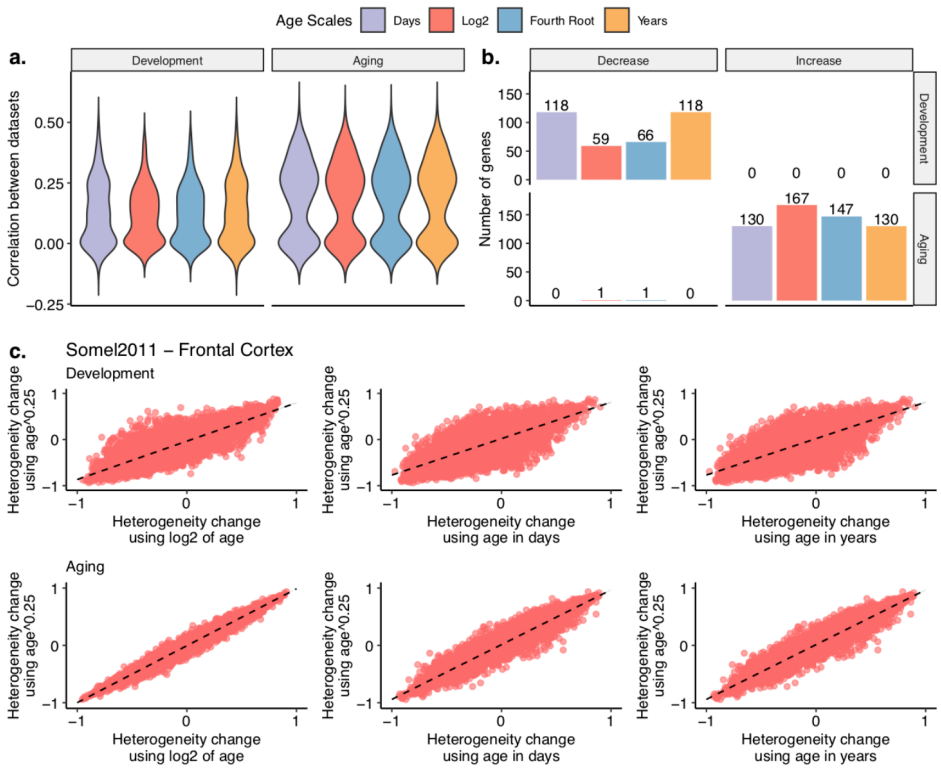

Figure F.2: Confirmation of the results using different age scales. (a) The distribution of Spearman's correlation coefficients in heterogeneity changes among development and aging datasets. (b) The number of genes showing a consistent decrease (left panels) or increase (right panels) among all 19 development (upper panel) and aging (lower panel) datasets. (c) The scatter plots showing heterogeneity changes calculated with different age scales for Somel2011_PFC dataset during development (upper panel) and aging (lower panel). Adapted from (Isildak et al., 2020).

.

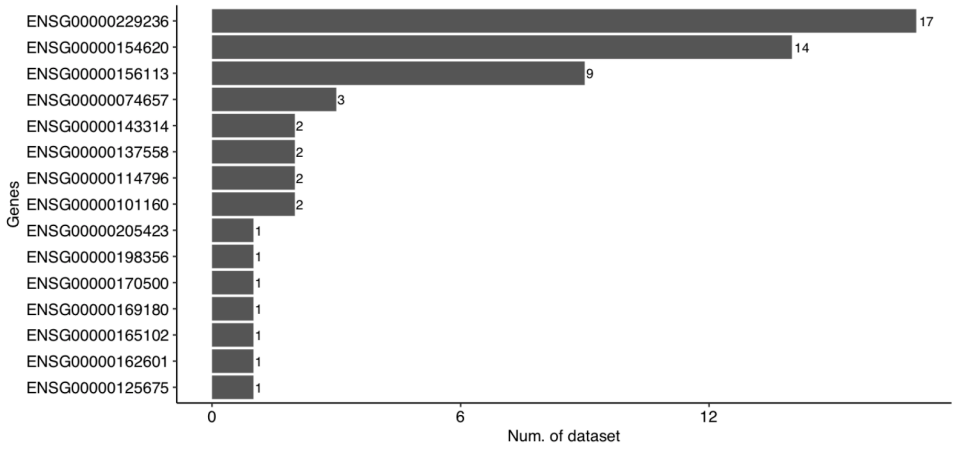

Figure F.3: The effect of sex-specific difference. The bar plot shows the number of datasets (x-axis) in which consistent genes (y-axis) show a significant difference between sexes. Adapted from (Isildak et al., 2020).

.

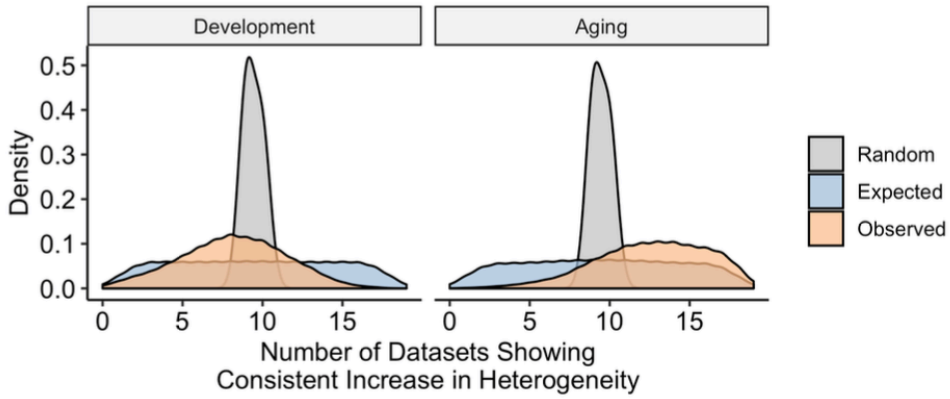

Figure F.4: The distribution of expected consistencies under the random permutations (gray) and the permutation scheme I used (blue), and of observed consistency (orange). Adapted from (Isildak et al., 2020).

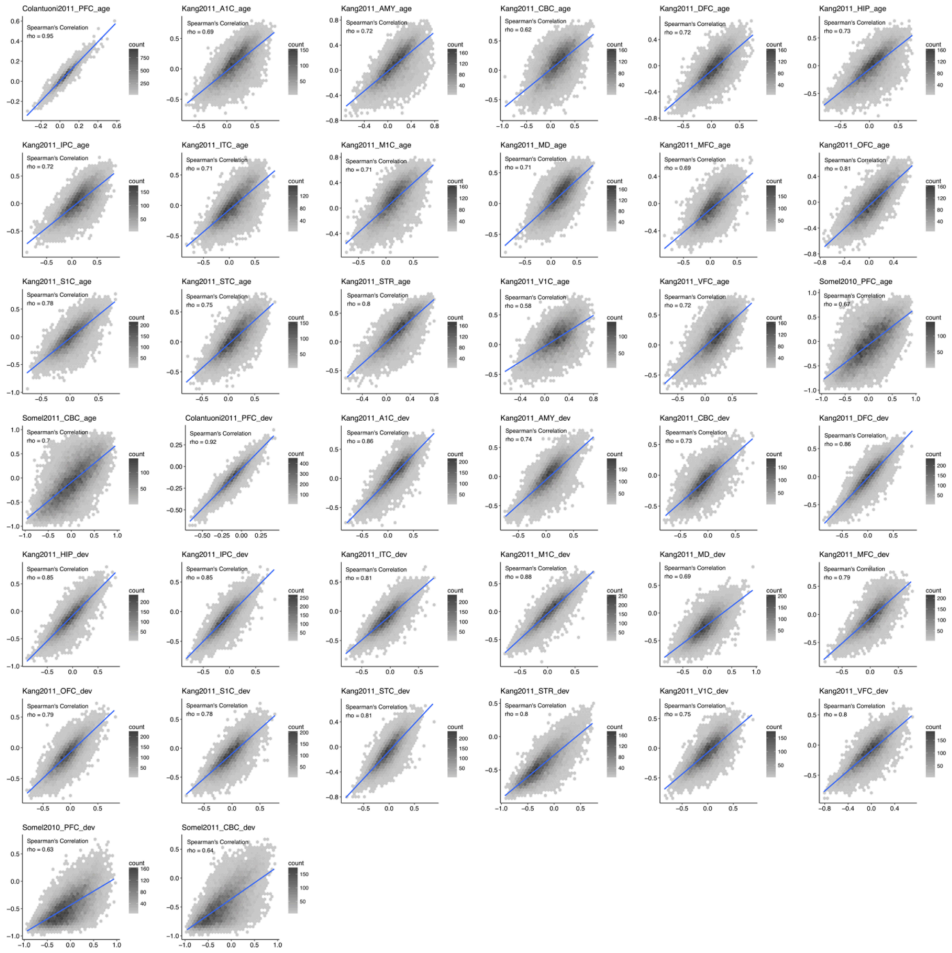

Figure F.5: Density maps showing the heterogeneity changes calculated with absolute residuals from linear regression (x-axis) and loess regression (y-axis). The rho values shown on the figures were calculated by Spearman's correlation test. Adapted from (Isildak et al., 2020).

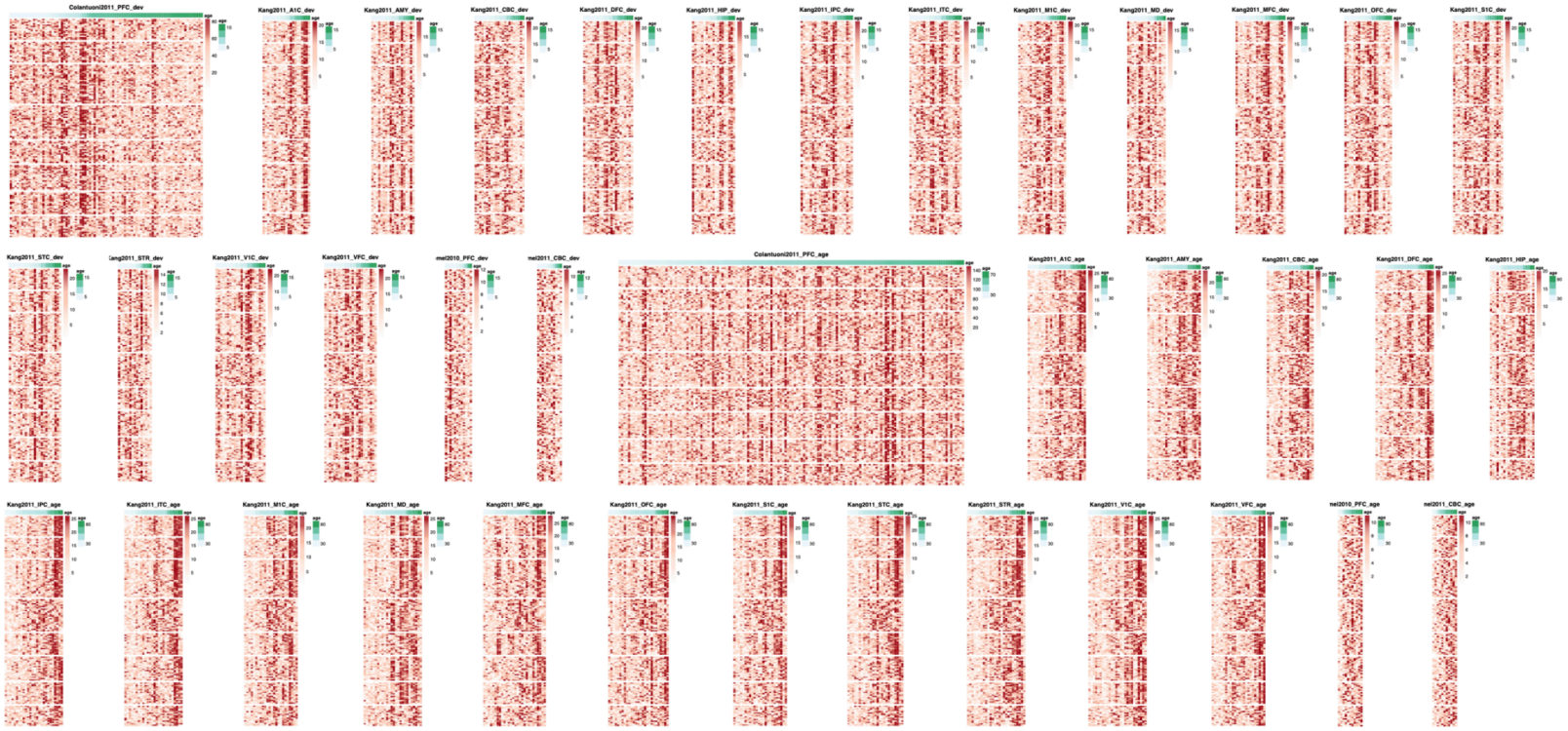

Figure F.6: Heatmaps showing the absolute value of residuals (colors) for each 147 consistent genes (rows) for each sample (columns). Adapted from (Isildak et al., 2020).