Genetic Basis of Longevity and Age-Related Diseases: Evidence from Genetic Association Studies

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy

By

Jingwei Wang

February 2019

Acknowledgement

I would like to thank my supervisors, Dr. João Pedro de Magalhães and Prof. Andy Jones for their advice and guidance during my study. It is such an honour to work with you.

I would like to thank all the members of Integrative Genomics of Ageing Group and many friends who have supported me over the four years. I would like to thank Dr. Aoife Doherty for her generous help in writing up this thesis.

Finally, I would like to thank my wife Zhi for her support. The daily lunchbox brought me courage to face any challenge. To my parents for their unwavering support.

Abstract

Ageing is a complex process that happens in almost all organisms. Many factors are involved in the ageing process. Age is a major risk factor for the onset of many diseases that severely affects life quality and lifespan in almost all known organisms, including humans. Studies focusing on ageing revealed that both biological and non-biological factors can affect the ageing process through direct or indirect manners. For example, in humans, the clustered distribution pattern of centenarians and supercentenarians in families and the plasticity of lifespan due to genetic manipulations and diet in model organisms further support the theory that ageing is a complex, multifactorial phenotype.

Among the factors that could affect ageing and longevity in model organisms and in human populations, genetic factors are of prime importance. As the fundamental element that distinguishes one from another, on a per-species level as well as on an individual level, genetic make-up determines the style of growth, metabolism, and adaptation to external environment of organisms. The existence of genetic variation among species and individuals shaped the differentiation in metabolic pathways and phenotypes such as ageing.

In this thesis, genetic factors were compiled and analysed to reveal their relationship

with longevity and ageing. In this regard, an introduction of ageing theories and ageing research is described in Chapter 1. Following this, Human Longevity-Associated Genes (HLAGs) from hundreds of published longevity-genetic association studies were manually curated and implemented into a user-friendly database – the LongevityMap (`http://genomics.senescence.info/longevity/`). The process of implementing the LongevityMap is described in Chapter 2.

In the following two chapters, the features (attributes) of those HLAGs collected in the LongevityMap were analysed. Functional enrichment analysis, which is a powerful tool to gather common functions from a list of genes, was utilised in analysing the HLAGs in the LongevityMap. The functional enrichment analysis of HLAGs revealed enriched clusters of important metabolic and cell signal pathways. Additionally, the metadata, such as the involvement of pathways, of those HLAGs, which represents the attributes of the gene set of HLAGs in LongevityMap, was also investigated. The analysis of this metadata revealed novel perspectives for ageing research. The results showed evidence of how candidate genes were selected for longevity-genetic association studies by researchers, as well as how researchers typically submitted and published the results. These explorations are described in Chapter 3 and Chapter 4.

Although thousands of genes have been examined for their association with longevity, very few of them have been consistently observed in different studies. Based on this, perhaps genetic heterogeneity could affect our understanding of the process of ageing, including longevity and age-related diseases. Through this concept, we investigated the relationship between genetic heterogeneity and traits/diseases that has been proven to be ageing related. A measurement of nucleotide changes on the gene level was defined and termed as "Genetic Diversity (GD)" (described in section 5.3.2.3) to represent the genetic heterogeneity on gene level. The analyses showed there was consistent correlation between gene length and the number of traits associated with the gene in

Genome-Wide Association Studies (GWASs), but not between the GD and the number of traits associated with the gene. The GD of human Age-Related Traits/Disease (ARTD) associated genes, some cancers associated genes and Early Onset Disease(EOD) genes were also investigated. Results showed genetic heterogeneity in EOD genes were significantly higher than in ARTD or EOD genes. These analysis and results are described in Chapter 5.

In conclusion, HLAGs identified by genetic association studies are a valuable resource for ageing research. Organising those HLAGs into the LongevityMap database further facilitates the usage of HLAGs data, even though publication/study biases may exist. The results from functional enrichment analysis and pathway analysis not only verified the importance of some key biological functional pathways in affecting lifespan but also gave some hits on other pathways that could contribute to ageing/longevity. Finally, correlation analyses showed GWAS results are affected by gene length or GD. GD is different in ARTD, cancer and EOD associated genes.

Contents

List of Figures

List of Tables

Abbreviations

1KGP	The 1000 Genomes Project
ANOVA	Analysis of Variance
ARTD	Ageing-Related Trait/Disease
bp	base pair
CGAS	Candidate Gene Association Study
CNV	Copy Number Variation
CR	Calorie Restriction
DAVID	Database for Annotation, Visualization and Integrated Discovery
EFO	Experimental Factor Ontology
eQTLs	expression Quantitative Trait Loci
FDR	False Discovery Rate
FWER	Family-wise Error Rate
GHG	GWAS-Hit Gene, *ie* Gene was associated with trait in GWAS-Catalog
GO	Gene Ontology
GSA	Gene Set Analysis
GWAS	Genome-Wide Association Study
HAGR	Human Ageing Genomic Resources
HLAG	Human Longevity-Associated Gene
HWE	Hardy–Weinberg Equilibrium
kb	kilobase, 1000 bases of DNA or RNA

LAG	Longevity-Associated Gene
LGAS	Longevity Genetic Association Study
LLI	Long-Lived Individual
MAF	Minor Allele Frequency
MT	Mapped Trait (in GWAS-Catalog)
mtDNA	Mitochondrial DNA
mTOR	Mammalian Target of Rapamycin
NMT	Number of Mapped Trait
non-GHG	non-GWAS Hit Genes, *ie* Genes was not associated with trait in GWAS-Catalog
OMIM	Online Mendelian Inheritance in Man
PPI	Protein-Protein Interaction
PMID	PubMed unique Identifier
RNS	Reactive Nitrogen Species
ROS	Reactive Oxygen Species
rs	Reference SNP
RT	Reported Trait (in GWAS-Catalog)
SE	Study Effect
SNP	Single Nucleotide Polymorphism
SRRGP	Sequential Removal of Rarely reported Genes Procedure
VCF	Variant Call Format

Highlights

1. The LongevityMap database, which is a collection of manually curated human longevity-associated genes, was constructed following an extensive mining of the literature.

2. The genetic attributes of the whole set of Human Longevity-Associated Genes (HLAGs) were analysed.

3. Candidate Gene Association Studies (CGASs) present in the LongevityMap were investigated for publication biases.

4. The relationship between genetic heterogeneity and various traits, including ageing related diseases, was investigated.

Chapter 1

Introduction

1.1 Ageing and ageing research

1.1.1 Ageing

> *"Ageing is an intrinsic process of loss of viability and increase in vulnerability."* (Comfort, 1964)

> *"Ageing is usually defined as the progressive loss of function accompanied by decreasing fertility and increasing mortality with advanced age."* (Kirkwood et al., 2000)

> *"Ageing is a universal, intrinsic, progressive and deleterious process."* (Viña et al., 2007)

The current definition of ageing is usually quite flexible and changes from time to time with the advance of medical and ageing research. Different research areas may use a slightly different definition of ageing. In spite of this, all the definitions of ageing do share some common concepts: ageing is universal, intrinsic, progressive and deleterious (Viña et al., 2007).

In *Animalia Kingdom*, organisms age differently, both in the ageing rate and the longevity. For example, Roundworm (*Caenorhabditis elegans* or *C. elegans*), a common organism used in ageing research, has a typical lifespan of 2 to 3 weeks, while Ocean quahog clam (*Arctica islandica*) can live more than 507 years (Butler et al., 2013). *Hydra* does not show signs of ageing (Martínez, 1998). Much shorter lifespans were usually observations in vertebrate organisms. To date, the documented most long-lived vertebrate is a Greenland shark (*Somniosus microcephalus*). With an estimate of 392 ± 120 years (Nielsen et al., 2016), the lifespan of Greenland shark is ∼2 to 4 fold of human species. In primates, humans have the longest lifespan, with the documented maximum lifespan is ∼122.5 years old. Chimpanzee, the closest relative of human species, only has the documented maximum lifespan ∼65 years, which is roughly half of the longest lifespan of human(for an up-to-date list, please see the AnAge database (`http://genomics.senescence.info/species/index.html`). The variation of maximum lifespan across the species suggests ageing is determined by many factors. Apparently, genetic factors are among those.

Global human average lifespan increased steeply in the past two centuries. Notably, in the latest half century, the global life expectancy at birth has been extended by roughly 20 years (*Source: Life expectancy at birth* `http://www.worldbank.org`). In this background, the global median age was also improved roughly by 20 years. Several factors contributed to this shift in global age structure. The first part of contribution can be accounted for the combination of fertility decline in recent years and the "Baby

Boom" after World War II (Centers for Disease Control and Prevention, 2003; Sander et al., 2014). Others including the improved living environment including clean drinking water, improvement of nutrients and the decrease of premature death also contributed to the shift in global median age (Fries, 1980; Bunker, 2001). In addition to those above, the advance of medical assistance in prolong the lifespan of Cardiovascular Disease and Cancer patients also played an important role in supporting the lifespan extension in the last decades (Passarino et al., 2016). In contrast to the huge improvement in global median age, the maximum lifespan fluctuated only in a much smaller scale. Therefore, it was proposed that maximum lifespan is predetermined and unchangeable. Individuals may achieve an age close to the predetermined maximum lifespan but hardly push any further (Thatcher, 1999). The expectation of curing ageing completely (*i.e.* obtain immortal lifespan) in human is still premature (Vijg et al., 2008).

Nonetheless, investigating and understanding the ageing process in order to delay or alter it for a better, healthier lifespan is still of importance. Individuals sharing a similar lifespan could have a different quality of life due to the time point of disease onset (Figure 1.1). A better understanding of the ageing process and applying appropriate interventions targeting postpone the onset of age-related diseases and prolong lifespan could improve the life quality and the outcome of ageing.

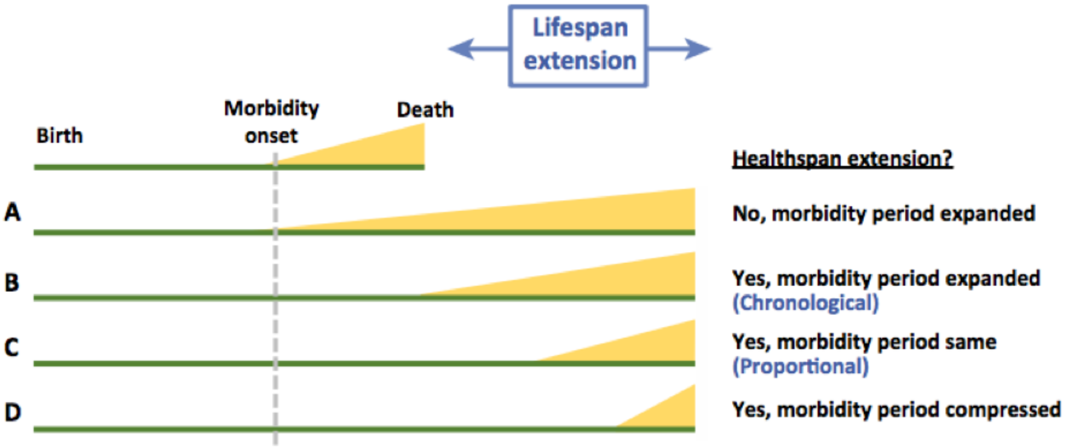

Lifespan extension
Birth
Morbidity onset
Death
Healthspan extension?
A
No, morbidity period expanded
B
Yes, morbidity period expanded (Chronological)
C
Yes, morbidity period same (Proportional)
D
Yes, morbidity period compressed

Figure 1.1: **Illustration of lifespan extension and health-span extension.** Taken from (Hansen et al., 2016)

.

1.1.2 A brief history of ageing research

Whether you like it or not, *"we are all destined to age"* (Mori et al., 2009). It has been a long history since human started aware of ageing and kept seeking interventions. About 5,000 years ago, *The Epic of Gilgamesh* enthusiastically searched for the *"Fountain of Youth"* (Magalhães et al., 2004; Vijg et al., 2008). Later, around 220BC, *Qin Shi Huang*, the first emperor of *Qin Dynasty* in ancient China, was believed to seek magic pills that could make him live eternally. They were not alone. Several hundred years later, it is believed that *Cleopatra VII Philopator of Egypt*, who was the last active pharaoh of Ptolemaic Egypt, tried everything in her power to keep her beauty (Mori et al., 2009). These efforts, including those that focused on magical power and not supported by scientific evidence, can be considered as the earliest sprout of ageing exploration.

In modern times, scientifically study of ageing dates back to as early as 1932, when senescence was first described as *"the after-result of the mechanism which secures specific size"* (Bidder, 1932). Based on the evidence from *"Giant trees, cultures of chick cells and of paramecium, measurements of plaice and of sponges"*, it was believed *"the indefinite growth is natural"* (Bidder, 1932). Decades later, in 1961, the cellular senescence was discovered by Hayflick et al. They found normal human fibroblasts can only divide a finite number of times *in vitro* before entering a state of irreversible cell-cycle arrest, which was defined as cellular senescence (Hayflick et al., 1961). This senescence state of cell was later considered as one of the contributors to organismal ageing phenotypes and age-associated chronic disorders (Collado et al., 2007; Magalhães et al., 2018; Kang et al., 2017).

In the light of the description of senescence, ageing research started to grow vigorously in the following years. With the accumulation of observations and experimental data from ageing related phenotype changes (such as described in (Wiesner, 1932)), many hypothetical theories tried to explain ageing mechanisms emerged during this time. Because these theories were based on very limited experimental data, many of them were redundant and unclassified in today's view. It was not unusual to see one ageing theory overlaps or contradicts with another. Some of them were even flawed. Even so, some ideas stood out of the crowd and formed the cornerstone of future ageing research.

In the final quarter of last century, ageing research progressed considerably. Those miscellany of ageing theories were simplified with the aid of rapid development of modern molecular biology and sequencing technology coupled with the bioinformatics approach (Magalhães, 2015). Several hundreds of ageing theories from previous studies were explained and rationally classified (Medvedev, 1990). In the meantime, data in ageing research was growing at an unforeseen speed. High-throughput data together with integrative methods from multiple disciplines further promoted the understanding of ageing and the building concepts of ageing theories in both model organisms and humans (Kirkwood, 2011).

1.1.3 Ageing research in non-human model organisms

Model organisms have been aiding scientific research for a long history (Müller et al., 2010). In addition to the most well-known advantages including the easy-accessibility, high reproducing ability and amenability to experimental studies, model organisms have other specific advantages in ageing research.

Firstly, model organisms are normally kept in constrained laboratory environments, which facilitates the experimental conditions manipulation and data collection. Ageing is a complex process that involves environmental, genetic factors and the intense interactions between them (Passarino et al., 2016; Benayoun et al., 2015; Dato et al., 2017). The constrained laboratory environment provides the feasibility of observing the effect of single factor alternation. Secondly, model organisms are much easier and cheaper to reproduce a large number of offspring, which allows large-scale observations of the population dynamics at a relative economic cost (Wiesner, 1932). Last but not least, widely-used model organisms normally have a much shorter lifespan compared to human, which makes the observation of mortality and longevity changes between generations become feasible. For example, *Caenorhabditis elegans (C. elegans)*, one of the most popular model organisms in ageing research, has a lifespan of several weeks (Uno et al., 2016) but shares many common biological features with human (Kenyon, 2010; Horvitz, 2003; Brenner, 1974). On the other hand, a typical reproduction interval in human is 20∼30 years. Comparing to *C. elegans*, it is much more difficult to carry out observations and gather a large amount of data in human subjects within a relatively short time. These advantages make model organisms are popular in ageing research. As a result, ageing research in model organisms brought not only the boom of data but also new concepts in human ageing research. Mechanisms and theories of ageing postulated from model organisms have been proposed for intervention in human population (Heilbronn et al., 2003). One example is caloric restriction (Holloszy et al., 2007), which is obtaining extended lifespan by reducing the intake of caloric with out being malnutrition.

Although we do share some common ageing characters with model organisms (Jones et al., 2014), we cannot neglect variations exist between human and model organisms. A recently systematic analysis of ageing and age-related disease genes across several organisms, including *M. Musculus*, *D. melanogaster*, *C. elegans* and *S. cerevisiae*, in ageing research confirmed human longevity-associated genes only modestly overlapped with other model organisms. A further investigation in the overlapped genes between ageing associated genes and age-related disease genes verified the difference in overlapped genes in individual organisms, and this overlap decreased with the increase of evolutionary distance with human (Fernandes et al., 2016).

1.1.4 Ageing research in human

The difference between model organisms and human brings extra barriers in translating results obtained in model organisms to human population. The worldwide growth of elderly populations raises both social concerns and academic interests in ageing research. As age is a risk factor for many chronic diseases, including cancers (Kennedy et al., 2014), there is also an urgent need to progress ageing research in human. Deciphering the mechanisms of ageing could extend health-span by improving the morbidity and reducing mortality in elder populations (Figure 1.1).

Studies have been designed to isolate factors that could contribute to or impair human longevity. For environmental factors, by comparing the lifestyle of long-lived individuals to other cohorts, scientists isolated several beneficial lifestyles such as appropriate amount exercise (Gremeaux et al., 2012), calorie restriction (Bordone et al., 2005) and Mediterranean diet (James et al., 1989; Trichopoulou et al., 2000).

For the genetic basis of longevity, genetic association studies identified many risk alleles that could affect human longevity (see Chapter 2). Even many alleles have been reported, they are not easy to be verified by another study (discussed in Chapter 4).

1.2 Mechanistic theories of ageing

Many theories of the ageing process have been proposed in the past few decades(Medvedev, 1990). Yet, none of them is capable of describing the whole picture (Kirkwood, 2011; Davidovic et al., 2010). The failure in trying to explain ageing process by single theories revealed the complexity of ageing process and the limited knowledge in ageing. Nonetheless, some of those theories were consistent with the experimental verifications in the later years and spread in the field. Herein, several widely-accepted theories in explaining ageing were summarised and reviewed in order to describe the latest knowledge in ageing.

1.2.1 Accumulation of DNA damage

Nuclear DNA damage

DNA damage is defined as abnormal structural changes on DNA strands. DNA changes occur in all cells regularly. It has been estimated that DNA damage occurs roughly 10,000 times a day in a single cell (Bernstein et al., 2013). DNA damage is susceptible to many factors from both inside the organism and outside environment. External factors, such as ultraviolet radiation and chemical toxins, and intrinsic factors, such as Reactive Oxygen Species (ROS) and Reactive Nitrogen Species (RNS), contribute to DNA damage. While most of these damages can be corrected by DNA repair mechanisms in nuclear DNA, some of the damages may be corrected improperly. Therefore, the total damage accumulates with the increasing of age. As nuclear DNA takes a big proportion in the total DNA(approximately $> 99\%$). It is accused the ageing process is the exhibition of accumulated damage in nuclear DNA. This time-dependent accumulation of damage will eventually alter the homeostasis of cells and causes senescence (Freitas et al., 2011). For example, in less replicating active cells such as neurons in central nervous system and myocytes in cardiac muscle, the adverse effect resulted from accumulated damage is more obvious(Hoeijmakers, 2009; Holmes et al., 1992).

Mitochondrial DNA damage

Mitochondrial DNA (mtDNA) is more susceptible to damage factors compared to nuclear DNA. Naturally, it is closer to the Reactive Oxygen Species (ROS) in the cell than nuclear DNA, therefore more exposed to ROS and therefore more susceptible than nuclear DNA. mtDNA lacks protection from histone protein and suffers a weaker repair mechanism (Freitas et al., 2011). All of these factors make mitochondrial DNA more exposed, and therefore more susceptible, to be oxidised than nuclear DNA does (Richter et al., 1988; Shigenaga et al., 1994). Because mitochondrial supplies energy for the maintenance of cell functions, the impaired mtDNA gradually compromise cell functions through compromised energy supply and cause senescence (Jin, 2010).

1.2.2 Free radical theory

The free radical theory of ageing was firstly described in the 1950s by Harman (Harman, 1956; Harman, 2009). Free radicals such as superoxide anion O_2^- and hydroxyl radical (‘OH), are generated as by-products of aerobic respiration and various catabolic processes in living organisms (Halliwell, 1991). Free radicals exert their effect on easily oxidized substances and the cellular constituents nearby where they were produced. eventually, cellular functional efficiency, reproductive ability and potentially genes will be impaired (Harman, 1956).

As mitochondria are the main place where aerobic respiration takes place (also where free radicals were produced), they are susceptible to oxidative attacks from free radicals. Because mtDNA encodes important oxidative phosphorylation machinery and it has a much weaker repair mechanism, the damage to mtDNA accumulates much faster than nuclear DNA (Taanman, 1999; Freitas et al., 2011). The damaged mtDNA impairs the efficiency of the respiratory chain, which lead to the accumulation of free radicals (Hiona et al., 2008; Harman, 1972). The excess of free radicals could damage DNA or membrane along with other cellular structure in the cell, therefore cause the cell senescent (Holmes et al., 1992). Ageing is a result of damage from the free radicals produced from oxygen metabolism (Cui et al., 2012). This theory is supported by the experiments conducted in *Drosophila* and *mice*. In both organisms, introducing antioxidant substances into daily diet prolonged average and maximum lifespan comparing to the controls (Ernst et al., 2013). Lastly, this theory is also proposed for the mechanisms operating underneath the facts of prolonged lifespan and delayed ageing in model organisms under Caloric Restriction (CR). In animals under CR, the damage caused by ROS was slowed because ROS was generated in a slower pace (Weindruch, 1996).

1.2.3 Telomere erosion theory

Telomeres are repetitive DNA sequences at the ends of linear chromosomes. Studies have suggested that the average length of telomeres in both human and other organisms is inversely correlated with the counts of DNA replication events (Harley et al., 1990). When the telomeres are depleted, the cell loses its ability to divide. This initiates the deficiency of regeneration in cells, then tissues and finally the whole body. This ongoing telomere depletion process, which happens inside cells, couples with outside ageing traits on whole body level (Magalhães, 2011). Experimental data revealed consistent correlation between the length of telomeres and chronological age. Therefore, the length of telomeres can be used to reflect biological ageing (Benetos et al., 2001).

1.3 Evolutionary based explanations

Aside from all of the theories above, some researchers claim that the process of ageing is predetermined; that ageing process follows a programmed manner (Jin, 2010). Survival rate, endocrine system and immunological response are programmed to decline with age (Davidovic et al., 2010; Heemst, 2010). With the advance of chronological age, the capacity of maintaining homoeostasis (i.e. Organ Reserve) gradually declines, therefore increasing the vulnerability of diseases and thus mortality (Viña et al., 2007). The observations from prolonged average and maximum lifespan were just parallel results from decreased premature mortality such as traumatic death. Manipulations that increased lifespan only push the boundary closer to programmed maximum lifespan. One of the major evidence that supports this hypothesis is the disproportional increasing in average and maximum lifespan in the latest centuries.

The programmed ageing theory emphasises on the optimal balance between survival and metabolic cost. Because the resource that an organism can access is limited, the organism has to allocate the energy between reproduce next generation and the maintenance of the organism himself. The programmed ageing accelerates the generation turnover rate and limits the total population size, which is beneficial for maintaining the species within a limited resource environment. In the scope of this theory, any individual lives longer beyond than successfully breeding next generation would be a waste of resource (Kirkwood, 2011). However, this theory does not take social factors into account. For example, an elder individual could have more knowledge and experience which might be useful in helping the population survival. Long-lived grandparents could help in breeding their grandchildren successfully.

1.4 Genetic basis of ageing

1.4.1 Genes, environments and traits

It is clear that genetic sequence partially determines traits/phenotypes in humans. For a given trait, the contribution can come from genes themselves, the complex interactions between genes and environmental factors, interactions between genes, or alternative status of genes, such as epigenetics. How epigenetics could affect traits is beyond the scope of this thesis, herein, only contributions from genetic factors are discussed.

During meiotic cell division and gamete formation, genetic information from both parents are passed on to the successive generation, and therefore contributes to the traits/phenotypes of offspring. Based on the relationship between different genotypes in affecting the phenotype, several patterns in how genotypes could contribute to phenotypes have been proposed:

- Dominance pattern: where individual carriers of heterozygote genotype exhibit indistinguishable phenotype as homozygote individual carriers.
- Partial dominance pattern: where the heterozygote genotype individuals exhibit intermediate phenotypes between the two homozygotes.
- Codominance pattern: where offspring simultaneously exhibit phenotypes from

both parents. For example, in human ABO blood group system, Individuals with blood type AB have both A protein and B protein on the surface of red blood cells (Stratton, 1952).

- Recessive pattern: where a trait is only exhibited when the genotype is homozygous. The effect of recessive genes is likely to be masked by other genes even they co-exist in the heterozygous genotypes.
- Overdominance pattern: where the heterozygote is better adapted than either homozygote.

In addition to the above described relative dominate and recessive properties of genes, other properties could also be involved in determining phenotypes. For example, expression Quantitative Trait Loci (eQTLs) could contribute to phenotypes by affecting gene expression levels (Nica et al., 2013). The distance between eQTLs and the gene that is related can be physically close or far away (even in anther chromosome). Sometimes, a trait of an individual is not simply determined by a single gene or allele, but rather by a group of genes/alleles. Each gene/allele independently contributes to a small portion of the overall trait. In this case, it called additive allelic effects, because the combined effect of those genes/alleles can be estimated by adding together their separate effects (Ashton, 2013). While not all the traits exhibit in a discrete manner, some traits reveal as a continuous distribution of phenotypes, such as height or blood pressure. These traits are named as quantitative traits. Normally, these traits are determined by the cumulation effects of many genes/alleles and their interaction with environmental factors.

1.4.2 Effects of evolution

Natural selection acts on traits through environmental factors. The variation of traits, which is usually the representation of variation of individual genotypes underneath, will face different selection pressure in the same environment. The pressure that comes from environment exerts on the individual through traits will affect the fitness, which describes individual reproductive success, of the individual and therefore has effects on the genotypes in the offspring gene pool. Based on the result of genotype/allele frequency change in the next generation after natural selection, natural selection can be classified into the following three classes.

- positive selection: when an allele that determines traits that has better fitness in a given environment, the frequencies of the allele will increase in the offspring generations. If the environment does not change, the frequency of the allele will increase until all the individuals in the population carry the same allele in which case, we say this allele is fixed in the population.
- purifying selection happens as the opposite of positive, alleles determines deleterious traits are being selected against. As a result, the frequencies of those deleterious alleles (in a given environment) will decrease in the offspring populations until they are completely removed in the population.
- stabilizing selection often happens with the overdominance pattern, that is heterozygote has better fitness than either homozygous genotype. Therefore, the frequencies of the two alleles are maintained by natural selection. One good example of stabilizing selection is the sickle-cell anaemia related alleles. The alternative allele codes sickle-shaped red blood cell due to the deletion of an amino acid in haemoglobin. Sickle-shaped red blood cells can “collapse” around the parasites in a malarial infection and therefore help to remove them out of

> blood. However, sickle-shaped red blood cells are not as efficient as normal red blood cells in transporting oxygen. Therefore, in the malaria risk areas, those two alleles were maintained by pressure from the environment. Neither of the two alleles gaining better fitness over the other allele.

Not all the allele have effects on the phenotypic traits, a big portion of alleles does not have direct influences on the phenotypic traits. They are functionally neutral. These neutral alleles account for a big proportion of total genetic variation. Because they are functionally neutral, natural selection cannot exert on them through phenotypic traits, therefore their variation is determined by other processes. Genetic drift, which refers to the drastic change of allele frequencies in the population due to chance, could lead to a loss of genetic variation. Because genetic drift is a stochastic process and the gene pool of "drifted" population is a subset of the original population, the allele frequencies in "drifted" population could differ from the original population due to the random sampling. The rare allele in the original population could become dominate allele in the "drifted" population. The genetic drift is dominated by the smallest population size (bottleneck) in a fluctuate-sized population (Masel, 2011). One of the most famous genetic drift in human history, the *Out of Africa* bottleneck severely reduced the genetic variation in the human population and lead to the prevalence of the oldest alleles (alleles originated in Africa) (Cavalli-Sforza et al., 2003; McClellan et al., 2010).

Genetic linkage happens when two alleles are close enough on a chromosome. During the recombination in gametes formation, these two alleles are more likely to stay together. As a result, a much higher co-occurrence of both two alleles is observed in population compared to if the two alleles are co-existing purely by chance. This is called Linkage Disequilibrium (LD). If two or more alleles are routinely observed in a population as a result of LD, then, the combination of those alleles formed a haplotype. The existence of LD has an effect on the genetic variation of the population.

A functional neutral (or even slightly deleterious) allele can be indirectly selected by the environment due to another allele has a strong fitness and in LD with it. This phenomenon is called genetic hitchhiking. Overall, the change of the human gene pool over time is a result of the action of many factors (mutation, migration, genetic drift and natural selection) (Arnold, 2001).

1.4.3 Genetic variation and genetic diversity

From the perspective of genetics, the underlying differences that distinguish one individual from another are the differences of genome DNA sequence. The subtle difference in DNA sequence contributes many aspects of human life, from very obvious skin colour to the less obvious risk of any disease or the ability to live survival until a late age. Several types of genetic variation in human genome (Figure 1.2) could affect phenotypic traits in direct or indirect ways. Genetic association studies, usually used to detect the potential causal loci, can examine SNPs, Copy Number variation (CNV) or even haplotype for the potential risk loci. In longevity and disease association studies, the most commonly used type of locus is SNP (Budovsky et al., 2013).

In the scope of this thesis, Genetic diversity(GD) refers to a measurement to describe the base pair difference of SNPs between genes (Chapter 5). GD was calculated based on the allele frequencies of a SNP locus in a given gene pool (usually a population cohort). The values of GD directly reflects the nucleotide change of genes in a population (See subsubsection 5.3.2.3 for method description). By examining genetic diversity, we could obtain the most straightforward information in nucleotide changes of a gene. As genes determine traits and traits are under the selection of environment, investigating the difference of GDs between genes could potentially reveal information in how selection pressure (comes from environmental factors) has been exerting on the phenotypic traits linked genes. Then, variants within a gene can be further examined to reveal the contribution to certain traits. In the following sections, only genetic variation in human population was discussed.

1.4.4 Complex traits, diseases in human population

Genetic variants and environmental factors governed almost all the phenotypic traits, which determines the outside appearance and inside susceptibility to diseases in human population. However, The long existed selection together with genetic patterns blurred the causation between genotype and phenotype. Although methods, such as Genome-wide association study (discussed in section 1.4.5), have been developed to recover the connections between genotypes and phenotypes, they are facing low explanation rate. Therefore, which SNPs contribute to complex traits and disease is still under debate. At present, there are two major theories on how genetic variants affect traits or diseases: “Common Diseases- Rare Variants theory (CDRV)” and “Common Diseases-Common Variants (CDCV) theory”.

In CDRV theory, it is proposed that rare alleles are the causal factor of common diseases/traits: “multiple rare alleles with high penetrance collectively contribute to a common phenotype in the general population” (Cohen, 2004). Because rare alleles are in low frequency in the population, GWAS scanning method, which usually only scans common SNPs in population, is not able to detect those variants (Goldstein, 2009).

The other theory, however, proposed that the common traits/diseases are displayed as a result of contributions of many common variants ($>$ 5%) (Pritchard, 2001). Each individual variant contributed a small fraction of total results. Because of this, the small contribution from single variant is easily escaped from being captured by GWAS.

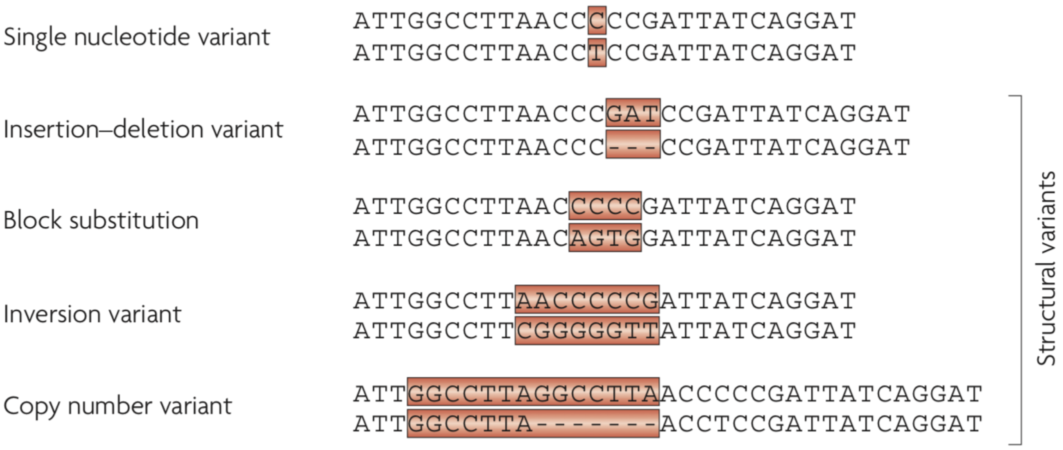

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA-------ACCTCCGATTATCAGGAT
Structural variants

Figure 1.2: **Classes of human genetic variants.** (Taken from (Frazer et al., 2009).)

1.4.5 Genetic association studies

Genetic association studies are effective for screening the substantial relationship between genetic variants (for example SNPs) and phenotypic traits. It identifies the risk variants by calculating the odds of risk between case group and control group. Risk alleles for many diseases and traits, such as cancer, Alzheimer's diseases, body mass index as well as ageing, have been successfully identified by genetic association studies (Easton et al., 2007; Saunders et al., 1993; Schächter et al., 1994; Speliotes et al., 2010). As a powerful tool to reveal association facts (not necessarily biological causation) between genotype and phenotype, it is especially useful in disclosing the genetic architecture of trait and screening candidate loci for functional validation (Korte et al., 2013).

The scope and capacity of genetic association study design have been changing over the time. Earlier studies generally focused on a very limited number of loci. Some of them investigated as few as one single locus. This is mainly due to the limited number of known human genetic variants loci as well as the restrictions from experimental throughput capacity. The variants being tested are usually supported by evidence from other sources, such as model organism or *in vitro* experiments. Therefore, the design is also known as Candidate Gene Association Study design (CGAS). Things have been changed greatly since the release of The 1000 Genome Project (1KGP) pilot phase data. Thousands of new loci were identified from the global population. The newly discovered variant loci, together with the Chip-based variants calling technology, facilitated genetic association study workflow in a more cost-effective and high-throughput manner. Chip-based sequencing technology has made the whole genome scanning for trait-associated loci became feasible. The number of loci under investigation in a single study increased enormously. It is no longer uncommon to see hundreds of variation loci were investigated in a single study.

As the number of variants being tested in Genetic Association Studies increasing, a new method that can examine variants genome-widely emerged. This method is named as Genome-Wide Association Study (GWAS). Unlike CGAS, which heavily relay on the supportive evidence of candidate variants, GWAS design does not need prior knowledge of candidate variants before testing. It searches the associations between genetic variants and traits by simply testing almost all the known tagging variants genome-widely. Because no biological evidence behind the association is needed in advance, it is hypothesis-free (Table 1.1).

GWAS has obtained a great achievement in discovering the connections between phenotypic traits to potential risk variants. Through finding a small set of risk variants, the scope of where true causal variants are likely to located in is scaled down from whole genome-wide to a small set of identified risk variants, which can then by used for experimental verification. As association is not causation, GWAS is useful in a way of capturing a set of "high-risk" variants (where the casual variant would be) (Korte et al., 2013) rather than pinpoint individual casual variant.

Until September 2016, 24,218 unique SNP-trait associations have been identified by GWAS (MacArthur et al., 2016). Although these potential connections between loci and traits provided new targets for exploring the underneath mechanisms, GWAS does have its intrinsic limitations.

Table 1.1: Comparison of GWAS and Candidate Gene Association Study (CGAS)

	GWAS	CGAS
Scope	Whole genome	Any interested loci
Typical number of loci involved	Millions	1 to hundreds
Hypothesis-free	Yes	No

One major problem is the discrepancy in reproducing the results. GWAS results are neither easily being reproduced from one population to another population nor not easily being reproduced from one study to another (Frazer et al., 2009). At present, the majority of GWAS were performed in European and Asian populations, which could be a major limitation on the detection of variants (Manolio et al., 2009). Another issue that GWAS has to face is the weak explanation rate. Only less than 10% of genetic variants were involved in explaining complex traits. The rest almost 90% of the genetic variants were unexplained common variants that barely contribute to the explanation of complex traits (Plomin et al., 2009). In addition to this low explanation rate, most (>80%) of the identified SNPs were located in intergenic region. Even the regulate role of these intergenic-region loci have been proposed, few of them have been experimentally verified (Hindorff et al., 2009).

Many reasons could lead to the fact of why such big number of variants were unexplained. One proposed reason is lacking statistical power in detecting variants. Another reason could be due to the contribution comes from rare variants. Because tagging SNPs in GWAS cannot capture those rare variants, there contribution did not get revealed. To date, genetic variants that have been identified by GWAS only explains a small fraction of the susceptibility comes from inherited factors, even for those well-identified diseases such as Crohn's disease (Barrett et al., 2008).

1.4.6 Achievements

Many factors could affect process and outcome (healthy lifespan and/or longevity) of ageing. Successful ageing is closely related to the environment, medical support as well as genetic factors. Among all the factors that could affect ageing, genetic factors constitute a large portion of total ageing effects. Results from studies on twin siblings suggested genetic variation can explain around 25% of all the differential of longevity. Family-based follow-up studies further confirmed this (Caselli et al., 2006; Herskind et al., 1996; Hjelmborg et al., 2006; Skytthe et al., 2003). Progress has been made in finding genes that associated with longevity.

One type of extreme example that genetic factor contributes to ageing is the mechanisms operating underneath Progeroid Syndromes. Mutations in single genes lead to accelerated ageing in those progeroid syndrome patients, such as Werner Syndrome RecQ Like Helicase (*WRN*) gene in Werner's Syndrome and lamin A/C (*LMNA*) gene in Hutchinson-Gilford Progeroid Syndrome (Agrelo et al., 2006; Dreesen et al., 2011). The biological basis of these syndromes are mainly due to impaired function of DNA repair proteins. Patients affected by these syndromes reveal older appearance than their actual age should have.

Besides those examples that single mutation can have a remarkable impact on ageing, it is commonly accepted that ageing is a complex trait and regulated by multiple genes. As described in programmed ageing theory, the duty of soma finishes after reproduction. Therefore, it is very much likely the longevity is controlled by multiple genes (Kirkwood, 2011). Also, evidence supports the hypothesis that lifespan is plastic and can be responsive to interventions of nutritional, pharmacological as well as genetic factors (Magalhães, 2011; Vijg et al., 2008; Wilkins et al., 2003).

In contrast, some researchers argue that ageing is ineluctable and genetics should not control it because obvious ageing-related traits normally appear after typical active reproduction ages, therefore, it escapes natural selection (Johnson, 2002). However, experiments carried out on model organisms demonstrated that genes do control lifespan. For example, the Mammalian Target of Rapamycin (mTOR) pathway and insulin signal pathway do have impacts on longevity (Johnson et al., 2013). The discovery of mutant genes in different biochemical pathways in model organisms that could affect lifespan validated the point that lifespan can be affected by genetic variation (Kenyon, 2005; Passarino et al., 2016).

As of July 2013, 328 genes that associated with longevity have been identified in human by both GWAS and candidate gene association designs (discussed in section 2.5.1) (Budovsky et al., 2013). 99 genes were identified by CGASs and 243 genes were identified by GWASs. 14 genes were identified by both methods, they were *APOC3*, *NR3C1*, *SOD2*, *LMNA*, *PPARG*, *KL*, *APOC1*, *AKT1*, *IGF2*, *MLH1*, *APOE*, *TOMM40*, *FOXO1*, *FOXO3*. Among those 14 genes above, only *LMNA* is a single gene disorder gene that leads to a progeroid syndrome. Others are all metabolic functional related genes. For example, *APOC1*, *APOC3*, *APOE* are apolipoprotein metabolism regulating genes therefore correlated with the onset of cardiovascular diseases and neurodegenerative diseases and eventually affects human lifespan. Actually, the decrease of death in cardiovascular diseases patients hugely contributed to the extension of global average lifespan (Passarino et al., 2016).

1.5 Challenges in ageing research

Ageing has been studied for a very long time. Results obtained from both model organisms and human (see section 1.2) have been contributing our understanding in ageing. Massive progress has been achieved in ageing research over the past several decades. However, we are still far from understanding the whole picture of ageing, particularly in human. Obstacles still exist in deciphering ageing.

1.5.1 Of model organisms

One of the major challenges is the transformation of achievements obtained from model organisms to human. Model organisms do have their advantages in research (see section 1.1.3). Majority of current knowledge in ageing was derived from experimental data in popular model organisms, namely yeast, nematode, fruit flies and rodents (Cohen, 2018). However, most of these popularly used model organisms are not phylogenetically closely related to human. It is clear that demographic trajectories such as relative mortality and fertility, survivorship varies a lot along the "tree of life" (Jones et al., 2014). The difference between model organisms and human could introduce errors when trying to explain the results obtained in the model organisms to human. For example, mice have separate receptors for insulin and IGF-1, however, in worms and flies, there is only one single insulin/IFG-1-like receptor (Kenyon, 2005). Another example is cancer suppression mechanisms. As age-related diseases, cancers severely impair life quality and longevity. In human and other mammals, cell apoptosis and senescence prevent cancer from occurring, however, nematodes and files rarely develop cancer (Vijg et al., 2008). Even though we share aspects of the ageing process with these model organisms, it takes a huge amount of work to build connections and do the translation work between human and model organisms. Not to mention the potential idiosyncratic conclusions drawn from model organisms due to the high genetic homogeneity (Cohen, 2018; McClearn, 1999). In this sense, how directly we can transfer the knowledge obtained from animal-based studies to human population is unclear (Tissenbaum, 2015; Vanhooren et al., 2013).

Apart from the organism itself, lacking of environmental variation is another barrier in transferring lab experimental results to the real world. By far, the majority of experiments were conducted in laboratories or controlled conditions that are different from real-world complex environments where natural populations live in and lack of the ecology context (Cohen, 2018). Ageing has its intrinsic properties, however, it also operates closely with the environment. Different species owns their private characteristics and ways of interacting with the environment they live upon with. The integration of external environmental factors could alter the expected patterns in mortality, fertility and survivorship (Baudisch, 2011; Jones et al., 2014).

It is not easy to directly transfer the knowledge we obtain about ageing from model organisms to benefit human species. Due to the different metabolic mechanisms involved in ageing between human and other species, transformation and summarization on animal-based data should be carried out. We cannot limit our understandings to what we obtained from model organism studies, instead, we should try to summarise from what we learn from the accumulating results. Building universal models that are valid in both human and other organisms should be targeted, rather than just revealing what happened in each model organism separately. A system biology approach should be introduced into the pipeline of ageing research (Kirkwood, 2011). In the meantime, introducing a wider variety of species into ageing research (Cohen, 2018) and gathering experimental/observational data directly from human population should never stop, even it is sometimes challenging. It is expected that better-designed experiments, more sophisticated statistical methods and availability of more data will cast light on the ageing research.

1.5.2 Of traits: longevity and ageing

The limitation in phenotyping imposes another level of burden in ageing research. Complex trait, such as ageing, usually involves many signal pathways and metabolic pathways and integrative with the environmental effects. Therefore, it is a huge challenge to distinguish causes (reasons) from effects (results) in ageing research (reviewed in Magalhães, 2005).

Many researchers argue that too much focus has been placed on research in longevity rather than ageing itself (Magalhães, 2005; Jones et al., 2014; Cohen, 2018). It is true. Simply using an organism's lifespan as a representation of ageing is not precise, because a longer lifespan does not necessarily mean better life quality or a slowed ageing pace (Figure 1.1).

Although the concept of using healthy lifespan instead of longevity as markers in ageing research is attractive, it is not easy to translate it into reality because ageing is a complex process with many factors involved in. Some researchers use metabolic indexes as indications of "how healthy" an individual is (i.e. health-span), while others use physical characters such as Grip Strength to represent it. But none of them is fully capable as a gold marker in ageing research, nor to combine both. Not every aged people will exhibit each particular characteristic that has already been widely using in ageing research. The ambiguousness in defining age-related traits could potentially impair the progress in ageing research.

In recent years, there is an increased interest in studying healthy ageing (Kennedy et al., 2014). Healthy ageing research is a new concept and an elevation of current ageing research. In contrast to current lifespan focused experimental design, healthy ageing emphasises the independent living ability and disease-free lifespan. The continuous introduction of new concepts brings new challenges into the ageing research field.

To date, it is still not possible to inclusively represent ageing with all traits that have been used as markers for ageing research. In this sense, using longevity as a proxy in ageing research is a reasonable compromise.

1.5.3 Of cohorts

From the angle of intrinsic factors, the extent of genetic diversity in a species is determined by the *de novo* mutation rate and the reproduction cycle. *De novo* mutation occurs randomly across the genome. Through reproduction, germline *de novo* mutations get the opportunity of being integrated into the lineage and passed to offspring (Kimura et al., 1969). If a mutation happens in the genic region or regulatory region of a gene, with the interference of environment, the *de novo* mutation is either kept at a certain frequency or completely "swept out" in the population. The stochastic mutation rate is low in nature. In human, the mutation rate is as low as $\sim 1.2 \times 10^{-8}$ per nucleotide per generation (Kong et al., 2012).

Considering the current global population size and typical human generation intervals, these SNPs need a long time to reach the current frequency in populations since their *de novo* mutations. Therefore, the SNPs are relatively old and common variants. One of the most well-known bottlenecks happened in human history, the *Out of Africa* bottleneck, further reduced the genetic variability in non-African human population (Chakravarti, 1999). Alternatively, to describe this from another perspective, the *Out of Africa* bottleneck reduced the genetic variability carried by common ancestors of current non-Africa populations. When common ancestors of current non-Africa populations dispersed around the world, the genetic exchange between subgroups was quite limited until long-distance transports are invented.

On global population level, human population exhibits a low divergence in terms of genetic variability (Barbujani et al., 2013). Of this low genetic variability, 83% comes from genetic variability between individuals (Lewontin, 1972). The relatively low global genetic variability, as well as high genetic variability between individuals, bring hurdles in the genetic association studies. In the scope of global population, lower genetic variability needs more samples to compensate in achieving a given strength of statistical power (Tan et al., 2014). If results that universally apply to the global human population are expected, huge sample sizes are necessary. In the scope of this particular thesis, where human longevity and ageing related traits/diseases were focused, long-lived individuals (LLI, i.e. centenarians and supercentenarians, who live to or beyond the age of 100 and 110, respectively) were needed for investigations. However, the number of LLI is quite limited, which obviously will impair the statistical power.

On the other hand, the high genetic variability between individuals introduced excess noises to the results of GWAS. Because the genetic variability between individuals is high at the basal line, it will make the difference between groups hard to be detected. In other words, When comparing the difference of genetic variability between LLI cohorts and younger-aged cohorts, the actual difference contributes to longevity may be immersed in the basal line genetic variability difference.

1.6 Summary

Ageing research has advanced obviously in the past decades. To date, it is commonly accepted that ageing as an irreversible, intrinsic, universal and complex biological process that cause deleterious effect, including gradual decline in function and increase of vulnerability, with the increase of chronological age. Regardless of the achievements and obstacles in ageing research mentioned above, there are still many questions need to be answered. For example, in the human population:

1. It is known genetic factors contribute to the ageing process and outcome. But,

what are those genes, do they share anything in common?

2. Ageing is a complex trait and many genes involved in the ageing process. However, those genes may have other effects apart from involving the ageing process. How could those longevity associated genes couple with the underneath metabolic processes?

3. Genetic association studies identified many genetic variants that could contribute to the longevity/ageing process, how much reliable are they?

4. In terms of the genes themselves, what could affect the ability of getting a hit by genetic association design? What is the difference of genic attributes between longevity associated genes and that of other complex traits, such as cancer, associated genes?

The development of genomic technology sheds lights on the genetic basis of ageing, especially in human. In this thesis, I described a full workflow from gathering HLAGs and populating them into a database, through looking into the functional enriched clusters, assessing the data quality delivered by genetic associated study design, to the examination of genetic diversity of those HLAGs and how that differs from the genetic diversity of other complex trait associated genes. By doing this, it is expected to have a better understanding in the genetic basis of ageing and age-related diseases from the perspective of genetic diversity and contribute to the ultimate goal, deciphering ageing.

1.7 Aims of the thesis

Ageing is inevitable, universal but variable from one species to another. Genetic factors are undoubtedly important in affecting the pace and outcome of ageing in humans. Herein, an exploration on the genetic basis of ageing and age-related traits/diseases was conducted in order to achieve those following aims:

1. To gather a dataset of HLAGs, then construct a database for HLAGs called LongevityMap.
2. To analyse the longevity-associated genes as a whole and look for new biological functions or pathways that could contribute to human longevity.
3. To assess data quality and analyse the attributes of the LongevityMap dataset.
4. To investigate the relationship between the Genetic Diversity (GD) of HLAGs and ageing related traits.

Chapter 2

The LongevityMap Database

2.1 Introduction

The ageing process is the result of interactions between genetic and environmental factors. Genetic variants undoubtedly affect the process and outcome of ageing. Many efforts have been put into the discovery of genetic variations. Some of the biggest international collaborative projects, such as The 1000 genomes project (The 1000 Genomes Project Consortium, 2010), The ENCODE Project (ENCODE Project Consortium, 2004), The International HapMap Project (International et al., 2003) have successfully discovered numerous variants in human genome. Those above projects have also been promoting usage of the data by making the data freely accessible. Besides the effort comes from the scientific community, the development of chip technology further accelerated data accumulation. New emerged methods can genotype SNPs in an accurate, high-throughput and low-cost manner. The reduced average cost for genotyping single locus promotes the increase of loci capacity in individual study (Magalhães, 2015). This leads to rapid accumulation of both human genomic data and the results from trait-genetic association studies in ageing research. In order to meet the urgent need for a tool to manage and index this enormous amount of data, we built the LongevityMap database to cope with the updating data. My main role in this project was a database curator. Therefore, only literature review, data preparation and the LongevityMap website interface will be described in this thesis. While the underneath of the database, such as how the database was implemented, will not be covered.

2.1.1 Demands of building the LongevityMap

Genetic association studies have been playing an important role in ageing research since the 1990s when the first longevity-genetic association study was published. Looking for potential ageing-causal variants through identifying longevity/ageing associated genetic variants became feasible and easier with the aid of high-throughput sequencing technologies. In the past decades, overwhelming novel information of longevity associated DNA loci were delivered to research community. The rapid increasing publications brought a huge amount of data in ageing research and also the issue of managing and indexing these data, which brought urgent demand for a database.

While the data was accumulating in an un-foreseen speed, the corresponding solution that can handle the data was yet ready in the ageing research community. Back to early 2013 when I started to work in human genetic ageing research, there was no single repository or any other equivalent tool-kit functions as a central hub available for storing or indexing the huge amount data. Gathering small pieces of information in ageing research was not easy, not mention to have an overview or obtain information in a more effective way. The shortage of essential tools brought the inspiration of building a longevity-associated-gene themed database.

We expected the database can integrate all the up-to-date published studies in longevity research. In the aim of indexing and promoting the usage of genetic data more efficiently, the LongevityMap database was implemented in 2013. The managed pooled information in the LongevityMap database has been serving ageing research effectively since then.

2.1.2 Databases in genetic research

Databases are important in bioinformatics research. They are useful in integrating, indexing and transferring information within the research community. Many popular databases such as GenBank, dbSNP have been serving the research community and promoting the use of these data for years (Teufel et al., 2006).

Well-designed databases can be user-friendly and functionally sophisticated. On one hand, databases can index and export specific information as the end user requested. On the other hand, it can keep itself updated by communicating and exchanging information with other outer resources. Databases seem to be the best solution for managing that information complex biomedical data. Having a well-designed database not only benefits the current research field but also promotes spreading information to other related disciplines. In addition, the centralised and structured data facilitate the exerting new information by other contemporary technology such as system biology methods (Kirkwood, 2011).

2.1.3 Rationale for the LongevityMap

To fill the gap of indexing human genetic association studies and storing the outcomes from those, we sketched up the LongevityMap database with following aims:

1. It should be theme-focused, inclusive and up to date. What we need is a database that collects all the human longevity-genetic association studies in healthy people (without obvious disease or morbidity). Therefore, it should be designed only focusing on this topic. In addition to this, the database should be updated regularly to keep abreast of the latest outcomes in the field.

2. It needs to be concise and functionally sophisticated. As the data is already overwhelming, the database should not be an overcrowded harbour that just simply populating all the available data. It should work as an index by providing key information of each individual study and pointing to its source.

3. It needs to be accurate. As the database is designed in the aim of being a central hub, we want the concise information provided by the LongevityMap is reliable. Therefore, we manually curated the data for constructing the database, followed with manually re-examination by another professional curator.

4. It needs to be user-friendly. Building a database needs some specific professional skills, however, accessing certain information does not have to be. In the consideration of potential users may consist of researchers with various backgrounds, we want the database is easy-accessible and convenient to use. A user-friendly interface should be supplied to researchers who do not involved many programming skills in their every research or just want a quick enquire in the database, However, for researchers who want to use the LongevityMap data in a programmatic way, a downloadable source file should be available.

5. It needs to be community involved. This project was inspired by the lacking of an essential database in the community. It is necessary to make this database publicly available to anyone interested in the ageing research. We want this database not only aid our own research but also support the community. With this regard, the database should be distributed with least restriction on the usage of data.

With targeting the above aims, we built the LongevityMap database in 2013.

2.2 Data preparation

2.2.1 Literature review

In the literature review phase, candidate journal articles that were published in the major journals in the field were obtained by searching key words, such as "*ageing*", "*association study*" and "*longevity*", in *PubMed*, *Web of Science* and *Google Scholar*. In this step, different spell format of the same word, like "*ageing*" vs. "*aging*", or the combinations of key words, such as "*longevity" AND "association study*", were also enquired.

Besides the results returned from online enquires, key journals in the ageing research field were also carefully checked to make sure not missing any studies. Additionally, the reference section of literature also provided a good resource for candidate literature. With this approach, we further boosted the coverage of our database. The candidate articles were collected for further examination.

In the next step, a filtering process was applied to those candidate articles. Only papers that focused on longevity-genetic association studies were selected, regardless of the involved sample size in the study. This means both large and small studies have equal opportunities of being included in the database. Although inclusion of small sample size studies could potentially introduce biases (discussed in Chapter 4) it is still important to do so because the LongevityMap is designed to be "inclusive" by capturing all the available information and honestly reflect the current status of the field (*i.e.* longevity-genetic association studies). However, as the focus of building this database is to provide information on the genetic basis of natural longevity and healthy ageing, studies that focused on morbid cohorts such as cancer patients were excluded from the current database ("exclusive" to irrelevant studies).

Additionally, the LongevityMap database follows another baseline "inclusive" criterion in data processing. The "exclusive" criterion refers to the standard that was followed when selecting the literature. In contrast, the "inclusive" criterion applied to those literature that have been selected. For a research paper that meets the selection criteria, we want more "inclusive" in retaining information from the paper. This means we want to provide users with sufficient details of selected literature. With respect to these above standards, information was manually curated from the literature and organised into the database.

In summary, the first "inclusive" criteria guaranteed the scope of candidate studies while the second asserted coverage depth in a given study. Additionally, the "exclusive" criteria filtered out the irrelevant information from being included. Moreover, manually curation process offered maximum data accuracy.

2.2.2 Data curation

Studies that meet the literature selection criteria were passed down for data curation. From each study, the meta-data such as PubMed unique Identifier (PMID), studied population, study design, results (significant/non-significant) were directly retrieved and recorded. In addition, a brief description of major outcome from the author's original conclusion was summarised for the outcomes of each study.

The statistical usage and criteria in the literature were also examined. Before carrying out a statistical test, the null hypothesis H_0 representing no difference between two groups was defined. An alternative hypothesis H_1 representing there is difference between two groups under examination was also defined. Usually, researchers draw conclusions based on the comparison between p values and the predefined significant threshold α. If p value less than α, then the H_0 was rejected and H_1 was accepted. This will result a significant/positive outcome. Otherwise when p value bigger than α, H_0 will be accepted and therefore a non-significant/negative conclusion will be drawn. In most of the cases, α is defined as 0.05. However, α can be adjusted to meet the actual need. For example, in Bonferroni correction, which is one of the most used multiple test correction method, an adjusted $\alpha = 0.05/n$, where n is the number of independent statistical tests, is often used to minimise the multiple test effect. Overall, three types of α were observed in the candidate studies:

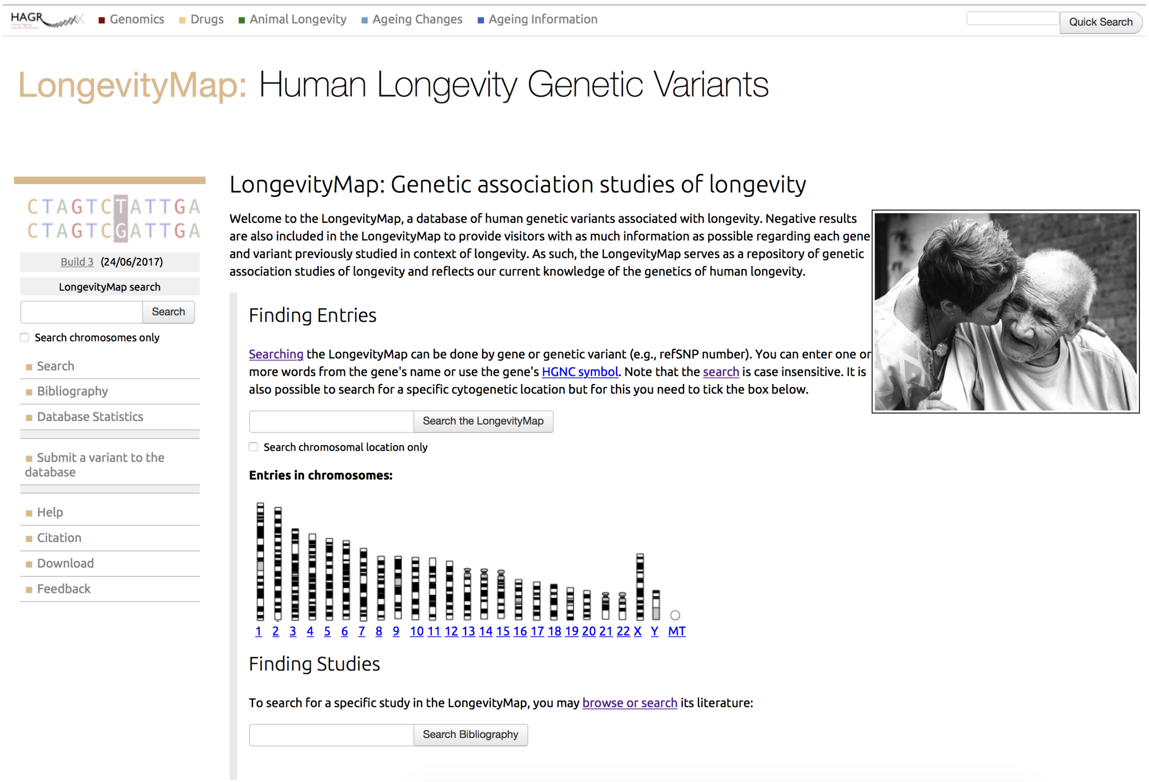
1. A cut-off of $\alpha = 0.05$ was used in majority of the CGASs.
2. Some CGASs use adjusted α values when multiple testing is involved.
3. In all GWAS studies, the default cut-off (normally $\alpha = 5 \times 10^{-8}$) was used.

Each variant was highlighted with either "Significant" or "Non-significant" based on the author reported results. In many studies, only a subset of studied genes, or variants in the same gene, were significantly associated with longevity while others were not. In these cases, multiple entries were created for the same study according to genes or their significant/non-significant outcomes. As a result, each entry only include one significant or non-significant result. All these efforts were targeting on clarifying the major outcomes.

GWAS and CGAS are two basic types of genetic association studies. The scale of targeted loci differs very much although they share similar statistical methods and study design underneath. GWAS design evaluates tens of thousands of genetic variants in a single study, however, the number of genetic variants that were investigated in a typical non-GWAS (CGAS) is usually quite limited (see Table 1.1). Given the different features of these two study types, we designed two templates to present data from these two type studies. The application of different templates based on the intrinsic properties of study made the database concise and robust. In short, only significant genes/variants were displayed in each GWAS. While in CGASs, all the candidate loci were listed regardless of whether the result was significant or not.

2.3 Data organisation and web interface

All the curated data as described in the previous section was integrated into a database, named as LongevityMap (`http://genomics.senescence.info/longevity/`). A user-friendly interface was provided for the ease of use (Figure 2.1). Since LongevityMap serves as an information hub of associations between longevity and genetic variants, the database was implemented with two types of the entry page, study-centric page and gene/variant-centric page.



LongevityMap: Human Longevity Genetic Variants

LongevityMap: Genetic association studies of longevity

Welcome to the LongevityMap, a database of human genetic variants associated with longevity. Negative results are also included in the LongevityMap to provide visitors with as much information as possible regarding each gene and variant previously studied in context of longevity. As such, the LongevityMap serves as a repository of genetic association studies of longevity and reflects our current knowledge of the genetics of human longevity.

Finding Entries

Searching the LongevityMap can be done by gene or genetic variant (e.g., refSNP number). You can enter one or more words from the gene's name or use the gene's HGNC symbol. Note that the search is case insensitive. It is also possible to search for a specific cytogenetic location but for this you need to tick the box below.

Entries in chromosomes:

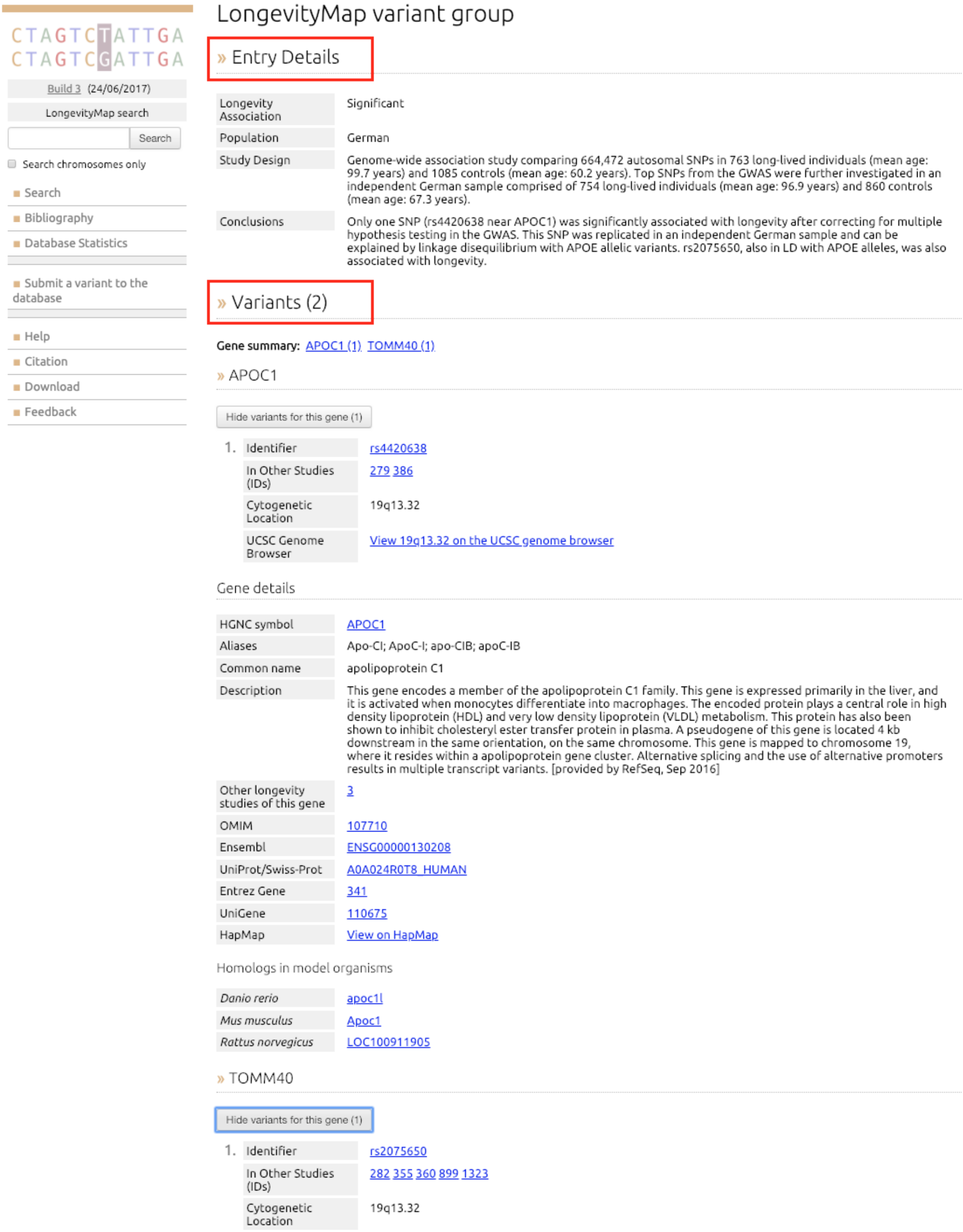
Finding Studies

To search for a specific study in the LongevityMap, you may browse or search its literature:

Figure 2.1: **The landing page of the LongevityMap database.** Available at `http://genomics.senescence.info/longevity/`.

In study-centric pages, information on the outcomes of the current study, the studied cohort, the design of a study as well as a brief conclusion was described in the “Entry Details” section. The variants that were reported in the current study were listed in following “Variants” section (Figure 2.2). The genes that harbours the variants were also presented on the page.

In gene-centric pages, the basic information such as cytogenetic location and description of the gene were displayed on the page (Figure 2.3). Further information was provided through additional links to other public databases, such as *Ensembl* and *Online Mendelian Inheritance in Man (OMIM)*, were also implemented into the page. Importantly, a list of studies that including the currently displayed gene was listed. Users can access all the studies including a given gene through cross-links in any gene-centric page. Similarly, users can retrieve all the studied genes from a listed study. It is also easy to navigate to other outer sources databases through the implemented links. For instance, Reference SNP (rs) numbers to the *dbSNP* database and cytogenetic locations to *UCSC genome browser*. All these features make retrieving information from the LongevityMap very efficient.



LongevityMap variant group

» Entry Details

Longevity Association	Significant
Population	German
Study Design	Genome-wide association study comparing 664,472 autosomal SNPs in 763 long-lived individuals (mean age: 99.7 years) and 1085 controls (mean age: 60.2 years). Top SNPs from the GWAS were further investigated in an independent German sample comprised of 754 long-lived individuals (mean age: 96.9 years) and 860 controls (mean age: 67.3 years).
Conclusions	Only one SNP (rs4420638 near APOC1) was significantly associated with longevity after correcting for multiple hypothesis testing in the GWAS. This SNP was replicated in an independent German sample and can be explained by linkage disequilibrium with APOE allelic variants. rs2075650, also in LD with APOE alleles, was also associated with longevity.

» Variants (2)

Gene summary: APOC1 (1) TOMM40 (1)

» APOC1

Hide variants for this gene (1)

1.	Identifier	rs4420638
	In Other Studies (IDs)	279 386
	Cytogenetic Location	19q13.32
	UCSC Genome Browser	View 19q13.32 on the UCSC genome browser

Gene details

HGNC symbol	APOC1
Aliases	Apo-CI; ApoC-I; apo-CIB; apoC-IB
Common name	apolipoprotein C1
Description	This gene encodes a member of the apolipoprotein C1 family. This gene is expressed primarily in the liver, and it is activated when monocytes differentiate into macrophages. The encoded protein plays a central role in high density lipoprotein (HDL) and very low density lipoprotein (VLDL) metabolism. This protein has also been shown to inhibit cholesteryl ester transfer protein in plasma. A pseudogene of this gene is located 4 kb downstream in the same orientation, on the same chromosome. This gene is mapped to chromosome 19, where it resides within a apolipoprotein gene cluster. Alternative splicing and the use of alternative promoters results in multiple transcript variants. [provided by RefSeq, Sep 2016]
Other longevity studies of this gene	3
OMIM	107710
Ensembl	ENSG00000130208
UniProt/Swiss-Prot	A0A024R0T8_HUMAN
Entrez Gene	341
UniGene	110675
HapMap	View on HapMap

Homologs in model organisms

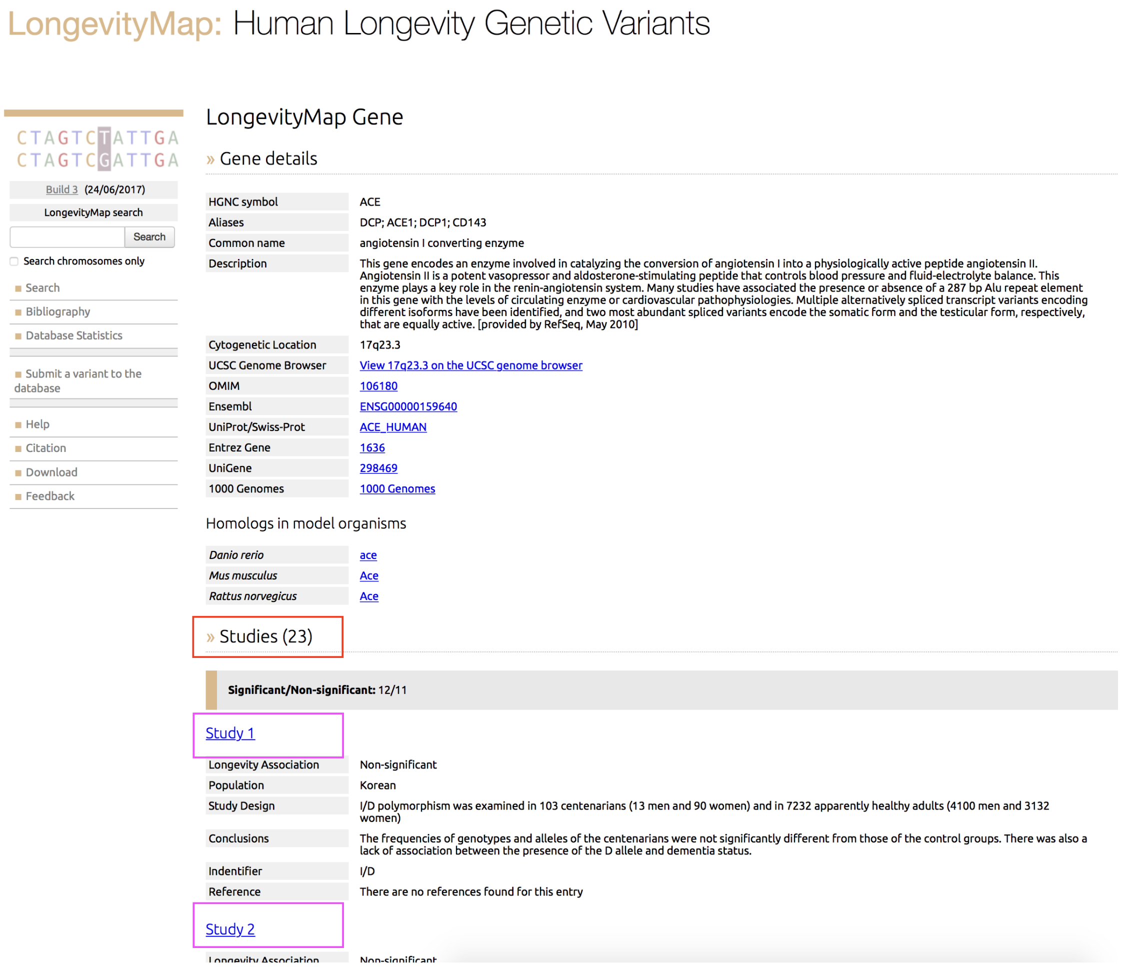
Danio rerio	apoc1l
Mus musculus	Apoc1
Rattus norvegicus	LOC100911905

» TOMM40

Hide variants for this gene (1)

1.	Identifier	rs2075650
	In Other Studies (IDs)	282 355 360 899 1323
	Cytogenetic Location	19q13.32

Figure 2.2: **Partial of a study-centric page in the LongevityMap database.** Red boxes indicate the "Entry Details" and "Variants" sections in the page.



LongevityMap: Human Longevity Genetic Variants

LongevityMap Gene

» Gene details

HGNC symbol	ACE
Aliases	DCP; ACE1; DCP1; CD143
Common name	angiotensin I converting enzyme
Description	This gene encodes an enzyme involved in catalyzing the conversion of angiotensin I into a physiologically active peptide angiotensin II. Angiotensin II is a potent vasopressor and aldosterone-stimulating peptide that controls blood pressure and fluid-electrolyte balance. This enzyme plays a key role in the renin-angiotensin system. Many studies have associated the presence or absence of a 287 bp Alu repeat element in this gene with the levels of circulating enzyme or cardiovascular pathophysiologies. Multiple alternatively spliced transcript variants encoding different isoforms have been identified, and two most abundant spliced variants encode the somatic form and the testicular form, respectively, that are equally active. [provided by RefSeq, May 2010]
Cytogenetic Location	17q23.3
UCSC Genome Browser	View 17q23.3 on the UCSC genome browser
OMIM	106180
Ensembl	ENSG00000159640
UniProt/Swiss-Prot	ACE_HUMAN
Entrez Gene	1636
UniGene	298469
1000 Genomes	1000 Genomes

Homologs in model organisms

Danio rerio	ace
Mus musculus	Ace
Rattus norvegicus	Ace

» Studies (23)

Significant/Non-significant: 12/11

Study 1

Longevity Association	Non-significant
Population	Korean
Study Design	I/D polymorphism was examined in 103 centenarians (13 men and 90 women) and in 7232 apparently healthy adults (4100 men and 3132 women)
Conclusions	The frequencies of genotypes and alleles of the centenarians were not significantly different from those of the control groups. There was also a lack of association between the presence of the D allele and dementia status.
Indentifier	I/D
Reference	There are no references found for this entry

Study 2

Longevity Association	Non-significant

Figure 2.3: **Partial of a gene-centric page in the LongevityMap database.** All studies include *ACE* gene were listed in the "Studies" section. Links to corresponding "study-centric" page were provided through those "Study 1", "Study 2" ... texts. The red box indicates the "Study" section. The pink boxes indicate links to "study-centric" pages.

2.4 Major update

Genomic research is a rapidly developing field. With the aim of LongevityMap being the central repository of longevity genetic association studies, keeping the data up to date and accurate are essential. Until 2016, almost three years after the initial release, many new studies emerged (Vanhooren et al., 2013). Newly published papers includes both Genome-Wide Association Study (GWAS) and Candidate Gene Association Study (CGAS or non-GWAS). Some of the new studies are relatively large-scaled in the number of candidate loci and sample sizes (Dato et al., 2014; Debrabant et al., 2014; Deelen et al., 2014; Raule et al., 2014). The organisation of LongevityMap database is loci/gene-oriented rather than individual study oriented, which means a single study could correspond to several entries in LongevityMap according to the number of studied genes. The large-scale studies aggravate the latency of data collection in the database. There was an urgent requirement to update the database at the time.

Due to the decreasing cost of high-throughput sequencing and the increasing of publicly accessible data from previous studies, individual study tends to include more sample (Figure 2.4) and more loci (not showed). Two extra large-scale CGAS studies (Dato et al., 2014; Debrabant et al., 2014), involving 311 SNPs from 38 genes in 1089 individuals and 592 SNPs from 77 genes in 1825 individuals respectively, were observed in the newly published studies. New study designs by gathering, combing and re-analysing previous data for potential new discoveries (Deelen et al., 2014) also emerged. The latest update brought 135 new entries from 12 studies into the database.

Corrections were also made to errors and incomplete information of entries in the first release. Twelve new entries split from initial entries were added and 28 loci with incomplete information were improved (see Table 2.1, Table 2.2).

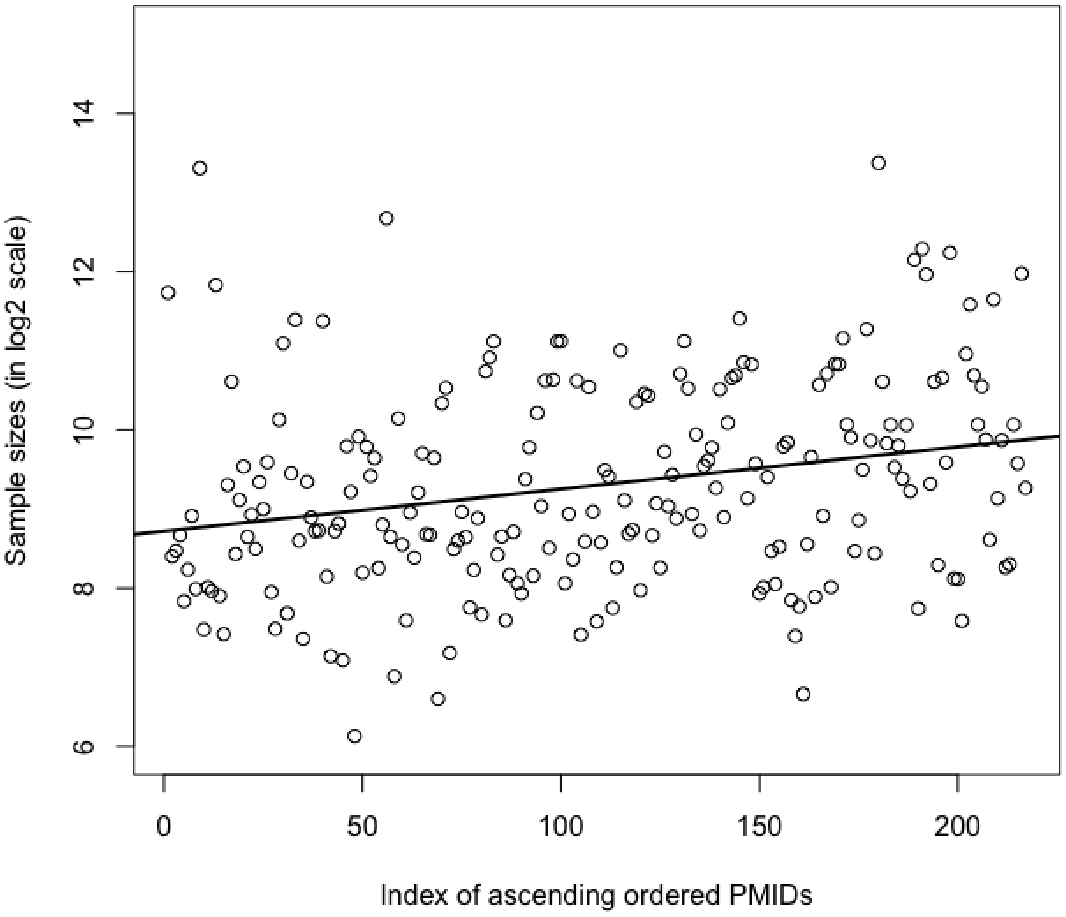


Figure 2.4: **Sample sizes involved in single CGAS increases over years.** (*Pearson correlation,* $r = 0.234, p = 0.0005$. The ascending ordered PMID on the x-axis represents time. Each data point in the figure represents single individual Candidate Gene Association study(CGAS) in the LongevityMap. The fitted line indicates the increasing trend of sample size.)

Table 2.1: Summary of the first LongevityMap update

Information type	Numbers
Studies	12
Entries	135
Missing entries added	12
Errors revised	28

Table 2.2: Comparison of two releases of LongevityMap database

	Release 1	Release 2
	(Jul 2013)	(Feb 2016)
Entries	512	551
Studies	255	267
Genes	755	860
Variants	2005	3025
Significant entries	257	275
Non-significant entries	255	276

2.5 Summary and discussion

2.5.1 Overview of the database

The latest LongevityMap release covers longevity-genetic studies from three decades since 1987. Since all literature was collected following an in-depth literature survey method and included both large and small studies, users should feel confident for the coverage of the database, in terms of depth-coverage and time-span coverage.

The initial release of the database was on 26th July 2013. In the first release, there were 755 genes that have been investigated in the 255 studies. Of the 755 genes, 328 of them were significantly associated with human longevity. With an update in the early of 2016, the number of studies and studied genes/loci were increased. As of 14th Feb 2016, the second release of LongevityMap presented 3025 variants came from 859 genes, which is approximately 1.5-fold of the original (described in section 2.4). The new included genes and variants will certainly bring more information capacity for the genetic basis of ageing research.

Sufficient statistical power is vital for detecting risk alleles in genetic association studies (both CGASs and GWASs). Genetic models, minor allele frequency and sample sizes and many other factors affect the statistical power. For example, less sample is needed under dominant model than other genetic models. Previous study has estimated the required sample size to achieve sufficient statistical power (Table 2.3) (Hong et al., 2012). Comparison between require sample size with actual involved cohort sizes in LongevityMap studies confirmed studies that included in the LongevityMap are well powered for both CGASs and GWASs.

Table 2.3: Number of cases needed to achieve sufficient statistical power

Number of loci to be tested	Sample size
1	248
500,000	1206
1,000,000	1255

The required sample sizes were estimated under the assumption of $odds\ ratio = 2$, $disease\ prevalence = 5\%$, $MAF = 5\%$, $case/control = 1 : 1$, $Error\ rate_{allelic\ test} = 5\%$, complete linkage disequilibrium. Data was taken from (Hong et al., 2012).

2.5.2 Data access

A user-friendly interface was designed for users who want to have a quick check a desired gene or a study. Users can use the query box to quickly retrieve information by keywords, such as gene names or PMIDs. It is also possible to browse through cytogenetic regions from the presented chromosome figure (Figure 2.1). A built-in filter system is always there to help in retrieving information more efficiently.

Alternatively, it is also possible to download the whole database as a single file (`http://genomics.senescence.info/download.html#longevity`) for users want to integrate the whole LongevityMap data into their own working pipeline and analysis the LongevityMap data locally.

The LongevityMap database was distributed under *Creative Commons Attribution 3.0 Unported Licence*, which provided most convenient for the usage of the data. In short, it is free for all the purposes, including educational, academic and even commercial purposes.

2.5.3 Limitations

The LongevityMap database was built as a new member of the Human Ageing Genomic Resources (`http://genomics.senescence.info/`) (Tacutu et al., 2013), and a sister database of GenAge database (`http://genomics.senescence.info/genes/`), which is mainly focused on the genes that associate with longevity and progeroid syndromes in model organisms. Therefore, similar rigid data collection criteria as in its predecessors were obeyed.

Even so, it is worth to mention the limitations of the database. Curators made a great effort to make sure the information loyalty to the original study. However, any errors introduced before the curation will be kept and included into the LongevityMap. Misuse statistical methods and potential publication biases (discussed in Chapter 4) could inflate the type I error (false positive finding) in the database. Users should be aware of the limitations while using the database.

Another limitation is the time effectiveness. As the LongevityMap is manually curated, there will be latency between release cycles. Any new literature published after the release date maybe still in the curation process and will not be available in the LongevityMap. Therefore, checking the release date of current version is highly recommended before use.

2.5.4 Conclusion

The LongevityMap database is the first database that harbours all the human longevity-genetic association studies and the corresponding results in one place. It not only serves as an invaluable central hub for looking into studies and achievement of those studies, but also an excellent portal for exploring other related information collected in HAGR through built-in cross-links.

Genomic research is a fast-moving field, the LongevityMap database is maintained and updated on a regular basis, depending on the publication density and breakthroughs in this field. Error corrections are carried out as soon as they are spotted. All of these efforts are targeting at providing the community with a freshest and reliable data resource in ageing research.

Chapter 3

Data Analysis of the LongevityMap

3.1 Introduction

Thanks to the implementation of the LongevityMap database, we have a central resource of longevity genetic association studies and the outcome of studied variants. Using this database, we gained the chance to investigate the genetic basis of human longevity. However, a collection of HLAGs will provide nothing more than the database itself.

With the aim of having a clearer view of the properties of longevity genetic association studies in LongevityMap, and to have a better understanding of the genetic basis of longevity, analyses targeting at the meta information of the LongevityMap were carried out.

3.2 Pathway-based analysis methods and tool selection

3.2.1 Gene Ontology (GO) based methods

Gene Ontology (GO) is a list of standardised words that describing attributes of genes and gene functions (Harris et al., 2004). These standardised words are named as GO terms. They are useful in presenting information from different studies in standardised vocabularies. Because the terms are standardised, they are easier to be processed by computers in an automated way.

GO enrichment analysis is one of the most commonly used methods for Gene Set Analysis (GSA) (Subramanian et al., 2005). It statistically compares the observed occurrence of ontology terms with the expected purely random occurrence. Through these GSAs, a set of ontology terms that overrepresented in a given themed gene set can be obtained. These ontology terms contain biological information of the gene products. By examining the scope that covered by ontology terms, one can have an overview of the functions that were covered by those themed gene sets as well as the potential new functions that not yet known before the analysis (Jensen et al., 2003). However, the results from a GO based analysis will not reveal or imply any details regards to the biological interactions (Mooney et al., 2015).

In the current study, we mainly focused on the discovery of novel functional clusters based on the Longevity-themed gene sets. Therefore, GO analysis was carried out aiming to reveal any un-observed functional clusters from ontology terms. Functional enrichment analysis tools are useful in obtaining an overview of the functional clusters for a given set of genes (Huang et al., 2009b). By comparing the results obtained from functional enrichment analysis with the current knowledge of longevity, it is possible to discover new functional enrichment clusters that contribute to the ageing process.

In this context, DAVID Functional Annotation Tools (DAVID 6.7, `https://david-d.ncifcrf.gov`) was employed in analysing the functionally enriched clusters of GO terms that were associated with the longevity gene set. DAVID tools use an EASE(enrichment) score system, which is a modified, more conservative version of Fisher Exact test, to estimate whether the occurrence of GO terms associated with certain group is higher than by random chance (Huang et al., 2009a). If the observed occurrence is statistically higher than that of by random chance, then an enrichment is observed. A *p* value associated with the statistical test is also reported. In addition to the *p* value, an enrichment score, which is obtained by calculating the geometric mean (in - log scale) of the *p* values in a cluster, will also be reported. Higher ranked enrichment score of a functional enriched cluster means the smaller aggregated *p* values of the members in the cluster, therefore, the functional enriched cluster is more representative. The introduction of enrichment scores and more stringent Fisher's exact test made results generated by DAVID is reliable and easy to understand. The DAVID is also fast, reliable and user-friendly for gene functional annotation and classification (Jiao et al., 2012).

3.2.2 Physical interaction analysis

Physical interaction analysis methods focus on the details of how proteins translated from a set of themed genes interacted with each other in a given background. By providing the information of Protein-Protein Interactions (PPIs), the functions of genes can be characterised(Yook et al., 2004; Safari-Alighiarloo et al., 2014). One major limitation of this method is that the detected interactions are affected by the background. Variables existed in experimental design, material could result differences of the physical interaction analysis. The other limitation of physical interaction analysis is the results will not reveal any information on what biological role the whole set of themed genes act, or how the whole set of gene fit into any biological functions.

Since we wanted to know how biological basis could contribute and affect longevity, the physical interaction analysis is not suitable to the current longevity data. At the moment, studies collected in the LongevityMap comes from different studies examined from different geographic locations with distinct genetic backgrounds, the variation among studies is very high. Given the physical interaction analysis method is sensitive to the variations in experimental factors, it is still too early to apply these physical interaction based methods to LongevityMap data. Nevertheless, the physical interaction analysis will play an import role in deciphering the molecular basis of longevity and ageing once smaller functional clusters, such as pathways, has been pinpointed.

3.2.3 Pathway based analysis

Pathway based analysis discovers the most basic, common shared biological interactions in a given set of genes. Normally, how those genes interconnected with each other, how a set of genes as a whole fits into the biology functions can be revealed from pathway analyses. It explains the biological information of a set of genes from well-studied pathway interactions, therefore, it is useful in connecting a set of genes to the associated biological functions and revealing the potential explanations of biological processes.

Many resources can provide the most updated information of pathway analysis, such as KEGG pathways (Kanehisa et al., 2000), PANTHER (Mi et al., 2009). Among all those data resources, Reactome database suits our needs best because a), it is a manually annotated database, which provided the best data accuracy. b), it only focuses on the single species, *Homo sapiens* (Croft et al., 2014; Fabregat et al., 2018), which guarantees the results obtained from Reactome analysis can be translated to the explanation the biological basis of human longevity with minimum variation.

Cytoscape is an open-source software to visualising, modelling and analysing different types of biological data by integrating biological networks (Shannon et al., 2003). The core algorithm of Cytoscape only provides the function of layout and query the network, to visualise the network and to link the network to function annotation network (Shannon et al., 2003).

Even the basic function of Cytoscape is quite limited, new functions can be easily obtained by installing plugins through the *Cytoscape App Store* (`https://apps.cytoscape.org`)(Saito et al., 2012). For example, scientists can extend the functions to expression enrichment analysis by installing the BiNGO plugin (Maere et al., 2005), or install the KEGGscape plugin (Nishida et al., 2014) to analysis and visualise KEGG pathway (Kanehisa et al., 2014). In this study, the ReactomeIFIViz plugin (Wu et al., 2014), which is distributed by Reactome, was used to rendering the layout of Reactome analysis results following the steps described in (Cline et al., 2007).

3.3 Results

3.3.1 DAVID functional enrichment analysis

The initial release of LongevityMap collected 755 genes in total (Table 2.2). Of the total 755 genes, 328 genes have been positively (significantly) reported at least once. In order to see how those 328 longevity-associated genes were functionally clustered against the human genome background, the first DAVID Functional Enrichment analysis was performed.

Two sets of genes, the "query genes" and the "background genes", are needed for running DAVID Functional enrichment analysis. Users will need to provide a set of "query genes" to the DAVID tool. For the "background genes", users can either select the built-in default background, which is the human genome genes, or provide their own gene set as a customised background.

In the first analysis, those 328 genes were submitted to the DAVID tools as the "query genes". However, due to the update intervals of DAVID tools, 29 genes failed to be identified by DAVID tools. Therefore in the following analysis, the "query genes" list only contains the rest 299 genes that were identified by DAVID tools. The "background genes" was set as the default human genome genes. The results show that those 299 genes were clustered into 190 functional clusters with the enrichment scores ranged from 14.09 to 0.03. Of the 190 clusters, 108 clusters showed an enrichment score greater than 1.3, which met the significance threshold (Huang et al., 2007; Huang et al., 2009a). The enrichment score (E score) of top five percentile of the 190 clusters (the top 9 functional annotation clusters) stopped at an enrichment score of 5.55. Considering the reality of using those data, a more reasonable (applicable) cut-off of enrichment score of 2.5 was used to reduce the noise in the results. After applying the new cut-off score of 2.5, 49 clusters were left (Appendix 1). The functional annotation clusters mainly related to the vital processes or key mechanisms that affect lifespans, such as cell-programmed death, cell locomotion, ion binding, and signal pathways et al (Table 3.1).

Functional enrichment analysis is background-dependent. The same set of "query genes" could reveal different enrichment clusters against different sets of "background genes". In order to see how these genes were enriched against the background of all the LongevityMap genes, all 755 genes that collected in the LongevityMap were set as the "background genes" for the second run. Again, in these 755 genes, some genes failed to be recognised by the DAVID tool due to the DAVID tool did not keep up-to-date with the latest known genes. 706 genes out of the total 755 candidates were recognised by the DAVID tools. The "query genes" list was kept the same as those 299 genes in the first run. In short, the second run was performed with those 299 genes as "query genes" and 706 genes as "background genes". The result showed that the enrichment scores of the new 188 functional annotation clusters were ranged from 18.23 to 0.11, and the top 5% of the 188 clusters scored from 7.86. There were 126 clusters enriched over the significance threshold. After filtering the results with an enrichment score of 2.5 as the threshold, 62 functional enrichment clusters were obtained (Appendix 2). Major clusters consisted of regulation of apoptosis, regulation of phosphorylation, response to the environment, regulation of locomotion and response to hormone stimulus (Table 3.2, analysed on 21th Jan 2015). By comparing the results of the two analysis, a similar functional annotation clustering pattern as the first run was observed (Figure 3.1).

Table 3.1: Annotation clusters with the human genome background

Annotation Cluster	Enrichment Score	the most representative term	n	FDR*
ACDB 1	14.09	Regulation of cell death	64	5.50E-22
ACDB 2	10.81	positive regulation of signal transduction	32	6.40E-14
ACDB 3	8.88	regulation of response to external stimulus	24	3.00E-13
ACDB 4	8.69	regulation of locomotion	23	1.40E-10
ACDB 5	7.77	response to hormone stimulus	33	4.30E-12
ACDB 6	6.98	response to extracellular stimulus	23	2.10E-09
ACDB 7	6.09	regulation of phosphorylation	36	2.40E-11
ACDB 8	5.98	regulation of cell size	20	1.60E-07
ACDB 9	5.55	response to oxidative stress	16	6.80E-06
ACDB 10	5.31	regulation of transferase activity	27	1.00E-07
ACDB 11	5.04	regulation of protein kinase B signalling cascade	8	2.00E-07
ACDB 12	5.02	cell fraction	45	8.00E-06
ACDB 13	4.78	regulation of lipid metabolic process	22	2.00E-14
ACDB 14	4.59	mTOR signalling pathway	19	3.10E-14
ACDB 15	4.53	behavior	26	3.60E-05

*DAVID reports FDR as percentage, therefore, the above $FDR = FDR_{\text{DAVID}}/100$.

n: Number of genes linked to the current term.

ACDB: Annotation Clusters with Default DAVID tools Background.

Table 3.2: Annotation clusters with the LongvityMap background

Annotation Cluster	Enrichment Score	the most representative term	n	FDR*
ACLB 1	18.23	membrane-enclosed lumen	49	2.80E-21
ACLB 2	16.11	regulation of cell death	64	4.50E-25
ACLB 3	14.8	cell fraction	45	1.90E-18
ACLB 4	11.8	non-membrane-bounded organelle	48	1.00E-13
ACLB 5	10.02	protein dimerization activity	29	3.90E-12
ACLB 6	9.29	regulation of cellular protein metabolic process	36	2.20E-14
ACLB 7	8.79	regulation of locomotion	23	1.30E-10
ACLB 8	8.56	macromolecular complex subunit organization	31	3.40E-13
ACLB 9	7.86	cation binding	78	1.10E-12
ACLB 10	7.64	nucleus	72	1.30E-19
ACLB 11	7.42	response to organic substance	45	4.30E-13
ACLB 12	7.39	positive regulation of molecular function	35	7.60E-13
ACLB 13	7.32	plasma membrane	88	3.90E-26
ACLB 14	7.26	cell death	30	3.80E-09
ACLB 15	7.14	regulation of cellular component size	22	4.60E-10

*DAVID reports FDR as percentage, therefore, the above $FDR = FDR_{\text{DAVID}}/100$.

n: Number of genes linked to the current term. ACDB: Annotation Clusters with LongevityMap genes as Background.

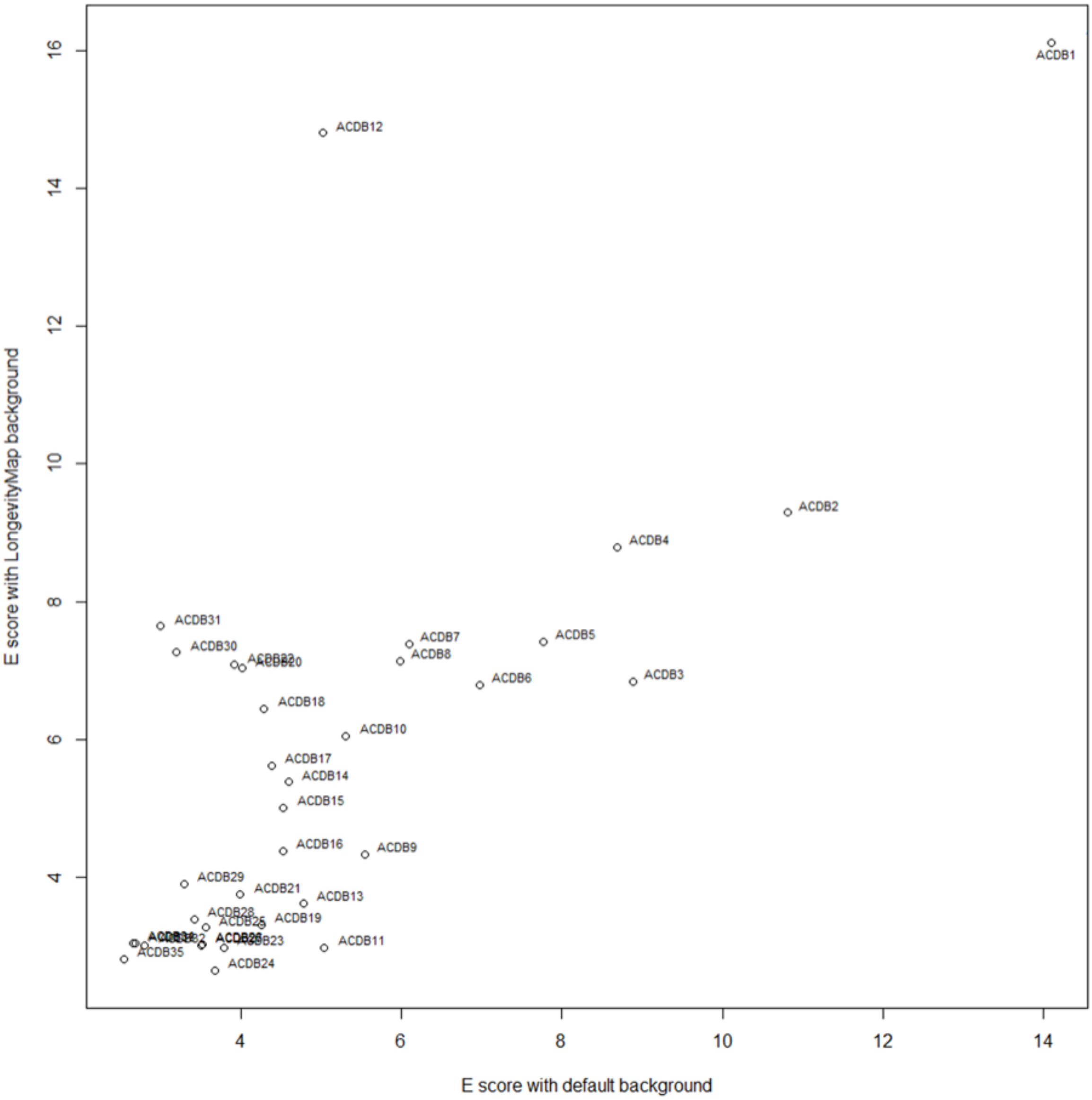


Figure 3.1: **Top ranked functional annotation clusters between two backgrounds.** Higher ranked cluster obtained from default background also ranked higher when using LongevityMap genes as background.

This may reflect how researchers choose candidate genes for longevity association studies. In CGASs, researchers tended to select candidate genes that play important roles in human lifespan, or in severe pathology processes that can significantly impair lifespan for their studies. Therefore, when longevity-associated genes were clustered by the functional annotation, these genes were enriched in the key pathology processes or mechanisms. Similar functional clustering pattern was also observed when using the DAVID tools to analyse different ethnic subgroups, like Americans (data not shown).

3.3.2 Pathway analysis

Pathway analysis was performed in the *Cytoscape Reactome FI plugin*. *Cytoscape Reactome FI plugin* calculates p values and FDR by binomial test and Benjamini-Hochberg method, respectively. The results showed many enriched pathways from the input longevity-associated genes. After filtering the results with $FDR < 0.001$, 161 clusters were left. Several longevity-associated pathways, such as mTOR pathway and Insulin Signalling pathway were listed in the results (Table 3.3).

Table 3.3: Top 15 enriched pathways detected by *Reactome FI plugin*

Pathway	Number of hit genes	p-values	FDR
AMPK signalling pathway(K*)	27	1.11E-16	1.02E-14
Generic Transcription Pathway(R)	40	1.11E-16	1.02E-14
HIF-1 signalling pathway(K)	22	1.11E-16	1.02E-14
Longevity regulating pathway(K)	31	1.11E-16	1.02E-14
Longevity regulating pathway - multiple species(K)	18	1.11E-16	1.02E-14
mTOR signalling pathway(K)	34	1.11E-16	1.02E-14
mTOR signalling pathway(N)	31	1.11E-16	1.02E-14
PI3K-Akt signalling pathway(K)	38	1.11E-16	1.02E-14
signalling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)(R)	27	4.89E-15	4.01E-13
Proteoglycans in cancer(K)	24	2.19E-14	1.62E-12
signalling by Insulin receptor(R)	27	3.46E-14	2.32E-12
FoxO signalling pathway(K)	20	4.69E-14	2.86E-12
Macroautophagy(R)	15	5.88E-14	3.35E-12
Insulin signalling pathway(K)	20	1.04E-13	5.51E-12
PIP3 activates AKT signalling(R)	18	2.36E-13	1.16E-11
EGFR tyrosine kinase inhibitor resistance(K)	16	2.87E-13	1.32E-11

*Indicate data source: C - CellMap, R - Reactome, K - KEGG, N - NCI PID, P - Panther, and B - BioCarta

Network Cluster function in the Cytoscape uses spectral partition based network clustering algorithm to detect community structure in the networks (Newman, 2006). In our case, the network clustering analysis is helpful in revealing modules in the longevity genes network. Eight Network Clusters with modularity of 0.492 were obtained after applying the *Cluster FI Network* function (Figure 3.2). Modularity is the over-represented fraction of edges (connections) observed within clusters compared to the expected fraction if the edges are randomly distributed. It was designed to describe the structure of networks. Networks with high modularity indicate nodes that are more densely connected together within modules than to the rest of the network (Subelj et al., 2011). The value of modularity ranges from -0.5 to 1 (Brandes et al., 2008). Higher modularity represents stronger connections within network modules than between network modules. Unsurprisingly, the clusters were connected through several hub genes, like *TP53*, *NFKB1*, *MTOR*.

High inner-cluster connections were observed in a cluster consisted of 30 genes (coloured in purple in Figure 3.2). Further investigating this cluster showed its main components were from *mTOR signalling pathway(K/N)* (Figure 3.3, Figure 3.4) and *Longevity regulating pathway - multiple species(K)* (Figure 3.5). High overlapped genes between the two pathways also supports the hypothesis that *mTOR pathway* plays an important role in regulating ageing process and longevity. *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* were clustered together with loose connection in the left lower corner. Two separate clusters, one consisted of *LMNA*, *SYNE1* and *POT1*, the other consisted of *TRIM25* and *NLRC5*, were isolated from the rest of gene network.

The cluster *LMNA*, *SYNE1* and *POT1* mainly contributes to the pathway of regulation of telomerases, apoptosis, which are important to the ageing (see section 1.2.3).

HLA-DQA1, *HLA-DQB1* and *HLA-DRB1* genes belong to the human leukocyte antigen(HLA) family. The protein complex, formed by binding protein products of *HLA-DQA1*, *HLA-DQB1* together, in vital to trigger immune response by present foreign peptides to the immune system. The product of *HLA-DRB1* also plays similar roles when binding to another protein produced by *HLA-DRA* gene. They play a critical role in human immune system response.

One of the most interesting facts wa1s the most enriched pathway from *TRIM25* and *NLRC5* cluster. The most enriched pathway is *Influenza A* pathway from KEGG (`http://www.genome.jp/kegg/pathway/hsa/hsa05164.html`). This cluster, together with the *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* cluster, indicate evidence of longevity/ageing is also affected by environmental factors.

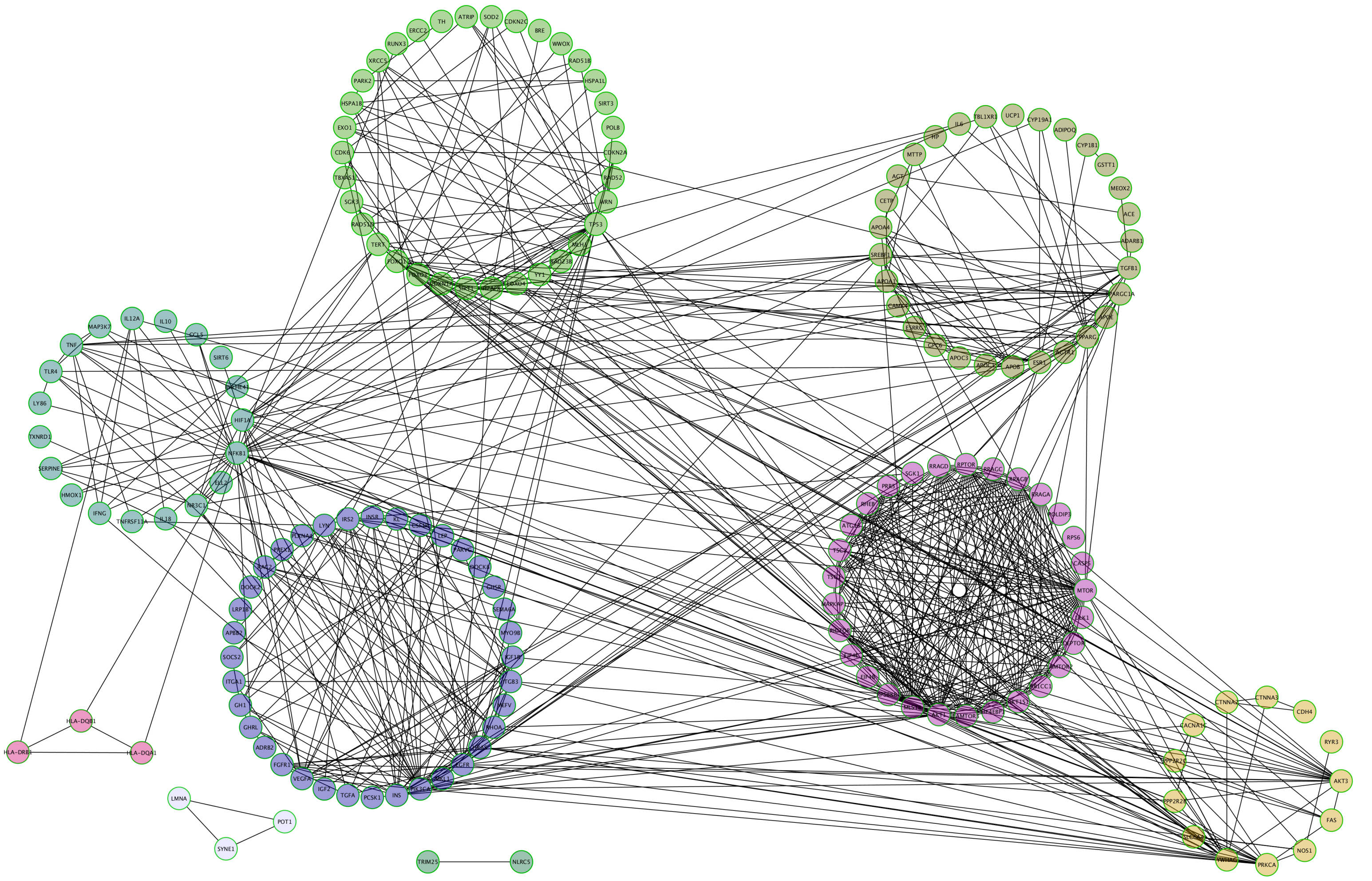


Figure 3.2: **Network clustering analysis use Reactome FI plugin in Cytoscape.** Modularity:0.492, generated by Reactome FI plugin in Cytoscape 3.61. The purple cluster in the right middle indicates the *mTOR-Longevity regulating* cluster.

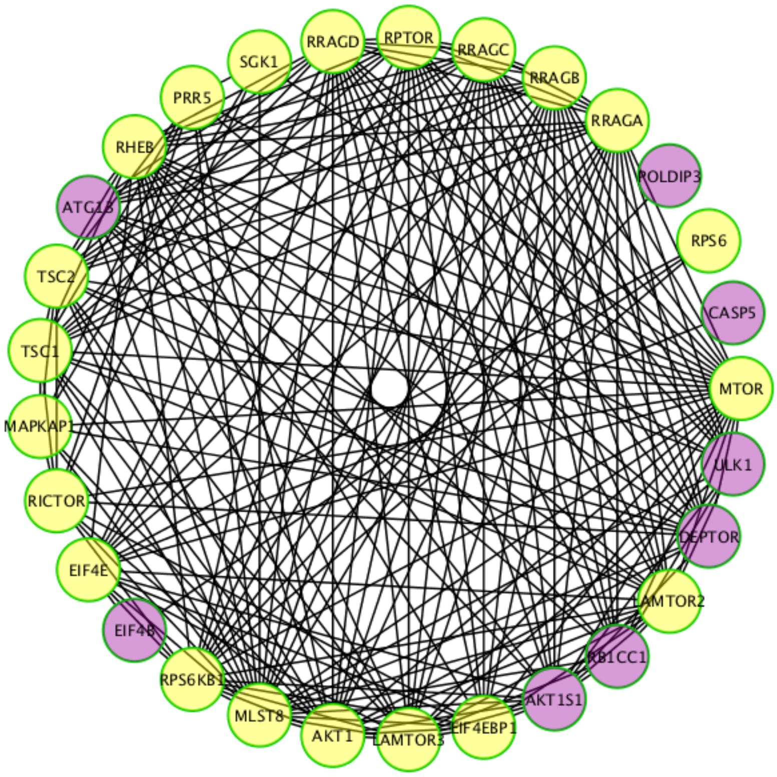


Figure 3.3: **mTOR pathway genes in the *mTOR-Longevity regulating* cluster.** Genes belong to mTOR pathway were labelled with yellow backgrounds. Figure was generated by *Reactome FI plugin* in Cytoscape 3.61.

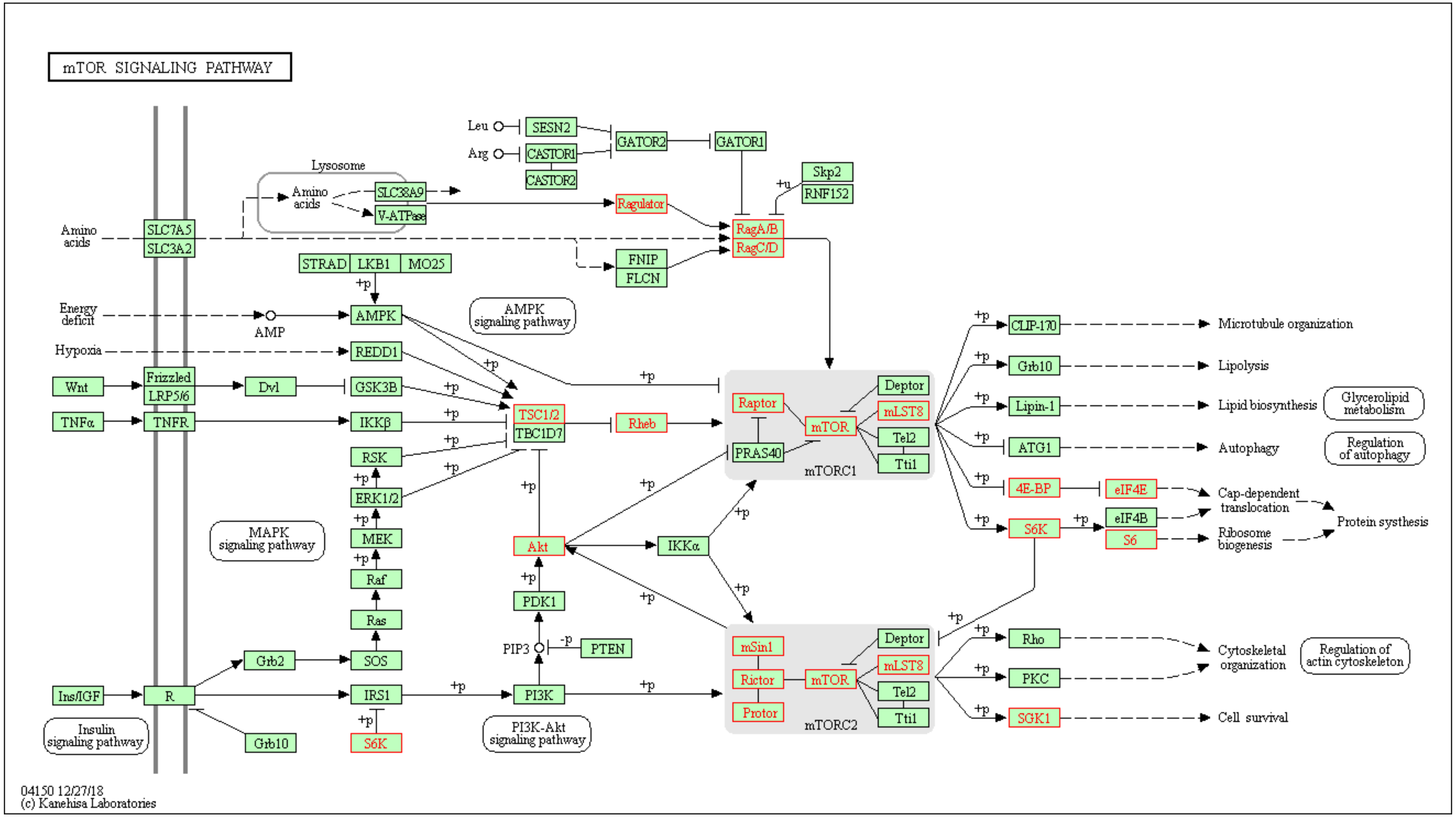


Figure 3.4: **Identified mTOR cluster genes in the mTOR signalling pathway.** The identified genes are highlighted in red colour in the above figure. The figure was rendered by *NCBI BioSystems* mTOR signalling pathway (KEGG: hsa04150) in *Homo sapiens* (`https://www.ncbi.nlm.nih.gov/biosystems/83059`).

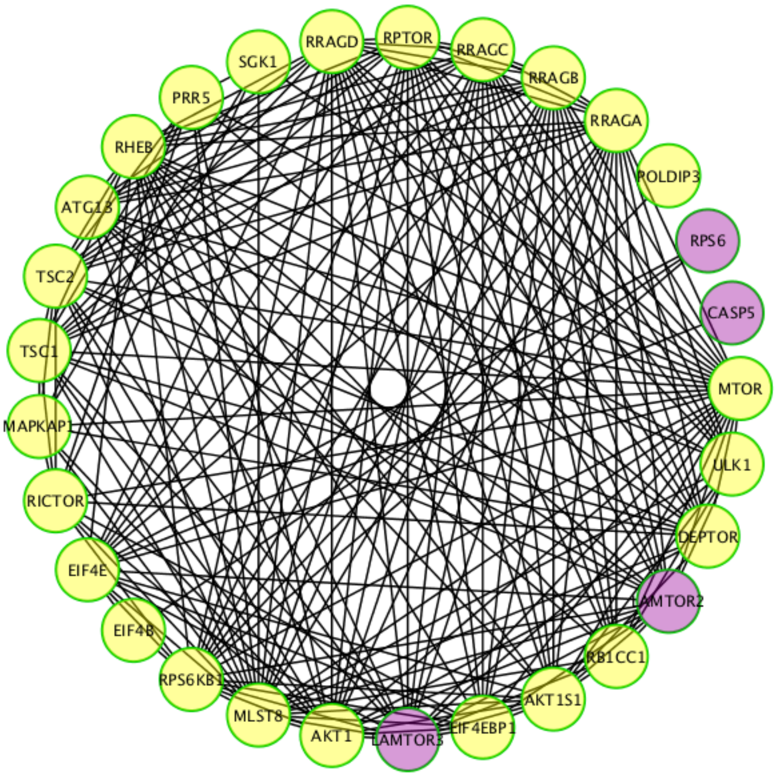


Figure 3.5: **Longevity regulating pathway genes in the Cluster.** Genes belong to Longevity regulating pathway were labelled with yellow backgrounds. The figure was generated by *Reactome FI plugin* in Cytoscape 3.61. The corresponding figure that showing the positions of those genes in KEGG pathways (like Figure 3.4) was not shown due to the identified *longevity regulating pathway - multiple species (K)* involves multiple species. It is difficult to show multiple species in a single figure.

3.4 Discussion

The huge number of functional enrichment clusters suggests that the factors that manipulating ageing and ageing process are quite disperse, which verified the common sense that ageing is a complex trait (Johnson, 2002; Johnson et al., 2013) and controlled by many internal factors as well as external factors such as environment (Caselli et al., 2006).

Given the potential study biases existing in the LongevityMap database (discussed in Chapter 4), the actual biological factors involved in ageing could be more than what has been shown here. Even in the current small set of longevity genes that has been repeatedly studied, there are plenty of modestly enriched clusters. If more genes/loci are integrated in the future, for instance, in the scope of all human genes, the results could be much richer.

Moderate overlap was observed between known longevity-associated genes from human-based studies and other model organism-based studies (Fernandes et al., 2016). This inconsistency reflects not only the difference between human and other organisms, but also the difference in experimental designs. Researchers may prefer to play on “safe-side” when carrying out human-based studies by verifying an identified longevity-associated locus in another population rather than looking for new longevity-associated loci. It should not be blamed because human population is quite mixed. However, in the model organism based studies, researchers showed the “courage” of exploring new continent in the genome, maybe due to they have more control of experimental factors and less variation exists in model organisms.

Investigation of contributions from small enriched clusters is also necessary. Even some of the most studied pathways have been revealed in the current analysis, the biological mechanisms of the majority modest enriched pathways still yet to be discovered. As survival to extremely old age is rare, it is likely the effect is due to contributions from rare alleles or private pathways (Brooks-Wilson, 2013). Further investigation of those middle or lower enriched pathways could provide new insight in ageing research.

Other data could help to identify the biological basis of ageing. For example, RNAseq data from long-lived cohorts can be integrated into the current analysis to see if any expression changes contribute to living into old age, therefore help in locating the key mechanisms in the ageing process.

Finally, network analysis is based on the current knowledge in the field. It provides a new way to combine existing data and suggesting the hypothesis. However, it lacks of mechanistic explanations (Cho et al., 2012). With biological data accumulation in the field, results shown here could change in the future when more data is available. Results shown here only represents the current understanding to ageing research. The predicted pathways are reflections of where the true contribution to ageing could come from. Over time, some of the predictions could be supported or rejected by future studies. Each support or rejection will help in shaping the correct understanding of ageing. Eventually, the key to ageing will be deciphered.

Chapter 4

Publication Biases investigation in Longevity Association Studies

4.1 Introduction

Publication bias is *“the tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings”* (Dickersin, 1990). It is a phenomenon that researchers tend to submit, and publishers tend to publish significant/positive results over non-significant/negative results. Every year, research papers reporting significant outcomes are overwhelmingly published in academic journals comparing to non-significant/negative results (Fanelli, 2012).

Many reasons could contribute to this fact. One of the most common reasons is the preference comes from readers. It is likely those “breakthrough” researches attract more attention than studies showing nothing has been found. Researchers are unlikely to spend time on negative results papers, although the negative results are equally important as positive ones (Matosin et al., 2014). The unequal reporting of significant versus negative results has been noticed by the community. Opinion and discussions on how to improve publication biases have been given on this topic (Dickersin, 1990; Easterbrook et al., 1991; Ioannidis et al., 2007).

Nevertheless, the fact is we are still in lack of effective ways to detect publication biases. Those potentially biased results could bring illusions and make research less objective if we do not have a clear view of this issue. Thus, a process of assessing data quality in the LongevityMap by detecting publication biases was described in the following sections.

4.1.1 Statistical terms

Statistical test provides mechanisms for making quantitative decisions about one or more processes. With the supporting evidence from statistical tests, we can make more reasonable decisions.

Before applying a statistical test, one question should be asked is what decisions we want to make. In other words, what hypothesis we want to accept or reject? The hypothesis is termed as null hypothesis in statistics. Null hypothesis, as indicated by its name, is an assumption that no relationship between two accessed variables. By observing something that is not consistent with the null hypothesis, we reject the null hypothesis and accept the opposite hypothesis, termed as alternative hypothesis, and draw a conclusion.

How confident of the conclusion is indicated by the p values from statistical test. p value, ranges from 0 to 1, is the probability of the null hypothesis is true. In the context of null hypothesis, a p value returned by statistical method indicates the strength of evidence if the null hypothesis is rejected. Smaller p values suggest strong evidence against the null hypothesis, which means the null hypothesis is unlikely. A commonly accepted cut-off of p values is 0.05. If p less than 0.05, then it is normally confident enough to say the null hypothesis is unlikely to happen as there is only 5% chance of observing the observed results plus more extreme results under a given null hypothesis (Goodman, 2008).

No matter how sophisticated an experimental design is or how careful a statistical test was used, there is always a chance that the null hypothesis was wrongly rejected when it is true. Or the opposite, the null hypothesis was accepted when it was actually false. The former one, which wrongly rejected the null hypothesis and lead to false positive findings, is also referred as type I errors. While the latter one is referred as type II errors, which brings false negative findings by wrongly accepted the null hypothesis (Banerjee et al., 2009).

Care should be taken when testing the same null hypothesis with a set of different statistical inferences. As each statistical inferences has a potential to bring evidence to reject the null hypothesis, type I errors (false positive discoveries) are likely to happen when the number of statistical inference is big enough. In other words, the results will be less reliable due to the inclusion of type I error if no corrections were applied.

To estimate the adverse effect caused by multiple tests, Family-wise Error Rate (FWER) was introduced. FWER is defined as the probability that at least one type I error occurs after a series of statistical tests. In order to reduce the inflation of type I errors, FWER must be controlled. Methods have been developed to correct the inflated false discovery. One of the most used methods is Bonferroni correction. Bonferroni correction simply compare p values from each statistical test with an adjusted cut-off threshold α, where $\alpha = p_{\text{cut-off}}/n_{\text{number of tests}}$. Null hypothesis can only be rejected if $p < \alpha$. Bonferroni correction applied a more stringent standard to control type I errors. Although the Bonferroni method is easy to use and widely accepted, it also has been criticised for reducing the statistical power and increasing type II errors.

4.1.2 *p* values in scientific publications

The concept of p values has been applicable to scientific research field for a very long time. p value, by itself, was born as an indication of how likely errors occur for a given statistical test. It is a continuous vector without any breaking point. There is no compulsory rule of how to discriminate results by p values (Wasserstein et al., 2016). In practice, researchers tend to use a threshold of 0.05, which is suggested by Fisher (Fisher, 1926), to decide whether accepting or rejecting the null hypothesis and therefore conclude whether an outcome is "significant" or not.

The introduction of p value cut-offs brings a convenient way of describing the outcome of an experiment. However, it is flawed. Too much focus has been put on the threshold suggested by Fisher. When a threshold/cut-off is artificially introduced, the attribute of p values was converted from a continuous vector to a dichotomous factor, so changed the meaning. The term "low error probability" can be seen as an indication of "positive outcome" when p value is less than a predefined threshold. The ambiguous boundary of describing the outcome of a study as "the error chance of this study is low" or "the outcome of a study is true" cause non-standard use, sometimes misuse, of these two concepts.

The artificially introduced threshold brought marginal non-significant/negative p values. Two algebraical similar p values could be classified into opposite categorises because of the existence of p thresholds. Some study with p values slightly greater than the threshold will be tagged as "negative outcome study" and, as a consequence, less likely be published in the current p value dominated publish system. However, in the perspective of statistic, the chance of error does not differ very much if the two p values are comparable (Biau et al., 2010).

4.1.3 Impacts of overweighting p values

The prevalence of using p values in academic publication propels the emphasis of p values. Even though there is much debate on the imperfection of p values and other statistical methods such as Bayes statistical or confidence interval has been suggested in the publications, using p values to evaluate research results is still prevalent in academic publications (Nuzzo, 2014).

p values from statistical analysis are important in evaluating the outcome of an experiment. In most circumstances, whether a result is positive or negative (significant or non-significant) is simplified as if the p value has fallen into the significant zone. It is common to see in a published paper that the author described the outcome of an experiment was "significant", then a p value from the statistical test was followed. To some extent, this common format in reporting an experimental outcome promotes the emphasis on the conclusion ("significant" or "non-significant") rather than the statistical meaning. Because positive results are in the favour of being published, p value somewhat determines whether a study is publishable.

The favour of significant p values in the publication system potentially promotes scientists pursuing smaller p value. The smaller the p value is, the stronger the result seems to be (Sterne, 2001). There are a number of ways by which to get artificially low p values. For example, it is common to see a study did not do multiple tests correction even when it was essential (Streiner et al., 2011). Multiple statistical tests increase the chance of getting a smaller p value, therefore, correcting for multiple tests is necessary when multiple statistical tests were involved. When a multiple-tests correction is omitted where necessary, an error occurs. Another common circumstance is the researcher did multiple testing correction, but the corrected result did not get emphasised as expected (Bellavia et al., 1999; Naumova et al., 2004). These descriptions often seen when the uncorrected p value gives a significant outcome while the corrected p value does not. Sometimes the hints come from the uncorrected p values are emphasised and exaggerated. This is clearly wrong because unadjusted p values will inflate type I error rates and therefore should be discarded.

p values are important, sometimes vital, in affecting the possibility of publication. The underneath connections between p values and the chance of being published potentially encourage some researchers to pursuit p values less than the 0.05 threshold in order to increase the chance of being published. The behaviour of obtaining significant p values by manipulating data or dishonestly reported result is termed as p-hacking, which is also the main component of publication biases (Simmons et al., 2011).

Several common tricks have been reported relating to p-hacking. One of the most common tricks is selective including experimental data to obtain a significant result. Researchers selectively include data leads to significant outcomes and/or excludes data leads to non-significant outcomes (Head et al., 2015). Because these operations happen before presenting results, they are not easy to be detected. Another common trick is by manipulating the statistical methods. Several statistical methods were tried, but only the one giving significant result was reported. This conflicts with statistical rules because appropriate statistical method should be decided before carrying out a study, rather than picking up a statistical method that can give significant results after data collection (Nayak et al., 2011).

The favour of significant p values harms scientific research very much. Experiments presenting significant results are in the favour of being published. It potentially encourages researchers to submit the significant results and "file-drawer" non-significant (usually $p > 0.05$) ones (Scargle, 1999). For those published papers, whether they are trustworthy is still under debate. The significant-results-favoured publishing system potentially reduces the reliability and quality of papers (Smith, 2006). It makes the low probability event looks like a high probability event because the negative results did not get the equal chance to be displayed. This could be harmful to scientific research. Firstly, it misleads other researchers by assuming a low probability incidence as a general fact. The referential value is impaired. Any work or any hypothesis deduced from biased reports will have a much lower chance to be successful. Secondly, it makes some of the studies hardly to be replicated in some way. Last but not least, it prevents the truth from being discovered by providing a false illusion (Begley et al., 2012).

Luckily, research community has noticed the publication bias issue in recent years and efforts have been put into to reduce this adverse effect (Colhoun et al., 2003). For example, a journal named as Journal of Negative Results in Biomedicine has been created to publish the negative results specifically (Pfeffer et al., 2002).

4.1.4 GWAS and non-GWAS (CGAS) studies in publications

The differences in experimental design and the scope of an experiment between GWAS and non-GWAS brings many different aspects between one and the other. The typical format in reporting results is a good example.

The GWAS experimental design employs majority human genetic variants in a single study. With the advantage of current chip technology, GWAS screens variants genome-widely for any hit (risk allele) statistically associated with targeting traits, such as longevity. Since the scope of GWAS is the whole human genome and no assumptions were made before experiments, GWAS designs are hypothesis-free, and they are not biased (Frazer et al., 2009). in GWAS designs, researchers do not usually consider if there is any causal relationship between the variants and the traits in advance. It more like observing the results first, then seeking for a proper explanation for the association. Therefore, GWAS is important to identify potential causal alleles by scaling down the scope from genome-wide to several risk alleles.

In the CGAS designs, the selection of candidate gene(s)/variant(s) is knowledge-based. The experiment is used to verify the suspect relationship between target gene/variant and the traits. There is an expectation before performing the experiments. The expectation is core difference between GWAS and CGAS, as there is no hypothesis involved in GWAS designs (see Table 1.1).

The difference between these two types of experiment design brings the difference between these two typical formats in reporting the results. The wide scope of GWAS makes reporting results (p values) of all the variants very redundant. Therefore only positive or marginally positive variants are reported in GWAS outcomes. In contrast, candidate gene based design focuses much fewer variants, which makes it relatively easy to report results for all the variants. As a result, in the candidate gene based design, normally we can find p values of all the variants. Also, as there is an expectation in CGASs, publication biases are more likely to exist. Therefore, in the following estimation of publication bias, only candidate gene based studies were involved.

Although the above tricks are relatively private and implicit, they are not undetectable. It may be not easy to pinpoint whether a single study has been manipulated by any of the above tricks. However, if several similar studies under a same topic were gathered as a group, methods such as meta-analysis and *p-curve application* are available to detect the trace of publication bias (Macaskill et al., 2001; Peters, 2006; Simonsohn et al., 2014).

4.2 Inspiration for this project

As a central repository of longevity genetic association studies, the LongevityMap database is an up-to-date collection of longevity genetic association studies. Since publication bias has deleterious effects on scientific research, it is necessary to evaluate publication biases in the LongevityMap database before carrying out any further analysis. By integrating the information of publication biases into the projects, we can deliver more precise results with more objective and reasonable explanations. In this project, two categories of publication bias, p-hacking and file-drawer effect, are being examined in the CGASs in the LongevityMap.

4.3 Methods

4.3.1 *p* value selection criteria

The design and aims differences of an experiment could affect how results are reported. In global view of the LongevityMap, there were three types of variables being studied: alleles, genotypes and haplotypes. Allele-based study design tests single locus at a SNP position in the chromosome, like rs107251 in *SIRT6* gene (Soerensen et al., 2013). Genotype-based study design only seen in early literature, when SNP information was incomplete, researchers have to report the nucleotides on both strands at a position to refer the variation being examined. For instance, -438 A/A in *TAFI* gene (Reiner et al., 2005). Haplotype-based study tests the combinations of several alleles, it usually represents the cumulative effects of a group of small effect SNPs. For example, rs405509, rs440446 and rs769449 in *APOE* gene (Soerensen et al., 2013).

According to the subjects being studied (allele, genotype or haplotype), different assumptions on which genetic models were followed were made and statistical methods were used under the assumptions. For example, logistic regression and linear regression tests were used when additive allelic effects were assumed. While in a Chi-squared test or ANOVA test, an assumption of alleles being studied are functioning independently is usually made. Therefore, the alleles were treated as categorical variables.

Due to the heterogeneity of statistical methods and assumptions that authors have made in those studies, and the aim of this chapter is examining publication biases in CGASs, we are not going to evaluate the validity of statistical tests used in the article or make judgements on the correctness of assumptions in different genetic models. We only want to test if the authors are honestly reported their results without any modification or not. In this aim, p values reported in each individual study were manually extracted following the rules below:

- Discard studies do not meet the *Hardy–Weinberg Equilibrium* (HWE).
- If p values exist in the article, we take the p values.
- Where states as "significant" or "non-significant", but no p values can be found in the article, we excluded those studies as it was impossible to recover p values from articles.
- If p values for both of alleles and haplotype exist, take the p values for allele over the ones for haplotype.
- Take p values of genotype over haplotype.
- Gender specific p values were collected where exist.

For CGASs, the default cut-off is 0.05. The p value cut-offs in each study were also examined. Only one publication (PMID: 15621215) used an adjusted p value (Bonferroni corrected: $p_{cut-off} = 0.05/11$ in the article) as cut-off instead of 0.05. Therefore, the uncorrected p value was back-computed ($p_{uncorrected} = p \times 11$) in order to matches cut-offs from other studies. After these steps, a list of studies with corresponding reported p values was obtained.

4.3.2 Skewness of p values distribution by plotting

In statistics, the distribution of p values from a set of studies are tightly connected with the Study Effects (SEs). If there was no SE, each p value shares the same probability of being observed, which made p values following uniform distribution between 0 and 1. This means the probability of a p value falling between 0.01 and 0.02 should be identical as it falling between 0.04 and 0.05 (red dashed line in Figure 4.1). If SE does exist, the probability of obtaining a small p value is much higher than obtaining a modest or a large p value. It means the p value has more chance falling between 0.01 and 0.02 than falling between 0.04 and 0.05. Therefore, the distribution of p values from a series of studies should display a right-skewed pattern in the whole range between 0 and 1 (blue dashed line in Figure 4.1) (Masicampo et al., 2012; Simonsohn et al., 2014; Wallis, 1942). If the number of studies is big enough and all the p values were honestly reported, the number of p values in each 0.05-scale should gradually decrease from 0 to 1 (as indicated by the blue dashed line in Figure 4.1).

Since $p = 0.05$ is the most commonly used cut-off in longevity genetic association studies, if p values from "p-hacked" studies were included in the significant result studies, the existence of hacked p values would alter the skewness of the distribution. By detecting the sign of this alternation, we can trace the evidence of publication bias (green solid line in Figure 4.1).

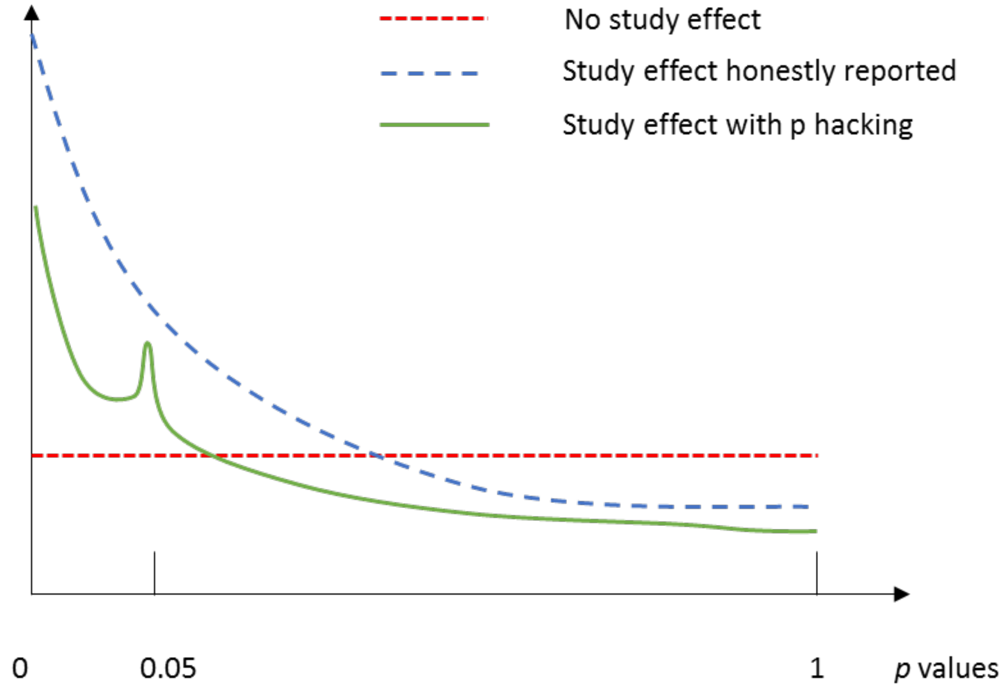


Figure 4.1: **Study effects, publication bias and *p* values distribution.** Red dashed line demonstrates uniform distribution, where no study effect exists. Blue dashed line demonstrates right-skewed distribution where study effect exists and results were honestly reported. Green solid line demonstrates altered right-skewed distribution, where study effect and publication biases co-exist.

4.3.3 *p-curve* analysis

To statistically test whether publication bias exists in the LongevityMap, a further investigation on the positively reported studies and the corresponding significant p values were carried out by the *p-curve application*.

p-curve application is an online tool implemented by Simonsohn et.al. in 2014 (Simonsohn et al., 2014). It assesses the reliability of published research by evaluating the distribution of p values. Since it checks the existence of evidential value, it only focuses on the significant studies that were reported with p values smaller than 0.05 (Simmons et al., 2011; Simonsohn et al., 2014).

The *p-curve application* not only asks for p values, it also needs the statistical method and the relevant parameters, such as the degree of freedom and the result of a Chi-Squared test. In this case, all the positively reported papers in the LongevityMap have been re-reviewed to retrieve raw statistical data. Surprisingly only approximately 29% in the positive studies reported raw data (Table 4.1).

A set of 183 formatted raw statistical data entries (one LongevityMap entry may contain several loci. For each locus, there could be one raw statistical data entry) was fed into *p-curve app online tool* (V3.01, `http://www.p-curve.com/app3/pcurve3.php`) for analysis.

Table 4.1: Summary of significant/non-significant entries and the number of raw data reported entries

Entries	Significant entries	Non-significant entries	Total
Total	177	232	409
Raw data reported	51	40	91
Percentage	$\sim$ 29%	$\sim$ 17%	$\sim$ 22%

4.3.4 *D'Agostino skewness test*

As *p-curve analysis* can only analysis *p* values with raw statistical data, *D'Agostino skewness test* in *RStudio Statistics Package* (RStudio Team, 2015) were used to test the skewness of *p* values from studies did not report raw statistical data.

4.4 Statistical hypotheses

In the current project, the null hypothesis(H_0) is defined as no publication biases in the LongevityMap. The alternative hypothesis(H_1) is publication biases exist in the LongevityMap. If analyses from *p-curve* and *D'Agostino skewness test* detect any evidence of publication biases, then we can reject H_0 and accept H_1. Otherwise, we will accept H_0.

To be more specific, plotting of all the available *p* values can visualise the distribution of *p* values. If no publication biases exists in the LongevityMap, the *p* plotting will show a smooth right-skewed distribution(blue dashed line in Figure 4.1). If publication biases exists, the plotting of *p* values will display discontinuity around $p = 0.05$. The discontinuity of *p* value can attributes to "*p*-hacking", "file-drawer effect" or both.

p-curve Application only tests significant *p* values. If the missing marginal *p* values were due to *p*-hacking, it will be detected by *p-curve analysis.* If the missing marginal *p* values disappears and no evidence of *p*-hacking were observed, it is likely due to "file-drawer effect". *D'Agostino skewness test* was used as an supplement of *p-curve analysis* to test the skewness of *p* values did not report raw statistical information.

4.5 Results

4.5.1 Sample size effects with *p* values

Undoubtedly, the sample size of a study has impacts on the results. Larger sample size increases statistical power, which allows the low study effects factors can be detected. Given the same study effects, increasing sample size also increases the reliability of study by decreasing the sampling bias. Therefore, generally speaking, studies with larger sample size are more reliable compared to the ones with fewer samples involved. In this context, the sample sizes of CGASs entry was also extracted.

In the scope of statistical sampling, where assume samples are a subset of individuals that can partially reveal properties of the whole population, larger sample size can better inherit the properties of the population. To see if there is any relationship between reported p values and sample sizes, a scatterplot of p values against sample sizes in each CGAS (non-GWAS) entry was generated by R statistics package (`http://www.R-project.org/`). No clear relationship between sample sizes and p values was revealed (Figure 4.2).

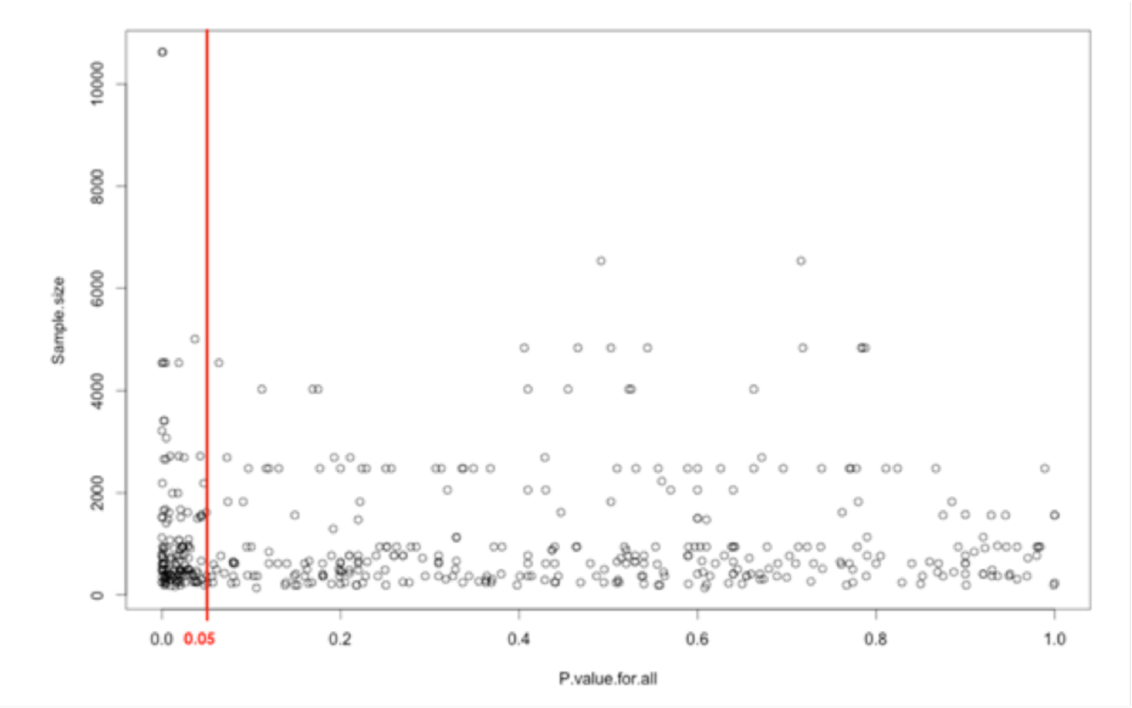


Figure 4.2: **Scatterplot of p values against sample sizes for CGASs.** The red vertical line indicates the widely-used significant threshold of $p = 0.05$.

4.5.2 Evidence from p values plotting

As discussed before, p values are important in affecting the outcome of a study. The distribution of p values could reveal some hints in the existence of publication bias (see Figure 4.3). Therefore, a histogram plot of p values from CGASs could provide the most direct impression of the p value distribution. Overwhelming p values were observed in the commonly defined significant range ($p < 0.05$) when author reported p values from all the CGASs were plotted (Figure 4.4). The extremely high number of p values in the most right-side bin ($0.99 < p \leq 1.00$) was due to a single study reported many p values of 1.00. Other bins, including the bin ($0.00 < p \leq 0.01$), were consist of p values from multiple studies with small number of p values reported. Therefore, in the following analysis, the most right-side bin ($0.99 < p \leq 1.00$) was considered as an outlier and will not be discussed.

The distribution of p values showed a drop between the two bins around $p = 0.05$ (Red solid line in Figure 4.4). By comparing the distribution of p values with the theoretical p value distribution (Figure 4.3), it is likely the p values could be influenced, as it showed signs of chance like Figure D in the Figure 4.3. The irregularly distributed p values around 0.05 could indicate the existence of publication bias. However, it is difficult to tell whether it was due to "file-drawer" or "p-hacking". For example, if some researchers "file-drawered" their non-significant results ($0.05 < p < 1$). The base level of p values in the non-significant area ($p > 0.05$) will be decreased. This will cause the discontinuity of p frequencies around $p = 0.05$ (Figure D in Figure 4.3).

If enough non-significant p values ($0.05 < p < 1$) were "hacked" into significant area, the distribution of p values will show an altered distribution pattern (like Figure B in Figure 4.3) This was based on the assumption of most p-hacking activity stops after obtaining a p value just below the significant threshold (Simonsohn et al., 2014).

The modest tense of p-hacking can be detected by observing the over-accumulation of positive p values near $p = 0.05$. This accumulation of marginal significant results could lead to a left-skewed distribution of p values in the significant area (Masicampo et al., 2012; Simonsohn et al., 2014). Even in the current p value distribution plot did not show sign of increased p frequency in the bin of $0.04 < p \leq 0.05$, it is still impossible to make conclusion of no "p-hacking" because the shape of histogram changes with the bin-sizes. To find out which reason is more likely, other more sophisticated methods, *p-curve Application* and *D'Agostino Skewness Test*, were used to detect publication bias.

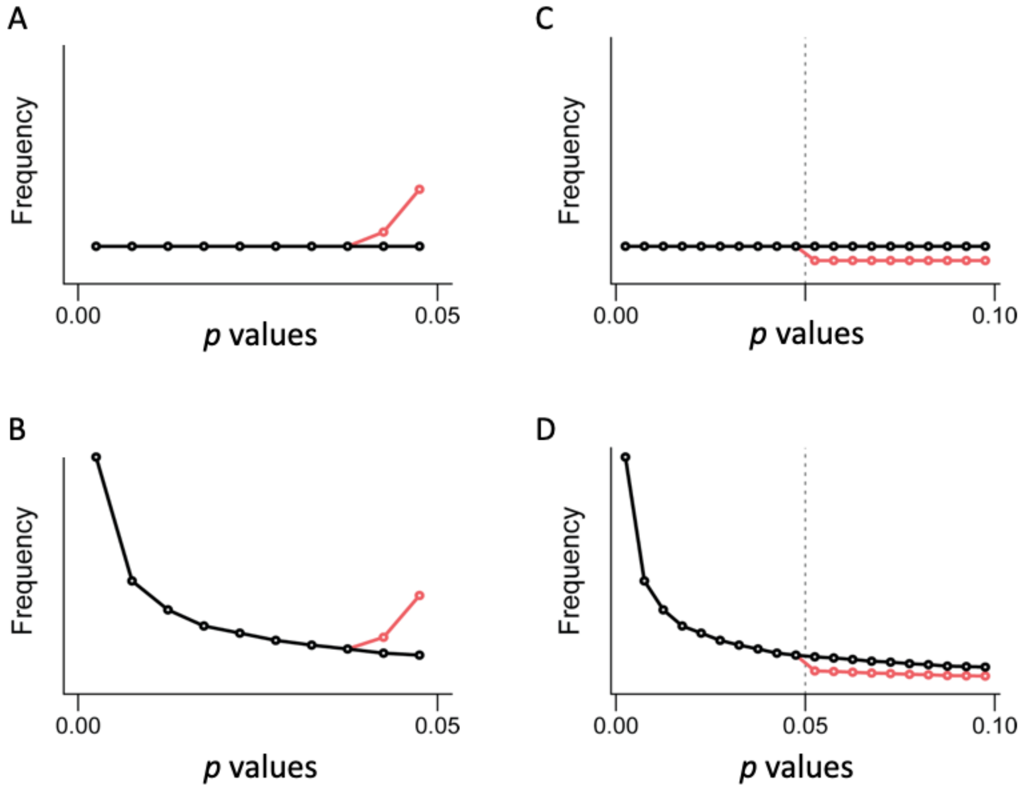


Figure 4.3: **The effect of file-drawer and *p* hacking on the distribution of *p* values around the significance threshold of 0.05.** A) The black line shows the distribution of *p* values when there is no study effect and no *p* hacking, the red line shows how *p* hacking influences the distribution. B) The black line shows the distribution of *p* values when there are evidential value and the red line shows how *p* hacking influences this distribution. C) The black line shows the distribution of *p* values when there is no study effect and no file-drawer. Red line shows the effect of file-drawer on the distribution of *p* values. D) The black line shows the distribution of *p* values when there is no file-drawer. Red line shows the distribution of *p* values influenced by file-drawer effect. Taken from (Head et al., 2015)

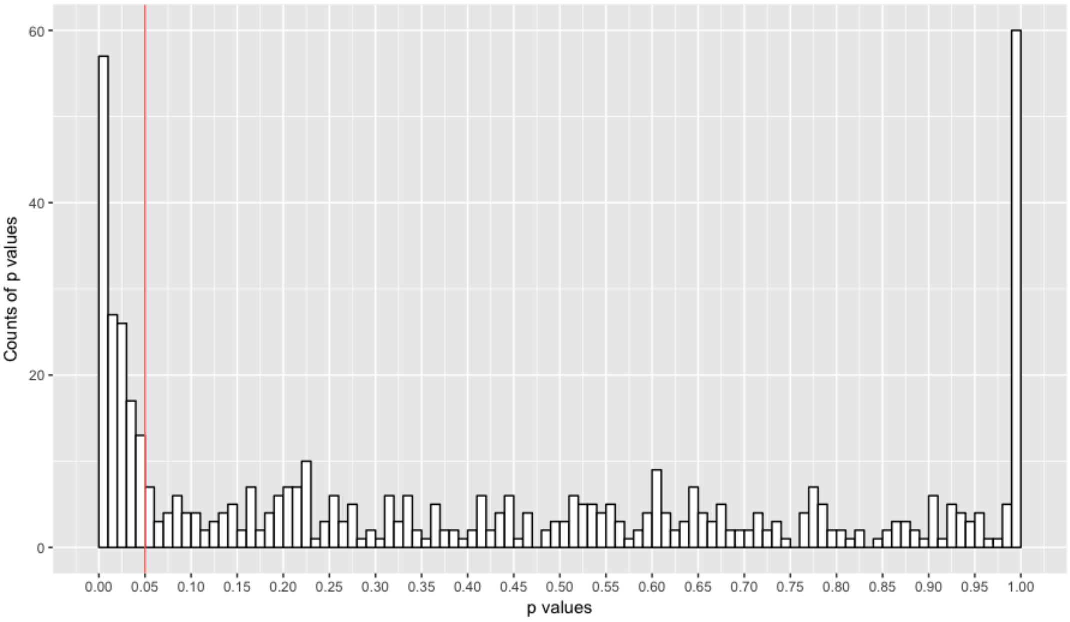


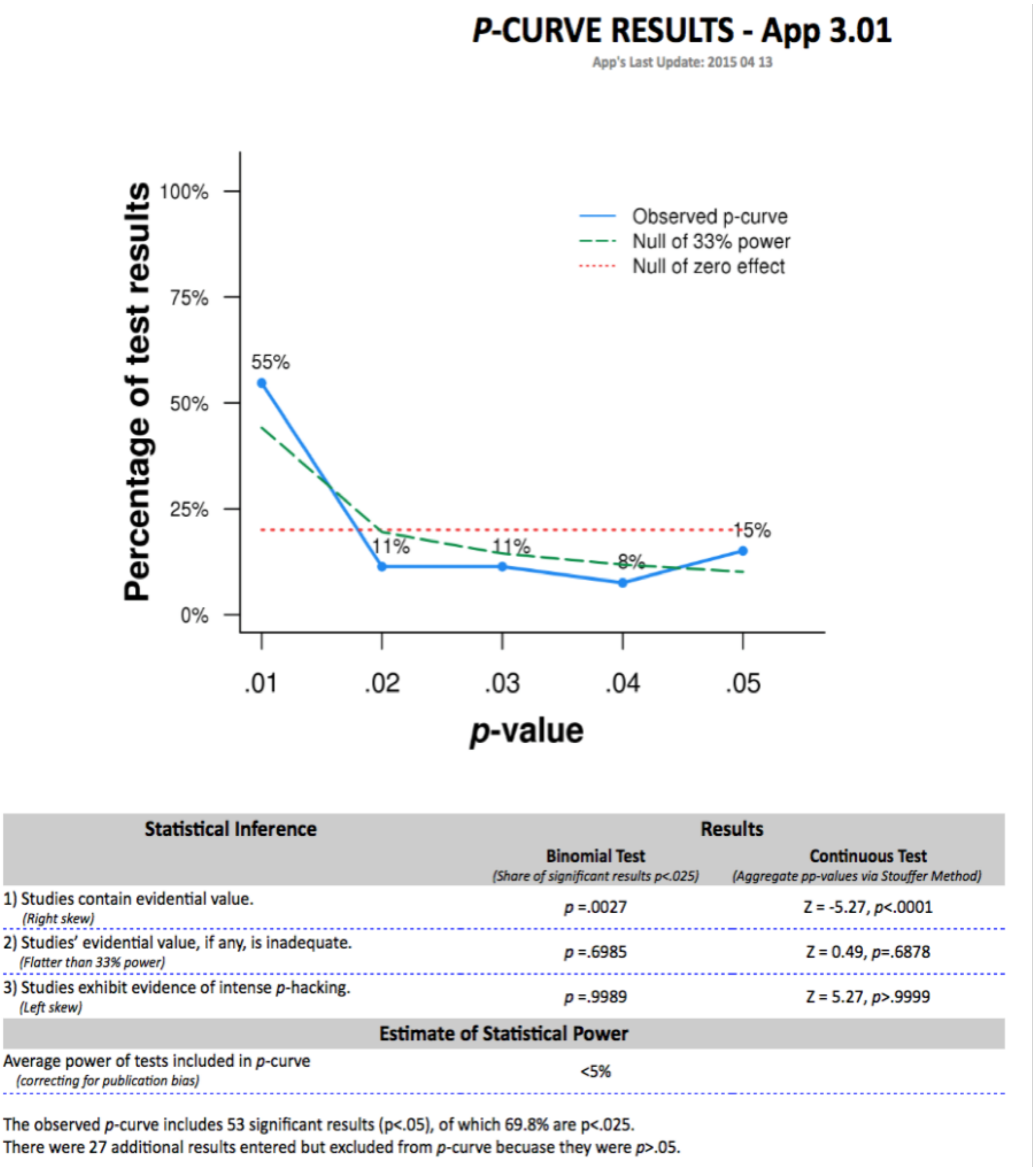
Figure 4.4: **Distribution of p values from CGASs in the LongevityMap (bin size = 0.01).** The red vertical line indicates the significance threshold of 0.05. The visible drop of p value counts between the two adjacent bins across 0.05 (bins of $0.04 < p \leq 0.05$ and $0.05 \leq 0.06$) indicates publication bias could exist (see Figure 4.3). The very high counts of p values in the bin of $0.99 < p \leq 1.00$ was due to a single study reported many p of 1.00. As p values in other bins were from multiple studies, the study reported many p value of 1.00 can be considered as an outlier.

4.5.3 *p*-curve analysis

Results were displayed below (Figure 4.5). The blue solid line in the figure demonstrates the observed p-curve, which is from entries that with raw statistical data provided. The green dash line demonstrates the expected p-curve's position if the studies have 1/3 power. The red dotted line demonstrates the expected p-curve position when there was no study effect (in this case, the p values distribute uniformly). The binomial test here compares the observed proportion of significant results that are $p < 0.025$ to the expected proportions of 33% or no study effect.

The Continuous Test, on the other hand, computes the p value of each p value (pp-value) and converts them to Z scores (Simonsohn et al., 2014). Then the obtained Z scores were added together before being divided by the square root of the number of inputted tests. After this, the Z score reported in the p-curve results is obtained.

In this case, the binomial test results showed those p values contain evidential values. The curve is right skewed ($p = 0.0027$) and they are unlikely to contain any inadequate evidential value ($p = 0.6985$) or to be p-hacked ($p = 0.9989$). The Continuous Test verified these results with Z scores by different p values (Figure 4.5). It is worth to mention that the p-curve analysis results gave an estimated statistical power of less than 5%. This was probably due to very limited number of p values were involved to run the analysis.



P-CURVE RESULTS - App 3.01

App's Last Update: 2015 04 13

Statistical Inference	Results	
	Binomial Test *(Share of significant results p<.025)*	**Continuous Test** *(Aggregate pp-values via Stouffer Method)*
1) Studies contain evidential value. *(Right skew)*	*p* =.0027	Z = -5.27, *p*<.0001
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p* =.6985	Z = 0.49, *p*=.6878
3) Studies exhibit evidence of intense *p*-hacking. *(Left skew)*	*p* =.9989	Z = 5.27, *p*>.9999
Estimate of Statistical Power		
Average power of tests included in *p*-curve *(correcting for publication bias)*	<5%	

The observed *p*-curve includes 53 significant results (p<.05), of which 69.8% are p<.025.
There were 27 additional results entered but excluded from *p*-curve becuase they were *p*>.05.

Figure 4.5: **Results of *p-curve* analysis.** No publication bias was observed. Both Binomial Test and Continuous Test gave $p < 0.05$. However, the statistical power was very low ($< 5\%$). The blue solid line in the figure demonstrates the observed p-curve, from entries that with raw statistical data provided. The green dash line demonstrates the expected p-curve's position if the studies have $1/3$ power. The red dotted line demonstrates the expected p-curve position when there was no study effect. The figure was generated by *p-curve application v3.01* (`http://www.p-curve.com/app3/pcurve3.php`).

Statistical power is inversely correlated with type II error (false negative findings) rates. It is determined by sample size and the strength of study effect. Same statistical power can be achieved by bigger sample size with smaller study effect or vice versa. Lower statistical power means higher probability of making type II errors, which results the null hypothesis is less likely to be rejected. Given the statistical power is low ($< 5\%$) and p values supported the existence of publication biases, we cannot make a conclusion unless the type I error rate has been calculated. Therefore, no conclusion regards to the publication bias can be made at the moment based on the *p-curve* analysis results.

Apolipoprotein E (*APOE*) gene is a popular gene in longevity research, and it is the only gene that has been replicated by large-scale in longevity genetic association studies among different cohorts (Deelen et al., 2014). In LongevityMap there were many *APOE* related significant results, which could potentially contribute to the excessive significant p values in the significant area. I did another p-curve analysis of 44 genes (*APOE* gene was excluded). The new results were similar to the previous ones (Binomial Test $p = 0.0012$ for right skewness, $p = 0.8819$ for inadequate evidence value and $p = 0.9996$ for p-hacking, statistical power $< 5\%$. Figure not shown). Again, this low statistical power prevented a conclusion of no publication bias from being made.

4.5.4 *D'Agostino skewness test*

D'Agostino skewness test was used to test the skewness of all the p values. The null hypothesis here was defined as the data does not have skewness, and the alternative hypothesis is the data has skewness. Firstly, the skewness of p values from p-curve studies (significant studies with raw statistical results reported, e.g. $\chi^2(1) = 8.54$, $p = 0.003$) was tested. The results showed the skewness was right-skewed (*skew* $= 3.5885$, $z = 6.5559$, $p = 5.532e\text{-}11$). This result was similar to the p-curve analysis. Secondly, the skewness of p values from all significant studies were tested. This includes all studies that claimed to have a significant result, no matter the raw statistical data has been reported or not. The results showed the data was not skewed (*skew = 0.595, z = 1.851, p = 0.064*). It is worth to mention that a widely-used cut-off of $p = 0.05$ in the current publishing system was used when deducing the above conclusion. As discussed before, the introduction of cut-offs could bring some problems. However, even we do not view the result as either significant or not by imposing a threshold on the p values, the skewness was still much weaker than the results came from p-curve analysis. In short, the skewness of p values was much weaker in either case.

A summary of how p-curve analysis and skewness test were carried out was demonstrated in Figure 4.6. The results from p-curve analysis were consistent with the results from *D'Agostino skewness test* on studies with raw statistical data reported. However, with the inclusion of studies that without raw statistical data reported, the skewness of the p values distribution was changed from right-skewed to left-skewed.

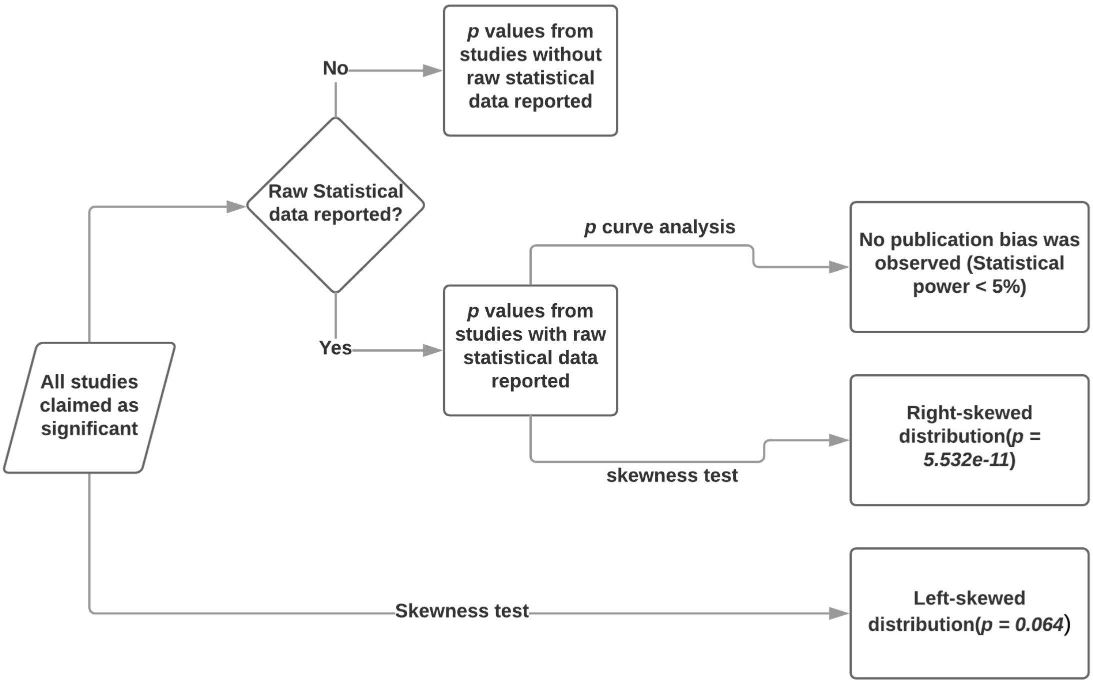

p values from studies without raw statistical data reported
No
Raw Statistical data reported?
p curve analysis
No publication bias was observed (Statistical power < 5%)
p values from studies with raw statistical data reported
Yes
All studies claimed as significant
Right-skewed distribution(*p = 5.532e-11*)
skewness test
Skewness test
Left-skewed distribution(*p = 0.064*)

Figure 4.6: **A brief view of the statistical methods and results.** The inclusion of studies without raw statistical data changed the total right-skewness. This indicates publication bias may exist in those new included studies.

4.6 Discussion

Sample sizes can affect the results of genetic association studies. For a given study effect of a random variant, there is a corresponding effect size with the locus. If sample size involved in a genetic association did not meet the requirement of the effect size, the association test is unlikely to detect the relationship. The impact of a variant is inversely correlated with the required sample size for detecting the effect. Bigger impact variants needs smaller sample size for a successful genetic association study. In contrast, smaller impact variants will require larger sample size to compensate. Longevity, as a complex trait, can be affected by many factors. Even solely talking about genetic factors, different variants may have different strength in affecting longevity. Some variants have strong impacts on longevity (e.g. *WRN*), while some other variants may act in a more subtle way. In either of the above cases, once the required sample size is achieved, the relationship between the impact of a variant and the ability of being detected is no longer linear. Further increasing sample sizes will not increase the power in detecting the association.

For variants that have been identified from genetic association studies, how the variants affect, and how much a variation contributes to, longevity is still unclear. Therefore, it is not possible to obtain a clearer signal by removing those noises came from non-related loci. In the case of non-related loci signal exists, the distribution of p values for those loci involved studies will distribute uniformly. *i.e.* no matter how many individuals involved, the p values will not be affected by that (Figure 4.1). Association does not necessarily mean causation. Due to lacking of sound explanation on the causal relationship between each association-study-identified locus and longevity, we cannot filter out the non-related loci. Too much noise in those p values prevented the pattern from being discovered.

The change of skewness might indicate the existence of publication bias. Only $\sim 29\%$ of significant entries reported the raw statistical data. This percentage is surprisingly low, and it is even lower in non-significant entries (Table 4.1). The potential explanations to this low rate of reporting raw statistical includes researchers did not think it is important, or due to a protective concern of their data. However, what concerns us most is another possibility. That is some researcher did not report all the raw statistical data maybe just because they are not confident enough to do so. As there is no compulsory requirement of reporting all the statistical raw data in genetic association studies, if someone obtained a significant p value by playing tricks on the data or on statistical methods, the last thing he/she wants to do might be showing out the raw statistical data. The non-transparent part in statistics could be a shelter for those data-manipulated studies. In the above analysis, right-skewed distribution disappeared after the inclusion of p values from studies without raw statistical data ($p_{skewness}$ changed from 5.532e-11 to 0.064). The alternation of p values skewness suggested that there was a high probability that newly included entries contain left-skewed p values. The left-skewed p values are a clear indication of publication bias and they are likely from p-hacking.

Although we cannot estimate to what extent the publication bias exists in the LongevityMap and it is even still premature to conclude that publication bias exists in the LongevityMap entries, the current study still provided some clues in this issue. With data accumulates over time (see Figure 2.4) and breakthroughs in methodology, it is very optimistic to assess the publication bias in longevity genetic association studies in the future.

Chapter 5

Genetic Diversity and its Impact on Genetic Association Studies of Ageing, Ageing-related Diseases, Cancers and Early Onset Diseases

5.1 Introduction

It is clear that genetic factors contribute to diseases and phenotypic traits, such as longevity, in human population. Gene functions can be inferred based on the genetic variability in accordance with the respective phenotypic variability. Mapping gene functions to phenotypic traits is easier for some simple diseases, as they are usually high penetrance and inherited following specific patterns. While in complex diseases, it is usually tricky to identify disease-causal genes under this way, because the amount of contribution to the complex trait come from each locus could be too small to be detected (CDCV) or the actual causal variants are too rare to be detected (CDRV) (see section 1.4.4).

The main difference between those two theories is where the contribution to complex diseases comes from. CDCV theory proposed that the common variants, which are shared in population, contributed to the onset of complex diseases. In contrast, CDRV theory proposed that complex diseases are attributed to the rare alleles. Both of the two theories have evidence and lack of in supporting the genetic basis of complex diseases. Neither of them could cover the whole picture and explain the complex disease well.

GWAS design is useful in connecting the phenotypic traits with the known variants, It is especially useful in screening connections between risk variants and traits (see section 1.4.5). However, we cannot simply rely on GWAS without careful consideration and further examination. GWAS has its own limitations in identifying the actually causal variants in the framework of either CDCV or CDRV. In CDCV theory, as the number of common variants is huge in populations, therefore, the contribution from each individual common variant is too small to be detected by GWAS. While in CDRV theory, the rare variant loci simply do not get covered by typical GWAS chip design. With this uncertainty, a positive GWAS hit could be just a proxy rather than the actual causal variant itself. Results from longevity genetic association studies further confirmed the ambiguity of GWAS hits in explaining the complex traits from the aspect of genetic basis. In the scope of LongevityMap, inconsistent outcomes were observed across loci/genes in different genetic association studies (GWASs and CGASs) (Figure 5.1) (Budovsky et al., 2013). Even new studies get published on an unforeseen speed, to date, few gene has been consistently reported.

Arguably, *APOE* and *Forkhead Box O3 (FOXO3)* are the only two genes that were constantly associated with longevity (Deelen et al., 2014; Morris et al., 2015). They were verified by ever-since large-scale study. Even for these two genes, there were failure replications in some early CGASs. Given the total numbers of human genes and traits that have been investigated, those outcomes from longevity-genetic association studies appeared to be disappointing.

There have been voices accusing the over-stringent multiple testing corrections has been applied in GWAS (Kenyon, 2010). This leads to the limited finding from GWAS results, as the true-causal loci could fail to pass the over-stringent statistical test. Apart from the methodologies reason, the intrinsic properties could also affect the GWAS hits. A statistically successful GWAS-hit was obtained by calculating the allele frequency differences between case and control groups. Bigger difference usually means better chance of getting a successful GWAS-hit. This has nothing to do with whether they are functionally connected to a studied trait or not.

Based on this, in the current thesis, the normalised nucleotide change on gene level, which we termed as "*Genetic Diversity* (GD)", were calculated based on The 1000 Genomes Project (1KGP) data (see section 5.3.2.3). The genetic diversity represents the genetic variation on the gene level. Estimating the relationship between GWAS outcomes and GD could give hints in better understanding the limitations of current GWAS design and help in explaining why few SNP gets constant GWAS hits.

In addition, examining the genetic diversity in Age-Related Traits/Diseases(ARTDs) associated genes and cancer-associated genes could help in revealing the genetic basis of complex diseases. Herein, a pipeline was attempted to reveal the relationship between human genetic diversity and its ability in affecting phenotypic traits, especially ageing-related traits and diseases.

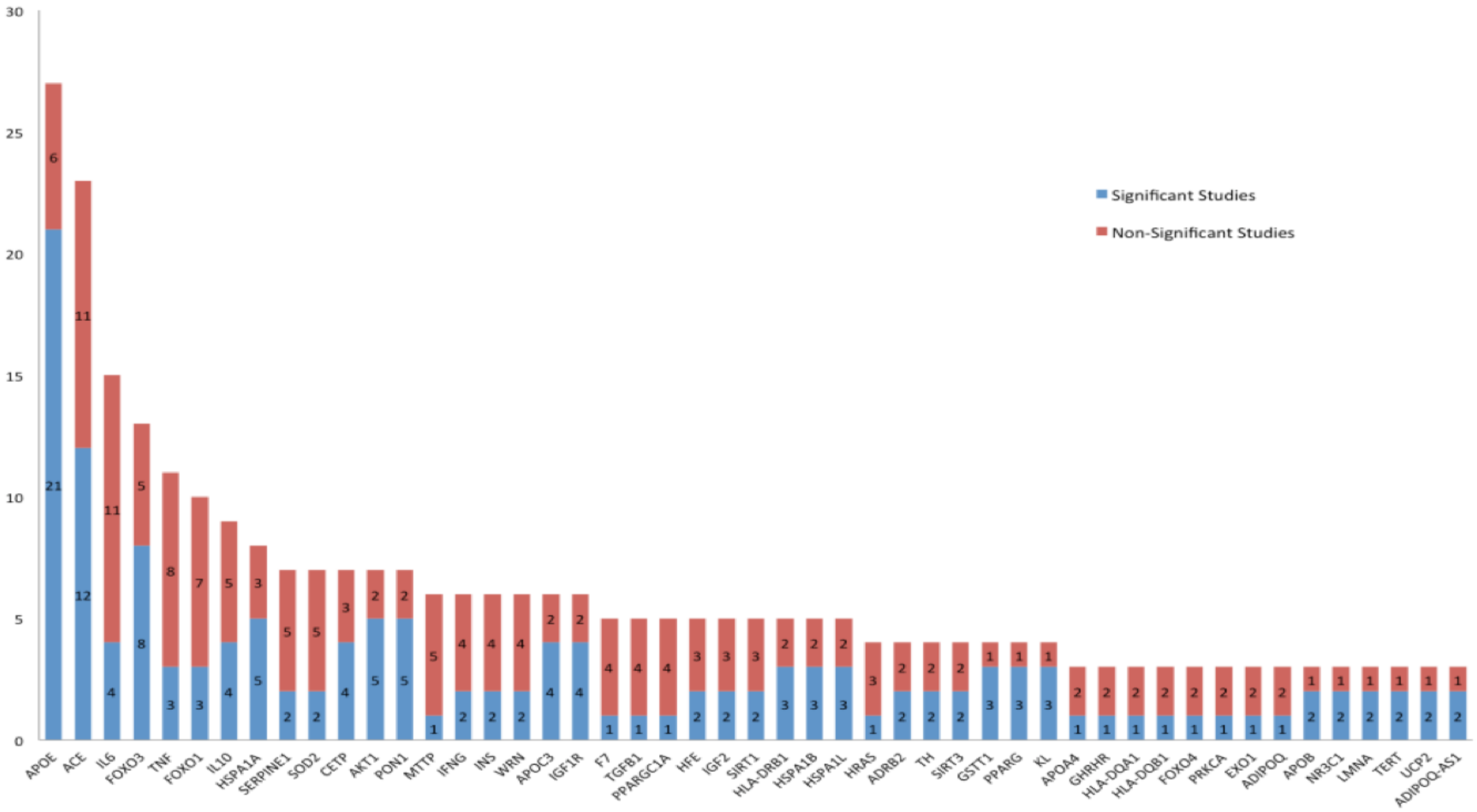


Figure 5.1: **Genes with conflict findings in the LongevityMap.** Red bars indicate non-significant studies, and blue bars indicate significant studies. The numbers in each individual coloured area are the counts of genes in that category. Only genes that have been reported more than two times were included.

5.2 Hypothesis and aims

Since GWAS calculates the frequency difference of genetic variants distributed in case and control groups, the characters of a locus are the main factors that affect the outcome of a GWAS besides the impact of statistical methods involved in GWAS.

In the CDCV theory, the common variants will unlikely be captured by GWAS because the contribution from each variant is too small. Therefore, GWAS should get random hits because no study effect exists. In CDRV theory, the true causal variants will not be captured by GWAS either, because the true causal variants are not in the tag SNP sets. GWAS could capture the tag SNPs that in LD with the casual SNPs. Cohorts from different genetic background likely to have different rare alleles in LD with different tag SNPs. Therefore, GWAS should identify different risk loci. However, we identified genes/loci that associated with longevity but lacking of biological explanations in different populations. This result contradicts with the prediction above. Therefore, we propose there might be something else that could affect the probability of getting a GWAS hit and it is likely to be the basic characters of genes.

In this context, an investigation of whether or not the genetic diversity affects the outcome of GWASs became necessary. GWAS hits could be biased towards to high diverse genes, in which case the true causal will be excluded from risk variants if the genetic diversity of true causal variants are low. In this case, the following experimental verification of biological functions between risk allele and traits will be very difficult because the biased GWAS hits will unlikely including true causal variants from risk variants. A better understanding of how genetic properties could potentially affect GWAS outcome would help in better understanding and interpreting GWAS results.

To illustrate, *APOE* is the only consistently reported longevity-association gene to date in large-scale GWAS (Magalhães, 2014). We could propose that the fact that *APOE* can be consistently detected may be due to its variability on common loci are different from other genes. It is possible the variability on common loci are higher than that in other genes, because low variability on common loci within a gene is likely to be missed by GWAS statistical test. By comparing the genetic variability of *APOE* to other genes could reveal some new evidence that can contribute to the interpreting of this topic.

To the current knowledge, ageing is still a complex trait. It is a phenotype reveals the combination of many different sub-traits, such as reduced mobility, decreased ability to maintain homeostasis or weakened cognitive abilities in elder age. The current limitations in accurately defining ageing brought hurdles in ageing research. For example, many ageing related studies use longevity as a proxy of ageing rates. The underneath hypothesis is longer lifespan somewhat means slower ageing rates in a given organism. Some ageing related risk genes/variants were actually identified from some age-related traits. This is acceptable but we are still far away from seeing the full picture of ageing.

In this aim, we used a collection of well studied Age-Related Traits/Diseases (ARTDs) to maximumly represent the current knowledge of ageing traits. The genes associated with the collection of traits can be recognised as ageing-associated genes. In order to keep concise, we termed the collection above as ARTDs class. Similarly, a Cancers class and an Early-Onset Diseases (EODs) class were also constructed. Cancers class were used because it is age-related and has clear characters. EODs class served as a control group because those traits onset in early age, which is different from ARTDs class or Cancers class. Environmental factors exert on those traits and leave traces in the genome sequence. These traces can be reflected in the genetic diversity. By comparing the genetic diversity between those three trait classes, it is expected to reveal the genetic basis of ageing.

5.3 Data sources and methods

5.3.1 Data sources

5.3.1.1 The 1000 Genomes Project data

The 1000 Genomes Project was employed as the source data for calculating the genetic diversity across all genes. Since the first release, 1KGP provides complete, unbiased, reliable, high-quality and open accessed human genome variation data to the scientific community (The 1000 Genomes Project Consortium et al., 2012; The 1000 Genomes Project Consortium, 2010). In the latest release, the genetic variation data of 2504 individuals from all of the major populations across all the continents was released. In addition to its high geographical coverage, the high chromosomal coverage rate and the high sequencing depth guaranteed the quality of data. Lastly, the public availability, as well as its unbiased sequence, made this repository an ideal resource for the current project (The 1000 Genomes Project Consortium et al., 2012).

5.3.1.2 GWAS-Catalog data

GWAS-Catalog is a manually curated database that collected all the published human GWASs. New individual GWAS were retrieved from PubMed on a daily basis through an automatic tool (Welter et al., 2014). Then the literature was examined and key information was extracted by a group of trained researchers. New data is added to the database on weekly basis after a double curation process (MacArthur et al., 2016). As of 31st Jan 2018, 3279 GWASs were collected in GWAS-Catalog database.

In the current project, the GWAS-Catalog data (available at: `www.ebi.ac.uk/gwas`. Accessed [31st Jan 2018]), were retrieved for calculating the total number of associated traits of each gene.

5.3.1.3 *Ensembl* data

Other generic information such as locations of genetic variants and gene length were retrieved from *GRCh37 Ensembl BioMart* (available at `http://grch37.ensembl.org/biomart/martview/`).

5.3.2 Methods

5.3.2.1 Overview

1KGP phase 3 genetic variants data was retrieved from 1KGP FTP server (`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) (1000 Genomes Project Consortium et al., 2015). Then, numbers of variant loci (both synonymous and non-synonymous) were counted and normalised at per 1000 bp length of DNA at whole population, subpopulation and small cohort levels. This normalised nucleotide change on gene level was defined as the GD of a gene. On the other hand, the total number of GWAS hits of each gene were counted from GWAS-Catalog data. After obtaining both GD data and GWAS hits count, correlation tests were carried out to see the relationship between genetic variability and the outcome of association studies.

In the later sections, age-related diseases and cancer-associated genes were examined to see if there was any difference in GD between those diseases/cancers-associated genes.

5.3.2.2 SNPs mapping to genes

Since the latest major release of 1KGP data (2$^{\text{nd}}$ May 2013) was based on GRCh37 assembly, all the analyses involved in the current project were referenced to GRCh37 assembly. The location of each individual SNP on the chromosome was obtained from 1KGP Phase 3 data (available from EBI FTP site `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`). The spans of genes were retrieved from *Ensembl BioMart GRCh37* base assembly. Considering the regulatory regions around a gene may be involved in affecting the functions of a gene, an inclusive flank length of 1000bp was applied on both upstream and downstream sequences. Any SNP located in the flank regions was considered a genic SNP (Figure 5.2). Where a SNP is located in the overlapped regions of two or more adjacent genes, the SNP was counted separately in each gene.

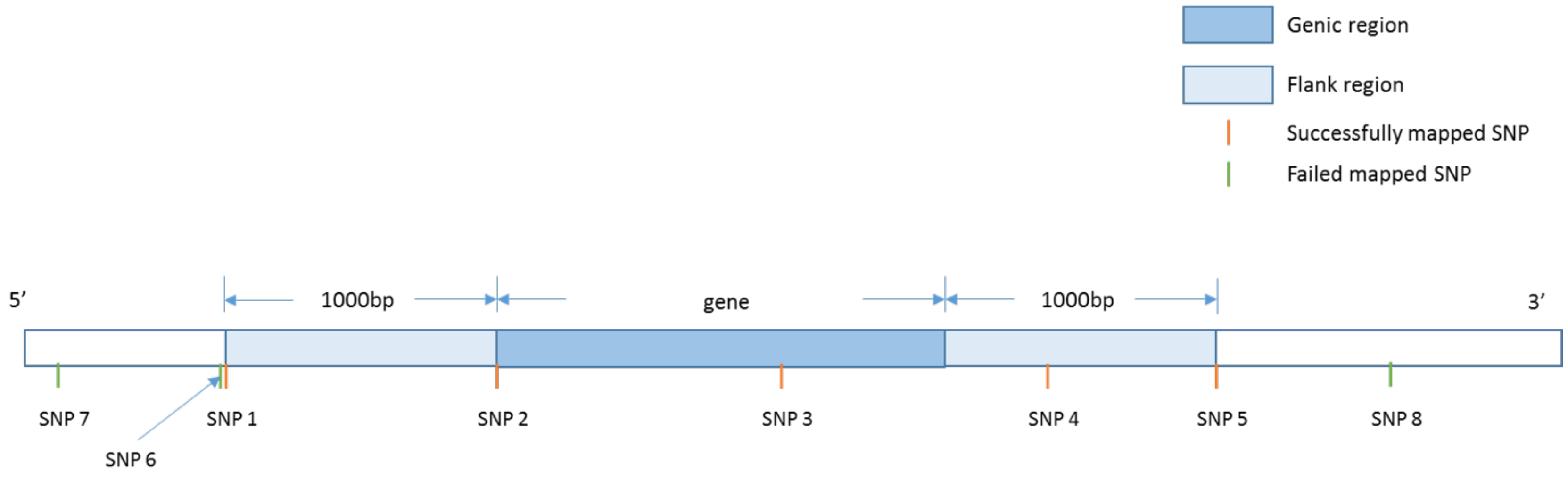


Figure 5.2: **An indication of how SNPs were mapped to genes.** Blue colour indicates genic region as shown in GRCh37 base assembly. Light blue indicates 1kb flanks around a gene. Vertical bars indicate SNPs. An orange coloured indicates a successful mapped SNP (e.g. SNP 1-5), while a green one indicates a failed mapping (e.g. SNP 6 - 8). Any SNP located exactly on the border were counted as a successful mapping (e.g. SNP 1 and SNP 5).

5.3.2.3 Genetic Diversity (GD) calculation

The basis of genetic variation between two population or even two individuals is the variability represented by the sequence of DNA base pairs. Here, we used the normalised nucleotide change at the individual gene level, defined as Genetic Diversity(GD), as a representative of genetic variability across all the genes(see equation 5.4). The calculation of GD was performed as follows:

1. Consider a cohort consisting of n individuals, for a given SNP with a minor allele frequency MAF) at position i, the total number of minor alleles (representing the nucleotide changes) at position i in the population is:

$$n \times MAF_i \tag{5.1}$$

2. If there are m SNPs located in the genic region and the corresponding two flank regions, the total number of minor alleles within the genic and flank regions can be calculated by:

$$\sum_{i=0}^{m} n \times MAF_i \tag{5.2}$$

3. As n is a constant, the above formula can be simplified as:

$$\sum_{i=0}^{m} MAF_i \tag{5.3}$$

4. Then the normalised nucleotide change (defined as GD) in every 1000bp DNA can be represented as:

$$GD = \frac{\sum_{i=0}^{m} MAF_i \times 1000}{length_{genic} + length_{upstream} + length_{downstream}} \tag{5.4}$$

Where the lengths of upstream and downstream flanks were arbitrary chosen as 1000 bp DNA. GD is the measurement of genetic diversity at the level of individual genes. MAF is the Minor Allele Frequency of SNPs located in the same gene and upstream downstream flanks (see Figure 5.2).

This method calculate the genetic diversity in the most straightforward way. It only takes into account the number of nucleotides change in a given gene without making any assumptions. The normalised nucleotide change made comparison between genes become feasible. With the aid of population coverage from 1KGP, it is expected calculating GD across human genes could provide new insights in understanding human genetic variation and phenotypic traits. This methods was verified by comparison the GDs between protein-coding genes and non-protein-coding genes.

The major limitation of this method is it only reflects the nucleotide change on the genes level. The genetic variation information at the allele level was not included in the GD calculation. For example, two genes with the same GD could have different genetic composition. One might consist of a high number of low-MAF loci while the other might have a low number of high-MAF loci. As long as those two genes share the same number of total nucleotide changes, they will have the same GD value and there is no way to distinguish one from another in cases like this in the above described GD calculation method. Another limitation is about the missing information in the low-frequency alleles. When defining SNPs, 1KGP embedded cut-offs of MAF $> 0.1\%$ and $> 1\%$ for coding regions and the rest of genome, respectively. Therefore, any genetic variation with MAFs less than the above thresholds was not considered as polymorphism and would not contribute to the GD in the following analyses. This means individual private variants or rare alleles will unlikely to be included in the analyses, even they contribute to the overall individual phenotypic traits.

5.3.2.4 Reported traits and mapped traits

Two types of traits were recorded in GWAS-Catalog data, Reported Traits(RTs) and Mapped Traits(MTs). Reported traits were taken from the original paper. They were the terms that were used by the author of literature to describe the traits. The same referred trait can be described in different words between authors due to language habit or different emphasis. For example, "Alzheimers Disease" and "Alzheimer's Disease" are referring to the same disease based on common sense. Similar with "type-2 diabetes" and "type II diabetes mellitus". Therefore, words used in RTs can have variation. For human reading, it is not difficult to understand these cases and probably will not cause confusion. But to computer and software, they are different from each other in either group. The variation in describing a same trait could cause errors for automated data processing.

In contrast, mapped traits were ontology based-terms (Welter et al., 2014). They are managed, standardised terms that mapped from curated traits description based on Experimental Factor Ontology (EFO) (Malone et al., 2008). Through this ontology mapping method, different descriptions of the same trait or disease were mapped to the same ontology-based term, which is MT in the GWAS-Catalog. Take the above two cases as an example, either "Alzheimers Disease" or "Alzheimer's Disease" is mapped to "Alzheimers disease (`http://www.ebi.ac.uk/efo/EFO_0000249`)". Similarly, either "type-2 diabetes" or "type II diabetes mellitus" is mapped to "type II diabetes mellitus (`http://www.ebi.ac.uk/efo/EFO_0001360`)". In this way, confusion was minimised. Also, the introduction of ontologies facilitates the integration and processing data obtained from multiple sources (Smith et al., 2007; Welter et al., 2014). Therefore, in the current project, MTs were utilised for software analysis. By using MTs, data consistency and the quality of results generated from software can be assured.

5.3.2.5 Correlation analyses

Correlation analysis is a statistical method that investigates the extent to which two variables tend to change together. There are two typical correlation methods, one type examines the linear relationship between two variables and the other examines monotonic relationship between two variables. Pearson's correlation test measures the linear dependency between two variables while Spearman's correlation test and Kendall's correlation test examines the monotonic relationship (rank-order) between two continuous or ordinal variables. Rank-order correlation analysis measures the degree of similarity between the two rankings from the two variables that being tested.

In this project, the monotonic correlation method were selected over linear correlation method because 1) linear correlation result is susceptible to outlier data points. 2) the relationships between genetic diversity of a gene and other attributes being tested are unlikely to be linear due to the variation in gene effects.

For Spearman's correlation and Kendall's correlation, the limitation of Spearman's correlation is it cannot deal with a "tie" situation. Whereas, on the opposite, Kendall's Correlation test can handle this problem. Therefore, in the following sections where correlation tests were employed, Kendall's Correlation test was used.

5.3.2.6 Sequential Removal of Rarely successfully reported Genes Procedure (SRRGP)

As a hypothesis-free method, GWAS design examines each individual variant evenly in different studies. However, the outcomes from those GWAS studies varies very much from study to study. In fact, the majority of genes has been positively reported very few times. These low-frequent positively reported genes potentially bring bias into correlation analysis due to lacking replicating ability. In this case, a Sequential Removal of Rarely reported Genes(SRRGP) method was introduced into the following analyses to minimise the effects from rare reported genes.

5.3.2.7 Mapping traits to EFO terms

As GWAS-Catalog uses Experimental Factor Ontology(EFO) mapping as a dictionary to convert author reported traits to standardised ontology terms. Each traits in ARTDs, Cancers and EODs trait classes were manually mapped to the EFO terms. Then any significantly reported SNP associated with the related EFO term in GWAS-Catalog were extracted and recorded. Following this, the author reported most significant SNP(s) was/were mapped to GRCh37 assembly genes using 1kb flanks. After this, three lists of genes corresponding to individual three trait classes disease (ARTDs, Cancers, EODs) were obtained.

5.3.2.8 Statistical tests

Mann–Whitney *U* tests were performed in comparing GDs between groups. Bonferroni correction was introduced to control type I errors when comparing GDs between genes from Age-Related Traits/Diseases (ARTDs), Cancers and Early-Onset Diseases (EODs) classes. All the statistical tests were performed in RStudio (RStudio Team, 2015).

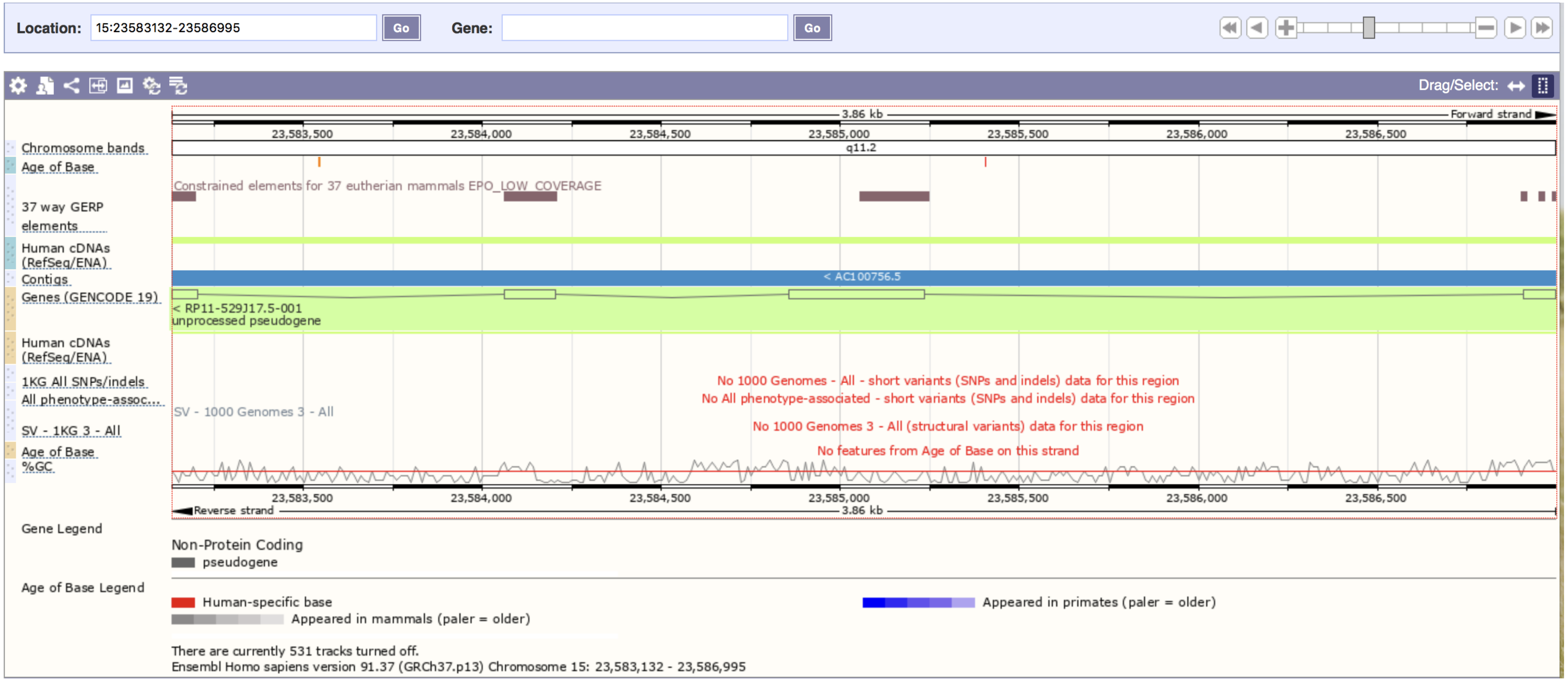
5.4 Results

5.4.1 Summary of the data

In the human genome GRCh37 assembly, there are 54849 genes in the autosomes of GRCh37 assembly, including protein-coding gene, lincRNA, antisense, misc_RNA, snRNA, pseudogene. To clarify, "gene" here and below refers to all the types of genes listed above. 54592 of the total 54849 genes have been successfully calculated GD. The rest 257 genes did not have GD information (discussed in section 5.5). Figure 5.3 demonstrates one of the 257 genes (ENSG00000260399) in ensembl genome browser. See appendix 3 for the gene types of those 257 genes. Of the 54592 genes, 11869 genes have at least one GWAS hit in the GWAS-Catalog. These genes were named as GWAS-Hit Genes(GHGs).

About 68% of the 11869 GHGs were protein-coding genes (Table 5.1). The minimum GD calculated in the genome-wide was zero. It was due to the reported allele frequency was zero in the 1KGP data. The summary information of GDs were listed in Table 5.2.

Those GHGs were reported at different frequencies. The most reported gene in GWAS-Catalog was *GCKR* (ENSG00000084734), which appeared 83 times (Table 5.3). On the contrary, the majority of those genes were only shown less than three times in GWAS-Catalog. Among those 11869 genes, 4640 ($\sim$ 39.1%), 2546 ($\sim$ 21.5%) and 1441 ($\sim$ 12.1%) of them have been reported only once, twice and three times, respectively. Only 1827 and 628 of those genes have been reported more than five and more than 10 times (Figure 5.4), respectively.



Location: 15:23583132-23586995 Go

Gene: Go

Drag/Select:

3.86 kb

Forward strand

23,583,500 23,584,000 23,584,500 23,585,000 23,585,500 23,586,000 23,586,500

Chromosome bands q11.2

Age of Base

37 way GERP elements: Constrained elements for 37 eutherian mammals EPO_LOW_COVERAGE

Human cDNAs (RefSeq/ENA)

Contigs: < AC100756.5

Genes (GENCODE 19): < RP11-529J17.5-001 unprocessed pseudogene

Human cDNAs (RefSeq/ENA)

1KG All SNPs/indels: No 1000 Genomes - All - short variants (SNPs and indels) data for this region

All phenotype-assoc...: No All phenotype-associated - short variants (SNPs and indels) data for this region

SV - 1KG 3 - All: SV - 1000 Genomes 3 - All — No 1000 Genomes 3 - All (structural variants) data for this region

Age of Base: No features from Age of Base on this strand

%GC

23,583,500 23,584,000 23,584,500 23,585,000 23,585,500 23,586,000 23,586,500

Reverse strand 3.86 kb

Gene Legend

Non-Protein Coding

▬ pseudogene

Age of Base Legend

🟥 Human-specific base

▬ Appeared in mammals (paler = older)

▬ Appeared in primates (paler = older)

There are currently 531 tracks turned off.

Ensembl Homo sapiens version 91.37 (GRCh37.p13) Chromosome 15: 23,583,132 - 23,586,995

Figure 5.3: **An example of genes without GD information (ENSG00000260399).**

Table 5.1: Number of genes with GD in GRCh37 assembly

	Protein coding genes	non-protein coding genes	Total
GWAS-hit genes	8196	3673	11869
Non-GWAS hit genes	11204	31519	42723
Total	19400	35192	54592

Table 5.2: Summary of GD

Gene class	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Genome-Wide	54592	0	0.00057	0.00092	0.00110	0.00138	0.03352
GHGs	11869	0.000024	0.00072	0.00101	0.00116	0.00136	0.03352
non-GHGs	42723	0	0.00053	0.00089	0.00108	0.00139	0.01902

n :Number of genes; GHG: GWAS-Hit Genes

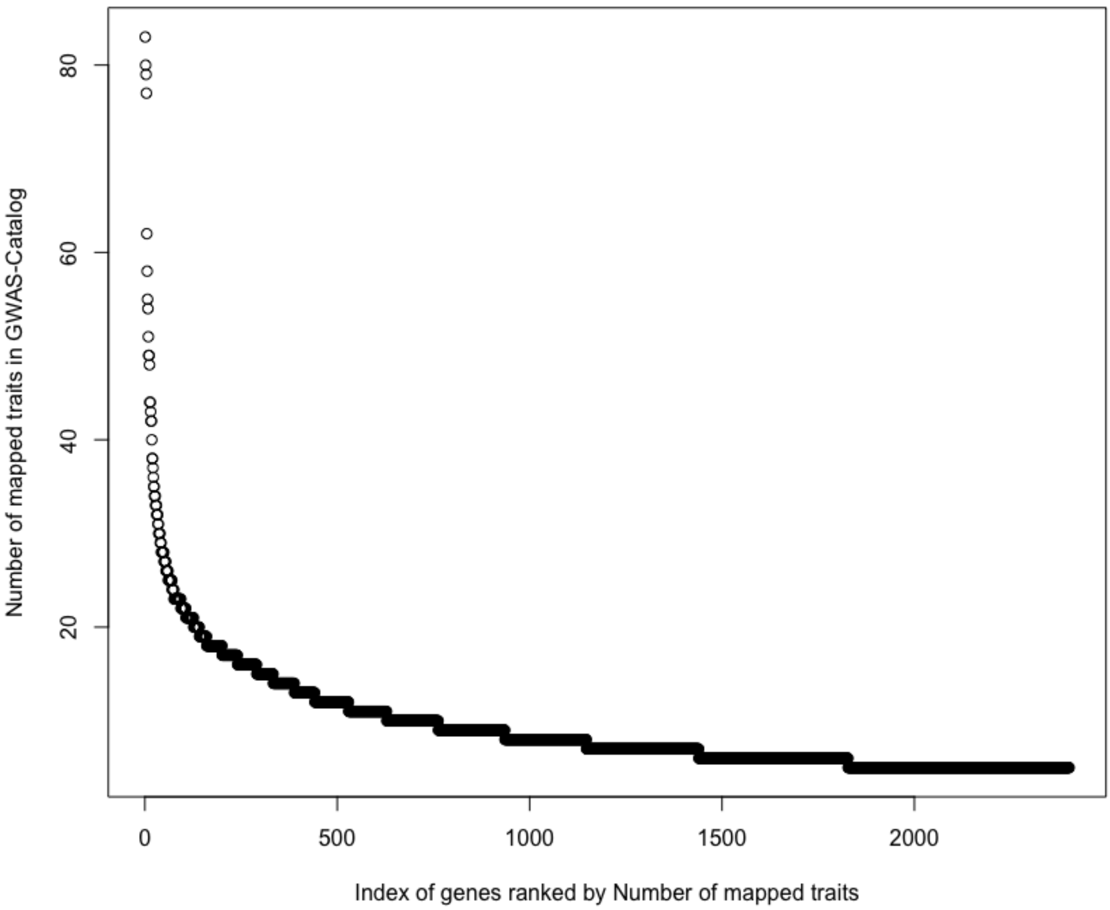


Figure 5.4: **Scatter plot of the number of mapped traits for each gene.** 2402 genes with $NMTs \geq 5$ were included. Genes were ordered by NMTs in GWAS-Catalog. Each circle in the figure represents an individual gene. NMT: Number of Mapped Traits in GWAS-Catalog

Table 5.3: The top ten most reported genes in the GWAS-Catalog.

ENSG ID	GD*	Gene name	NMTs
ENSG00000084734	0.00056	*GCKR*	83
ENSG00000183117	0.00292	*CSMD1*	80
ENSG00000134824	0.00063	*FADS2*	79
ENSG00000175164	0.00327	*ABO*	77
ENSG00000149485	0.00039	*FADS1*	62
ENSG00000140945	0.00197	*CDH13*	58
ENSG00000153707	0.00145	*PTPRD*	55
ENSG00000253111	0.00161	*RP11-136O12.2*	54
ENSG00000206337	0.00331	*HCP5*	51
ENSG00000109917	0.00074	*ZNF259*	49

*indicates the calculated Genetic Diversity for all the 1KGP data based on the formula 5.4; ENSG, Ensembl Gene IDs; NMTs: number of mapped traits. (as of Jan/2018)

5.4.2 *Kendal correlation tests*

5.4.2.1 Genome-wide: gene length vs. GD

A correlation test was carried out between gene length and GD to see if genes length could potentially affect GD. In genome-wide ($n = 54592$, all genes in Table 5.1), there was significant correlation between gene length and genetic diversity (*Kendall Tau = 0.021, z = 7.416, p = 1.206e-13*). The correlation between Gene length and GD was positive, which means longer genes have higher GD. However, the strength of correlation was weak ($Tau = 0.02$).

5.4.2.2 Within GWAS-Hit-Genes (GHGs): gene length vs. GD

Considering there were many genes that have been rarely reported (Figure 5.4), and these less reported genes could have biases in affecting the correlation test results, the correlation was tested after SRRGP (see section 5.3.2.6). Significant positive correlations were observed when $NMT \geq 3$, $NMT \geq 4$ and $NMT \geq 6$ in GWAS-Catalog (Table 5.4).

5.4.2.3 Gene length vs. NMTs

The correlation analyses results showed there were significant positive correlations between Gene Length and NMTs in GHGs ($Kendall's\ Tau = 0.3075,\ p < 2.2e - 16$). The results demonstrated that longer genes are more likely to have more NMTs in GWAS-Catalog.

Table 5.4: Correlation test: gene length vs. GD

NMT(n)	No. of genes	Kendall's Tau	Z score	p
$n \geq 1$	11869	0.00424	0.6922	0.4888
$n \geq 2$	7229	0.0127	1.613	0.1068
$n \geq 3$	4683	0.0208	2.1329	0.03293*
$n \geq 4$	3242	0.02996	2.5563	0.01058*
$n \geq 5$	2402	0.02575	1.8909	0.05864
$n \geq 6$	1827	0.04533	2.9026	0.003701*

*indicates significant p values. For global correlation between gene length and NMT, see section 5.4.2.1.

5.4.2.4 GD VS. NMTs

To test whether or not higher genetic variation could bring more GWAS hits, another correlation test between GD and the NMTs in GWAS-Catalog was examined. Of all the genes that have been reported in GWAS-Catalog, Kendall Correlation test between the number of mapped traits and GD showed a significant positive correlation (*Tau = 0.05018, Z = 7.4985, p = 6.456e-14*). This means genes with higher GD are more likely to have more NMTs in GWAS-Catalog.

In the current data from GWAS-Catalog, many genes were reported only once in the GWAS-Catalog. As GWAS design usually use a stringent, maybe over-stringent, statistical test threshold, the chance of getting false positive hits is very low (Kenyon, 2010). Therefore, the big proportion of GWAS-hits with $NMT = 1$ indicates the pleiotropy of genes.

Nevertheless, to minimize the potential false positive effect comes from genes with the least reported numbers of times, SRRGP was applied here (see section 5.3.2.6). Correlation tests were performed between GD and NMTs in SRRGP until $NMT \leq 6$ times. Positive significant correlations were observed when $NMTs \geq 1$, $NMTs \geq 2$, $NMTs \geq 3$, $NMTs \geq 4$. No correlation was observed when $NMTs \geq 5$ and $NMT \geq 6$ (Table 5.5).

Table 5.5: Correlation test: GD vs. NMTs

NMT	No. of genes	Kendall's Tau	Z score	p	GD Min.	GD 1st Qu.	GD Median	GD Mean	GD 3rd Qu.	GD Max.
$n \geq 1$	11869	0.05018	7.4985	6.456e-14*	0.000024	0.00072	0.00101	0.00116	0.00136	0.03352
$n \geq 2$	7229	0.04996	5.8907	3.846e-09*	0.00006	0.00076	0.00103	0.00117	0.00136	0.03352
$n \geq 3$	4683	0.03769	3.6165	0.0002986*	0.00006	0.00078	0.00105	0.00119	0.00136	0.03352
$n \geq 4$	3242	0.03469	2.7937	0.005211*	0.000115	0.000804	0.001065	0.001228	0.001348	0.033518
$n \geq 5$	2402	0.00818	0.56972	0.5689	0.00012	0.00082	0.00108	0.00124	0.00136	0.03352
$n \geq 6$	1827	0.01840	1.1227	0.2616	0.00013	0.00083	0.00107	0.00127	0.00135	0.03352

*indicates significant p values.

5.4.3 GD comparisons

5.4.3.1 GHGs vs. non-GHGs

Mann-Whitney test showed there was significant difference between GHGs and non-GHGs in the global population ($p < 2.2e - 16$, see Table 5.2). GD in GHGs were significantly higher than that in non-GHGs.

5.4.3.2 Protein-coding genes vs. non-protein-coding genes

In GRCh37 assembly, there are 19400 protein coding genes and 35192 non-protein coding genes. Genome-widely, GD of protein-coding genes was significantly lower than that in non-protein-coding genes (*Mann–Whitney test,* $p = 9.001e - 09$, Table 5.6). Similar results were observed when comparing GDs in GHGs (Table 5.7) or non-GHGs (Table 5.8, *Mann–Whitney test,* $p < 2.2e - 16$ in both tests).

In summary, GD of protein-coding genes was significantly lower than that of non-protein-coding genes in GHGs, non-GHGs or genome-wide.

Table 5.6: GD: Protein-coding genes vs non-protein-coding genes

Gene class	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Genome-Wide	54592	0	0.00057	0.00092	0.00110	0.00138	0.03352
Protein-coding Genes	19400	0	0.00061	0.00091	0.00102	0.00127	0.03105
non-protein-coding genes	35192	0	0.00055	0.00093	0.00114	0.00146	0.03352

Table 5.7: GD in GHGs: Protein-coding genes vs non-protein-coding genes

Gene class	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All GHGs	11869	0.000024	0.000721	0.001009	0.001158	0.001362	0.033518
Protein-coding Genes	8196	0.000061	0.000701	0.000976	0.001073	0.001302	0.031054
non-protein-coding genes	3673	0.000024	0.000773	0.001087	0.001347	0.001518	0.033518

Table 5.8: GD in non-GHGs: Protein-coding genes vs non-protein-coding genes

Gene class	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
All non-GHGs	42723	0	0.00053	0.00089	0.00108	0.00139	0.01902
Protein-coding Genes	11204	0	0.00054	0.00085	0.00098	0.00124	0.01343
non-protein-coding genes	31519	0	0.00052	0.00091	0.00112	0.00145	0.01902

5.4.3.3 GD comparison between genes in ARTDs, Cancers and EODs trait classes

In addition to the above analyses, the genetic characters of those genes that are associated with age-related traits/diseases and cancers were also investigated. Different traits/diseases have different gene/loci or a set of genes/locus involved. Apparently, some traits/diseases have more genes/SNPs involved than others. In the following analysis, we focused on the genetic diversity character of those commonly regarded complex traits/diseases. In particular, those well-known ARTDs, Cancers and EODs (See Table 5.9, Table 5.10 and Table 5.11).

Each disease and cancer entries that met the targets were manually extracted and recorded (see material and methods). Then a collection of author reported SNPs from those selected entries were mapped to genes with 1kb flanks upstream and downstream according to GRCh37 assembly. Subsequently, the GD of those genes under each age-related traits and cancers were investigated. The summary of GD data was showed in Table 5.12.

Mann-Whitney tests with Bonferroni correction ($p_{corrected} = 0.0167$) were selected in comparing GDs between genes in three trait-classes (Table 5.12). The results showed GDs of Cancers class genes were significantly lower than that of EODs class genes ($p = 8.53e - 6$). The GDs of ARTDs class genes were also significantly lower than that of EODs class genes ($p = 0.0019$). However, no significant difference was observed between ARTDs class genes and Cancers class genes ($p = 0.0421$).

Table 5.9: Age-Related Traits and Diseases (ARTDs) class

trait	Ontology term	number of associated genes
Alzheimers disease	EFO_0000249	338
age-related macular degeneration	EFO_0001365	69
cardiovascular disease	EFO_0000319	8
hypertension	EFO_0000537	89
metabolic syndrome	EFO_0000195	44
obesity	EFO_0001073	84
Parkinson's disease	EFO_0002508	114
stroke	EFO_0000712	78
type 2 diabetes mellitus	EFO_0001360	379

Table 5.10: Cancers class

trait	Ontology term	number of associated genes
breast cancer breast carcinoma	EFO_0000305	522
colorectal cancer	EFO_0005842	175
ovarian carcinoma	EFO_0001075	56
pancreatic carcinoma	EFO_0002618	69
prostate carcinoma	EFO_0001663	190

Table 5.11: Early Onset Diseases class (EODs)

trait	Ontology term	number of associated genes
asthma	EFO_0000270	273
type 1 diabetes mellitus	EFO_0001359	111
testicular cancer	EFO_0005088	17

Table 5.12: Summary of GD across the three trait classes

Trait class	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ARTDs	1111	0.00013	0.00076	0.00103	0.00121	0.00133	0.03352
Cancers	898	0.00011	0.00073	0.00100	0.00107	0.00130	0.00712
EODs	389	0.00014	0.00084	0.00112	0.00143	0.00140	0.03105

5.5 Discussion

The genome-wide correlation analysis between gene length and genetic diversity indicated that there was significant positive correlations between them (see section 5.4.2.1). However, the correlation did not maintain during the SRRGP. Additionally, Kendall's correlation analysis showed consistent significantly positive correlation between NMTs and gene lengths ($p < 2.2e - 16$, see section 5.4.2.1). Longer genes are more likely to get GWAS-hits and therefore more NMTs mapped to them. On the other hand, the correlation analysis between GD and NMTs showed fluctuate correlations depending on which data have been excluded in the SRRGP. The correlation was maintained until $n \geq 4$ but disappeared from $n \geq 5$. The observation of the strongest correlation between GD and NMTs only exists when low NMTs are included indicates the trace of natural selection. Generally speaking, if a gene has multiple phenotypic traits mapped to it (pleiotropy), it becomes functionally important and therefore tends to be conserved across generations.

GD of GHGs was significantly higher than that of non-GHGs, which means genes with higher GD could potentially have better chance of getting a GWAS-hit (see section 5.4.3.1). GD of protein-coding genes was significantly lower than that of non-protein-coding genes. This can be explained as protein-coding genes normally functionally related and therefore likely to be conserved. Unsurprisingly, protein-coding genes have lower GD in both GHGs and non-GHGs group (see section 5.4.3.2).

The above findings of both gene length and GD of genes are positively correlated with NMTs suggest GWAS hits are likely to be found in longer genes and/or genes with higher GD. One straightforward explanation would be longer genes and high-GD genes are more likely to be genes that determine more phenotypic traits or diseases.

Therefore, those genes (longer and/or with higher GD) are easy to be detected by GWAS. Our correlation tests just revealed the true fact.

An alternative explanation is the GWAS hits could be biased to longer genes and/or genes with higher GD. If this is the case, then the loci identified by GWAS will contain false positive hits and the trait associated gene will be less accurate because they are biased towards to genes that are longer and/or with higher GD. In this context, the true causal variants and the true functional genes may not be captured by GWAS if they are not long enough or display high GD.

Similar level of GDs were observed between ARTDs class genes and Cancers class genes. However, GDs of EODs class genes were significantly higher than GDs of either ARTDs class genes or Cancers class genes (see section 5.4.3.3). These results indicate EODs class genes may experiencing different selection process from genes in the other two classes.

Chapter 6

General Discussion

In this work, we presented a work-flow in exploring the genetic basis of longevity and ageing. The project started from building the LongevityMap database by curating information from current available longevity genetic association studies, to exploring the functional clusters in the longevity, then assessing publication biases in those study and finally testing the genetic diversity of genes in different trait classes. The outcomes from the whole work-flow have helped in developing a better understanding of how genetic components contribute to the complex biological process of ageing, and how the variability of genes could affect the discovery of potential causal SNPs.

The manually curated LongevityMap database is a reliable data repository for HLAGs. It is the first database that presenting the latest knowledge of human longevity associated genes as a whole. This facilitates the integration of new technologies into analysing the data in a systemic way (like described in Chapter 3). The following analyses of LongevityMap data with functional enrichment tools and reactome pathway tools provided new insights in understanding the LongevityMap data. The most enriched functional clusters and pathways revealed the current limitations in selecting candidate genes for CGASs. Researchers tend to select genes/variants that known to play vital roles in deciding human lifespan for their CGASs design. The preference in selecting candidates was reflected by functional enrichment analyses.

On one hand, those results verified the ageing process could be decided by effect from all those vital biological processes. On the other hand, those modest to least enriched clusters and pathways could also suggest some new directions in the future work. For example, the small pathway clusters consisted by *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* genes could be an indication of contributions from environmental factors. The cluster consisted of *LMNA*, *SYNE1* and *POT1* explains the importance of telomerases, through which the length of telomeres are maintained, in affecting human lifespan (see section 1.2.3). Similarly, each of those small clusters should be examined for any potential contribution to ageing.

To assess the data quality in the LongevityMap, the publication biases in the LongevityMap were investigated. By examining the skewness of p values with *p-curve application* and *D'Agostino skewness test*, a skewness change was found when including p values from studies that did not report raw statistical data. This suggests there is high probability of existing publication biases.

One issue that arises from GWAS analyses is the identified loci cannot be consistently replicated in other studies. One possible explanation, which is out of the scope of current thesis, is the over-stringent multiple test correction criteria. Another possible explanation includes impacts from genetic diversity. Therefore, in Chapter 5, we investigated the relationship between GWAS hits and the GD of genes. Results showed longer genes and higher GD genes are likely to get GWAS-hits. This could due to longer genes or higher GD genes are more likely to be the genes that determines more phenotypic traits or diseases. Or, GWAS-hits could be biased towards to those above genes. Results from other studies also suggested the impact of gene length. Longer genes are more likely enriched in the cancer related pathways (Sahakyan et al., 2016). Investigating the strength of biases from gene length and GD should be planned in the future work.

In most cases, it is not known if it is a single SNP itself or a set of SNPs that is working with the external environment in shaping a phenotypic trait. Therefore, only considering individual SNPs for causal variants may lead to incomplete conclusions. Luckily, some researchers have already started to account for this issue. For example, in 2015, Kim et al. reported longevity traits associated with chromosome regions rather than individual SNPs (Kim et al., 2015).

Although old and new methods have generated new data into academia on daily basis, genetic association study design is not flawless. A certain number of SNPs that have been identified by genetic association study located in the intergenic regions, which are inferred "regulatory regions" but rarely been verified (Hindorff et al., 2009). Further to this, compared to accurate genotyping, the relatively ambiguous phenotyping brings another layer of noise in remapping the causal relationship between SNPs and phenotypes. As stated by Altshuler et al. *"The ability to measure genotype now far exceeds our ability to measure phenotype, plus the environment exposures play a larger role in human phenotypic variation than does genetic variation".* (Altshuler et al., 2008).

GD comparison between ARTDs, Cancers and EODs class showed GD was significantly higher GD in EODs class than that in Cancer class or ARTDs class. This indicated the potential difference of selection pressures exert on genes in different classes.

In summary, our work contributed to the genetic basis of ageing research in several aspects:

1. The LongevityMap database is the first repository providing the human longevity-genetic association studies and the outcomes to the community.
2. Functional enrichment analyses and pathway analyses on longevity-associated genes provided many potential biological functions that could contribute to ageing.
3. Publication biases investigation firstly provided new perspective on how to objectively view the data.
4. Genetic diversity analyses provided some clue in connections between GWAS-hits and gene properties as well as the GD difference between different trait classes.

Although some achievements have been made in the thesis, there are many interesting questions worth to be answered in the future. For example, many modest enriched clusters and pathways should be explained and experimentally tested when possible. More detailed information on the processes and mechanisms of environmental factors affect GD in different trait classes should be explored.

Ageing is a complex trait and not yet well defined in terms of phenotype. Many risk alleles have been identified by GWASs and CGASs. However, the true causal relationship still to be discovered. Potential publication biases brings extra difficulties in to the field. Even so, we should be encouraged by the achievements have been made in ageing research in the past several decades. Genome research is a fast-moving field, new experimental methods and new statistical algorithm emerge quickly. With the increasing number of centenarians and global average lifespan, we have the opportunity of applying the most advanced technologies and methods to deciphering ageing.

Chapter 7

Appendices

1. Functional annotation of longevity-associated genes with whole genome as background

Annotation Cluster	Enrichment Score	The most representative term	Number of genes linked to the current term	FDR*
ACDB 1	14.09	Regulation of cell death	64	5.50E-22
ACDB 2	10.81	positive regulation of signal transduction	32	6.40E-14
ACDB 3	8.88	regulation of response to external stimulus	24	3.00E-13
ACDB 4	8.69	regulation of locomotion	23	1.40E-10
ACDB 5	7.77	response to hormone stimulus	33	4.30E-12
ACDB 6	6.98	response to extracellular stimulus	23	2.10E-09
ACDB 7	6.09	regulation of phosphorylation	36	2.40E-11
ACDB 8	5.98	regulation of cell size	20	1.60E-07

ACDB 9	5.55	response to oxidative stress	16	6.80E-06
ACDB 10	5.31	regulation of transferase activity	27	1.00E-07
ACDB 11	5.04	regulation of protein kinase B signaling cascade	8	2.00E-07
ACDB 12	5.02	cell fraction	45	8.00E-06
ACDB 13	4.78	regulation of lipid metabolic process	22	2.00E-14
ACDB 14	4.59	mTOR signaling pathway	19	3.10E-14
ACDB 15	4.53	behavior	26	3.60E-05
ACDB 16	4.52	response to abiotic stimulus	25	1.50E-06
ACDB 17	4.38	homeostatic process	53	8.20E-16
ACDB 18	4.29	response to wounding	26	3.00E-04
ACDB 19	4.25	protein dimerization activity	29	2.30E-05
ACDB 20	4.25	regulation of foam cell differentiation	8	4.40E-06
ACDB 21	4.01	neuron projection	23	5.80E-06
ACDB 22	3.99	positive regulation of DNA metabolic process	9	1.40E-04
ACDB 23	3.91	Glioma	13	3.00E-06
ACDB 24	3.79	diabetes mellitus	7	1.10E-03
ACDB 25	3.74	response to reactive oxygen species	10	1.70E-04
ACDB 26	3.67	regulation of lipid transport	13	1.40E-12
ACDB 27	3.56	regulation of monooxygenase activity	8	4.40E-06
ACDB 28	3.56	regulation of vasodilation	6	4.60E-04
ACDB 29	3.52	regulation of secretion	21	1.80E-08
ACDB 30	3.51	cellular response to extracellular stimulus	9	4.00E-04
ACDB 31	3.49	regulation of foam cell differentiation	8	4.40E-06
ACDB 32	3.45	negative regulation of lipid transport	6	9.60E-05

ACDB 33	3.42	positive regulation of cellular component organization	17	4.40E-06
ACDB 34	3.29	regulation of hormone levels	14	8.10E-05
ACDB 35	3.22	positive regulation of multicellular organism growth	7	8.90E-05
ACDB 36	3.2	cell death	30	1.10E-03
ACDB 37	3.07	regulation of interleukin-6 production	7	8.30E-04
ACDB 38	3	positive regulation of macromolecule metabolic process	49	5.80E-11
ACDB 39	2.96	regulation of smooth muscle cell proliferation	8	3.50E-04
ACDB 40	2.93	macromolecular complex subunit organization	31	3.50E-04
ACDB 41	2.8	peptide binding	13	7.60E-03
ACDB 42	2.72	regulation of interleukin-6 production	7	8.30E-04
ACDB 43	2.71	regulation of hormone levels	14	8.10E-05
ACDB 44	2.68	blood vessel development	14	1.20E-02
ACDB 45	2.66	regulation of neurological system process	12	2.30E-03
ACDB 46	2.65	obesity	6	1.20E-03
ACDB 47	2.54	immune system development	19	6.80E-05
ACDB 48	2.51	regulation of protein secretion	10	1.90E-05
ACDB 49	2.5	regulation of behavior	7	3.40E-03

*DAVID reports FDR as percentage, therefore, the above $FDR = FDR_{\text{DAVID}}/100$.

2. Functional annotation of longevity-associated genes with LongevityMap as background

Annotation Cluster	Enrichment Score	The most representative term	Number of genes linked to the current term	FDR
ACLB 1	18.23	membrane-enclosed lumen	49	2.80E-21
ACLB 2	16.11	regulation of cell death	64	4.50E-25
ACLB 3	14.8	cell fraction	45	1.90E-18
ACLB 4	11.8	non-membrane-bounded organelle	48	1.00E-13
ACLB 5	10.02	protein dimerization activity	29	3.90E-12
ACLB 6	9.29	regulation of cellular protein metabolic process	36	2.20E-14
ACLB 7	8.79	regulation of locomotion	23	1.30E-10
ACLB 8	8.56	macromolecular complex subunit organization	31	3.40E-13
ACLB 9	7.86	cation binding	78	1.10E-12
ACLB 10	7.64	nucleus	72	1.30E-19
ACLB 11	7.42	response to organic substance	45	4.30E-13
ACLB 12	7.39	positive regulation of molecular function	35	7.60E-13
ACLB 13	7.32	plasma membrane	88	3.90E-26
ACLB 14	7.26	cell death	30	3.80E-09
ACLB 15	7.14	regulation of cellular component size	22	4.60E-10
ACLB 16	7.08	Pathways in cancer	28	1.70E-19
ACLB 17	7.03	cell projection	31	1.50E-11
ACLB 18	6.84	regulation of response to external stimulus	24	2.40E-11

ACLB 19	6.79	response to extracellular stimulus	23	5.80E-09
ACLB 20	6.62	cell-cell signaling	27	4.80E-10
ACLB 21	6.44	defense response	27	1.50E-07
ACLB 22	6.05	regulation of transferase activity	27	6.10E-10
ACLB 23	5.62	homeostatic process	53	2.30E-16
ACLB 24	5.38	mTOR signaling pathway	19	1.60E-16
ACLB 25	5.16	reproductive process in a multicellular organism	22	1.90E-07
ACLB 26	5.15	negative regulation of macromolecule metabolic process	26	7.00E-08
ACLB 27	5.01	behavior	26	1.60E-07
ACLB 28	4.92	Hypertrophic cardiomyopathy (HCM)	8	3.80E-06
ACLB 29	4.81	vesicle	22	8.00E-06
ACLB 30	4.67	Pancreatic cancer	12	6.00E-08
ACLB 31	4.53	cell projection organization	19	1.00E-06
ACLB 32	4.38	response to abiotic stimulus	25	5.20E-07
ACLB 33	4.34	negative regulation of biosynthetic process	25	3.40E-08
ACLB 34	4.33	response to oxidative stress	16	3.30E-04
ACLB 35	4.09	cell adhesion	20	1.50E-04
ACLB 36	3.91	vesicle-mediated transport	16	3.30E-05
ACLB 37	3.9	regulation of hormone levels	14	1.10E-05
ACLB 38	3.64	cytoskeleton	14	6.60E-06
ACLB 39	3.84	cytoskeleton organization	15	2.00E-05
ACLB 40	3.75	regulation of DNA metabolic process	12	1.40E-04
ACLB 41	3.63	regulation of lipid metabolic process	22	6.40E-10
ACLB 42	3.55	cell leading edge	9	4.60E-04
ACLB 43	3.53	protein localization	21	4.70E-07

ACLB 44	3.39	positive regulation of cellular component organization	17	1.50E-06
ACLB 45	3.37	endoplasmic reticulum part	11	6.20E-05
ACLB 46	3.31	regulation of foam cell differentiation	8	2.30E-04
ACLB 47	3.27	blood circulation	14	9.00E-05
ACLB 48	3.13	Toll-like receptor signaling pathway	10	2.20E-06
ACLB 49	3.05	regulation of neurological system process	12	6.80E-04
ACLB 50	3.04	blood vessel development	14	5.10E-03
ACLB 51	3.03	regulation of cellular localization	22	4.50E-07
ACLB 52	3.02	peptide binding	13	8.20E-06
ACLB 53	3.01	cellular response to extracellular stimulus	9	1.30E-03
ACLB 54	2.98	regulation of protein kinase B signaling cascade	8	7.30E-04
ACLB 55	2.98	diabetes mellitus	7	6.40E-04
ACLB 56	2.85	nucleotide binding	57	5.20E-08
ACLB 57	2.82	cell activation	18	2.90E-05
ACLB 58	2.7	p53 signaling pathway	7	6.80E-04
ACLB 59	2.67	regulation of cell cycle	23	1.60E-07
ACLB 60	2.65	regulation of lipid transport	13	4.30E-08
ACLB 61	2.65	cytoplasmic vesicle part	8	9.20E-03
ACLB 62	2.51	Systemic lupus erythematosus	8	2.70E-05

*DAVID reports FDR as percentage, therefore, the above $FDR = FDR_{\text{DAVID}}/100$.

3. Frequencies of genes in each gene type category

Type of gene	Frequency in whole genome	Frequency in genes without GD
3prime_overlapping_ncrna	20	0
antisense	5160	7
IG_C_gene	14	0
IG_C_pseudogene	9	0
IG_D_gene	37	0
IG_J_gene	18	0
IG_J_pseudogene	3	0
IG_V_gene	138	2
IG_V_pseudogene	187	9
lincRNA	6932	27
miRNA	2847	14
misc_RNA	1936	12
polymorphic_pseudogene	45	0
processed_transcript	499	2
protein_coding	19430	30
pseudogene	12745	120
rRNA	497	17
sense_intronic	723	0
sense_overlapping	194	0
snoRNA	1391	3
snRNA	1814	14
TR_C_gene	5	0
TR_D_gene	3	0
TR_J_gene	74	0
TR_J_pseudogene	4	0

TR_V_gene	97	0
TR_V_pseudogene	27	0

Chapter 8

Published Works

- Budovsky A*, Craig T*, **Wang J***, Tacutu R, Csordas A, Lourenco J, Fraifeld VE, de Magalhaes JP. (2013) "LongevityMap: A database of human genetic variants associated with longevity." *Trends in Genetics* **29**:559-560.

- Fernandes M, Wan C, Tacutu R, Barardo D, Rajput A, **Wang J**, Thoppil H, Thornton D, Yang C, Freitas A, de Magalhães, JP. (2016). Systematic analysis of the gerontome reveals links between aging and age-related diseases. *Human Molecular Genetics* **25**:4804–4818.

- Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, Diana E, Lehmann G, Toren D, **Wang J**, Fraifeld VE, de Magalhães JP. (2018). Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Research* **46(D1)**:D1083–D1090.

* indicates first co-authorship

Bibliography

1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis (2015). "A global reference for human genetic variation." In: *Nature* 526.7571, pp. 68–74.

Agrelo, Ruben, W.-H. Cheng, Fernando Setien, Santiago Ropero, Jesus Espada, Mario F Fraga, Michel Herranz, Maria F Paz, Montserrat Sanchez-Cespedes, Maria J Artiga, David Guerrero, Antoni Castells, Cayetano von Kobbe, Vilhelm A Bohr, and Manel Esteller (2006). "Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer". In: *Proceedings of the National Academy of Sciences* 103.23, pp. 8822–8827.

Altshuler, David, Mark J Daly, and Eric S Lander (2008). "Genetic mapping in human disease". In: *Science* 322.5903, pp. 881–888.

Arnold, J (2001). "Genetic Drift". In: *Encyclopedia of Genetics.* Ed. by Sydney Brenner and Jefferey H Miller. New York: Elsevier, pp. 832–834.

Ashton, Michael C (2013). "Chapter 6 - Genetic and Environmental Influences on Personality". In: *Individual Differences and Personality (Second Edition).* Ed. by Michael C Ashton. Second Edi. San Diego: Academic Press, pp. 123–151.

Banerjee, Amitav, U B Chitnis, S L Jadhav, J S Bhawalkar, and S Chaudhury (2009). "Hypothesis testing, type I and type II errors." In: *Industrial psychiatry journal* 18.2, pp. 127–31.

Barbujani, G., S. Ghirotto, and F. Tassi (2013). "Nine things to remember about human genome diversity." In: *Tissue antigens* 82.3, pp. 155–64.

Barrett, Jeffrey C, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M Michael Barmada, Alain Bitton, Themistocles Dassopoulos, Lisa Wu Datta, Todd Green, Anne M Griffiths, Emily O Kistner, Michael T Murtha, Miguel D Regueiro, Jerome I Rotter, L Philip Schumm, A Hillary Steinhart, Stephan R Targan, Ramnik J Xavier, NIDDK IBD Genetics Consortium, Cécile Libioulle, Cynthia Sandor, Mark Lathrop, Jacques Belaiche, Olivier Dewit, Ivo Gut, Simon Heath, Debby Laukens, Myriam Mni, Paul Rutgeerts, André Van Gossum, Diana Zelenika, Denis Franchimont, Jean-Pierre Hugot, Martine de Vos, Severine Vermeire, Edouard Louis, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Lon R Cardon, Carl A Anderson, Hazel Drummond, Elaine Nimmo, Tariq Ahmad, Natalie J Prescott, Clive M Onnie, Sheila A Fisher, Jonathan Marchini, Jilur Ghori, Suzannah Bumpstead, Rhian Gwilliam, Mark Tremelling, Panos Deloukas, John Mansfield, Derek Jewell, Jack Satsangi, Christopher G Mathew, Miles Parkes, Michel Georges, and Mark J Daly (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." In: *Nature genetics* 40.8, pp. 955–62.

Baudisch, Annette (2011). "The pace and shape of ageing". In: *Methods in Ecology and Evolution* 2.4, pp. 375–382.

Begley, C. Glenn and Lee M. Ellis (2012). "Drug development: Raise standards for preclinical cancer research." In: *Nature* 483.7391, pp. 531–3.

Bellavia, D, G. Frada, P Di Franco, S Feo, C Franceschi, P Sansoni, and M Brai (1999). "C4, BF, C3 Allele Distribution and Complement Activity in Healthy Aged People and Centenarians". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 54.4, B150–B153.

Benayoun, Bérénice A, Elizabeth A Pollina, and Anne Brunet (2015). "Epigenetic regulation of ageing: linking environmental inputs to genomic stability." In: *Nature reviews. Molecular cell biology* 16.10, pp. 593–610.

Benetos, A, K Okuda, M Lajemi, M Kimura, F Thomas, J Skurnick, C Labat, K Bean, and A Aviv (2001). "Telomere length as an indicator of biological aging: the gender effect and relation with pulse pressure and pulse wave velocity." In: *Hypertension (Dallas, Tex. : 1979)* 37.2 Pt 2, pp. 381–5.

Bernstein, Carol, Anil R., Valentine Nfonsam, and Harris Bernstei (2013). "DNA Damage, DNA Repair and Cancer". In: *New Research Directions in DNA Repair*. InTech, pp. 413–466.

Biau, David Jean, Brigitte M. Jolles, and Raphaël Porcher (2010). "P value and the theory of hypothesis testing: an explanation for new researchers." In: *Clinical orthopaedics and related research* 468.3, pp. 885–92.

Bidder, G P (1932). "SENESCENCE." In: *British medical journal* 2.3742, pp. 583–5.

Bordone, Laura and Leonard Guarente (2005). "Calorie restriction, SIRT1 and metabolism: understanding longevity." In: *Nature reviews. Molecular cell biology* 6.4, pp. 298–305.

Brandes, U., D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner (2008). "On Modularity Clustering". In: *IEEE Transactions on Knowledge and Data Engineering* 20.2, pp. 172–188.

Brenner, S. (1974). "The genetics of Caenorhabditis elegans." In: *Genetics* 77.1, pp. 71–94.

Brooks-Wilson, Angela R (2013). "Genetics of healthy aging and longevity." In: *Human genetics* 132.12, pp. 1323–38.

Budovsky, Arie, Thomas Craig, Jingwei Wang, Robi Tacutu, Attila Csordas, Joana Lourenço, Vadim E. Fraifeld, and João Pedro de Magalhães (2013). "LongevityMap: a database of human genetic variants associated with longevity." In: *Trends in genetics : TIG* 29.10, pp. 559–60.

Bunker, John P (2001). "The role of medical care in contributing to health improvements within societies". In: *International Journal of Epidemiology* 30.6, pp. 1260–1263.

Butler, Paul G., Alan D. Wanamaker, James D. Scourse, Christopher A. Richardson, and David J. Reynolds (2013). "Variability of marine climate on the North Icelandic Shelf in a 1357-year proxy archive based on growth increments in the bivalve Arctica islandica". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 373, pp. 141–151.

Caselli, Graziella, Lucia Pozzi, James W. Vaupel, Luca Deiana, Gianni Pes, Ciriaco Carru, Claudio Franceschi, and Giovannella Baggio (2006). "Family clustering in Sardinian longevity: A genealogical approach". In: *Experimental Gerontology* 41.8, pp. 727–736.

Cavalli-Sforza, L Luca and Marcus W Feldman (2003). "The application of molecular genetic approaches to the study of human evolution." In: *Nature genetics* 33 Suppl, pp. 266–75.

Centers for Disease Control and Prevention (2003). "Public Health and Aging: Trends in Aging—United States and Worldwide". In: *JAMA* 289.11, p. 1371.

Chakravarti, Aravinda (1999). "Population genetics—making sense out of sequence". In: *Nature Genetics* 21.january, pp. 56–60.

Cho, Dong-Yeon, Yoo-Ah Kim, and Teresa M. Przytycka (2012). "Chapter 5: Network Biology Approach to Complex Diseases". In: *PLoS Computational Biology* 8.12. Ed. by Fran Lewitter and Maricel Kann, e1002820.

Cline, Melissa S, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, Kristina Hanspers, Ruth Isserlin, Ryan Kelley, Sarah Killcoyne, Samad Lotia, Steven Maere, John Morris, Keiichiro Ono, Vuk Pavlovic, Alexander R Pico, Aditya Vailaya, Peng-Liang Wang, Annette Adler, Bruce R Conklin, Leroy Hood, Martin Kuiper, Chris Sander, Ilya Schmulevich, Benno Schwikowski, Guy J Warner, Trey Ideker, and Gary D Bader (2007). “Integration of biological networks and gene expression data using Cytoscape.” In: *Nature protocols* 2.10, pp. 2366–82.

Cohen, Alan A. (2018). “Aging across the tree of life: The importance of a comparative perspective for the use of animal models in aging.” In: *Biochimica et biophysica acta. Molecular basis of disease* 1864.9 Pt A, pp. 2680–2689.

Cohen, J. C. (2004). “Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol”. In: *Science* 305.5685, pp. 869–872.

Colhoun, Helen M., Paul M. McKeigue, and George Davey Smith (2003). “Problems of reporting genetic associations with complex outcomes”. In: *The Lancet* 361.9360, pp. 865–872.

Collado, Manuel, Maria A. Blasco, and Manuel Serrano (2007). “Cellular senescence in cancer and aging.” In: *Cell* 130.2, pp. 223–33.

Comfort, A (1964). *Ageing: The biology of senescence.* English. London: Routledge & Kegan Paul Ltd. Eev. ed., xvi + 365 pp.

Croft, David, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Rothfels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D’Eustachio (2014). “The Reactome pathway knowledgebase.” In: *Nucleic acids research* 42.Database issue, pp. D472–7.

Cui, Hang, Yahui Kong, and Hong Zhang (2012). “Oxidative stress, mitochondrial dysfunction, and aging.” In: *Journal of signal transduction* 2012, p. 646354.

Dato, S., M. Soerensen, V. Lagani, A. Montesanto, G. Passarino, K. Christensen, Q. Tan, and L. Christiansen (2014). “Contribution of genetic polymorphisms on functional status at very old age: A gene-based analysis of 38 genes (311 SNPs) in the oxidative stress pathway”. In: *Experimental Gerontology* 52, pp. 23–29.

Dato, Serena, Giuseppina Rose, Paolina Crocco, Daniela Monti, Paolo Garagnani, Claudio Franceschi, and Giuseppe Passarino (2017). “The genetics of human longevity: an intricacy of genes, environment, culture and microbiome.” In: *Mechanisms of ageing and development* 165.Pt B, pp. 147–155.

Davidovic, Mladen, Goran Sevo, Petar Svorcan, Dragoslav P Milosevic, Nebojsa Despotovic, and Predrag Erceg (2010). “Old age as a privilege of the "selfish ones".” In: *Aging and disease* 1.2, pp. 139–46.

Debrabant, Birgit, Mette Soerensen, Friederike Flachsbart, Serena Dato, Jonas Mengel-From, Tinna Stevnsner, Vilhelm A Bohr, Torben A Kruse, Stefan Schreiber, Almut Nebel, Kaare Christensen, Qihua Tan, and Lene Christiansen (2014). “Human longevity and variation in DNA damage response and repair: study of the contribution of sub-processes using competitive gene-set analysis”. In: *European Journal of Human Genetics* 22.9, pp. 1131–1136.

Deelen, Joris, Marian Beekman, Hae Won Uh, Linda Broer, Kristin L. Ayers, Qihua Tan, Yoichiro Kamatani, Anna M. Bennet, Riin Tamm, Stella Trompet, Daníel F. Guobjartsson, Friederike Flachsbart, Giuseppina Rose, Alexander Viktorin, Krista Fischer, Marianne Nygaard, Heather J. Cordell, Paolina Crocco, Erik B. Van Den Akker, Stefan Böhringer, Quinta Helmer, Christopher P. Nelson, Gary I. Saunders, Maris Alver, Karen Andersen-Ranberg, Marie E. Breen, Ruud van Der Breggen, Amke Caliebe, Miriam Capri, Elisa Cevenini, Joanna C. Collerton, Serena Dato, Karen Davies, Ian Ford, Jutta Gampe, Paolo Garagnani, Eco J C de Geus, Jennifer Harrow, Diana Van Heemst, Bastiaan T. Heijmans, Femke Anouska Heinsen, Jouke Jan Hottenga, Albert Hofman, Bernard Jeune, Palmi V. Jonsson, Mark Lathrop, Doris Lechner, Carmen Martin-Ruiz, Susan E. Mcnerlan, Evelin Mihailov, Alberto Montesanto, Simon P. Mooijaart, Anne Murphy, Ellen A. Nohr, Lavinia Paternoster, Iris Postmus, Fernando Rivadeneira, Owen A. Ross, Stefano Salvioli, Naveed Sattar, Stefan Schreiber, Hreinn Stefánsson, David J. Stott, Henning Tiemeier, André G. Uitterlinden, Rudi G J Westendorp, Gonneke Willemsen, Nilesh J. Samani, Pilar Galan, Thorkild I A Sørensen, Dorret I. Boomsma, J. Wouter Jukema, Irene Maeve Rea, Giuseppe Passarino, Anton J M de Craen, Kaare Christensen, Almut Nebel, Kári Stefánsson, Andres Metspalu, Patrik Magnusson, Hélène Blanché, Lene Christiansen, Thomas B L Kirkwood, Cornelia M. Van Duijn, Claudio Franceschi, Jeanine J. Houwing-Duistermaat, and P. Eline Slagboom (2014). "Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age". In: *Human Molecular Genetics* 23.16, pp. 4420–4432.

Dickersin, K (1990). "The existence of publication bias and risk factors for its occurrence." In: *JAMA : the journal of the American Medical Association* 263.10, pp. 1385–1389.

Dreesen, Oliver and Colin L. Stewart (2011). "Accelerated aging syndromes, are they relevant to normal human aging?" In: *Aging* 3.9, pp. 889–895.

Easterbrook, P J, J A Berlin, R Gopalan, and D R Matthews (1991). "Publication bias in clinical research". In: *Lancet* 337.8746, pp. 867–872.

Easton, Douglas F et al. (2007). "Genome-wide association study identifies novel breast cancer susceptibility loci." In: *Nature* 447.7148, pp. 1087–93.

ENCODE Project Consortium, The ENCODE Project (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." In: *Science* 306.5696, pp. 636–40.

Ernst, I. M A, K. Pallauf, J. K. Bendall, L. Paulsen, S. Nikolai, P. Huebbe, T. Roeder, and G. Rimbach (2013). "Vitamin E supplementation and lifespan in model organisms." In: *Ageing research reviews* 12.1, pp. 365–75.

Fabregat, Antonio, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and

Peter D'Eustachio (2018). "The Reactome Pathway Knowledgebase." In: *Nucleic acids research* 46.D1, pp. D649–D655.

Fanelli, Daniele (2012). "Negative results are disappearing from most disciplines and countries". In: *Scientometrics* 90.3, pp. 891–904.

Fernandes, Maria, Cen Wan, Robi Tacutu, Diogo Barardo, Ashish Rajput, Jingwei Wang, Harikrishnan Thoppil, Daniel Thornton, Chenhao Yang, Alex Freitas, and João Pedro de Magalhães (2016). "Systematic analysis of the gerontome reveals links between aging and age-related diseases." In: *Human molecular genetics* 25.21, pp. 4804–4818.

Fisher, Ronald Aylme (1926). "The arrangement of field experiments". In: *Journal of the Ministry of Agriculture of Great Britain* 33, pp. 503–513.

Frazer, Kelly A, Sarah S Murray, Nicholas J Schork, and Eric J Topol (2009). "Human genetic variation and its contribution to complex traits". In: *Nature Reviews Genetics* 10.4, pp. 241–251.

Freitas, Alex A. and João Pedro de Magalhães (2011). "A review and appraisal of the DNA damage theory of ageing." In: *Mutation research* 728.1-2, pp. 12–22.

Fries, James F (1980). "Aging, Natural Death, and the Compression of Morbidity". In: *New England Journal of Medicine* 303.3, pp. 130–135.

Goldstein, David B. (2009). "Common Genetic Variation and Human Traits". In: *New England Journal of Medicine* 360.17, pp. 1696–1698.

Goodman, Steven (2008). "A dirty dozen: twelve p-value misconceptions." In: *Seminars in hematology* 45.3, pp. 135–40.

Gremeaux, Vincent, Mathieu Gayda, Romuald Lepers, Philippe Sosner, Martin Juneau, and Anil Nigam (2012). "Exercise and longevity." In: *Maturitas* 73.4, pp. 312–7.

Halliwell, B (1991). "Reactive oxygen species in living systems: source, biochemistry, and role in human disease." In: *The American journal of medicine* 91.3C, 14S–22S.

Hansen, Malene and Brian K. Kennedy (2016). "Does Longer Lifespan Mean Longer Healthspan?" In: *Trends in cell biology* 26.8, pp. 565–8.

Harley, C B, A B Futcher, and C W Greider (1990). "Telomeres shorten during ageing of human fibroblasts." In: *Nature* 345.6274, pp. 458–60.

Harman, D (1956). "Aging: A Theory Based on Free Radical and Radiation Chemistry". In: *Journal of Gerontology* 11.3, pp. 298–300.

– (1972). "The biologic clock: the mitochondria?" In: *Journal of the American Geriatrics Society* 20.4, pp. 145–7.

Harman, Denham (2009). "Origin and evolution of the free radical theory of aging: a brief personal history, 1954–2009." In: *Biogerontology* 10.6, pp. 773–81.

Harris, M A, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan, H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz, P Tonellato, P Jaiswal, T Seigfried, R White, and Gene Ontology Consortium (2004). "The Gene Ontology (GO) database and informatics resource." In: *Nucleic acids research* 32.Database issue, pp. D258–61.

Hayflick, L. and P.S. Moorhead (1961). "The serial cultivation of human diploid cell strains". In: *Experimental Cell Research* 25.3, pp. 585–621.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions (2015). "The Extent and Consequences of P-Hacking in Science". In: *PLOS Biology* 13.3, e1002106.

Heemst, D van (2010). "Insulin, IGF-1 and longevity". In: *Aging and Disease* 1.2, pp. 147–57.

Heilbronn, Leonie K and Eric Ravussin (2003). “Calorie restriction and aging: review of the literature and implications for studies in humans.” In: *The American journal of clinical nutrition* 78.3, pp. 361–9.

Herskind, Anne Maria, Matthew McGue, Niels V. Holm, Thorkild I A Sørensen, Bent Harvald, and James W. Vaupel (1996). “The heritability of human longevity: A population-based study of 2872 Danish twin pairs born 1870-1900”. In: *Human Genetics* 97.3, pp. 319–323.

Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio (2009). “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.” In: *Proceedings of the National Academy of Sciences of the United States of America* 106.23, pp. 9362–7.

Hiona, Asimina and Christiaan Leeuwenburgh (2008). “The role of mitochondrial DNA mutations in aging and sarcopenia: implications for the mitochondrial vicious cycle theory of aging.” In: *Experimental gerontology* 43.1, pp. 24–33.

Hjelmborg, Jacob B., Ivan Iachine, Axel Skytthe, James W. Vaupel, Matt McGue, Markku Koskenvuo, Jaakko Kaprio, Nancy L. Pedersen, and Kaare Christensen (2006). “Genetic influence on human lifespan and longevity”. In: *Human Genetics* 119.3, pp. 312–321.

Hoeijmakers, Jan H J (2009). “DNA damage, aging, and cancer.” In: *The New England journal of medicine* 361.15, pp. 1475–85.

Holloszy, John O. and Luigi Fontana (2007). “Caloric restriction in humans.” In: *Experimental gerontology* 42.8, pp. 709–12.

Holmes, George E., Carol Bernstein, and Harris Bernstein (1992). “Oxidative and other DNA damages as the basis of aging: a review”. In: *Mutation Research/DNAging* 275.3-6, pp. 305–315.

Hong, Eun Pyo and Ji Wan Park (2012). “Sample size and statistical power calculation in genetic association studies.” In: *Genomics & informatics* 10.2, pp. 117–22.

Horvitz, H. Robert (2003). “Worms, life, and death (Nobel lecture).” In: *Chembiochem : a European journal of chemical biology* 4.8, pp. 697–711.

Huang, Da Wei, Brad T Sherman, Qina Tan, Jack R Collins, W Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki (2007). “The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.” In: *Genome biology* 8.9, R183.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki (2009a). “Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists”. In: *Nucleic Acids Research* 37.1, pp. 1–13.

Huang, Da Wei, Richard a Lempicki, and Brad T Sherman (2009b). “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.” In: *Nature Protocols* 4.1, pp. 44–57.

International, The and Hapmap Consortium (2003). “The International HapMap Project.” In: *Nature* 426.6968, pp. 789–796.

Ioannidis, J. P. and Thomas a Trikalinos (2007). “An exploratory test for an excess of significant findings”. In: *Clinical Trials* 4, pp. 245–253.

James, W P, G G Duthie, and K W Wahle (1989). “The Mediterranean diet: protective or simply non-toxic?” In: *European journal of clinical nutrition* 43 Suppl 2, pp. 31–41.

Jensen, L J, R Gupta, H-H Staerfeldt, and S Brunak (2003). “Prediction of human protein function according to Gene Ontology categories.” In: *Bioinformatics (Oxford, England)* 19.5, pp. 635–42.

Jiao, Xiaoli, Brad T Sherman, Da Wei Huang, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki (2012). “DAVID-WS: a stateful web service to facilitate gene/protein list analysis.” In: *Bioinformatics (Oxford, England)* 28.13, pp. 1805–6.

Jin, Kunlin (2010). “Modern Biological Theories of Aging.” In: *Aging and disease* 1.2, pp. 72–74.

Johnson, Simon C, Peter S Rabinovitch, and Matt Kaeberlein (2013). “mTOR is a key modulator of ageing and age-related disease.” In: *Nature* 493.7432, pp. 338–45.

Johnson, Thomas E. (2002). “A personal retrospective on the genetics of aging”. In: *Biogerontology* 3.1-2, pp. 7–12.

Jones, Owen R., Alexander Scheuerlein, Roberto Salguero-Gómez, Carlo Giovanni Camarda, Ralf Schaible, Brenda B. Casper, Johan P. Dahlgren, Johan Ehrlén, María B. García, Eric S. Menges, Pedro F. Quintana-Ascencio, Hal Caswell, Annette Baudisch, and James W. Vaupel (2014). “Diversity of ageing across the tree of life”. In: *Nature* 505.7482, pp. 169–173.

Kanehisa, M and S Goto (2000). “KEGG: kyoto encyclopedia of genes and genomes.” In: *Nucleic acids research* 28.1, pp. 27–30.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe (2014). “Data, information, knowledge and principle: back to metabolism in KEGG.” In: *Nucleic acids research* 42.Database issue, pp. D199–205.

Kang, Hyun Tae, Joon Tae Park, Kobong Choi, Yongsub Kim, Hyo Jei Claudia Choi, Chul Won Jung, Young Sam Lee, and Sang Chul Park (2017). “Chemical screening identifies ATM as a target for alleviating senescence”. In: *Nature Chemical Biology* 13.6, pp. 616–623.

Kennedy, Brian K., Shelley L. Berger, Anne Brunet, Judith Campisi, Ana Maria Cuervo, Elissa S. Epel, Claudio Franceschi, Gordon J. Lithgow, Richard I. Morimoto, Jeffrey E. Pessin, Thomas A. Rando, Arlan Richardson, Eric E. Schadt, Tony Wyss-Coray, and Felipe Sierra (2014). “Geroscience: linking aging to chronic disease.” In: *Cell* 159.4, pp. 709–13.

Kenyon, Cynthia (2005). “The plasticity of aging: insights from long-lived mutants.” In: *Cell* 120.4, pp. 449–60.

Kenyon, Cynthia J. (2010). “The genetics of ageing”. In: *Nature* 464.7288, pp. 504–512.

Kim, Sangkyu, David A. Welsh, Leann Myers, Katie E. Cherry, Jennifer Wyckoff, and S. Michal Jazwinski (2015). “Non-coding genomic regions possessing enhancer and silencer potential are associated with healthy aging and exceptional survival". In: *Oncotarget* 6.6, pp. 3600–3612.

Kimura, Motoo and Tomoko Ohta (1969). "The average number of generations until fixation of a mutant gene in a finite population". In: *Genetics* 61.692, pp. 763–771.

Kirkwood, T B and S N Austad (2000). "Why do we age?" In: *Nature* 408.6809, pp. 233–8.

Kirkwood, Thomas (2011). "Systems biology of ageing and longevity". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1561, pp. 64–70.

Kong, Augustine, Michael L Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S W Wong, Gunnar Sigurdsson, G Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson (2012). "Rate of de novo mutations and the importance of father's age to disease risk." In: *Nature* 488.7412, pp. 471–5.

Korte, Arthur et al. (2013). "The advantages and limitations of trait analysis with GWAS: a review". In: *Plant Methods* 9.1, p. 29.

Lewontin, Richard (1972). "The Apportionment of Human Diversity". In: *Evolutionary Biology* 6, pp. 381–398.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson (2016). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." In: *Nucleic Acids Research*, gkw1133.

Macaskill, Petra, Stephen D. Walter, and Les Irwig (2001). "A comparison of methods to detect publication bias in meta-analysis". In: *Statistics in Medicine* 20.4, pp. 641–654.

Maere, Steven, Karel Heymans, and Martin Kuiper (2005). “BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.” In: *Bioinformatics (Oxford, England)* 21.16, pp. 3448–9.

Magalhães, João Pedro de (2005). “Open-minded scepticism: inferring the causal mechanisms of human ageing from genetic perturbations.” In: *Ageing research reviews* 4.1, pp. 1–22.

– (2011). “The biology of ageing: a primer”. In: *An Introduction to Gerontology.* Ed. by Ian Stuart-Hamilton. Cambridge, UK: Cambridge University Press. Chap. The Biolog, pp. 21–47.

– (2014). “Why genes extending lifespan in model organisms have not been consistently associated with human longevity and what it means to translation research.” In: *Cell cycle (Georgetown, Tex.)* 13.17, pp. 2671–3.

– (2015). “The big, the bad and the ugly: Extreme animals as inspiration for biomedical research.” In: *EMBO reports* 16.7, pp. 771–6.

Magalhães, João Pedro de and Olivier Toussaint (2004). “Telomeres and Telomerase: A Modern Fountain of Youth?” In: *Rejuvenation Research* 7.2, pp. 126–133.

Magalhães, João Pedro de and João F. Passos (2018). “Stress, cell senescence and organismal ageing.” In: *Mechanisms of ageing and development* 170, pp. 2–9.

Malone, J, Tim F Rayner, Xiangqun Zheng Bradley, and Helen Parkinson (2008). “Developing an application focused experimental factor ontology: embracing the OBO Community”. In: *Proceedings of the Eleventh Annual Bioontologies Meeting.*

Manolio, Teri A, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy

F C Mackay, Steven A McCarroll, and Peter M Visscher (2009). “Finding the missing heritability of complex diseases.” In: *Nature* 461.7265, pp. 747–53.

Martínez, Daniel E. (1998). “Mortality patterns suggest lack of senescence in hydra”. In: *Experimental Gerontology* 33.3, pp. 217–225.

Masel, Joanna (2011). “Genetic drift.” In: *Current biology : CB* 21.20, R837–8.

Masicampo, E J and Daniel R Lalande (2012). “A peculiar prevalence of p values just below .05”. In: *The Quarterly Journal of Experimental Psychology* 65.11, pp. 2271–2279.

Matosin, Natalie, Elisabeth Frank, Martin Engel, Jeremy S Lum, and Kelly a Newell (2014). “Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture”. In: *Disease Models & Mechanisms* 7.2, pp. 171–173.

McClearn, G E (1999). “Exotic mice as models for aging research: polemic and prospectus by R. Miller et al.” In: *Neurobiology of aging* 20.2, 233–6; discussion 245–6.

McClellan, Jon and Mary-Claire King (2010). “Genetic heterogeneity in human disease.” In: *Cell* 141.2, pp. 210–7.

Medvedev, Z A (1990). “An attempt at a rational classification of theories of ageing.” In: *Biological reviews of the Cambridge Philosophical Society* 65.3, pp. 375–98.

Mi, Huaiyu and Paul Thomas (2009). “PANTHER pathway: an ontology-based pathway database coupled with data analysis tools.” In: *Methods in molecular biology (Clifton, N.J.)* 563, pp. 123–40.

Mooney, Michael A and Beth Wilmot (2015). “Gene set analysis: A step-by-step guide.” In: *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 168.7, pp. 517–27.

Mori, Ikue and Hiroyuki Sasakura (2009). “Aging: shall we take the high road?” In: *Current biology : CB* 19.9, R363–4.

Morris, Brian J., Donald Craig Willcox, Timothy A. Donlon, and Bradley J. Willcox (2015). “FOXO3: A Major Gene for Human Longevity–A Mini-Review.” In: *Gerontology* 61.6, pp. 515–25.

Müller, Bruno and Ueli Grossniklaus (2010). “Model organisms–A historical perspective.” In: *Journal of proteomics* 73.11, pp. 2054–63.

Naumova, Elissaveta, Anastasia Mihaylova, Milena Ivanova, Snejina Michailova, Kalina Penkova, and Daniela Baltadjieva (2004). “Immunological markers contributing to successful aging in Bulgarians”. In: *Experimental Gerontology.* Vol. 39. 4, pp. 637–644.

Nayak, BarunK and Avijit Hazra (2011). “How to choose the right statistical test?” In: *Indian Journal of Ophthalmology* 59.2, p. 85.

Newman, M E J (2006). “Modularity and community structure in networks”. In: *Proceedings of the National Academy of Sciences* 103.23, pp. 8577–8582.

Nica, Alexandra C and Emmanouil T Dermitzakis (2013). “Expression quantitative trait loci: present and future.” In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 368.1620, p. 20120362.

Nielsen, Julius, Rasmus B Hedeholm, Jan Heinemeier, Peter G Bushnell, Jørgen S Christiansen, Jesper Olsen, Christopher Bronk Ramsey, Richard W Brill, Malene Simon, Kirstine F Steffensen, and John F Steffensen (2016). “Eye lens radiocarbon reveals centuries of longevity in the Greenland shark Somniosus microcephalus”. In: *Science (New York, N.Y.)* 353.6300, pp. 702–704.

Nishida, Kozo, Keiichiro Ono, Shigehiko Kanaya, and Koichi Takahashi (2014). “KEGGscape: a Cytoscape app for pathway data integration.” In: *F1000Research* 3, p. 144.

Nuzzo, Regina (2014). “Scientific method: Statistical errors”. In: *Nature* 506.7487, pp. 150–152.

Passarino, Giuseppe, Francesco De Rango, and Alberto Montesanto (2016). “Human longevity: Genetics or Lifestyle? It takes two to tango”. In: *Immunity & Ageing* 13.1, p. 12.

Peters, Jaime L (2006). “Comparison of Two Methods to Detect Publication Bias in Meta-analysis”. In: *JAMA* 295.6, p. 676.

Pfeffer, Christian and Bjorn Olsen (2002). “Editorial: Journal of Negative Results in Biomedicine”. In: *Journal of Negative Results in BioMedicine* 1.1, p. 2.

Plomin, Robert, Claire M. A. Haworth, and Oliver S. P. Davis (2009). “Quantitative Traits”. In: *Genetics* 10.12, pp. 872–878.

Pritchard, J K (2001). “Are rare variants responsible for susceptibility to complex diseases?” In: *American journal of human genetics* 69.1, pp. 124–37.

Raule, Nicola, Federica Sevini, Shengting Li, Annalaura Barbieri, Federica Tallaro, Laura Lomartire, Dario Vianello, Alberto Montesanto, Jukka S. Moilanen, Vladyslav Bezrukov, H??l??ne Blanch??, Antti Hervonen, Kaare Christensen, Luca Deiana, Efstathios S. Gonos, Tom B L Kirkwood, Peter Kristensen, Alberta Leon, Pier Giuseppe Pelicci, Michel Poulain, Irene M. Rea, Jos?? Remacle, Jean Marie Robine, Stefan Schreiber, Ewa Sikora, Peternella Eline Slagboom, Liana Spazzafumo, Maria Antonietta Stazi, Olivier Toussaint, James W. Vaupel, Giuseppina Rose, Kari Majamaa, Markus Perola, Thomas E. Johnson, Lars Bolund, Huanming Yang, Giuseppe Passarino, and Claudio Franceschi (2014). “The co-occurrence of mtDNA mutations on different oxidative phosphorylation subunits, not detected by haplogroup analysis, affects human longevity and is population specific”. In: *Aging Cell* 13.3, pp. 401–407.

Reiner, Alexander P, Paula Diehr, Warren S Browner, Stephen E Humphries, Nancy S Jenny, Mary Cushman, Russell P Tracy, Jeremy Walston, Thomas Lumley, Anne B Newman, Lewis H Kuller, and Bruce M Psaty (2005). “Common promoter polymorphisms of inflammation and thrombosis genes and longevity in older adults: the cardiovascular health study.” In: *Atherosclerosis* 181.1, pp. 175–83.

Richter, C, J W Park, and B N Ames (1988). “Normal oxidative damage to mitochondrial and nuclear DNA is extensive.” In: *Proceedings of the National Academy of Sciences of the United States of America* 85.17, pp. 6465–7.

RStudio Team (2015). *RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.* Boston, MA.

Safari-Alighiarloo, Nahid, Mohammad Taghizadeh, Mostafa Rezaei-Tavirani, Bahram Goliaei, and Ali Asghar Peyvandi (2014). “Protein-protein interaction networks (PPI) and complex diseases.” In: *Gastroenterology and hepatology from bed to bench* 7.1, pp. 17–31.

Sahakyan, Aleksandr B and Shankar Balasubramanian (2016). “Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases.” In: *BMC genomics* 17, p. 225.

Saito, Rintaro, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker (2012). “A travel guide to Cytoscape plugins.” In: *Nature methods* 9.11, pp. 1069–76.

Sander, M., B. Oxlund, a. Jespersen, a. Krasnik, E. L. Mortensen, R. G. J. Westendorp, and L. J. Rasmussen (2014). “The challenges of human population ageing”. In: *Age and Ageing* 44.2, pp. 185–187.

Saunders, AM, WJ Strittmatter, Schmechel D, PH George-Hyslop, MA Pericak-Vance, SH Joo, BL Rosi, JF Gusella, and MJ Crapper-MacLachlan, DR Alberts (1993). “Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer’s disease.” In: *Neurology* 43.8, pp. 1467–1472.

Scargle, Jeffrey D. (1999). “Publication Bias (The "File-Drawer Problem") in Scientific Inference”. In: *Journal of Scientific exploration* 14.1, p. 31.

Schächter, François, Laurence Faure-Delanef, Frédérique Guénot, Hervé Rouger, Philippe Froguel, Laurence Lesueur-Ginot, and Daniel Cohen (1994). “Genetic associations with human longevity at the APOE and ACE loci.” In: *Nature genetics* 6.1, pp. 29–32.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Beno Schwikowski, and Trey Ideker (2003). “Cytoscape:

A software Environment for integrated models of biomolecular interaction networks". In: *Genome Research* 13.11, pp. 2498–2504.

Shigenaga, M K, T M Hagen, and B N Ames (1994). "Oxidative damage and mitochondrial decay in aging." In: *Proceedings of the National Academy of Sciences of the United States of America* 91.23, pp. 10771–8.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn (2011). "False-Positive Psychology 1 Running Head: FALSE-POSITIVE PSYCHOLOGY False-Positive Psychology:" in: *Psychological Science* 22, pp. 1359–1366.

Simonsohn, Uri, Leif D Nelson, and Joseph P Simmons (2014). "P-curve: a key to the file-drawer." In: *Journal of experimental psychology. General* 143.2, pp. 534–47.

Skytthe, Axel, Nancy L Pedersen, Jaakko Kaprio, Maria Antonietta Stazi, Jacob v.B. Hjelmborg, Ivan Iachine, James W Vaupel, and Kaare Christensen (2003). "Longevity Studies in GenomEUtwin". In: *Twin Research* 6.5, pp. 448–454.

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". In: *Nature Biotechnology* 25.11, pp. 1251–1255.

Smith, Richard (2006). "The trouble with medical journals." In: *Journal of the Royal Society of Medicine* 99.3, pp. 115–9.

Soerensen, Mette, Serena Dato, Qihua Tan, Mikael Thinggaard, Rabea Kleindorp, Marian Beekman, H Eka D Suchiman, Rune Jacobsen, Matt McGue, Tinna Stevnsner, Vilhelm A Bohr, Anton J M de Craen, Rudi G J Westendorp, Stefan Schreiber, P Eline Slagboom, Almut Nebel, James W Vaupel, Kaare Christensen, and Lene Christiansen (2013). "Evidence from case-control and longitudinal studies supports associations of genetic variation in APOE, CETP, and IL6 with human longevity." In: *Age (Dordrecht, Netherlands)* 35.2, pp. 487–500.

Speliotes, Elizabeth K et al. (2010). “Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index”. In: *Nature Genetics* 42.11, pp. 937–948.

Sterne, J. A C (2001). “Sifting the evidence—what’s wrong with significance tests? Another comment on the role of statistical methods”. In: *BMJ* 322.7280, pp. 226–231.

Stratton, F (1952). “The human blood groups.” In: *Nature* 170.4333, pp. 821–3.

Streiner, David L. and Geoffrey R. Norman (2011). “Correction for Multiple Testing”. In: *Chest* 140.1, pp. 16–18.

Subelj, Lovro and Marko Bajec (2011). “Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction.” In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 83.3 Pt 2, p. 036103.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov (2005). “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” In: *Proceedings of the National Academy of Sciences of the United States of America* 102.43, pp. 15545–50.

Taanman, Jan Willem (1999). “The mitochondrial genome: structure, transcription, translation and replication.” In: *Biochimica et biophysica acta* 1410.2, pp. 103–23.

Tacutu, Robi, Thomas Craig, Arie Budovsky, Daniel Wuttke, Gilad Lehmann, Dmitri Taranukha, Joana Costa, Vadim E. Fraifeld, and João Pedro de Magalhães (2013). “Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing.” In: *Nucleic acids research* 41.Database issue, pp. D1027–33.

Tan, Qihua, Jing Hua Zhao, Torben Kruse, and Kaare Christensen (2014). “Power Estimation for Gene-Longevity Association Analysis Using Concordant Twins”. In: *Genetics Research International* 2014, p. 154204.

Teufel, Andreas, Markus Krupp, Arndt Weinmann, and Peter R. Galle (2006). "Current bioinformatics tools in genomic biomedical research (Review)." In: *International journal of molecular medicine* 17.6, pp. 967–73.

Thatcher, A R (1999). "The long-term pattern of adult mortality and the highest attained age." In: *Journal of the Royal Statistical Society. Series A, (Statistics in Society)* 162.Pt. 1, pp. 5–43.

The 1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing." In: *Nature* 467.7319, pp. 1061–73.

The 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean (2012). "An integrated map of genetic variation from 1,092 human genomes." In: *Nature* 491.7422, pp. 56–65.

Tissenbaum, Heidi A (2015). "Using C. elegans for aging research". In: *Invertebrate Reproduction & Development* 59.sup1, pp. 59–63.

Trichopoulou, A and E Vasilopoulou (2000). "Mediterranean diet and longevity." In: *The British journal of nutrition* 84 Suppl 2, S205–9.

Uno, Masaharu and Eisuke Nishida (2016). "Lifespan-regulating genes in C. elegans". In: *npj Aging and Mechanisms of Disease* 2.August 2015, p. 16010.

Vanhooren, Valerie and Claude Libert (2013). "The mouse as a model organism in aging research: Usefulness, pitfalls and possibilities". In: *Ageing Research Reviews* 12.1, pp. 8–21.

Vijg, Jan and Judith Campisi (2008). "Puzzles, promises and a cure for ageing." In: *Nature* 454.7208, pp. 1065–1071.

Viña, Jose, Consuelo Borrás, and Jaime Miquel (2007). "Theories of ageing." In: *IUBMB life* 59.4-5, pp. 249–254.

Wallis, W A (1942). "Compounding probabilities from independent significance tests". In: *Econometrica* 10.3/4, p. 229.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). “The ASA’s Statement on p -Values: Context, Process, and Purpose”. In: *The American Statistician* 70.2, pp. 129–133.

Weindruch, Richard (1996). “The retardation of aging by caloric restriction: studies in rodents and primates.” In: *Toxicologic pathology* 24.6, pp. 742–5.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson (2014). “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.” In: *Nucleic acids research* 42.Database issue, pp. D1001–6.

Wiesner, B P (1932). “THE EXPERIMENTAL STUDY OF SENESCENCE.” In: *British medical journal* 2.3742, pp. 585–7.

Wilkins, Jon F. and David Haig (2003). “What good is genomic imprinting: the function of parent-specific gene expression”. In: *Nature Reviews Genetics* 4.5, pp. 359–368.

Wu, Guanming, Eric Dawson, Adrian Duong, Robin Haw, and Lincoln Stein (2014). “ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis.” In: *F1000Research* 3, p. 146.

Yook, Soon-Hyung, Zoltán N Oltvai, and Albert-László Barabási (2004). “Functional and topological characterization of protein interaction networks.” In: *Proteomics* 4.4, pp. 928–42.